Technische Universität München
Department of Electrical Engineering and Information Technology
Bio-Inspired Information Processing

Machine Learning Strategies for the Acoustic Interpretation of Snoring Noise

Christoph Janott

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation

Vorsitzender:
Prof. Dr.-Ing. Bernhard Seeber

Prüfende der Dissertation:
1. Prof. Dr.-Ing. Werner Hemmert
2. Prof. Dr.-Ing. habil. Björn Schuller

Die Dissertation wurde am 13.07.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 09.03.2021 angenommen.

Abstract

Snoring is a relevant social problem which can have a negative effect on the quality of life of the snorer's bed partner and it can be a cause for social disturbance. Snoring is multifactorial and can be generated in different locations in the upper airways. A targeted snoring treatment requires knowledge of its excitation locations and contributing anatomical structures. Existing diagnostic methods are expensive, time consuming and not tolerated by every patient. Acoustic methods to distinguish different snoring types can help to make the diagnosis more convenient and acceptable for the patient. An objective definition of snoring based on acoustic parameters does not exist until today. This thesis shows for the first time that snoring and loud breathing can be distinguished by acoustic features that are independent of the absolute sound pressure of the snoring or breathing event. Furthermore, a novel database of snoring events is described that have been annotated according to their sound excitation location based on an objective and independently verifiable ground truth. Based on this corpus, machine learning strategies have been applied to distinguish snoring sounds according to their source of excitation and the results are analysed and discussed. Derived from the VOTE classification, a popular anatomical scheme used in the diagnosis of sleep related breathing disorders, two simplified classification schemes are presented and compared with respect to their performance using machine learning strategies, and the results are discussed in view of their diagnostic usefulness. In summary, this work contributes to the development of clinical methods to improve snoring diagnosis which are cost effective, easily to apply, do not disturb natural sleep and can be used in the home environment and are therefore well tolerated by patients.

Zusammenfassung

Schnarchen ist ein relevantes gesellschaftliches Problem. Es kann sich negativ auf die Lebensqualität des Bettpartners auswirken und eine Ursache für Beziehungsprobleme sein. Schnarchen ist multifaktoriell und kann an unterschiedlichen Orten in den oberen Atemwegen entstehen. Für eine gezielte Behandlung des Schnarchens ist es erforderlich, die Entstehungsorte des Schnarchens und die beteiligten anatomischen Strukturen zu identifizieren. Bestehende Diagnosemethoden sind teuer, zeitaufwendig und werden nicht von allen Patienten toleriert. Akustische Methoden zur Unterscheidung verschiedener Schnarcharten können dazu beitragen, die Diagnose für den Patienten angenehmer zu gestalten und ihre Akzeptanz zu erhöhen. Bis heute existiert keine objektive Definition des Schnarchens anhand akustischer Parameter. Diese Arbeit zeigt erstmals, dass Schnarchen und lautes Atmen durch akustische Deskriptoren unterschieden werden können, die unabhängig vom absoluten Schalldruck des Schnarch- oder Atemereignisses sind. Weiterhin wird in dieser Arbeit eine neuartige Datenbank von Schnarchereignissen vorgestellt, welche basierend auf einer unabhängig verifizierbaren, objektiven Referenz entsprechend Ihres Schallentstehungsortes annotiert wurden. Auf der Basis dieses Korpus werden Strategien des maschinellen Lernens angewendet, um Schnarchgeräusche nach ihrer Anregungsquelle zu unterscheiden, und die Ergebnisse werden analysiert und diskutiert. Abgeleitet von der VOTE-Klassifikation, einem häufig verwendeten anatomischen Schema in der Diagnose schlafbezogener Atmungsstörungen, werden zwei vereinfachte Klassifikationsschemata vorgestellt, hinsichtlich ihrer Eignung für den Einsatz mit Methoden des maschinellen Lernens untersucht und die Ergebnisse im Hinblick auf ihren diagnostischen Nutzen diskutiert. Diese Arbeit soll einen Beitrag leisten für die Entwicklung klinischer Methoden zur Verbesserung der Schnarchdiagnostik, welche kostengünstig und einfach anzuwenden sind, den natürlichen Schlaf nicht stören, in der häuslichen Umgebung eingesetzt werden können und daher von den Patienten gut toleriert werden.

Relevant Publications

Journal Papers

Z. Zhang, J. Han, K. Qian, C. Janott, Y. Guo, and B. Schuller, "Snore-gans: Improving automatic snore sound classification with synthesized data," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 300–310, 2020

C. Janott, M. Schmitt, C. Heiser, W. Hohenhorst, M. Herzog, M. Carrasco Llatas, W. Hemmert, and B. Schuller, "Vote versus aclte: comparison of two snoring noise classifications using machine learning methods," *HNO*, vol. 67, no. 9, pp. 670–678, 2019

K. Qian, M. Schmitt, C. Janott, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller, "A bag of wavelet features for snore sound classification." *Annals of biomedical engineering*, vol. 47, pp. 1000–1011, 2019

C. Janott, M. Schmitt, Y. Zhang, K. Qian, V. Pandit, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller, "Snoring classified: The munich-passau snore sound corpus," *Computers in Biology and Medicine*, vol. 94, pp. 106–118, March 2018

K. Qian, C. Janott, Z. Zhang, J. Deng, A. Baird, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmer, and B. Schuller, "Teaching machines on snoring: A benchmark on computer auditionfor snore sound excitation localisation," *Achives of Acoustics*, vol. 43, no. 3, pp. 465–475, 2018

C. Janott, B. Schuller, and C. Heiser, "Acoustic information in snoring noises," *HNO*, vol. 65, no. 2, pp. 107–116, 2017

K. Qian, C. Janott, V. Pandit, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller, "Classification of the excitation location of snore sounds in the upper airway by acoustic multi-feature analysis," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1731–1741, 2017

Journal Papers (Continued)

C. Janott, W. Pirsig, and C. Heiser, "Acoustic analysis of snoring sounds," *Somnologie-Schlafforschung und Schlafmedizin*, vol. 18, no. 2, pp. 87–95, 2014

Conference Papers

C. Janott, C. Rohrmeier, M. Schmitt, W. Hemmer, and B. Schuller, "Snoring - an acoustic definition," in *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Berlin, Germany, 2019, pp. 3653–3657

K. Qian, C. Janott, J. Deng, C. Heiser, W. Hohenhorst, M. Herzog, N. Cummins, and B. Schuller, "Snore sound recognition: On wavelets and classifiers from deep nets to kernels," in *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Seogwipo, South Korea, 2017, pp. 3737–3740

B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 20–24

M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller, "A bag-of-audio-words approach for snore sounds' excitation localisation," in *Proceedings of ITG Speech Communication*, Paderborn, Germany, 2016, pp. 230–234

K. Qian, C. Janott, Z. Zhang, C. Heiser *et al.*, "Wavelet features for classification of vote snore sounds," in *2016 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 221–225

Contents

Chapter 1

Introduction

"Laugh and the world laughs with you.
Snore and you sleep alone."
(Antony Burgess, 1917 - 1993)

1.1 Motivation

Approximately one out of three adults in the western world snores (14; 15).

While primary snoring without the disposition of airway obstruction does not directly affect the health of the snorer, it can have a negative effect on the sleep structure and quality of life of the bed partner (16). Further, snoring can be a reason for social disturbance, e.g., when sleeping in dormitories or camping sites. Last not least, it can affect partnerships.

Snoring is a considerable social problem. A study on the sleep quality of parents revealed that 32% of mothers reported sleeping difficulties caused by the snoring of their husband. Interestingly, vice versa, only 17% of fathers felt disturbed by their snoring wives (17). Robin noted that snoring "can ruin a happy marriage and in some parts of the USA it is considered justification for divorce" (18). It has even been fatally concluded that "isolation may be the only effective measure" against snoring (19). Cure is often seeked by the bed partner. As Fabricant put it: "It is strange that the snorer himself is rarely distressed by the havoc he causes. Rather it is the victimised spouse, the sleepy person with bloodshot eyes, who leads the sheepish person to the doctor" (20).

Snoring can be a lead symptom of Obstructive Sleep Apnoea (OSA), a serious disease that can severely affect the health of the person concerned. It has been

shown that OSA can cause serious cardiovascular conditions and it is an independent risk factor for hypertension, heart attack and stroke. Besides, OSA disturbs the sleep structure and prevents a restful sleep. In consequence, it can lead to excessive daytime sleepiness affecting the quality of life, and it even can cause micro-sleep attacks during the day, a fact that is especially dangerous in certain professions such as truck or bus drivers. Still, studies have shown that OSA is undiagnosed in the vast majority of patients (21). Thus, there is a demand for easily accessible and cost-effective methods for the screening of OSA.

Snoring is caused by vibrations of slackening muscles in the upper respiratory tract at physiological bottlenecks (22). Due to the complex anatomy of the upper airways, snoring is multi-factorial and can be generated by several different anatomical structures. These include the soft palate and the uvula, the oropharyngeal area where the palatine tonsils are locared, the tongue base, the epiglottis, as well as laryngeal structures. Figure 1.1 gives an orientation of the location of the different structures within the upper airways. During sleep the muscle tone decreases and the soft tissue's tendency to vibrate increases. Constrictions in the upper airways that occur during sleep can also trigger tissue vibrations by increasing the respiratory flow velocity, which can lead to turbulent flows (Law of Hagen-Poiseuille), in turn causing snoring noise.

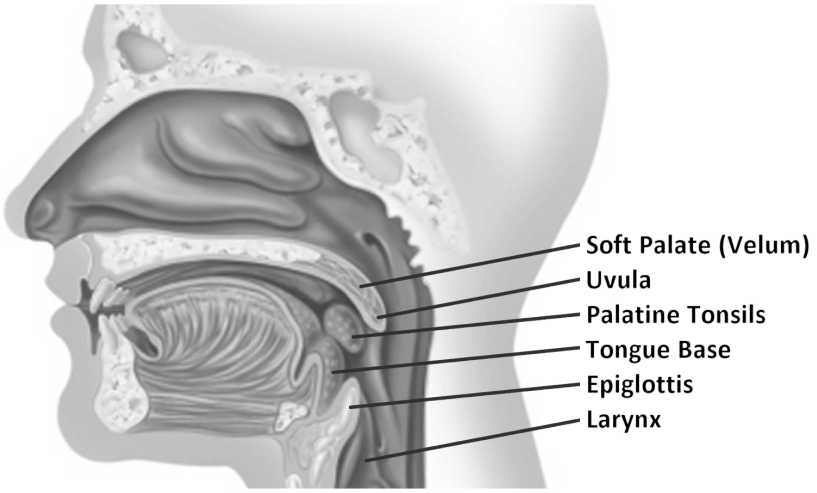

Figure 1.1: *Anatomical structures in the upper airways that can contribute to snoring*

For a targeted treatment of snoring and related sleep-related breathing disorders, it is of decisive importance to know the exact mechanisms in the upper airways that cause the snoring sound. Although diagnostic methods exist, these require the introduction of diagnostic equipment into the upper airways during

sleep, a requirement that not every patient tolerates, or which is not possible at all without disturbing the sleep. Acoustic methods to distinguish between different snoring types can help to make the diagnosis easier tolerable and therefore more acceptable for the patient. As acoustic measurements do not require invasive measures, they can be performed without the need for the patient to wear sensors attached to the body during sleep, and the availability of small and lightweight measurement equipment permits the measurements to be performed at home in the accustomed sleeping environment of the patient.

In the past decades, the acoustics of snoring have been a popular field of research. Much has been published on the difference in acoustic properties of primary snoring compared to snoring as a symptom of OSA. Also, the acoustic characteristics of different types of snoring have been investigated and specific differences of their acoustic peculiarities have been described.

However, the knowledge about significant differences in acoustic properties of different snoring types does not allow the assignment of a specific snoring subject to one of several snoring types with satisfactory accuracy. The relatively new technology of machine learning is a promising tool to develop strategies to solve this task. By training a machine classifier to assign snoring events to one of several defined classes, the type of snoring can be predicted solely based on its acoustic properties.

Although much work has been done on machine learning methods to distinguish between primary snoring and OSA and to predict OSA severity, little has been published on machine classification of different types of snoring. An obvious reason for the virginity of this specific field is the lack of available, objectively labelled training data. Successful and meaningful machine learning depends on the availability of sufficiently large amounts of structured data that is labelled based on an objective ground truth. Audio data that is objectively labelled according to OSA severity can be comparatively easily derived from polysomnographic examinations. Polysomnography is the gold standard for the diagnosis of sleep disorders and it is performed in a vast number of sleep laboratories throughout the world each single night. In contrast, an objective ground truth on the underlying excitation mechanisms of snoring sounds can only be obtained by visual observation of the upper airway structures at the exact moment when the snoring event occurs.

A possibility to obtain this simultaneous visual and acoustic information are drug induced sleep endoscopy (DISE) examinations, in which the upper airways are inspected by an experienced examiner by means of a thin, flexible endoscope that is introduced through the nose while the patient is in a state of artificial sleep. Such recordings are comparatively rare and not easy to obtain. Firstly, DISE examinations require skilled and experienced medical experts, therefore they are performed in selected Ear, Nose and Throat (ENT) centres only. Secondly, not

all medical centres routinely file the audio information. In many cases, only the video data is recorded. Thirdly, often only selected sections of a DISE video are filed for medical documentation purposes for a limited time in order to reduce the amount of data to be stored.

This work has been made possible by the cooperation with ENT experts from four medical centres who routinely record simultaneous video and audio information during DISE examinations and file recordings of complete DISE examinations for research purposes. In fact, the initial idea for this work was triggered by an incidental discussion during a routine work day in an ENT operating room where the attending medical personnel performed a 'quiz' whether they could 'hear' the findings from an ongoing DISE examination before looking at the endoscopic pictures. Experienced ENT experts performed impressively well, and the idea was born to find out if machine intelligence could cope with human intelligence in this task.

The author was able to resort to a treasure of historic data that has been recorded in clinical routine during a period of more than 10 years. For this reason, complete medical records of the respective patients were not accessible for all recordings, but essential patent data such as age and gender could completely be retrieved.

In total, more than 100 hours of video and audio data stemming from more than 2.500 patients were processed for this work in a combination of automated and manual steps in order to create a database of snoring sounds labelled by their excitation location. The yield of data that is suited for machine learning is relatively small, more than 85% of the raw material had to be discarded due to technical issues during recording, insufficient audio or video quality, or simply because the patient did not snore during the examination.

The scientific interest in the resulting Munich-Passau Snore Sound Corpus (MPSSC) shows that the effort was worthwhile. Since the first publication of the corpus in the year 2017, it has been licensed and used independently by more than 35 academic groups worldwide for machine learning experiments and therefore has become a popular database for research in this very specific scientific topic.

1.2 Research Questions and Contributions

This thesis addresses the following research questions.

1. What is snoring? Almost everybody has a instinctual idea how snoring sounds, but the delimination to other nocturnal breathing sounds is quite

individual and lies in the ear of the beholder. In addition, subjective judgement can be influenced by other, non-acoustic factors. The mild snoring of a beloved bed partner may be tolerated well, whereas pronounced breathing sounds of the person sharing the same bed can be rather disturbing in a complicated relationship. Although the acoustics of snoring have been extensively researched for more than a century, a satisfactory, objective definition of this acoustic phenomenon is still not available. An essential prerequisite for going about the discrimination of different snoring types is the identification of snoring events themselves and their distinction from other types of respiratory sounds. This problem is addressed by investigating a number of acoustic features representing the temporal and spectral properties of the underlying signal, in order to explore their sensitivity and specificity in the discrimination of snoring from its most similar acoustic counterpart, loud breathing. This is done utilising a corpus including snoring and loud breathing sounds, which were labelled by a comparatively large group of blinded annotators to reduce individual bias. In order to allow for a practical measurement setup that does not require calibration, the investigated features or feature sets are required to be independent of the absolute sound pressure.

2. Can different excitation locations of snoring sounds be distinguished solely by their acoustic properties? In order to address this question, a novel database of snoring events is presented that have been classified by their sound excitation location in the upper airways. Annotation of the snoring events has been carried out based on simultaneous endoscopic video recordings of the upper airways and is therefore objective and independently verifiable. To the author's knowledge, no other such database is publicly available to date. Based on this corpus, machine learning strategies are applied to train classifiers to distinguish snoring sounds according to their source of excitation.

3. How relevant and useful are the classification results achieved for conservative and surgical therapy decisions? Different anatomical schemes exist for the classification of respiration-dependent sounds and constrictions, distinguishing different levels and orientations of vibrations or airway narrowing. To address this question, two simplified classification schemes are compared for their performance using machine learning strategies, and the results are discussed in view of their diagnostic usefulness. Both schemes are derived from a classification that is frequently used and widely accepted for DISE-based diagnosis in clinical routine.

Overall, this work might contribute to find a consensus for an objective definition of snoring as opposed to (loud) breathing based on acoustic parameters, and

serve as a guidance for future applications in the automatic detection of snoring sounds. Further it might contribute to the development of clinical methods for the diagnosis of snoring which are cost effective, easily to apply, do not disturb natural sleep and can be used in the home environment of the patients.

1.3 Structure of this Thesis

The remainder of this thesis is structured as follows. Chapter 2 provides a background on sleep medicine. Section 2.1 contains a historical sketch on the development of sleep medicine and an overview of the different kinds of sleep disorders according to the current medical consensus, explaining in detail the established methods for the diagnosis of sleep disordered breathing. Section 2.2 presents the results of a comprehensive literature research on snoring sound analysis and gives an overview of the research contributions that the author of this thesis has authored or co-authored. Section 2.3 explains the acoustic particularities of snoring sounds and the features that can be used to describe its properties. Chapter 3 describes experimental work on machine learning strategies for acoustic snoring noise interpretation. Section 3.1 deals with the identification of suitable acoustic features to distinguish snoring and loud breathing, Section 3.2 describes in detail the development and the properties of the Munich-Passau Snore Sound Corpus and the results of baseline machine learning experiments, and in Section 3.3, two different snoring sound classification schemes are compared for their suitability for machine-learning-based snoring sound classification. Finally, a summary is provided in Chapter 4.

Some parts and results of this thesis are based on work that has been previously published by the author of this thesis as the first author. In particular, the literature research results in Section 2.2 were originally published in German language in (6), while parts of Section 2.3 were previously published in German language in (8). Section 3.1 is based on (9), the database in Section 3.2 was first introduced in (4), and Section 3.3 is based on (2), which was published in German language. Further, parts of the contents in Section 2.1 and 2.3 are due to appear in a chapter of a book titled 'Biomedical Signal Processing with Artificial Intelligence'.

Chapter 2

Background

“ ‘Sleep!’ said the old gentleman,
‘he’s always asleep. Goes on errands
fast asleep, and snores as he waits at table.’
‘How very odd!’ said Mr. Pickwick.”
(Charles Dickens, The Posthumous Papers of the Pickwick Club, 1836)

2.1 Sleep Disorders

2.1.1 Evolution of Sleep Medicine

In ancient cultures, sleep was considered a state of unconsciousness which was closely related to darkness, death, and the mystic phenomenon of dreaming.

In the Roman mythology, Somnus, the god of sleep, resided in the underworld. His sons, the Somnia, appeared in dreams in various shapes and forms. The Greek god of sleep, Hypnos, and his twin brother Thanatos, the god of a peaceful death, could relieve humans from suffering and help them to die a peaceful death during sleep. Of his four children, Morpheus shaped the dreams, while Ikelos represented people in dreams and Phantasus could transform into objects. Phobetor was responsible for nightmares by appearing in the shapes of monsters and animals. In ancient China, dreams were an important factor in the diagnosis of illnesses, whereas the Egyptians believed that dreams were predictors of the future.

The ancient myths live on in a number of ubiquitous English words. Somnus has lent his name to the field of sleep research, somnology. Hypnosis stems from Hypnos, whereas euthanasia is derived from Thanatos. Finally, morphine is

fittingly named after Morpheus, who used to rest in his dark cave among poppy flowers.

Today, it is well known that sleep is composed of a series of active processes of the human brain, rather than being a perceived as a passive state of the brain being shut down. Measuring these processes serves to detect abnormalities in the sleep structure, which can be telling symptoms of an underlying disease.

Figure 2.1: *The god of sleep*
Relief by Baumeister. Source: Sorbonne University

In the 19th century, the phenomenon of sleep was the subject of extensive research and it was understood that the brain was not simply turned off during sleep. Ernst Kohlschuetter in 1863 considered that "Sleep and waking are two opposing states of mental life. The same is not completely extinguished in sleep, which is proved by dreams and the possibility of awakening a sleeper by a strong local stimulus." (23). Kohlschuetter measured the awaking reaction of sleeping subjects at different times during the night by applying acoustic stimuli of defined strength. He found that the depth of sleep changes during the course of the night, being deeper in the beginning and becoming shallower towards the morning.

Several theories were debated about the physiological mechanisms inducing the state of sleep. A common misconception was that sleep was induced by a blood congestion in the head, with the increased pressure on the brain causing unconsciousness (24). Others thought that sleep is caused by increased cerebral

blood flow (25), or decreased cerebral blood supply. Durham in 1860 published his observations of the blood circulation in dog brains, the skull of which he had cut open and covered with a glass window (26). He reported that the blood-flow in the brain is reduced during sleep. Consequently, sleeplessness was treated by substances reducing the cerebral blood flow, such as potassium bromide.

The debate whether sleep disorders should be treated pharmacologically or psychologically was already in full progress. For centuries, wine and opium were used as effective sleeping aids. In stark contrast, Russell in 1861 had a strong opinion that sleeping problems have their cause in the mental state of the sufferers, and stated that the physician's "attention must be directed to regulating and strengthening the mind. ... Even in cases of organic disease, the chances of recovery depend quite as much on the state of the patient's mind as on that of his body" (27). Russell's opinion should prove quite visionary. Recent research shows that behavioural therapy is an effective means in the treatment of primary insomnia (28).

Figure 2.2: *Medieval surgical instrument for uvula resection invented by Abulcasis*

A milestone in the understanding of the structure of sleep was the invention of electroencephalography and its use for measuring the activity of the human brain in the early 20th century (29). For the first time, the brain could be directly monitored during sleep my means of its electric emissions. Not long after this discovery, distinct sleep states were observed and defined based on their electroecncephalographical wave properties. It was discovered that in frequent cases, body "movement was immediately followed by a change of state upward, occasionally downward, and occasionally a movement occurred just after a change of state. During sleep there was a continual shift in states upward and downward, sometimes associated with recognised stimuli, sometimes without any external stimulus but probably as a result of internal stimuli." (30). The sleep states defined are still the basis of today's definition of depth of sleep. The phenomenon of rapid eye movements (REM) during sleep and its relation to dreaming was discovered in the 1950s (31). Finally, in the 1960s, a consensus committee of sleep specialists and researchers under the chairmenship of Allan Rechtschaffen and Anthony Kales agreed on standardised criteria for sleep staging. With the so-called R&K manual,

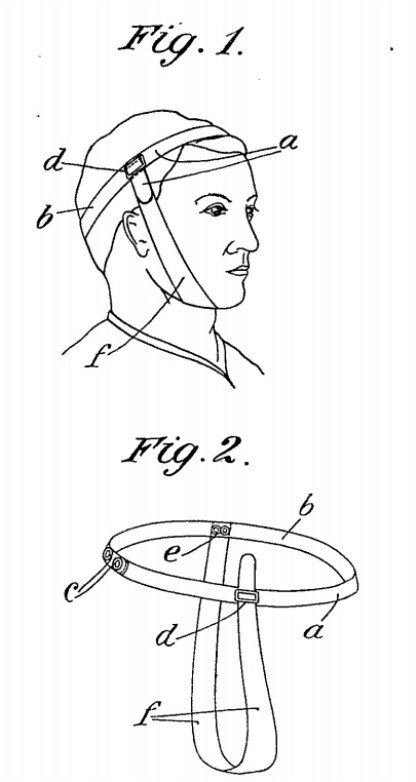

Figure 2.3: *Drawing of an anti-snore device*
Swiss patent no. 2077 from the year 1890

they have created a standard piece of work for an internationally accepted system of sleep scoring, which is, with slight adaptations, valid to date and a reference for sleep research and diagnosis (32).

Snoring as a symptom during sleep was known and treated since the antiques. Already in the medieval ages, Abu al-Qasim Khalaf ibn al-Abbas al-Zahrawi (also known as Abulcasis) described surgical instruments used for the resection or cauterisation of the uvula as a cure for snoring (33), see Figure 2.2. In the 16th century, Levinus suspected that sleeping in supine position while breathing through the mouth may lead to an unrestful sleep (34). In 1891, Catlin described the relationship between jaw position and snoring: "Few people can be convinced that they snore in their sleep, for the snoring is stopped when they awake; and so with breathing through the mouth, which is generally the cause of snoring - the moment that conciousness arrives the mouth is closed, and nature resumes her usual cause" (35). Around the same time, items preventing the mouth from opening were described as means to prevent snoring (see Figure 2.3).

Numerous methods for the treatment of snoring have been developed, ranging from established conservative measures such as oral appliances to advance the mandible during sleep (36), to rather unusual methods such as didgeridoo playing as a means of oral musculature training (37). It is well proven that weight loss effectively reduces the severity of snoring in the majority of overweight patients (38).

Japanese surgeon Ikematsu was one of the pioneers of modern snoring surgery and opened new therapeutic perspectives from the 1950s. In 1964, he proposed a combined surgery of the velum, the uvula and the palate for snoring, later becoming known as Uvulopalatopharyngoplasty (UPPP). His approach was quite successful with a reported improvement of snoring symptoms in 81% of patients (39). A decade later, In 1977, Simmons et. al. published on operations for sleep disordered breathing, targeting the nose, tonsils and the laryngeal area (40). Eventually, the First International Congress on Chronic Rhonchopathy was held in Paris in the year 1987. At this congress, Chabolle described that "anatomic causes of snoring involve on the one hand soft palate abnormalities and on the other, an obstacle of the upper aerial tracts." (41).

In the following decades, sleep surgery became more radical by more invasive versions of the UPPP. A new approach was established, the so-called multi-level surgery, involving the surgical treatment of several levels of the upper airways (42). While complication rates and morbidity strongly increased, success rates remained moderate (43; 44). It became apparent that improving surgical success rates while at the same time reducing morbidity to an acceptable level could only be achieved by more accurate diagnostic approaches (45). It is easy to apprehend that, for example, treatments targeting the soft palate prove to be more successful in patients where the reason for the snoring or the OSA lies in the velar area (46; 47) and has less effect when snoring is predominantly generated by the tongue base or the posterior pharyngeal walls (48). Vice versa, procedures primarily targeting the hypopharyngeal area might not be the first choice of treatment in purely palatal snorers (49; 50).

Therefore, knowledge of the mechanism and location of obstruction and snoring sound excitation within the UA is vital for targeted interventions. The following sections will give an overview on sleep disorders and diagnostic options, with a special focus on sleep disordered breathing.

2.1.2 Classification of Sleep Disorders

Today's key reference for the description of sleep anomalies is the International Classification of Sleep Disorders (ICSD), which is issued and regularly updated by the American Academy of Sleep Medicine (AASM) (51). Table 2.1 shows the

conditions that are described in this classification and their prevalence in healthy adults (52; 53; 54; 55; 56; 57; 58).

Diagnostic Section	**Disorder**	**Prevalence**
Insomnia	Chronic insomnia disorder	up to 10%
Central hypersomnolence	Narcolepsy Hypersomnia Insufficient sleep syndrome	up to 6%
Circadian rhythm sleep-wake disorders	Delayed sleep-wake phase disorder Advanced sleep-wake phase disorder Irregular sleep-wake rhythm disorder Non-24-h sleep-wake rhythm disorder Shift work disorder Jet lag disorder	up to 2%
Parasomnias	Non-REM-related parasomnias REM-related parasomnias Other parasomnias	<1%
Sleep-related movement disorders	Restless legs syndrome Periodic limb movement disorder Leg cramps Bruxism Rhythmic movement disorder	up to 15%
Sleep-related breathing disorders	Obstructive sleep apnoea syndromes Central sleep apnoea syndromes Sleep-related hypoventilation disorders Sleep-related hypoxemia disorder	up to 38%

Table 2.1: *Sleep anomalies according to the International Classification of Sleep Disorders*

- *Insomnia* is characterised by problems to initiate or maintain sleep as well as the lack of adequate circumstances or opportunities to sleep. It also comprises the daytime consequences of this disorder. Chronic insomnia, as opposed to short-term insomnia, occurs at least three times a week and lasts longer than three months (51).

- *Central Hypersomnolence* is characterised by a subjective complaint of excessive daytime sleepiness. A central hypersomnolence is diagnosed if this symptom cannot be related to other types of sleep disorders. This diagnostic section subsumes narcolepsy, characterised by sudden episodes of muscle weakness or loss of muscle tone, which is often triggered by strong positive or negative emotions. Narcolepsy can also go along with hypnagogic hallucinations and sleep paralysis. It is relatively rare with a prevalence of 0.03...0.05% (51; 59; 60).

  Hypersomnia is defined as an excessive duration of nighttime sleep in combination with daytime sleepiness (61), while the insufficient sleep syndrome is simply a consequence of regularly getting less sleep than the body needs.

- *Circadian rhythm sleep-wake disorders* are defined as recurrent disruptions of the sleep-wake rhythm by a mismatch of the endogenous circadian timing system (the inner clock) with external requirements or circumstances. Mild forms of the disorder include the typical 'early birds' or 'night owls', meaning people who tend to wake up early or go to sleep late. Non-24-h sleep-wake rhythm disorder often occurs in blind people, where the inner circadian timer is not triggered by the natural light and dark rhythm of day and night. By definition, circadian rhythm sleep-wake disorders must last more than three months to be clinically relevant. Shift work disorder and jet lag are diagnosed whenever the symptoms can be related to the respective cause (51).

- *Parasomnias* are disruptive sleep disorders that involve abnormal behaviour or perceptions during sleep. They are subdivided by their occurrence during either non-REM or REM sleep. Non-REM parasomnias include sleep-walking, confusional arousal and night-terrors. REM sleep parasomnia is often characterised by an absence of muscle atonia while dreaming. Consequently, patients can act out their dreams, with the risk of injuring themselves or others. In rarer cases, muscle atonia is still retained while waking up, leading to the symptom of incubus, the inability to move or speak whilst being awake and concious (51).

- *Sleep-related movement disorders* are abnormal movements related to sleep. The restless legs syndrome is characterised by the urge to move the legs during periods of physical inactivity primarily occurring at night time or in the evening. In contrast, periodic limb movements, rhythmic movements, or other movement disorders such as bruxism (teeth grinding), or myoclonus (brief, irregular jerks) occur involuntarily during sleep.

- *Sleep related breathing disorders* are abnormal breathing behaviours during sleep, which can have different causes and characteristics.

Obstructive sleep apnoea (OSA) is characterised by repeated episodes of decreased (hypopnoea) or completely halted (apnoea) airflow despite an ongoing effort to breathe. The reported prevalence in the general population varies. A recent systematic review reports a prevalence in the range of 9% and 38% for mild OSA and between 6% and 17% for moderate OSA (58). Other estimates for mild OSA are as high as over 60% in men, and 30% in women, respectively, in some countries (62). However, the latter research was funded by one of the leading manufacturers of OSA therapy devices and should therefore be considered carefully. It is indisputable that OSA prevalence is higher in men than in women and increases with age (63). Symptoms associated with OSA include daytime sleepiness, excessive fatigue, and morning headache. It is a serious health condition and an independent risk factor for cardiovascular diseases such as hypertension and myocardial infarction (14).

Precisely, an apnoeic event is defined as a complete cessation of airflow for at least ten seconds. Some stricter definitions require a drop in blood oxygen saturation by >3% as a result of the apnoeic event. A hypopnoeic event is defined as a drop of >30% in peak airflow compared to the pre-event baseline lasting for a minimum of ten seconds, and a drop in blood oxygen saturation by >3% or an arousal associated with the hypopnoeic event. The averaged number of apnoeas and hypopnoeas occurring per hour of sleep is measured by the Apnoea-Hypopnoea-Index (AHI). By common grading standards (64; 65), an AHI of 5...<15 events per hour is considered mild OSA, 15...30 events per hour is moderate, and > 30 events per hour is defined as severe OSA. A similar measure is the Respiratory Disturbance Index (RDI), which counts any respiratory event of a minimum of ten seconds if it is associated with an arousal, even if it does not meet the formal criteria of a hyponoea.

Central sleep apnoea (CSA) also involves repeated nightly breathing cessations. In contrast to OSA, however, these are caused by a central disorder of the ventilatory control system. In many sleep apnoea patients, both OSA and CSA episodes can be detected, a condition sometimes referred to as mixed apnoea or complex sleep apnoea. (51).

Sleep related hypoventilation and *hypoxemia disorders* are diagnosed by decreased oxygen and raised carbon dioxide levels in the blood, and they usually occur in overweight patients. Charles Dickens pointedly described a typical phenotype of a person suffering from the obesity hypoventilation syndrome in the character of Little Fat Joe in the Pickwick Papers (66). Originally, this condition was named the *Pickwickian syndrome* after this novel (67).

Primary snoring (simple snoring) is characterised by the absence of apnoeic or hypopnoeic episodes. According to the ICSD, snoring itself is not a sleep-related breathing disorder, but it can be an isolated symptom or normal variant of other sleep-related breathing disorders. Loud snoring is a typical symptom associated with OSA in more than 90 % of patients (68; 69; 70).

The designation of snoring events is inconsistent in literature. In this work, the noise that occurs within a single breathing cycle is referred to as a *snoring event*, while a sequence of snoring events, i.e. a period of continuous snoring, is referred to as a *snoring episode*. A snoring episode contains several snoring events. Figures 2.4 and 2.5 illustrate this. In natural sleep, episodes of quiet breathing and snoring episodes of varying duration usually alternate several times during the night.

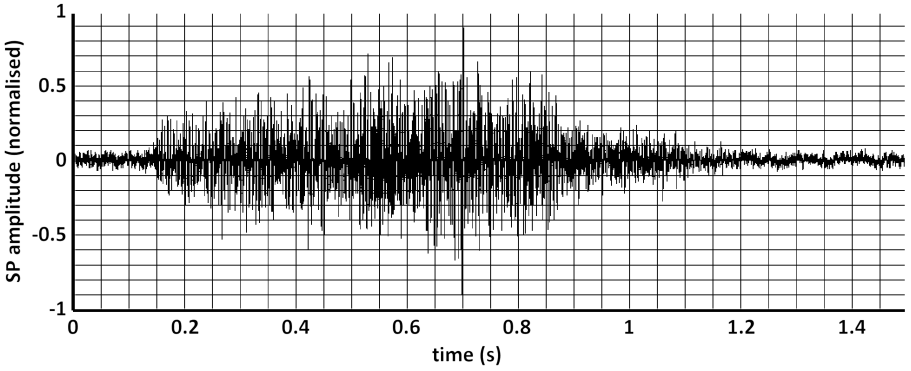

Figure 2.4: *Temporal course of the sound pressure (SP) amplitude of a single snoring event.*

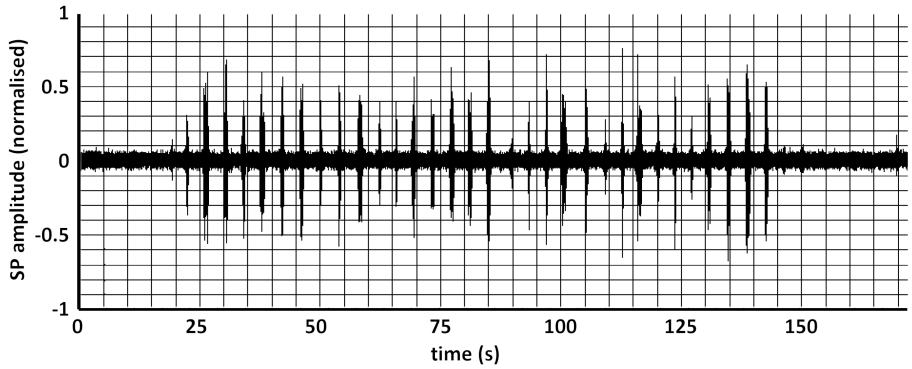

Figure 2.5: *Temporal course of the sound pressure amplitude of a snoring episode containing several snoring events.*

2.1.3 Diagnosis of Sleep Disordered Breathing

Polysomnography

Sleep structure and sleep events coincide with characteristic patterns of physiological activity that can be monitored through a variety of parameters. The most comprehensive method of sleep monitoring and diagnostic gold standard is Polysomnography (PSG). Full PSGs are carried out in sleep laboratories consisting of one or several bedrooms, fully equipped to record a patient's physiological parameters during sleep for a whole night. The definition of a whole night's sleep differs, but is generally considered a minimum sleep time of six hours. The physiological data is transferred to a central ward where the patents are monitored during the night and the somnogram is recorded for subsequent evaluation.

Figure 2.6 shows a typical bedroom in a sleep laboratory, while a central monitoring ward can be seen in Figure 2.7.

Figure 2.6: *Bedroom in a sleep laboratory*
Courtesy of Klinikum rechts der Isar, Munich, Germany

A standardised measurement recommendation for PSG was developed in the 1950s and further refined in the following decades. The current recommended parameters are as follows (71).

- The brain activity measured through several Electroencephalographical (EEG) derivations is indispensable to detect sleep stages, which coincide

Figure 2.7: *Monitoring ward in a sleep laboratory*
Courtesy of Klinikum rechts der Isar, Munich, Germany

with specific brain wave patterns. Derivations are unipolar against a central reference point, which is usually located at the ear lobe or mastoid bone.

- Eye movements are detected by electrooculography (EOG) and are important indicators to detect REM sleep. Two electrodes attached to the right corner of the right eye (right outer canthus), and the left outer canthus, respectively, record the muscle activity related to eye movements.

- Electromyograms (EMG) are derived to aid in sleep phase estimation and to record limb movements. An electrode located above the chin is used to measure general muscle tone. Further, the surface muscle tone of the right and left anterior tibial muscles are derived to detect leg movements.

- A one-channel Electrocardiogram (ECG) is derived to measure heart rate and to detect tachycardia and cardiac arrhythmias.

- A finger or and earlobe pulse oximetry sensor detects the course of oxygen saturation to determine apnoeas and hypopnoeas.

- A contact microphone is attached to the neck in the tracheal region to detect snoring, coughing and speaking activity.

- Nasal pressure sensors and oral airflow detectors record breathing by pressure changes and thermal differences between inspiration and expiration. The data is used to detect breathing irregularities during sleep.

- Chest and abdominal movements during inspiration and expiration are measured using two expansion belts applied above the bony thorax and above the umbilicus. Paradoxical thoracic and abdominal movements are a sign for breathing effort without airflow, signifying an obstructive apnoeic event, whilst simultaneous cessation of airflow and chest movements are a sign for a central apnoea.
- A positional sensor is fixed to the thoracic belt to record the sleeping position.
- Optionally, an infrared video recording of the sleeper is carried out to diagnose parasomnia and to allow for a differential diagnostic delimitation to certain forms of epilepsy (72).

Sample rates for EEG, EMG, ECG, and EOG data as well as the data from the microphone are recommended to be 500 Hz, while airflow, chest and abdomen movements are sampled at 100 Hz. Oxygen saturation is sampled at 25 Hz, while the body position is recorded with a resolution of 1 Hz.

Figure 2.8 shows a somnogram of several subsequent obstructive apnoeic events, whilst Figure 2.9 displays a series of central apnoeas, a condition also referred to as Cheyne Stokes breathing. Figure 2.10 shows an example of a hypopnoeic event. Finally, Figure 2.11 contains an example somnogram of a primary snorer. For intranasal pressure, thermal airflow, thoracic and abdominal movement, combined effort, and contact microphone, the specified measured values are those of the sensors used, representing the actual physiological parameters.

Until recently, sleep analysis based on polysomnographic recordings and identification of sleep-related events is largely performed manually by sleep experts with the help of semi-automated PSG monitoring systems, providing assistance by suggesting sleep stages, movement and respiratory events from the sleep recordings. However, the ultimate scoring is still made by a human expert. Recently, machine learning approaches have been evaluated for the automation of PSG analyses and data preprocessing, such as artifact detection (73). Using combinations of different feature sets and classifiers, promising results could be achieved.

Ambulatory Cardiorespiratory Screening

Full polysomnography has a number of disadvantages. It is expensive, limiting the number of nights that a subject can be investigated. What's more, it requires the subject to sleep in a dedicated sleep laboratory and therefore in an unfamiliar environment, which is likely to influence the sleep behaviour, sleep structure and patterns. This effect is especially pronounced when the subject sleeps in a sleep lab for the first time, called the first night effect.

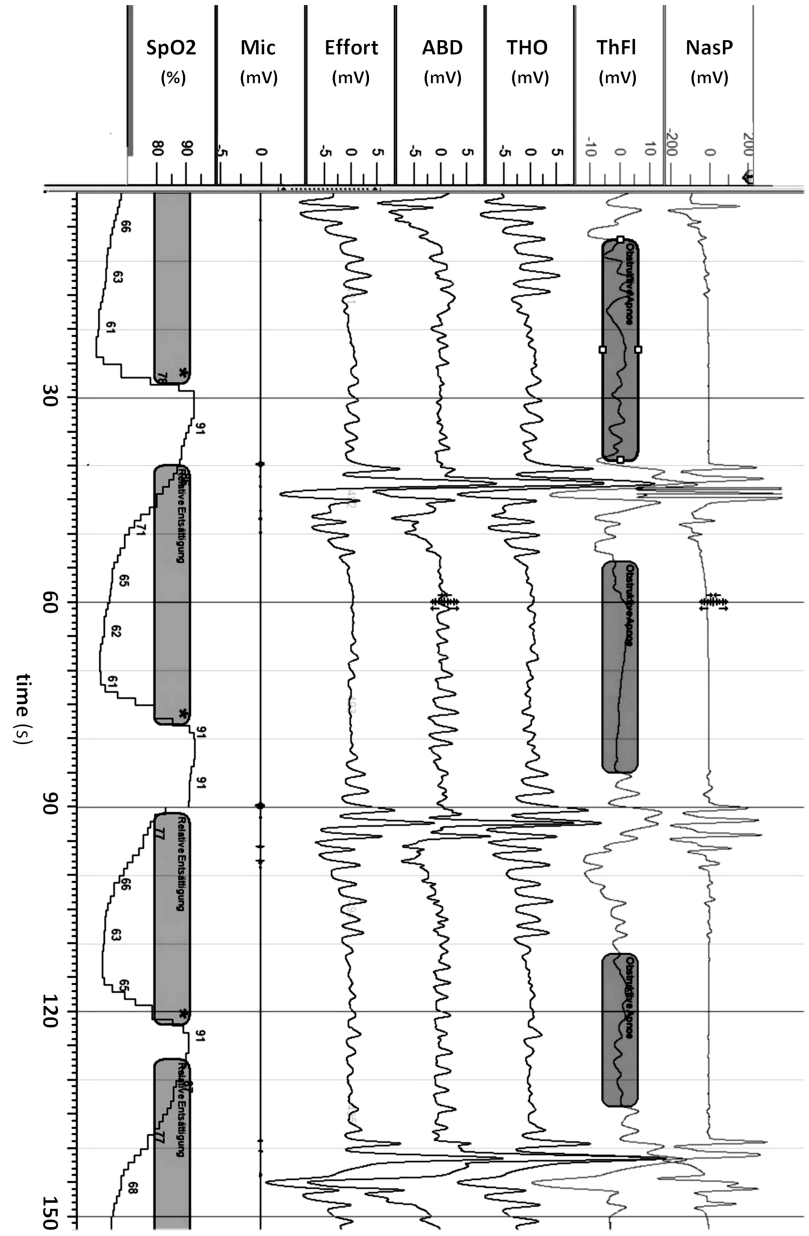

Figure 2.8: *Example somnogram of several subsequent obstructive events. NasP = intranasal pressure. ThFl = thermal airflow sensor, periods of airflow cessation are marked. THO = thoracic movement. ABD = abdominal movement. Effort = combined thoracic and abdominal effort, showing a paradoxical movement of thorax and abdomen during phases of apnoea. Mic = neck contact microphone, showing single snoring events when breathing is restored. SpO2 = blood oxygen saturation, phases of severe desaturation are marked.*

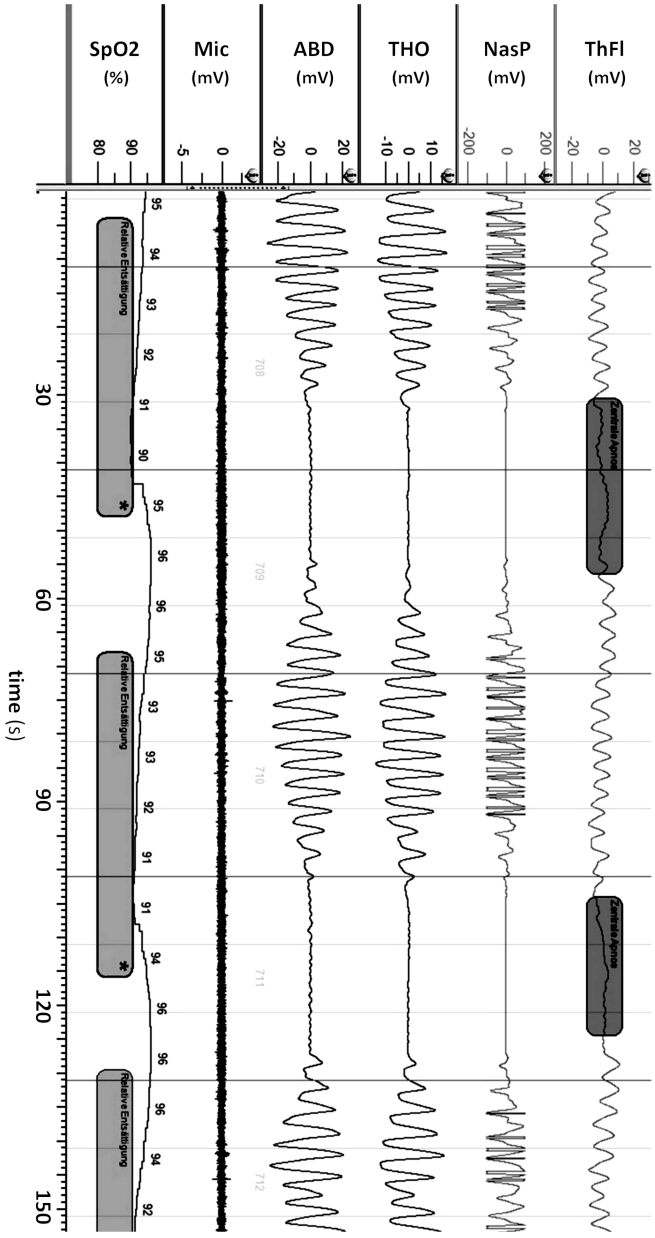

Figure 2.9: *Example somnogram of several subsequent central apnoeic events. ThFl = thermal airflow sensor, periods of airflow cessation are marked. NasP = intranasal pressure. THO = thoracic movement. ABD = abdominal movement, showing repeated cessations of breathing effort leading to apnoea. Mic = neck contact microphone. SpO2 = blood oxygen saturation, phases of desaturation are marked.*

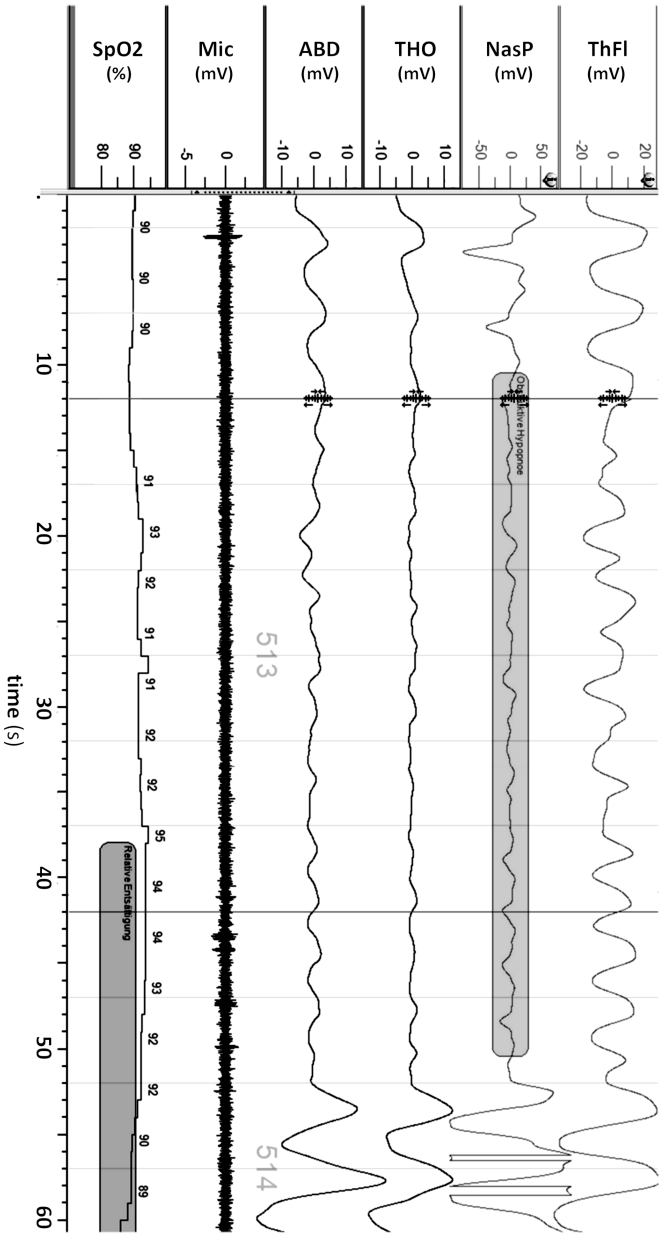

Figure 2.10: *Example somnogram of a hypopnoeic event. ThFl = thermal airflow sensor, showing a reduction, but not complete cessation of airflow. NasP = intranasal pressure. THO = thoracic movement. ABD = abdominal movement. Mic = neck contact microphone, measuring regular snoring noise during inspiration. SpO2 = blood oxygen saturation, showing time-delayed measurement of blood oxygen desaturation related to the hypopnoeic event.*

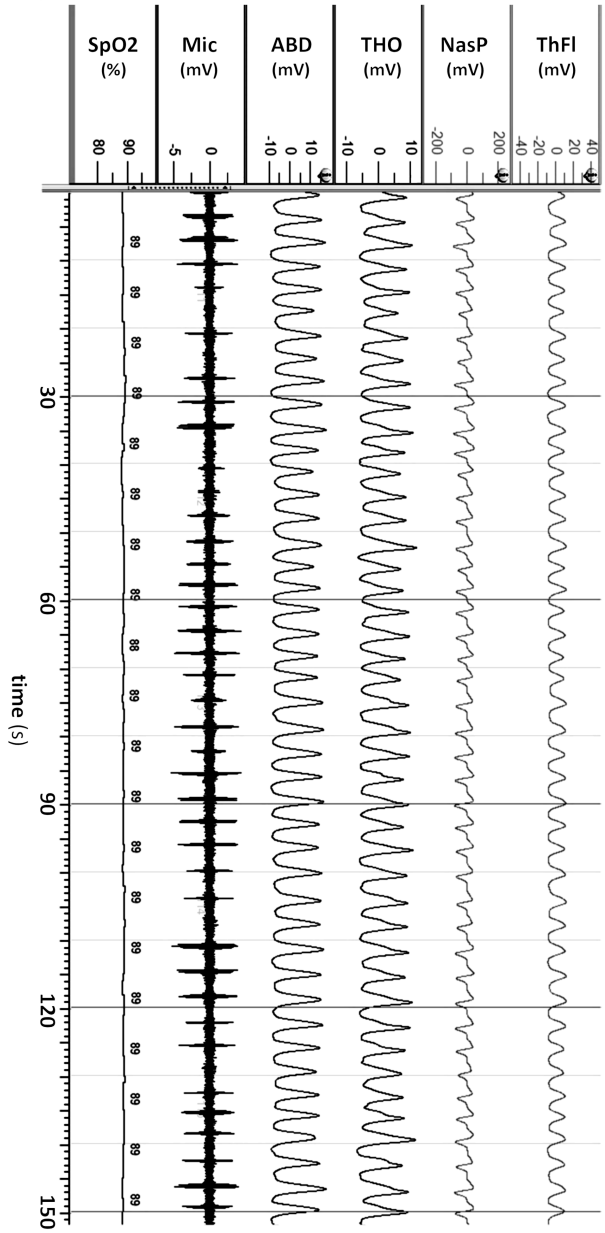

Figure 2.11: *Example somnogram of a primary snorer. ThFl = thermal airflow. NasP = intranasal pressure. THO = thoracic movement. ABD = abdominal movement, showing regular breathing. Mic = neck contact microphone, measuring regular snoring noise during inspiration. SpO2 showing normal blood oxygen saturation.*

In order to reduce the first night effect and to allow for sleep monitoring over longer periods of several nights, ambulatory cardiorespiratory screening allows for the recording of a reduced number of parameters with a portable recording device that can be used in the home environment. Usually, these systems allow recording of oxygen saturation levels, airflow, breathing effort, heart rate and body position. Since they do not record EEG data, their use is limited to the diagnosis of sleep related breathing disorders.

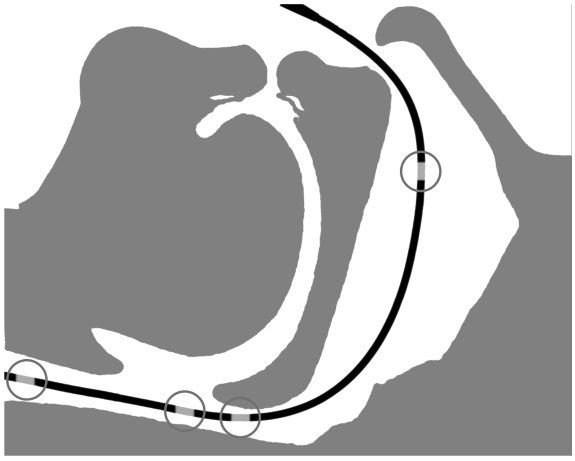

Figure 2.12: *Principle of multi-channel pressure measurement. A thin tube with multiple pressure sensors is introduced intranasally into the upper airways. Pressure differences between the sensors during inspiration provide information on the level of obstruction.*

Actigraphy

Introduced to clinical use in the 1990s, actigraphy plays an important role in the assessment of activity-rest-patterns. Usually, actigraphy measurements are carried out over a longer period of time covering several days and nights.

The core of an actigraphy recorder is a set of motion sensors measuring acceleration in different axes to detect gross motor activity, and a memory unit to record the motions over the duration of the examination. Actigraphy recorders are small devices usually worn on the wrist. In addition to motor activity, some devices additionally measure further parameters, such as light intensity, skin temperature, or noise volume level (74).

Actigraphy in combination with the measurement of tracheal respiratory noise using a contact microphone has been investigated for the monitoring of sleep structure (75). The authors found a moderate correlation of REM, non-REM, and

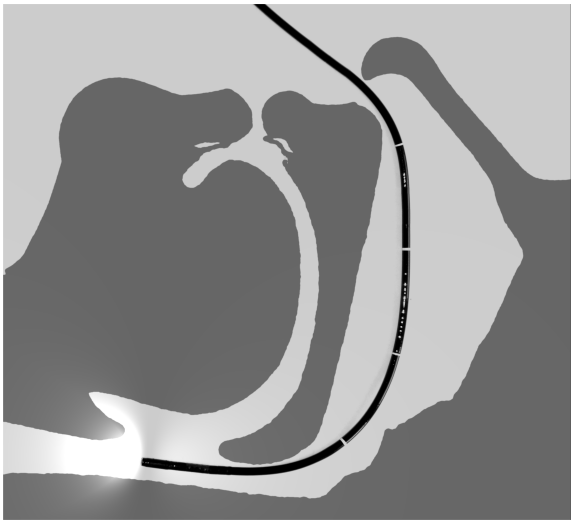

Figure 2.13: *Principle of a drug induced sleep endoscopy. A flexible nasopharyngoscope is introduced intranasally into the upper airways of the sedated patient. By moving the tip of the endoscope back and forth, different planes of the pharynx can be observed during sleep to identify the vibration and obstruction mechanisms. Video and audio signals are simultaneously recorded and stored for later evaluation.*

awake stages with polysomnographic sleep staging, suggesting that actigraphy, especially in combination with other parameters, can be useful for sleep screening.

Multi-Channel Pressure Measurement

Polysomnography and cardiorespiratory screening are valuable tools to identify the existence and severity of a sleep related breathing disorder. However, they do not reveal information about the actual mechanisms that lead to the breathing difficulties. Multi-channel pressure measurement helps to narrow down the levels within the upper airways in which constrictions or obstructions occur (76; 77; 78). A thin tube with multiple pressure sensors is introduced into the upper airways. The pattern of pressure changes during breathing of the different sensors allows for the determination of the obstruction location during an apnoeic or hypopnoeic event. This method can in principle be used in natural sleep. However, the tube within the upper airways is not tolerated by every patient.

Figure 2.12 shows the principle of multi-channel pressure measurement in the upper airways.

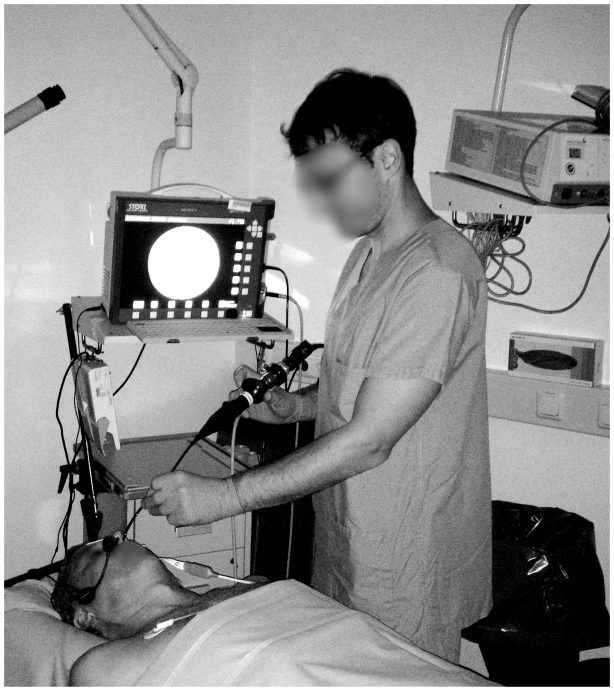

Figure 2.14: *Setting of a DISE procedure*

Drug Induced Sleep Endoscopy

Already in 1978, Borowiecki et. al. pioneered in video-endoscopy of the upper airways in patients during natural sleep, and the authors found that "the structures involved in production of airway obstruction in the patients with OSA syndrome are the muscles of velopharyngeal sphincter and tongue." (79).

The introduction of Drug Induced Sleep Endoscopy (DISE) by Croft and Pringle in 1991 was a milestone in more precise diagnosis of the underlying individual mechanics causing snoring and OSA. (80). DISE is increasingly used by surgical sleep specialists as a diagnostic tool in addition to PSG to identify the location of vibration and obstruction, especially in the targeted therapy planning for sleep disordered breathing patients (81).

For a DISE examination, a state of artificial sleep is induced using titrated doses of narcotic agents, such as propofol or midazolam. Being in a state of unconsciousness that resembles sleep, the upper airways are intranasally inspected using a flexible nasopharyngoscope. Sleep experts evaluate the video image for the location and mechanism of airway narrowing or obstruction. Snoring activity provides additional information for diagnosis and therapy planning (8).

DISE has a number of disadvantages. It is costly, requires the attendance of qualified medical personnel, availability of appropriate equipment for safe administration and monitoring of sedation, as well as sophisticated endoscopic equipment. Further, a DISE investigation often takes more than 20 minutes overall. Also, a subject would not remain asleep undisturbed through the introduction and movement of the endoscope during the procedure, so DISE cannot be performed in natural sleep.

Figure 2.13 illustrates the principle of a DISE examination, whilst Figure 2.14 shows a typical examination room setting during a DISE procedure.

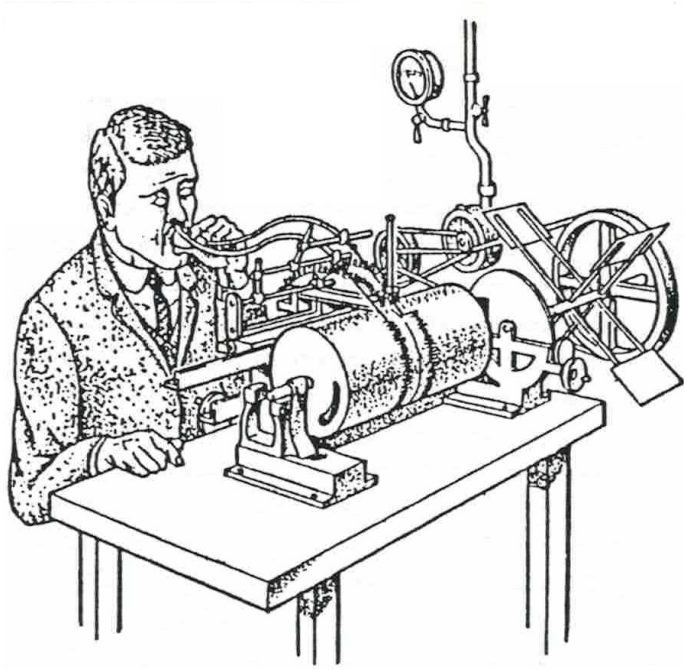

Figure 2.15: *The Kymograph, an early mechanical device for measuring pitch and intensity of the speech signal (source: (82)).*

Acoustic Analysis of Snoring

The acoustic properties of snoring have been analysed and investigated by researchers of several disciplines, including ENT medicine, acoustics engineering, audio signal processing, and last not least machine learning, with the aim of developing methods to complement or simplify the diagnosis of sleep disordered breathing (83). The acoustic snoring analysis is closely related to the analysis of speech, and methods deployed are largely based on those originally developed and used for the investigation of human speech.

An early example is the investigation of speech using a Kymograph, described by Rousselot at the beginning of the 20th century (84), see Figure 2.15. In this mechanical measuring device, speech is transmitted via a rubber tube to a membrane, which vibrates when excited by the sound waves. A pen attached to the drum transmits the vibrations to a sheet of lampblack, which is attached to a cylinder rotating at a constant speed. The drum acts as a low-pass filter which suppresses the higher frequency harmonics of the speech signal. Accordingly, no formants and thus no different phonemes could be distinguished from kymographic recordings; however, with a good magnifying glass, good eyes and a lot of patience, three prosodic parameters could be measured: Duration (total length of the recording), intensity (strength of the deflections of the curve) and course of fundamental frequency (number of oscillations within a period of time). Around the turn of the century and in the first decades of the 20th century, this method was regularly used by phoneticians conducting research.

A comprehensive overview on research done in the field of acoustic snoring noise analysis is given in the following Section 2.2.

2.2 Prior Work

2.2.1 Literature Research

Methodology

Research on the acoustic properties of snoring noise has been undertaken in the past decades pursuing different goals. This literature research is structured according to the following research objectives.

1. Differentiation between primary snoring and OSA or between mild, moderate and severe OSA.

   This information can be used to develop suitable screening systems to complement or partly replace polysomnographic examinations or ambulatory cardiorespiratory screening. A technology based purely on acoustic information can easily be used in the home environment and could even be implemented as a software application to be used in a smartphone without the need for additional, proprietary hardware. It is convenient for the subject under examination as no sensors are required to be attached to the body during the night, reducing the risk of measurement errors due to incorrectly fixed sensors. Further, it is easy to apply and does not disturb natural sleep.

2. Detection of respiratory events.

   In contrast to the first objective, in this case, the aim is the detection of the exact moment of occurrence of apnoeic or hypopnoeic events. This can provide an additional source of information for polysomnography systems or it can simplify ambulatory polygraphy devices. If apnoeic events can be reliably detected using acoustic parameters, the measurement of certain other physiological parameters may become redundant. On the other hand, existing measurement methods can be made more reliable and automatic analysis systems can be improved by adding the acoustically evaluated information.

3. Determination of the annoyance of snoring sounds using psychoacoustic parameters.

   The severity or annoyance of snoring is routinely assessed through rating by the bed partner using visual analogue scales, which is a purely subjective

measure. The aim of developing objective snoring severity determination methods is to make the acoustic impairment of the bed partner objectively measurable and to develop an unequivocal definition of snoring severity.

4. Classification of the location of sound generation.

   For a reliable surgical therapy planning it is crucial to know the mechanisms and locations of obstructions and vibrations in addition to the severity and type of the sleep disorder. Today, drug-induced sleep endoscopy is often performed in addition to polysomnographic examinations. The motivation for acoustic recognition of the snoring sound source is to obtain knowledge about the snoring mechanisms during natural sleep and thus to support sleep endoscopic examinations or make them obsolete in certain cases.

Database	Search string	# Results
Pubmed	(((snore[Title/Abstract]	69
	AND acoustic[Title/Abstract]))	
	OR (snore[Title/Abstract]	
	AND sound[Title/Abstract]))	
	OR (snore[Title/Abstract]	
	AND audio[Title/Abstract])	
Pubmed	((snore[Title/Abstract]	2
	AND annoyance[Title/Abstract]))	
	OR (snore[Title/Abstract]	
	AND psychoacoustic[Title/Abstract])	
IEEE Xplore	(((snore AND sound)	113
	OR snore AND acoustic)	
	OR snore AND audio)	
IEEE Xplore	((snore AND annoyance)	3
	OR snore AND psychoacoustic)	

Table 2.2: *Search strings and number of results of the literature research*

For each of the above objectives, a search was conducted in the literature databases Pubmed and IEEE Xplore, covering publications until June 2016. The search results were manually selected according to title and abstract, and those not dealing with one of the respective objectives were discarded. Table 2.2 lists the search strings used.

Endpoints and Performance Assessment

Some of the published research deals with the statistical description of differences of certain acoustic parameters between two cohorts, e.g. primary snoring and OSA-related snoring. Identification of meaningful differentiators is important in the quest for the underlying anatomical mechanisms that lead to varied forms of snoring or causes for OSA. Furthermore, it can be beneficial for the pre-selection of features for a machine learning task.

The statistical description of difference of the average across cohorts regarding a certain parameter allows no conclusion as towards the class-prediction performance based on this parameter. Using machine learning strategies, a classifier is trained to assign particular snoring events to one of several defined classes. In a binary classification task, the classification success can be assessed by sensitivity (proportion of instances correctly classified as positive to the total number of objects actually positive) and specificity (proportion of instances correctly classified as negative to the total number of objects actually negative), with the terms *positive* and *negative* describing membership of one of the two classes.

In classification tasks with more than two classes, the classification performance per class can be described by the class-specific recall, being defined as the number of correctly predicted samples out of the total number of samples of the respective class. A measure for the overall classifier performance is the unweighted average recall (UAR), which is the the average share of correctly assigned events over all classes, or the unweighted mean of the class-specific recall of all classes. In the weighted average recall (WAR), defined as the mean of class-specific recalls weighted by class sizes, the contribution of smaller classes is underweighted. Especially in an unbalanced dataset with strongly different numbers of samples between classes, the WAR is often higher compared to the UAR, as usually a classifier model performs better for the larger classes. The machine learning experiments in the following chapters are carried out based on a strongly unbalanced set of data. For that reason, performance will be assessed primarily based on the UAR.

Results

A total of 187 papers were found on Pubmed and IEEE Xplore. There were a total of 16 duplications, so that the abstracts of 171 papers were analysed with regard to the defined objectives. As a result of the analysis, 46 papers remained. This excludes publications by the author of this thesis, which are described separately in Section 2.2.2.

Of the 46 publications, the majority, namely 32 papers, deal with the relationship of snoring sound properties and OSA severity, or the distinction of OSA and primary snoring. Three papers aim to acoustically detect apnoeic events, while another five target annoyance measurement of snoring. Six papers were published on the distinction of snoring noise excitation locations.

Review

In the following, the researched literature is reviewed in detail structured according to the respective research goals.

1. Differentiation between primary snoring and OSA, classification of the OSA severity level.

   The aim of most of the identified publications on this topic, namely 27 papers out of the total of 32, is to group subjects into two classes above and below a specific AHI threshold. The class definition is mostly based on the classification of the OSA severity level according to the AASM into 'mild' (AHI 5...15), 'moderate' (AHI 15...30), and 'severe' (AHI>30) (64). Using this definition, an AHI threshold of five can be used to distinguish between primary snoring and OSA. Alternatively, a threshold of 15 can be applied to differentiate between the classes 'primary snoring or mild OSA' and 'moderate or severe OSA'.

   Methodically, this is a two-class classification task. The allocation of the respective subjects to one of the two classes is based on a number of different acoustic characteristics or a combination thereof. All publications describing machine learning methods use the principle of supervised learning. The labelling of the data, i. e. the a priori assignment to the correct class, is done based on an objective reference or ground truth. In all of the considered studies, the ground truth is derived from polysomnography data that has been recorded in parallel to the audio recording of the snore signals.

   The first publications with this objective date from the turn of the last century. The relationship between sound pressure level (SPL) of snoring sounds and the RDI has been investigated in a large cohort of 1139 subjects, indicating a significantly higher SPL in snorers with OSA compared to primary snorers (85). Sola-Soler et. al. in the year 2000 described significant differences of mean value, standard deviation and density of the pitch in snorers with OSA and primary snorers (86). In the same year, members of the same working group published for the first time the application of machine learning methods for OSA severity classification. Based on several acoustic parameters, a sensitivity of 82% was achieved differentiating between OSA and

primary snoring (87). Three years later, the same group first described the use of the spectral envelope as well as formant analysis for the classification of OSA and primary snoring and showed clear differences in the variability of formant frequencies of snorers with an AHI smaller or larger than ten (88). The use of pitch jitter analysis for the same purpose was already described in 2001 (89). Abeyratne et. al. reported a specificity of >90% in distinguishing OSA and primary snoring in 14 subjects.

In the following years several working groups dealt with the classification of OSA and non-OSA subjects on the basis of selected acoustic parameters. For example, sound intensity and spectral parameters (90), formants (91; 92), combinations of time- and frequency-based parameters (93; 94; 95), wavelets (96), spectral bandwidth (97), pitch contour (98), mean value of divergence curve parameters (99) were investigated. Nakano et. al. used the decrease of the average sound energy of snoring noises recorded by smartphones for RDI classification at different threshold values and achieved specificities between 70% and over 90% (100). Psychoacoustic parameters were also used for classification experiments (101). Ben-Israel et. al. demonstrated that a distinction between snorers with and without OSA is possible based on the stability of the fundamental frequency of snoring sounds recorded during polysomnography (102). Pitch discontinuities (jumps in the fundamental frequency within a snoring event) also indicate OSA snoring (103; 104).

However, the amount of independent data used (i. e. the number of subjects) is quite small in these publications. They range from eight subjects (95) to 41 subjects (94). With a small number of different snoring individuals, care should be taken to generalise the results. Also, the conclusions of some studies are contradictory in comparison: Based on the first formant (F1) as classification characteristic, Ng et. al. found that the frequency of F1 is higher in severe OSA than in primary snoring (sensitivity 88%, specificity 82%), whereas for the snore formants F2 and F3 no significant correlation with the severity of OSA could be found (92). In contrast, Sola-Soler et. al. showed a significantly lower frequency of F1 in patients with AHI>10 (105) in the same year based on a different database of 24 subjects.

With the technological evolution of machine classifiers and the availability of more extensive databases, remarkable results have been achieved in later years. In some studies the methodology was extended to a three-class problem. Herath et. al. report a sensitivity and specificity between 87% and 91% in the classification of subjects into the classes AHI<15, 15<AHI<30 and AHI>=30 using Mel Frequency Cepstral Coefficients (MFCC) and Hidden Markov Model-based classifiers (106). Ben-Israel et. al. used a combination of MFCC and parameters describing the characteristics of snoring episodes

(running variance, apnoea phase ratio, inter-event silence) and achieved a hit rate of >80% for AHI thresholds of 10 and 20 (107). Fiz et. al. investigated sound intensity, number of snoring events, standard deviation of the spectrum, sound power ratio in different frequency bands and symmetry coefficients and achieved sensitivities and specificities between 71% and 90% at AHI thresholds of five and 15 (108).

Some groups supplemented the acoustic features with other, non-acoustic features. For example, De Silva et. al. relied on multifeature vectors and logistic regression, using gender and neck circumference as additional non-acoustic parameters to distinguish between OSA and primary snoring. They achieved a sensitivity of more than 90% (109; 110). Abeyrathne et. al. used the neck circumference as an additional physiological feature in combination with the analysis of tonal components in the snoring signal, achieving a specificity of more than 93% for the classification of an AHI greater or smaller than 15 (111). It is not reported, however, to which extent the acoustic and non-acoustic parameters contributed to the classification results. It is known that neck circumference by itself is a reliable predictor for the occurrence of OSA (112; 113) therefore, it might well be the main contributing feature in these experiments.

Although the results of the aforementioned studies are not directly comparable with each other, they show promising results applying machine learning methods for this task.

2. Identification of obstructive events based on the acoustic properties of snoring noise

   Like in the studies described in the previous section, the desired result of the following work is the determination of the AHI. The approach, however, is different. While the work described in the previous section aims to determine an AHI severity level from the sound recordings during sleep over a certain period of time (usually a whole night), the following publications deal with the detection of the actual individual obstructive events on the basis of acoustic data. The objective ground truth in these studies is polysomnographic data that was simultaneously recorded.

   In an early work on this topic from 1993, Perez-Padilla et. al. showed, based on data from 10 volunteers, that the first post-apnoeic snoring event has a higher proportion of sound energy >800 Hz compared to pre-apnoeic events (114). Karci et. al. confirmed this finding 18 years later (95). However, Yang et. al. showed different results in 2012. According to their work, the maximum frequency, the peak frequency, the mean frequency and the central frequency

after an apnoeic event were higher, but the sound energy >800 Hz was lower than the average (115).

In principle, solely from the acoustic properties of individual snoring events, obstructive events can be identified only to a limited extent. The combination of these parameters with features describing the characteristics of a snoring episode could be more promising for this task. However, this approach would also bear a degree of uncertainty. Obstructive events that take place quietly, i. e. without accompanying snoring, would remain undiscovered by acoustic detection alone. This problem can be mitigated by considering both snoring and breathing noise. In addition, oxygen saturation, airflow, or other selected polysomnographic parameters could be included in the analysis as non-acoustic parameters (116).

3. Evaluation of snoring sound annoyance

   A standardised procedure for the objective evaluation of the disturbance of the bed partner by snoring noises does not exist to date. The snoring strength estimation, i. e. the subjective assessment by the bed partner, is usually done by means of questionnaires and visual analogue scales (VAS). This bears the risk that, in addition to the acoustic aspects, the non-acoustic quality of a partnership will also be assessed, which may influence and skew the results. For these reasons, there is a scientific interest in making the annoyance of the actual snoring sound objectifiable.

   The first work found on this topic was published in 1994 (117). Hoffstein et. al. compared the subjective snoring perception of the snorers themselves and of their bed partners and concluded that the perception is very different. Objectification based on sound volume parameters was not successful.

   In 2007, an interdisciplinary team of physicians and acoustic environmental engineers investigated sound pressure level and volume characteristics of snorers using parameters used for noise exposure assessments, namely rating level, maximum level, two percentile levels for frequent maxima, and snoring time (118). Based on the snoring noises of 19 test persons, they created a snoring score. In this context, the authors were able to show limits defined in the World Health Organisation (WHO) noise guidelines for sleeping environments were in some cases substantially exceeded.

   A few years later, Rohrmeier et. al. used psychoacoustic parameters to define an objectifiable measure of the annoyance of snoring noises (119). They found a clear correlation between subjectively perceived annoyance, the A-weighted sound pressure level, the 5th percentile of psychoacoustic loudness and the perceived annoyance, a combination of the psychoacoustic parameters loudness, sharpness, fluctuation and roughness. This study will be

described in more detail in Section 3.1. Fischer et. al. came to a similar conclusion, showing a clear correlation between subjective annoyance rated by the bed partners by means of VAS and a score combined of 5th percentile of loudness and average roughness (120).

4. Classification of the location of sound generation

   In contrast to the detection of obstructive events, the aim in the following studies is to find out where snoring noise originates in the upper respiratory tract. Some of these studies are based on the assumption that primary snoring events are originating in the velopharynx, whereas in OSA-related snoring, the tongue or the hypopharynx are involved. Based on this assumption, a lower fundamental frequency was found in velum snorers than in lingual snorers on average (121). Conclusions from these results must be drawn with caution, provided that the assumption has not been objectively verified in the patient population used.

   Using an objective ground truth, Quinn et. al. in 1996 found different waveform patterns in velum snorers and non-velum snorers in a group of eleven subjects who were examined by sleep endoscopy (122). Hill et. al. (123) describe the measurement of crest factors in snoring signals from eleven test subjects recorded during DISE. The excitation level of the snoring sounds was visually observed and classified according to the levels soft palate, hypopharynx and epiglottis. In the group of soft palate snorers, the crest factors were significantly higher compared to the groups of non-soft-palate snorers. The authors conclude that the type of snoring may change during the night and that sleep video endoscopy may not be representative of snoring behaviour during natural sleep (124). Because of the small number of included subjects (five snorers), however, these results are not unrestrictedly generalisable.

   Beeton et. al. showed distinct differences in the statistical properties of the amplitude pattern between pure velar snoring and other forms of snoring based on sleep endoscopic investigations (125).

   In a sleep endoscopic study of 16 test subjects, Agrawal et. al. found a direct correlation between acoustic properties and the location of sound generation. Statistical analysis revealed that the mean peak frequency, defined as the frequency of maximum sound energy when averaged over the duration of the snore sample, is below 200 Hz in velar snoring, around 500 Hz in epiglottic snoring and above 1 000 Hz in tongue snoring (126). Notably, these results could not be reproduced in the natural sleep of the same volunteers. Due to the small number of test subjects (only two tongue snorers), further conclusions should be drawn with caution.

In a recent study, Herzog et. al. investigated the correlation of the snoring sound source with psychoacoustic parameters. Determination of the location of the excitation location was based on DISE examinations. They found a higher loudness in obstructive snoring generated at hypopharyngeal level, a higher roughness in velar snoring and an increased sharpness in snoring initiated at the palatine tonsil level (127).

These studies show that selected acoustic parameters differ considerably by type and location of snoring noise generation. However, results of the above publications should be regarded with caution due to the small sample sizes.

2.2.2 Prior Publications by the Author

The author of this thesis has authored and co-authored several publications on the use of machine learning methods for the classification of snoring sounds according to their excitation locations in the upper airways, which are summarised in the following. Also included in the summary are results that have been published by other groups on machine learning experiments based on a snoring sound dataset developed by the author, namely the Munich-Passau Snore Sound Corpus (MPSSC) in the framework of the INTERSPEECH 2017 Computational Paralinguistics ChallengE (ComParE). The MPSSC is described in detail in Chapter 3.2.

On a small pilot snoring sound dataset consisting of 24 subjects, Schmitt et. al. used a Bag-of-Audio-Words approach (BoAW) to train a linear Support Vector Machine (SVM) (12) to classify snoring sounds by their excitation location. Instead of directly training the classifier with acoustic features representing an audio signal, the frequency of occurrence is counted for each representative feature vector prototype (called audio words). In order to find the corresponding audio word for a current feature vector, the one with the smallest distance is selected. Subsequently, the feature space is divided from the training data by unsupervised learning so that the audio words well represent the distribution of the data. Applying the BoAW approach to wavelet features, formants, and MFCC, a UAR of almost 80% could be achieved.

Using snoring sound data from 40 subjects, Qian et. al. evaluated the performance of different feature sets to train different kinds of machine classifiers for excitation location classification, finding that spectral features, such as MFCCs, subband energy rations, and wavelet-based energy features, outperform features that describe the temporal properties of the underlying snore signal (7). In a diligent performance analysis of different combinations of acoustic feature sets and machine classifiers on snoring sound data from 40 subjects, Qian et. al. found the

best performance with a UAR of more than 70% was achieved with a Deep Neural Network (DNN) trained with a feature set of 1 kHz subband energy ratios. However, performance of subset-classifier-combinations varies widely among the different subset permutations, which can be attributable to the relatively small corpus size used (5).

The MPSSC, a snoring sound dataset comprising labelled snoring sounds from 219 subjects, was first introduced as the *Snore* Sub-Challenge in the INTERSPEECH 2017 Computational Paralinguistics ChallengE (ComParE) (11). In the context of the Challenge, baseline experiments were carried out using the official INTERSPEECH ComParE baseline feature set, which includes low-level descriptors (LLD) related to energy, spectral features, Mel Frequency Cepstral Coefficients (MFCC), voicing, Harmonic-to-Noise Ratio (HNR), spectral harmonicity, temporal course of the Fundamental Frequency F0 (Pitch), and Microprosodic Features (Jitter and Shimmer). In addition to these LLDs, their first order derivatives, or Deltas, are computed. In a second step, statistics of the LLDs, the so-called Functionals, are obtained. They comprise statistical moments of different orders, percentiles and extrema. An exhaustive list and description of the ComParE feature set is found in (128) and (129). In addition, a BoAW approach as well as an end-to-end learning model was employed. The highest UAR of 58.5% could be achieved using the ComParE functionals in combination with a Support Vector Machine. In comparison, the end-to-end learning and BoAW models have yielded inferior results.

Seven contributions on classification experiments with the MPSSC were accepted in the context of the INTERSPEECH 2017 ComParE Snore Sub-Challenge.

Tavarez et. al. (130) used i-vector representations of MFCCs, Constant Q Cepstral Coefficients (CQCC) and Relative Phase Shift (RPS) features obtained at frame level combined with the music-related pitch class profiles, tonal centroid and spectral contrast features as well as suprasegmental spectral statistics, voice quality and prosodic features to train a cosine distance classifier on the MPSSC audio data. Late fusion of the MFCC and RPS feature sets obtained the best classification performance with a UAR of 54.3% on the development set and 50.6% on the test set, respectively.

Nwe et. al. (131) approached the snoring sound classification task by fusing the results of three sub-systems by majority voting. The first subsystem consists of a Bhattacharyya-based Gaussian Mixture Model (GMM) supervector in an SVM classifier, using the ComParE baseline set as input features. In the second subsystem, the ComParE baseline feature set is reduced to a subset of 53 out of the originally 6 373 features by a correlation feature selection step, subsequently training a random forest classifier. Thirdly, a Convolutional Neural Network (CNN) is

trained based on the log power spectrogram of the snoring sound. Fusion of the three models achieved a UAR on the test set of 51.7%, while the Bhattacharyya-GMM-SVM subsystem reached a UAR of 52.4%.

A dual source-filter model simulating the acoustic transfer function of the airways was applied by Rao et. al. for feature extraction (132). The model consists of two all-pole filters resembling the acoustic properties of two consecutive tubes. The first tube ranges from the lungs to the obstruction location in the upper airways, whereas the second one models the upper airways from the obstruction location to the lips. The first filter is excited by white noise at lung level, while the second one is excited by periodic impulses at the obstruction level resembling snoring. Parameters of the two filters are estimated in a multi-step process comprising detection of the snore beat cycle impulse location, construction of two windows to attenuate the effect of source and filter, and estimation of the filter coefficients from the windowed signal. The resulting feature set consists of the filter coefficients and their respective framewise means, variances and medians and is used to train SVM classifiers with linear and Radial Basis Function (RBF) kernel, respectively, achieving a UAR of up to 52.8% on the test partition. Interestingly, comparison of the confusion matrices reveals that the classification error between velum (V) and epiglottis (E) class samples is reduced compared to the ComParE baseline approach. Anatomically, soft palate and epiglottis, representing the excitation locations in the upper airways for these two classes, are farthest apart, which might result in distinctly different filter coefficient estimates. On the other hand, velum and oropharynx class instances are misclassified more often using the source-filter model approach, which can be explained by the close proximity of velum and oropharyngeal area, resulting in rather similar filter coefficient estimates.

Gosztolya et. al. (133) extracted features at frame level using a feature set first proposed in the INTERSPEECH 2013 ComParE challenge (128), consisting of 39-dimensional MFCCs, voicing probability, harmonics-to-noise-ratio, F0 and zero-crossing rate and their respective first and seconnd derivatives, as well as mean and standard deviation over nine neighbouring frames. Further, each instance is divided into ten equal-sized segments, and each of the above features is averaged out in each of the segments. An SVM model was trained with this feature set and the results eventually fused with those of a second SVM classifier trained on the original INTERSPEECH 2017 ComParE baseline feature set, achieving a UAR of 64.0% on the test set.

Kaya et. al. (134) particularly approached the unbalanced nature of the corpus by proposing a weighting scheme for kernel classifiers. The audio signal is represented by MFCCs and a RASTA Perceptual Linear Prediction (PLP) cepstrum, complemented by the first and second order derivatives, resulting in relatively small feature sets with a dimension of 75 and 39, respectively. Both feature sets are then

fused and represented in a Fisher vector for the classification task. In parallel, the original baseline openSMILE feature set is applied. For classification, an Extreme Learning Machine (ELM) and a Partial Least Squares (PLS) classifier with linear kernels are used. Using a weight matrix counter-balancing the under-represented classes, a 'Weighted Kernel Extreme Learning Machine' and a 'Weighted Kernel Partial Least Squares' classifier are introduced and their performance compared to the unweighted models by applying a two-fold cross-validation of the training and development partition. The weighted classification models clearly outperformed their unweighted counterparts in three of four combinations of feature sets and folds. Notably, distinct differences in performance could be observed between the two folds. Fusing the best four combinations of feature sets and classifiers, a UAR of 64.2% on the test set could be achieved.

The following contributions did not officially participate in the challenge, since some of the co-authors were part of the challenge organisers.

Amiriparian et. al. (135) generated feature vectors using deep image CNNs trained with spectrogram plots of the snoring audio data. The feature vectors, with a dimension of 4 096 features, were extracted from the first and second fully connected neuronal layers, respectively, and used to train linear kernel SVMs, achieving a UAR of 67.0% on the test set. Notably, the choice of colour map for the spectrogram plots had an essential impact on the classification performance. Further, best results were achieved extracting the features from the second fully connected neuronal layer of the 'AlexNet' CNN. Fusion of different colour maps and layers did not yield an improvement in classification performance in this model.

Freitag et. al. (136) used the same setup of spectrogram-fed CNNs and combined it with an evolutionary feature selection algorithm based on competitive swarm optimisation, which was trained using a wrapper algorithm with a linear SVM. Results show that the UAR increases during the feature selection process until the feature subset reaches a size of about 65% of the original feature set. Little improvement in UAR is achieved when the number of selected features is reduced further. With this approach, a UAR of 57.6% on the development set and 66.5% on the test set could be achieved, using a feature subset containing 55% of the features from the original deep spectrum feature set.

Unrelated to the INTERSPEECH 2017 ComParE challenge, Qian et. al. performed a series of experiments on a predecessor of the MPSSC dataset comprising snoring sounds from 24 subjects (13). Using different wavelet and wavelet-packet feature sets to train an SVM, a UAR of up to 71.2% could be achieved. Epiglottis snorers were most reliably distinguished from all other snoring types, while velum and oropharynx snore events were most frequently confused. Further, combining wavelet-based LLDs with a Bag-of-Audio-Words approach to train a Naïve Bayes classifier on the MPSSC dataset yielded a UAR of 69.4% (3).

Zhang et. al. used the MPSSC to train a semi-supervised conditional Generative Adversarial Network (scGAN) in an ensemble setting in order to learn a mapping strategy starting from random noise, achieving a UAR of up to 67.4% (1).

Finally, Section 3.3 summarises the machine learning results of an extended MPSSC corpus that has been classified using two different classification schemes.

2.3 Acoustic Information in Snoring Noise

A variety of acoustic features has been evaluated for their suitability and performance in measuring information in snoring noise. Most of these features are derived from automated speech analysis and speech recognition applications. This chapter gives an overview of the features that are relevant for snoring analysis.

2.3.1 Periodicity

Instead of analysing the actual acoustic properties of a single snoring event itself, the temporal relationship of successive snoring events within a snore episode can provide valuable information on the type of snoring. Primary snoring is considered to be more regular than than snoring in subjects with OSA. Herzog et. al. used periodicity as a measure to distinguish between snoring episodes with rhythmic and non-rhythmic character, finding that non-rhythmic snoring is correlated with a higher AHI (137). Jones et. al. quantified periodicity, defined as the ratio of the number of periodic events to the total number of events within a snoring episode. They found a statistically significant decrease in periodicity after palatal surgery in patients with primary snoring or mild OSA (138). Furthermore, the periodicity can be used for the distinction of snoring from non-snoring acoustic events. Figure 2.16 and Figure 2.17 show examples of snoring episodes with high and low periodicity.

2.3.2 Zero Crossing Rate

The zero crossing rate is a feature that is used frequently to support the identification of voiced and unvoiced phonemes in automatic speech recognition. In snoring analysis, it is routinely deployed as part of larger feature sets. It is defined as the number of sign conversions, or crossing of the zero line, of the original temporal signal per second:

$$ZCR = \frac{n_{zc}}{l}, \tag{2.1}$$

with ZCR being the zero crossing rate, n_{zc} the total number of sign conversions of the sample, and l the total length of the sample in seconds.

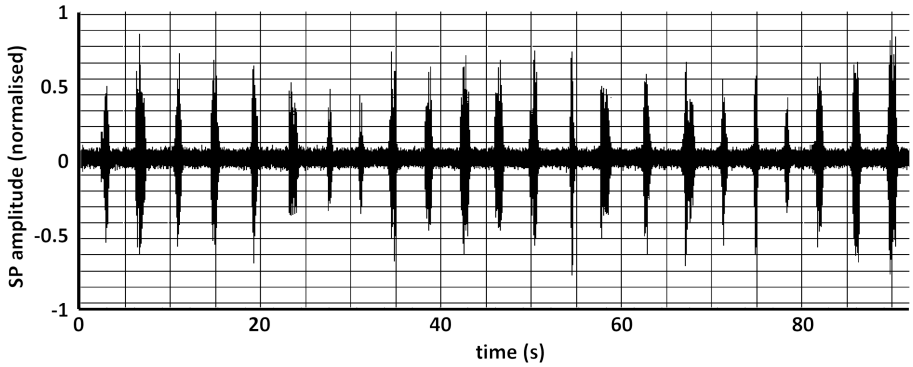

Figure 2.16: *Snoring episode of a primary snorer.*
All snoring events occur in regular intervals resulting in a high periodicity.

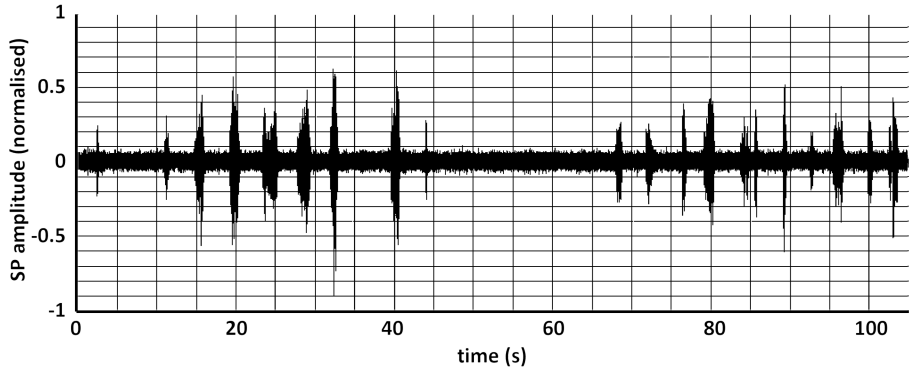

Figure 2.17: *Snoring episode of a snorer with OSA.*
The snoring episode is signified by several phases of irregularity, resulting in a lower periodicity compared to a primary snoring episode.

2.3.3 Fundamental Frequency and Harmonics

A sound consists of a fundamental frequency and a series of harmonics. The fundamental frequency determines the perceived pitch, while the structure of the harmonics determines the timbre. For example, the chamber note *A*, played on a piano and a guitar, has the same fundamental frequency, but a different harmonic spectrum which characterises the sound of the instrument. In contrast, noise contains no tonal components and therefore has no distinct harmonic structure (139).

Figure 2.18 shows a typical snore signal (a) and white noise (b) in the time domain, as well as their respective spectra (c) and (d).

Several studies conclude that primary snoring contains a more pronounced tonal component compared to OSA-related snoring. These investigations are based on the assumption that primary snoring is mainly generated on the velopharyngeal level, while snoring in OSA patients is predominantly excited in the retrolingual or hypopharyngeal level.

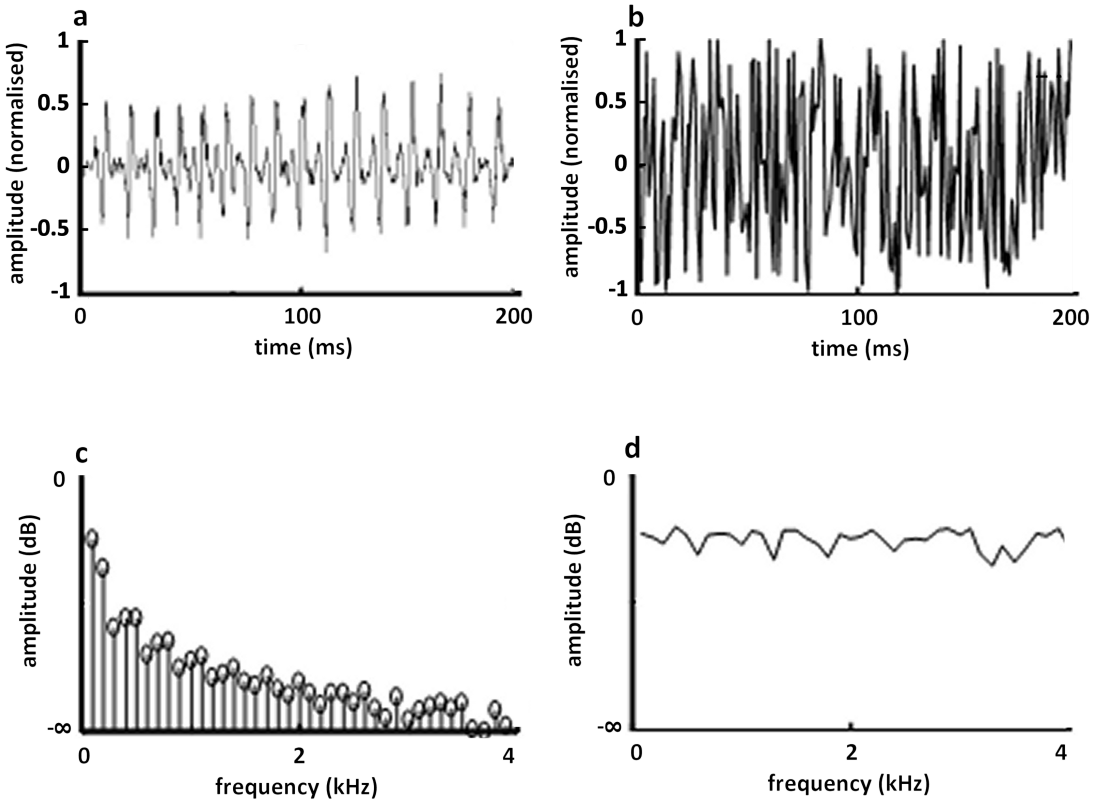

Figure 2.18: *Snoring signal and white noise*
Snoring signal in time domain (a) and spectrum (c). White noise in time domain (b) and spectrum (d).

2.3.4 Crest Factor

The crest factor is used to quantify the tonal fraction of a snoring sound. It describes the ratio of peak value to effective value (Root Mean Square, RMS value) of a signal amplitude.

$$C = \frac{|V_{peak}|}{V_{rms}}, \tag{2.2}$$

where C is the crest factor, V_{peak} is the peak value of the amplitude and V_{rms} is the root mean square of the amplitude in the considered signal period. Sometimes, the highest (above the 90^{th} percentile) and lowest (below 10^{th} percentile) components of the amplitude values are ignored in order to reduce the impact of random artefacts.

The crest factor is mainly used in telecommunications and audio engineering, for example to assess transmission quality. Generally, high crest factors signify a high fraction of harmonics, but can also be caused by by strong impulses in a signal (140). Crest factors have been applied to distinguish between different types of snoring. Using audio recordings from DISE investigations and calculating the crest factor for single snoring events, a significantly higher crest factor was found in palatal snoring compared to non-palatal snoring (123). In natural sleep, the crest factor of snoring noise can vary over time, which might be an indication that the anatomical characteristics of snoring change in the course of the night (124).

2.3.5 Spectral Energy Ratio

The Spectral Energy Ratio (SER) describes the ratio of energy content in snoring noise below and above a certain cut-off frequency.

$$SER_{f_c} = \frac{E_{0\ldots f_c}}{E_{f_c\ldots f_s}}, \tag{2.3}$$

where SER is the energy ratio, E is the spectral energy content, f_c is the cut-off frequency, and f_s is the sampling frequency.

The exact value of the cut-off frequency in snoring analysis applications varies in literature, but is usually in the order of 1 kHz. Alternatively, it is possible to determine the frequency above and below which the energy content of the sound signal is equal.

The SER can be used to differentiate between primary snoring and OSA-related snoring. Several studies conclude that primary snoring contains higher low-frequency energy components than snoring in OSA patients (114; 137; 141).

These investigations are based on the assumption that primary snoring is mainly caused by vibrations of the velum. The soft palate vibrates with a fundamental frequency well below 1 kHz and the harmonic's energy decreases with increasing frequency. Consequently, the signal's main energy content is in the low-frequency range. OSA-related snoring, in contrast, is supposed to occur at tongue base level, generating a higher portion of non-periodic, noise-like signals caused by turbulent air flow in the constricted areas. Hence, the energy spectrum of the resulting noisy signal contains a larger ratio of high-frequency energy components.

However, the results are based on statistical evaluation of the relationship between spectral energy ratio and AHI, assuming that a high AHI is related to tongue base snoring. A direct distinction of velar and non-velar snoring using the SER has not yet been demonstrated.

2.3.6 Pitch

In speech signal analysis, pitch refers to the course of the fundamental frequency of speech over time (142). In other words, pitch describes the speech melody. Accordingly, pitch can be used to describe the course of fundamental frequency over the duration of a snoring event, see Figure 2.19.

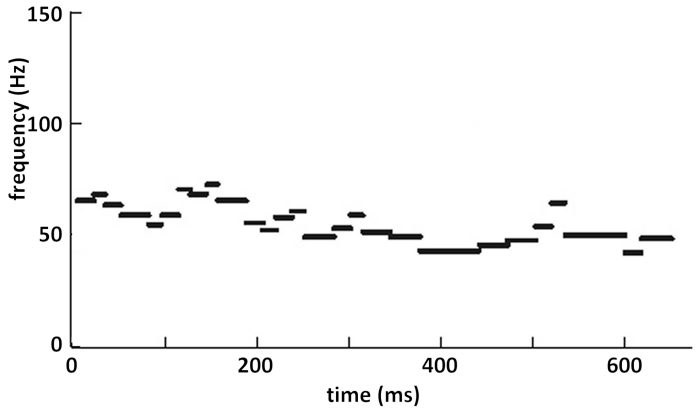

Figure 2.19: *Pitch diagram of a velum snoring event.*

Pitch analysis has been used for snoring analysis by several working groups, and differences in the stability of the fundamental frequency of snoring sounds as well as pitch discontinuities have been described to be suitable for the distinction of primary and OSA-related snoring (102; 103). Another measure used for snoring sound classification is pitch density, which is defined as the fraction of time containing a tonal component of the total duration of the snoring event. It allows

quantification of the observation that velum snoring events have a clearer tonal structure than tongue base snoring events (93).

2.3.7 Jitter and Shimmer

The prosodic features jitter and shimmer are regularly used in speech analysis and paralinguistic applications, and they can also be applied to analyse snoring sounds. Jitter describes the deviation from periodicity of the excitation signal, or (in spectral representation) periodic variations in the fundamental frequency, while shimmer describes periodic variations of the excitation signal's amplitude.

Jitter and shimmer are usually applied for snoring sound analysis as part of larger feature sets. Their isolated performance for the distinction of primary snorers and OSA patients has only been tested in small pilot settings with moderate results (89).

2.3.8 Sound Pressure

The sound pressure is the momentary deviation of pressure from the static local atmospheric pressure. For a point source, the sound pressure decreases inverse-proportionally to the distance from the sound source.

$$p(r) \propto \frac{1}{r}, \tag{2.4}$$

where p is the sound pressure, and r is the distance from the sound source.

The Sound Pressure Level (SPL) is the logarithmised sound pressure relative to a reference sound pressure p_0 of $20\,\mu Pa$.

$$L_p = 20\log_{10}(\frac{p}{p_0}), \tag{2.5}$$

where L_p is the RMS sound pressure level measured in Decibel (dB), and p_0 is the reference sound pressure.

The SPL has been used in numerous studies to quantify snoring. It should be noted, however, that the measurement of the absolute SPL in snorers can be very error-prone, as the distance between recording microphone and mouth and nose of the snorer has a considerable influence on the measured value. Additional potential error factors are the room acoustics (reverberation of the room) and the sleeping position. Especially in an environment that is not completely controllable over the course of the recording process, analysing absolute SPL values involves a high risk of uncertainty.

2.3.9 Source Filter Model and Formants

In voice production, the oscillating vocal folds in the glottis excite a glottis sound characterised by a fundamental frequency and a certain harmonic spectrum. On its way from the glottis to the openings of mouth and nose, the signal travels through the upper respiratory tract, consisting of larynx, pharynx, mouth, lips, paranasal sinuses and tongue. The frequency spectrum of the resulting signal exiting mouth and nose is shaped by a transfer function largely depending on the geometric properties (length and course of cross-section) of the upper respiratory tract from the location of sound excitation to the sound outlets. This effect is well known from speech analysis as Fant's source-filter model (143).

In an early publication from 1986, the source-filter model is applied to differentiate between different sound generation sites within the respiratory tract. The calculations are based on the acoustic data of one single subject with laryngomalacia. Differences in the frequency spectrum of the snoring sounds could be detected in prone and supine position (144).

In speech analysis, the frequencies at which the respiratory tract's transfer function shows maxima are called *formants* and are referred to in ascending order of their frequencies as F1, F2, F3 and F4. The frequencies and steepness of the formants are essentially influenced by the horizontal and vertical position of the tongue, the position of the soft palate and the position of the lips. The formants are independent of the stimulating signal. For example, they remain constant with changing pitch.

Similarly, snoring can be described acoustically by a system of sound excitation source and extension tube with a frequency-dependent transfer function. In contrast to vocal production, however, sound is not produced at a fixed location by the vocal folds but can be generated at different locations of the upper respiratory tract, such as soft palate, oropharynx, tongue base, epiglottis, or laryngeal structures, depending on the type of snoring. It is reasonable to assume that different types of snoring have a characteristic transfer function with resonance frequencies in the resulting spectrum determined by the typical positions of the tongue and soft palate as well as the different overall length of the acoustically active part of the upper respiratory tract, which could be referred to as *snoring formants*. See Figure 2.20 for an example of typical velum-snoring formants.

Formant properties can be calculated by linear prediction (LP). An all-pole filter is used to approximate the spectral properties of the upper respiratory tract. The filter parameters can be determined using Yule-Walker autoregression combined with Levinson-Durbin recursion (145). The formant frequencies correspond

with the poles of the absolute value of the complex roots of the filter model $H_{(Z)}$

$$H_{(Z)} = \frac{1}{1 - \sum_{i=1}^{n} \alpha_i z^{-i}}, \tag{2.6}$$

with α_i $(i = 1, 2, :., n)$ being the LP parameters.

While contact microphones attached to the neck at larynx level are routinely used in polysomnographic examinations to detect snoring, an airborne sound microphone is required to measure snoring formants, since the spectrum shaped by the upper respiratory tract becomes audible only when the sound exits the nose and mouth.

Snoring formant characteristics for the classification of OSA severity have been investigated in several studies with equivocal results (92; 146; 105). In a number of snoring sound machine learning experiments performed with the participation of the author of this thesis, feature sets based on formant characteristics yielded moderate classification results.

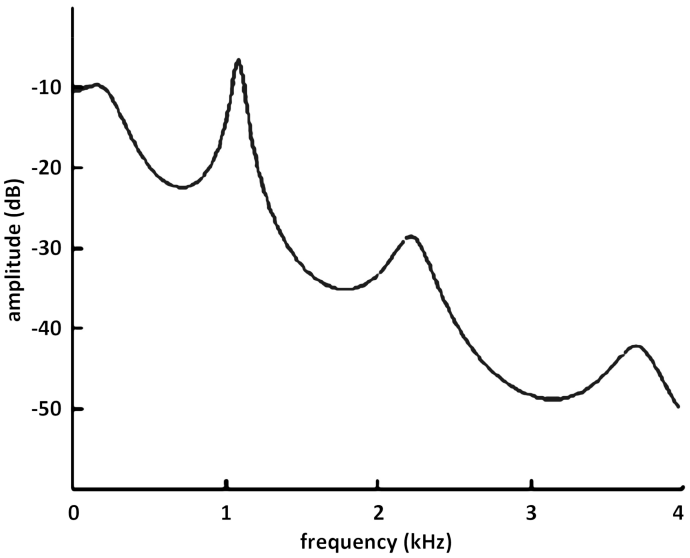

Figure 2.20: *Snoring formants of a velum snoring event, calculated by linear prediction (LP)*

2.3.10 Mel Frequency Cepstral Coefficients

Mel Frequency Cepstrum Coefficients (MFCCs) are a compact representation of a signal's spectral properties, and they are regularly used in speech recognition

applications because they allow a separation of the properties of excitation signal and transfer function.

MFCCs are calculated from the power spectrum of a discretely fourier transformed short-term segment of the original signal, usually with a length of 20...100 ms. The spectral coefficients of the fourier-transform are mapped in frequency blocks on a Mel scale by overlapping bandpass filtering. For speech analysis, usually a group of 12 to up to 20 frequency blocks are calculated, from which the coefficients are derived. In the experiments described in the following chapters, 14 coefficients were used. The Mel scale is a psychoacoustic representation of the frequency, the unit Mel describes the *tonality*, i. e. the perceived pitch of sine tones. For frequencies above approximately 500 Hz, equal frequency intervals are perceived as increasingly smaller pitch increments by the human auditory system (147). Consequently, the width of the Mel frequency blocks is increasing with the frequency. In a next step, the logarithmised energy of the Mel frequency blocks is discretely cosine transformed, resulting in a *cepstral* representation of the signal. Eventually, the magnitudes of the resulting values are the MFCCs.

In sleep laboratory studies of test subjects, the stability of MFCCs over the entire night was used to distinguish between primary and OSA snorers (102). MFCCs were more stable over the course of the night in primary snorers than in subjects with OSA. It is assumed that OSA is associated with an unstable muscular balance in the upper respiratory tract, which is acoustically expressed in a variability of its transfer function over time.

2.3.11 Perceptual Linear Prediction

PLP is a psychoacoustically adjusted approximation of the vocal tract's spectral transfer function. As in LP, an all-pole filter model is calculated in an order that is assumed to match the number of resonances of the transfer function in the frequency range considered. With the resonance frequencies corresponding to the formants of the articulation tract, LP represents the transfer function while to a considerable extent levelling out details of the excitation signal's harmonic structure. For PLP, the power spectrum is warped to a psychoacoustically adjusted scale (Mel-scale or Bark scale), then convoluted with the power spectrum of a critical-band masking curve, then pre-emphasised by an equal-loudness curve approximating the sensitivity of the human auditory system, and finally compressed in order to approximate the power law of hearing, before the all-pole filter model is calculated (148). Instead of using linear prediction, a cepstral transform can be performed, resulting in a set of Perceptual Linear Predictive Cepstral Coefficients (PLPCC).

2.3.12 Relative Spectral Transform Perceptual Linear Prediction

Relative spectral transform (RASTA) adds a step of linear band pass filtering to each energy band after the spectral transform, in order to reduce the impact of slowly varying and constant signal components, again simulating human auditory speech perception. RASTA filtering especially provides superior results over PLP for signals overlaid with static background noise (149).

RASTA-based analysis results of snoring sounds have not been published so far. In the experiments described in this thesis, RASTA features have been used as part of a larger feature set. Analysis of the performance of different feature subsets did not show superior results in the classification of isolated, single snoring events.

2.3.13 Wavelets

Short-Time Fourier Transform (STFT) generally results in a trade-off between length of the respective window length and the spectral resolution. A discrete short-time fourier transform F expresses a signal as a discrete spectrum of a time-domain signal $x(t)$ within a windowed timeframe:

$$F(t,\omega) = \sum_{n} x(n)w(n-t)e^{-j\omega n}, \tag{2.7}$$

where $w(n)$ is the window function and ω is the frequency.

With the window length being constant, a short window results in a coarser spectral resolution that is independent of the frequency to be displayed, whereas a fine spectral resolution can only be achieved at the cost of a longer window and therefore a reduced temporal resolution.

For audio signal analysis, a detailed spectral resolution of the low frequency range is desirable, whereas a fine resolution of the spectrum's temporal course is important for higher frequencies. Wavelet transform allows both a fine low frequency spectral and a detailed high frequency temporal resolution and therefore provides useful information for analysis of non-stationary audio signals (150).

A (non-windowed) fourier transform expresses a signal as a weighted sum of time-invariant sine and cosine functions and therefore loses any information for time localisation. Wavelets in contrast are oscillations that occur in a short period of time, and a wavelet transform represents the signal as a combination of scaled and translated wavelets which are defined in frequency and time. Figure 2.21 shows the graphical representations of a Morlet Wavelet, an often-used mother

wavelet which provides a good compromise between spectral and temporal resolution. Figure 2.22 shows the Mexican Hat Wavelet which provides an even higher spectral resolution.

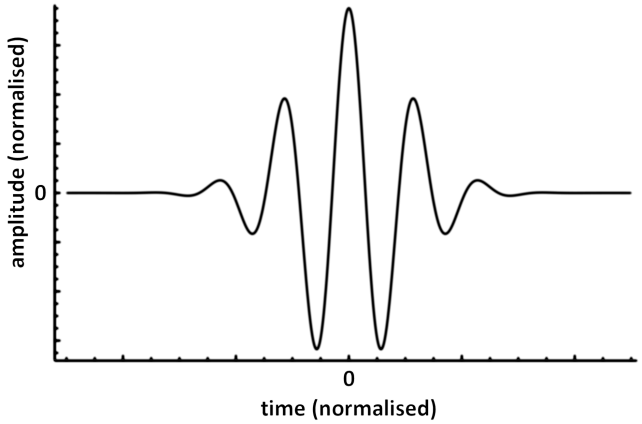

Figure 2.21: *Morlet Wavelet (real part)*

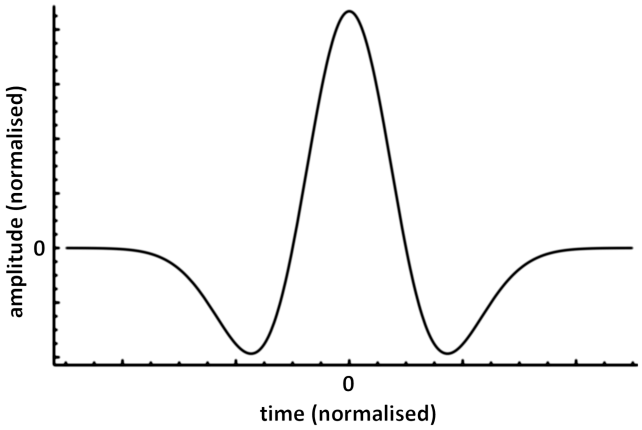

Figure 2.22: *Mexican Hat Wavelet (real part)*

The elementary functions of the wavelet transform are scaled and shifted versions of a time-local mother wavelet. A discrete wavelet transform W expresses a signal $x(t)$ as:

$$W_\psi x(a,b) = \frac{1}{\sqrt{a}} \sum_t \psi(\frac{t-b}{a}) x(t), \tag{2.8}$$

where ψ is the mother wavelet, a is the scaling parameter and b is the translation parameter.

A wavelet transform can be understood as a cascade of bandpass filters that decomposes the signal into a tree of orthogonal subspaces with an approximation part carrying the information of the low pass and a detail part that contains the high pass information. A wavelet packet transform (WPT), in turn, decomposes both the approximation and the detail part of the original signal. Or to express it differently, while a WT employs only low pass filters, a WPT uses both low pass and high pass filters (151).

The usefulness of wavelets has been explored for the identification of obstructive events in OSA patients (96). Using continuous wavelet transform, the spectral energy distribution in snoring events has been analysed immediately before, during and after an apnoeic event, revealing higher fractions of high-frequency energy components in the snoring event directly following the apnoeic event.

Qian et. al. used WT and WPT to create feature sets for the classification of snoring sounds based on the Munich-Passau Snore Sound Corpus and could achieve comparable results to large feature space extraction approaches (7; 10; 13).

Chapter 3

Own Contributions

"To date, a satisfactory definition of snoring is lacking"
(The Sleep Medicine Working Group of the German Society
of Otorhinolaryngology, Head and Neck Surgery, 2013)

3.1 Snoring and Breathing

3.1.1 Introduction

What is snoring? Almost everybody knows how snoring sounds and will recognise it when hearing it.

While such noises can also be simulated in the wake state, commonly, snoring is understood to occur involuntarily during sleep. Besides from natural sleep, snoring can also be forced artificially by titration of narcotising drugs such as propofol or midazolam, as used during drug-induced sleep endoscopy.

Other nocturnal respiration-related sounds are wheezing, puffing or whistling sounds, such as the stridor. In contrast to snoring, generated by vibrations of soft tissue in the pharynx, the stridor is caused by turbulence in the larynx or the tracheobronchial area (152), mostly due to subglottal stenoses or anomalies, e .g. because of a tumour disease. For a reasonably experienced listener, the two phenomena can be easily distinguished. A stridor is characterised by whistling noises with tonal components that have a fundamental frequency that is considerably higher than that of snoring. Other subglottal sounds are the rhonchus, which may contain low-frequency tonal components (humming rhonchus). This is often caused by floating mucus in the large airways and can be caused by asthma or Chronic Obstructive Pulmonary Disease (COPD).

Often, snoring sounds are defined as audible soft tissue vibrations in the upper respiratory tract, with a distinct tonal component in the resulting noise (153). In contrast, breathing noises are considered to be caused by turbulences in the nasal and oral airflow resulting in a noisy sound without pronounced tonal components.

Automatic algorithms distinguishing between breathing and snoring sounds, as used for example in machine classifiers, can only be as good as the data with which they are trained. Thus, the quality of the training data, annotated by human listeners, is a potential source of inaccuracies in classification tasks.

In 1990, the American Sleep Disorders Association (ASDA), the predecessor organisation of today's American Academy of Sleep Medicine (AASM), defined snoring as "loud upper airways breathing ... caused by vibrations of the pharyngeal tissues" (154). In 1996, Dasmasso et. al. noted that "snoring is a symptom of nasal obstruction ... however, its acoustic features in these disorders are not well-defined" (22). The authors defined a snoring index (numbers of snores per hour of sleep) and a snoring frequency (numbers of snores per minute of snoring time). Both definitions, however, refer to the frequency and severity of the snoring phenomenon, and do not consider the acoustic particularities of the snoring sound itself.

More recently, in 2017, Swarnkar et. al. described snoring as being "characterised by repetitive packets of energy that are responsible for creating the vibratory sound peculiar to snorers" (155). This definition is based on the fact that, in most subjects, snoring is generated in the inspirational phases during successive, regular breathing cycles.

Nevertheless, no definition exists to date that permits an objectively measurable distinction between snoring and loud breathing, which can occur at very similar temporal patterns. As the Sleep Medicine Working Group of the German Society of Otorhinolaryngology, Head and Neck Surgery puts it: "To date, a satisfactory definition of snoring is lacking" (65). Such a definition, however, is a fundamental prerequisite to develop algorithms that attempt to acoustically detect snoring events during natural or artificial sleep. Obviously, automated analysis and classification of snoring sounds is only possible after a reliable separation of inspiratory and expiratory snoring events from other background noise. Several different approaches, which were in part derived from speech signal processing, have been investigated for their suitability and show snoring sound detection specificities of of well over 95% (156; 157; 158; 102; 159; 160; 161; 146; 162; 163). Notably, none of these publications provide an objective definition of snoring in distinction to other respiratory or ambient noises. Rather, the distinction of snoring and non-snoring sounds has been made by the investigating authors themselves based on their own subjective judgement, making their findings not independently verifiable.

Rohrmeier et. al. (164) made efforts to overcome this lack of an objective distinction between snoring and loud breathing. In order to arrive at a reliable differentiation, they have created a corpus of nightly breathing and snoring sounds which was classified by 25 human raters as either breathing or snoring. Although still based on subjective judgement, the high number of independent raters provides a certain common ground. The sounds were analysed for sound pressure level as well as for the psychoacoustic parameters loudness, sharpness, roughness, fluctuation, and annoyance. Annoyance yielded a sensitivity and specificity of 76.9 % and 78.8 %, respectively.

The aim of this chapter is to find more selective and more robust objective acoustic descriptors and to deploy machine learning methods for the distinction between snoring and breathing. These findings can later be used to develop and improve applications for automatic identification of snoring events during sleep.

3.1.2 Materials and Methods

Database Properties and Data Preparation

The corpus created by Rohrmeier et. al. comprises 55 audio sequences of nightly breathing and snoring sounds from 23 subjects recorded during natural sleep in a sleep laboratory. The audio sequences are approximately ten seconds in length, each sample containing three complete, consecutive respiratory cycles (inspiration and expiration). Care has been taken to include sounds that cover the whole spectrum from 'normal' breathing to 'heavy' snoring. The sounds were classified by 25 human raters. An inter-rater agreement of 75% was used as a threshold to classify sounds as either 'snoring' or 'breathing'. 16 percent of the sound sequences could not be classified unequivocally (inter-rater agreement of less than 75%) and were labelled as 'unclear'. For details on subjects and annotation methods please refer to (164).

For the analysis, each of the 55 sequences were cut into three separate segments only containing the inspiratory phase of the respective breathing cycle, as, in the predominant number of cases, snoring occurs during inspiration. Two exceptions have been made: the subjects in two of the recordings showed pronounced snoring during expiration, with an acoustically unobtrusive inspiration phase. In these cases, the expiratory phase was selected for analysis. Further, four samples were excluded as they contained a level of distortion that might negatively affect the extraction of acoustic features.

All segments have been normalised and stored in wav PCM format at 48 kHz sampling rate and 16 bit resolution. In total, the resulting database comprises 161 snoring or breathing samples with an average length of 1.88 s (range 0.53 ... 2.88 s).

Of these, 95 samples were classified as 'snoring' (S), 39 samples as 'breathing' (B), and 27 samples as 'unclear' (U).

The S-class and B-class samples were stratified into two sequence-disjunctive partitions, namely, a training and a development set together containing the samples from 45 sequences. All samples stemming from one sequence have always been assigned to the same partition. Because of the lack of an unequivocal label, the U-class samples were not included in either the training or the development set.

Machine Learning Experiments

For feature extraction, the openSMILE (Speech & Music Interpretation by Large-Space Extraction) open-source audio feature extractor was used (165; 166). The INTERSPEECH ComParE feature set was deployed, which was developed for speech, linguistic and paralinguistic machine classification tasks. openSMILE and the INTERSPEECH ComParE feature set have been successfully used in a number of earlier projects on snoring analysis (128; 167; 7; 11; 10), hence is was selected it for these experiments.

The INTERSPEECH ComParE feature set is based on 65 low level descriptors (LLDs), describing temporal and spectral properties of the source signal. In addition, the first order derivatives (deltas) for each LLD are calculated and a set of nine statistical functionals is derived from the LLDs and their deltas, resulting in a total number of 6 373 features. Table 3.1 lists the groups of LLDs contained, while Table 3.2 shows the statistical functionals calculated for the LLDs. A detailed description of the feature properties can be found in (128) and (167).

The open-source support vector machine toolkit LIBLINEAR (168) was chosen to train a classifier. A support vector machine (SVM) determines a separator, a hyperplane, in a set of elements of different classes, which divides these classes in the best possible way. The hyperplane is arranged in such a way that the widest possible margin remains around the class boundaries. Figure 3.1 shows an example of a linear subdivision of two classes in a two-dimensional space $X \in \mathbb{R}^2$. In the experiments described here, the dimension of the space in which the support vector is calculated corresponds to the size of the feature vector, i.e. a 6 373-dimensional space.

Octave 3.6.1 with GCC 4.6.2 was used as programming platform.

The performance of two solver types was compared, dual L2-regularised L2-loss support vector classification, and dual L2-regularised logistic regression. Linear SVMs achieve good results especially with smaller data sets and a large number of features, as is the case in these experiments. Furthermore, their generalisation behaviour can be well controlled by the complexity parameter, allowing a certain

Subset	# LLDs	# Features	Description
pcm_RMSenergy	1	100	Root mean square energy
pcm_zcr	1	100	Zero crossing rate
F0final	1	83	Smoothed fundamental frequency contour
logHNR	1	78	Logarithmic ratio of harmonic signal energy to noise signal energy
voicingFinalUncl	1	78	Voicing probability
jitterLocal	1	78	Frame-to-frame period lengths differences between pitch periods
jitterDDP	1	78	First order derivative of jitter
shimmerLocal	1	78	Frame-to-frame amplitude differences between pitch periods
pcm_fftMag	15	1500	Magnitude of fast fourier transform coefficients
mfcc	14	1400	Mel-frequency cepstral coefficients
audSpec	26	2600	Mel frequency spectrum-generated perceptual linear predictive cepstral coefficients
audspec	1	100	Sum of the audSpec coefficients
audspecRasta	1	100	Relative spectral transform-style filtered auditory spectrum
ComParE	65	6373	All subsets combined

Table 3.1: *Feature subsets of the INTERSPEECH* ComParE *feature set # LLDs = Number of low-level descriptors; # Features = Number of features with first order derivatives and statistical functionals.*

Functionals
max, min, mean, range,
standard deviation, slope,
bias (linear regression approximation),
skewness, kurtosis

Table 3.2: *Statistical functionals of the INTERSPEECH* ComParE *feature set.*

percentage of errors during training and avoiding over-adaptation to the training data.

In a first experiment, a 45-fold cross validation using the S-class and B-class samples was performed, each time leaving the samples of one sequence out of the training, respectively the development set, and used for testing. The complexity parameter was set to 1, which has been experimentally determined in the range of $2^{-30}, 2^{-29}, \ldots, 2^{0}$ as providing the optimal UAR when optimised on the development partition. Training of the final model was performed fusing the training and development partition, in each case without the samples of the respective testing sequence.

The experiments were carried out 14 times. Besides the full INTERSPEECH ComParE feature set, the 13 subsets were deployed one by one in order to determine those feature classes which are most sensitive for the distinction of snoring and breathing sounds.

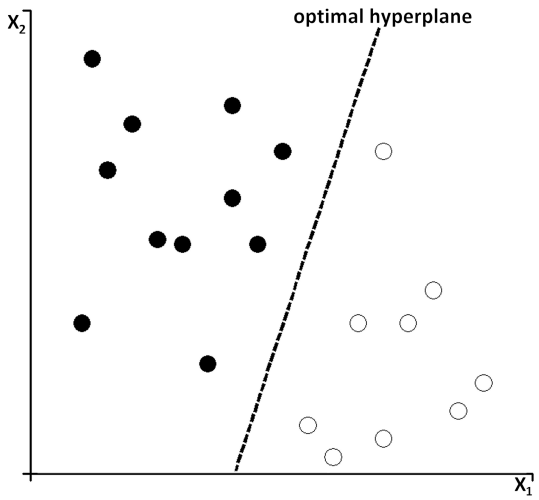

Figure 3.1: *Principle of a support vector classification in a two-dimensional space.*

In a second experiment, the probability values from the logistic regression solver type training results were compared with the level of agreement of the human raters, i. e., the percentage of raters that defined the sounds of the respective sequence as snoring. An agreement of >75 % was defined as snoring (S), <25 % as breathing (B), between 25 % and 75 % as unclear (U). In this experiment, the data for the S-class and B-class type samples were used which were generated as described above. In addition, the full combined training and development partitions of all S-class and B-class samples were used for model training and the model was tested on the U-class.

Ranking of Single Features

In order to evaluate the single most sensitive features for the distinction between snoring and breathing, the UAR was calculated for each of the 6 373 features, defined as the mean of the class-specific recalls for S-class and B-class samples. This exercise was done for all possible values of the respective feature, and considering the value yielding the maximum UAR as the ideal separator for this feature.

	SVM classification		Logistic regression	
Name of Set	% UAR	% WAR	% UAR	% WAR
pcm_RMSenergy	87.9	90.3	90.4	91.8
pcm_zcr	75.4	79.1	75.0	80.6
F0final	63.3	69.4	64.7	72.4
logHNR	81.9	85.1	79.6	82.8
voicingFinalUncl	84.5	86.6	81.1	82.8
jitterLocal	67.4	73.1	64.9	71.6
jitterDDP	69.5	73.9	73.1	76.7
shimmerLocal	79.4	83.6	81.9	85.1
pcm_fftMag	**93.0**	**93.3**	**93.0**	**93.3**
mfcc	88.0	85.0	89.0	86.6
audSpec	87.2	85.1	87.2	85.1
audspec	**92.2**	**93.3**	**91.7**	**92.5**
audspecRasta	62.3	67.9	60.2	67.2
ComParE	**92.9**	**91.0**	**92.4**	**90.3**

Table 3.3: *Classification results per feature subset of the S-class and B-class samples using two different solver types.*
UAR = unweighted average recall; WAR = weighted average recall.

3.1.3 Results

The results of the first experiment are summarised in Table 3.3, using the UAR as performance measure. The best classification performance could be achieved using the *pcm fftMag* feature subset, comprising 15 coefficients and their derivative and statistical functionals derived from the magnitudes (the real parts) of a fast fourier transform of the signal. Both SVM classification and logistic regression yielded a UAR of 93.0 %. The second best performance showed the *audspec* feature subset, with a UAR of 92.2 % using SVM classification, and 91.7 % using logistic regression. Interestingly, this subset is based only on a single LLD, which is the sum of 26 perceptual linear predictive (PLP) cepstral coefficients generated from the Mel frequency spectrum. The full INTERSPEECH ComParE feature set yielded a UAR of 92.9% using SVM classification, and 92.4% with logistic regression. Table 3.4 shows the confusion matrices using SVM classification for the three best-performing feature subsets.

Figure 3.2 shows a scatter plot of the probabilities from the trained logistic regression model versus the the percentage of raters that defined the sounds of the respective sequence as snoring (second experiment). Comparing the determination coefficient R^2 for all feature subsets, we found that the full INTERSPEECH ComParE set yielded the best result with an R^2 of 0.66.

audspec	pred ->	- S -	- B -
	- S -	97.4 %	5.3 %
	- B -	10.3 %	89.7 %

ComParE	pred ->	- S -	- B -
	- S -	88.4 %	11.6 %
	- B -	2.6 %	97.4 %

pcm_fftMag	pred ->	- S -	- B -
	- S -	93.7 %	6.3 %
	- B -	7.7 %	92.3 %

Table 3.4: *Confusion matrices of the best-performing feature subsets using SVM classification*

For the single features, Figures 3.3, 3.4, and 3.5 show scatter plots of inter-rater agreement versus value after openSMILE feature extraction of the three single features yielding the highest UAR. The x-axis shows the inter-rater agreement,

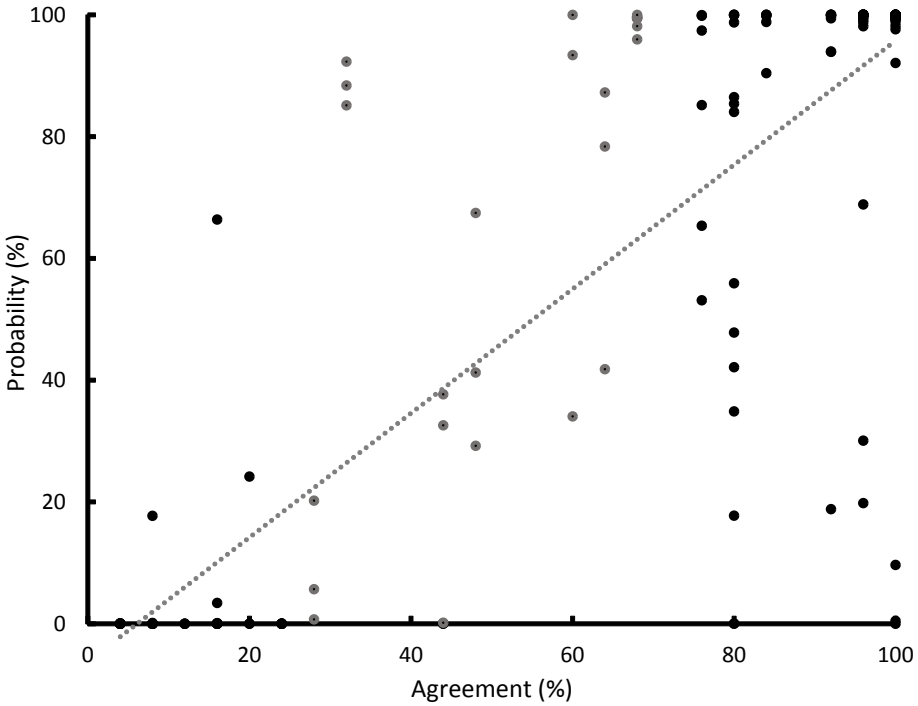

Figure 3.2: *Probabilities calculated by logistic regression versus inter-rater agreement.*
x-axis = inter-rater-agreement; y-axis = probability values for the snoring class of the logistic regression model. U-class samples are depicted in grey colour. The dashed line is the trendline.

the y-axis displays the respective feature value. The horizontal line denotes the ideal separator (value of highest UAR).

Table 3.5 summarises the UAR, sensitivity and specificity of the best-performing single features.

All of the three features are statistical functionals of a PLP cepstral coefficient generated from the mel frequency spectrum. Namely, the flatness of the second audspec coefficient (Figure 3.3), the 1%-percentile of the first coefficient (Figure 3.4), and the standard deviation of the first coefficient (Figure 3.5).

3.1.4 Discussion

The best-performing single features as well as the second-best-performing feature subset are based on Mel frequency spectrum-generated PLP cepstral coefficients.

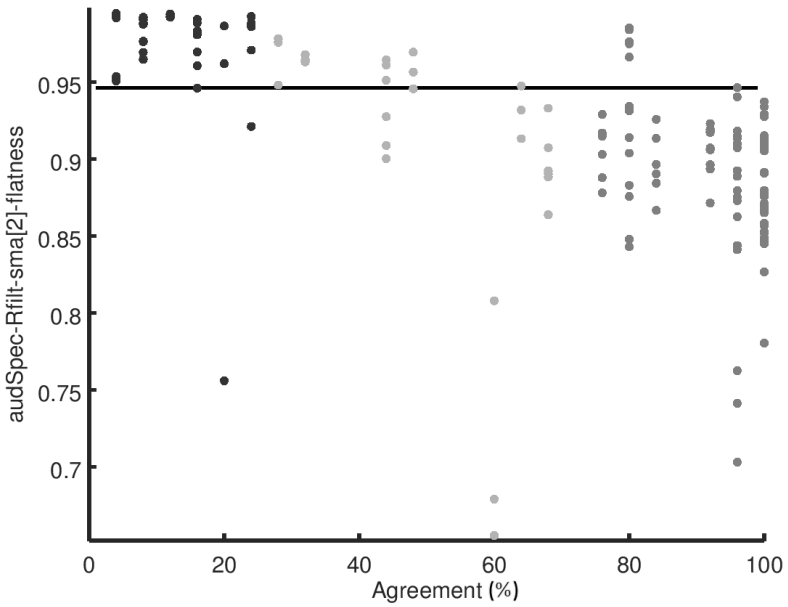

Figure 3.3: *Agreement versus feature value audSpec-Rfilt-sma[2]-flatness.*

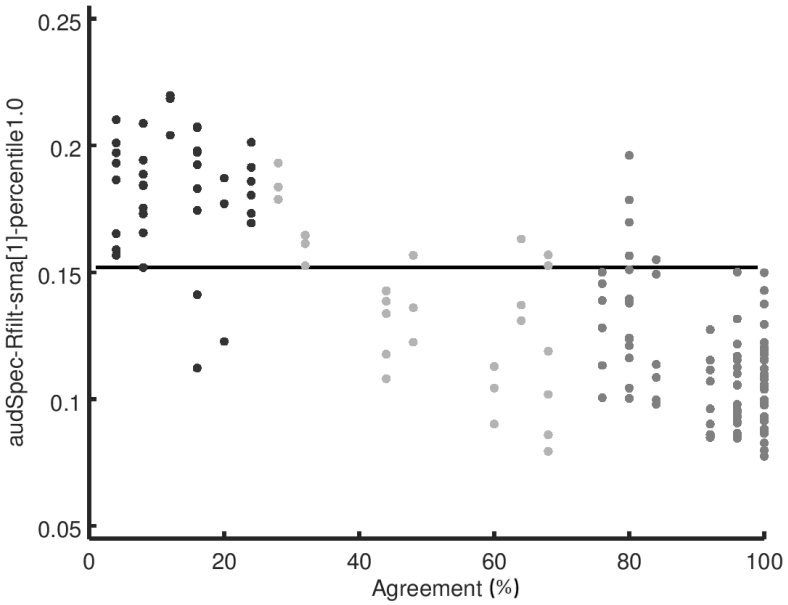

Figure 3.4: *Agreement versus feature value audSpec-Rfilt-sma[1]-percentile1.0.*

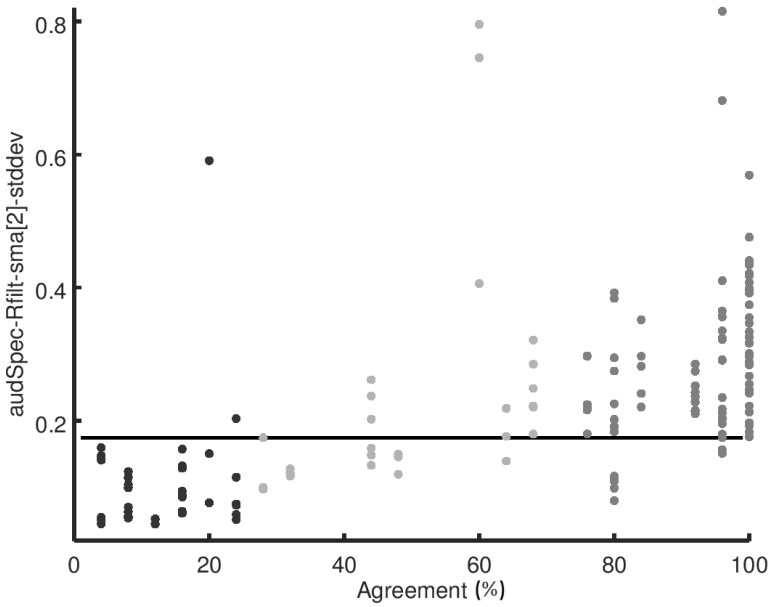

Figure 3.5: *Agreement versus feature value audSpec-Rfilt-sma[2]-stddev.*

The best performing feature subset is based on FFT-generated features, whereas single FFT-based features yielded a UAR of up to 91.3 %, which is the next best performance after the audspec-based features.

AudSpec and FFT are different representations of the signal's spectral properties. By comparison, features describing the temporal properties, such as jitter and shimmer, did not prove to be as good predictors for the difference of snoring and breathing.

Interestingly, based on this dataset, single features showed a performance that is comparable to models learnt on SVM-classification and logistic regression and based on complete feature sets. The generalisability of classifications based on a single feature remains questionable, however, and might well be worse than a feature set-based machine learnt model when applied to unknown, independent data.

In comparison, the results presented here are notably better than those reported by Rohrmeier et. al. using psychoacoustic parameters. Rohrmeier et. al. found that annoyance according to Zwicker's psychoacoustic annoyance model yielded the best distinction between loud breathing and snoring (sensitivity 76.9%, specificity 78.8%). Zwicker's annoyance model combines four acoustical parameters, particularly, loudness, sharpness, fluctuation strength and roughness (169). Loudness is a measure that aims to resemble the subjective perception of the volume of

Score	Name of Feature	% UAR	% Spec	% Sens
1	audSpec-Rfilt-sma[2] flatness	93.8	94.9	92.6
2	audSpec-Rfilt-sma[1] percentile1.0	93.5	92.3	94.7
3	audSpec-Rfilt-sma[2] stddev	93.2	94.9	91.6

Table 3.5: *Best-performing single features.*
UAR = Unweighted Average Recall; Sens = Sensitivity; Spec = Specificity

a sound. It is derived from the sound pressure, which in turn depends on the distance and position of the recording microphone relative to the sound source, i. e., the snorer's mouth and nose. This parameter therefore requires a careful setup and calibration of the recording situation. Differences in microphone positions, amplification settings of the recording equipment, and even sleeping positions of the snoring subject may result in differences of the annoyance value and therefore skew the results.

The PLP cepstral coefficients, in contrast, are based on the spectral properties of the signal and independent of the absolute SPL. Further, the amplitude of the audio samples has been normalised. This promises to yield more robust results when used in real life applications, where microphone positions and room conditions might not be precisely controllable.

A weakness of the experimental setup used here is that the experiments are based on a ground truth that is still subjective, although the high number of raters promises a certain level of consensus compared to classifications that are based on the evaluation of one single or a small group of raters. Further, the original classification by the raters was made by listening to all three snore cycles of the respective sequence. For the experiments described here, the sequences were separated these into single samples. Potential differences in sound between the three respiratory cycles of the same individual have therefore not been considered, a fact that potentially might have an influence on the results. Finally, the size of the corpus is small for a machine learning task. Therefore, the robustness and generalisability of these findings is yet to be confirmed by larger datasets.

3.2 The MPSSC Database

3.2.1 Introduction

Polysomnography and cardiorespiratory screening provide a reliable diagnosis as towards the type and severity of a sleep related breathing disorder. However, it is of very limited use to identify its underlying mechanisms. DISE is increasingly used by sleep surgeons and appreciated as a useful tool to identify the location of vibration and obstruction. Because it is cost intensive, time consuming, and cannot be performed in natural sleep, it is of interest to develop alternative methods for the identification of the excitation location of snoring sounds that do not have the mentioned limitations. A possible solution can be the acoustic analysis of snoring sounds.

It has been shown in principle that different excitation locations of snoring sounds are correlated with distinct acoustic characteristics (22; 8). Consequently, the determination of different types of snoring must be possible by acoustic analysis of the related sound. In order to test this hypothesis, the *Munich Passau Snore Sound Corpus* (MPSSC) has been developed. The corpus consists of audio recordings of separate snoring events that have been annotated using simultaneous endoscopic video recordings of the upper airways taken during DISE examinations. The labelling is therefore based on an objective and independently verifiable ground truth.

Applying machine learning strategies to distinguish snoring sounds according to their source of excitation may perspectively complement DISE investigations or even replace them by acoustic analysis of snoring sounds in selected patients, and thus decrease the physical strain for patients undergoing snoring diagnosis and reduce healthcare cost.

In contrast to earlier work, it is not the aim of the experiments described in this chapter to distinguish between primary snoring and OSA or to classify OSA severity, but to identify vibration locations, no matter if the snorer shows obstructive episodes or not.

3.2.2 Materials and Methods

Data Collection

The database is derived from original endoscopic recordings taken during DISE procedures. The material is available in mp4 format and contains simultaneous video and audio recordings. The recordings were made during DISE examinations of patients who had undergone previous PSG and were diagnosed with OSA. DISE was performed as an additional diagnostic measure in these patients for planning of subsequent surgical interventions, for pressure titration of a continuous positive airway pressure (CPAP) system, or for fitting of a mandibular advancement device (MAD). The material was obtained from the following three clinical centres which use DISE examinations as a routine diagnostic method in selected patients.

- Klinikum rechts der Isar (Technical University Munich), Munich, Germany: recordings from 38 subjects taken 2013 through 2014.
- Alfried Krupp Hospital Essen, Germany: recordings from 2090 subjects taken 2006 through 2015.
- University Hospital Halle/Saale, Germany: recordings from 46 subjects taken 2012 through 2015.

Table 3.6 shows the equipment and the microphone setups used for recording of the DISE videos.

Figure 3.6 displays screenshots taken from DISE recordings as examples of typical snoring events. The upper left image (V) shows a vibrating velum at the palatal level. In the lower left image (O), the oropharyngeal level can be seen with vibrating palatine tonsils. In the upper right image (T), the tongue base vibrates against the posterior pharyngeal wall. Finally, the lower right image (E) shows a vibrating epiglottis. The white arrows in the images mark the respective vibrating structures.

Pre-Processing

First, the audio signal was extracted from the mp4 files and stored in wav-format (16 bit, 44 100 Hz). Subsequently, audio events were identified using an automated algorithm. As in the experiments described in the previous Chapter 3.1, Octave 3.6.1 with GCC 4.6.2 was used for programming. The absolute value of the signal amplitude was averaged in non-overlapping 10 ms windows. The background noise level was determined by means of a histogram of the signal amplitude with

Centre	Recording equipment and setup
Munich	Storz flexible nasopharyngoscope, Storz Telepack X recording system (Storz, Tuttlingen), headset microphone Sennheiser ME3 (Sennheiser, Wedemark), recording position 5-10 cm to the side of the patient's mouth.
Essen	Olympus flexible nasopharyngoscope, Rehder/Partner rpSzene recording system (Rehder und Partner, Hamburg), hand-held microphone with recording distance approx. 1m in front of the patient's mouth, alternatively forehead microphone with recording distance approx. 30cm above the patient's mouth.
Halle	Storz flexible nasopharyngoscope, Storz AIDA recording system (Storz, Tuttlingen), stand-mounted condenser microphone NT3 (RODE, Silverwater, Australia), recording distance 30cm in front of the patient's mouth.

Table 3.6: *Recording setup at the clinical centres*

1024 equally spaced intervals, averaged in on-overlapping 10 s windows. The background noise level was defined as the respective maximum value of the histogram. All segments exceeding a level of two times the determined background noise level for a minimum duration of 300 ms were annotated. Adding 100 ms of signal before and after the actual onset and end of the event, the events were extracted from the original audio file, normalised, and saved as separate wav files (16 bit, 16 000 Hz). Figure 3.7 illustrates the segmentation procedure. All described values were experimentally optimised during the algorithm development based on a subset of the DISE audio recordings.

Pre-Selection: Snoring and Non-Snoring Sound Events

In a next step, an experienced human listener (the author of this thesis) listened to all selected events and classified them manually as either pure snoring (*snoring*) or other sounds (*non snoring*). Also, those events that contained a snoring event but

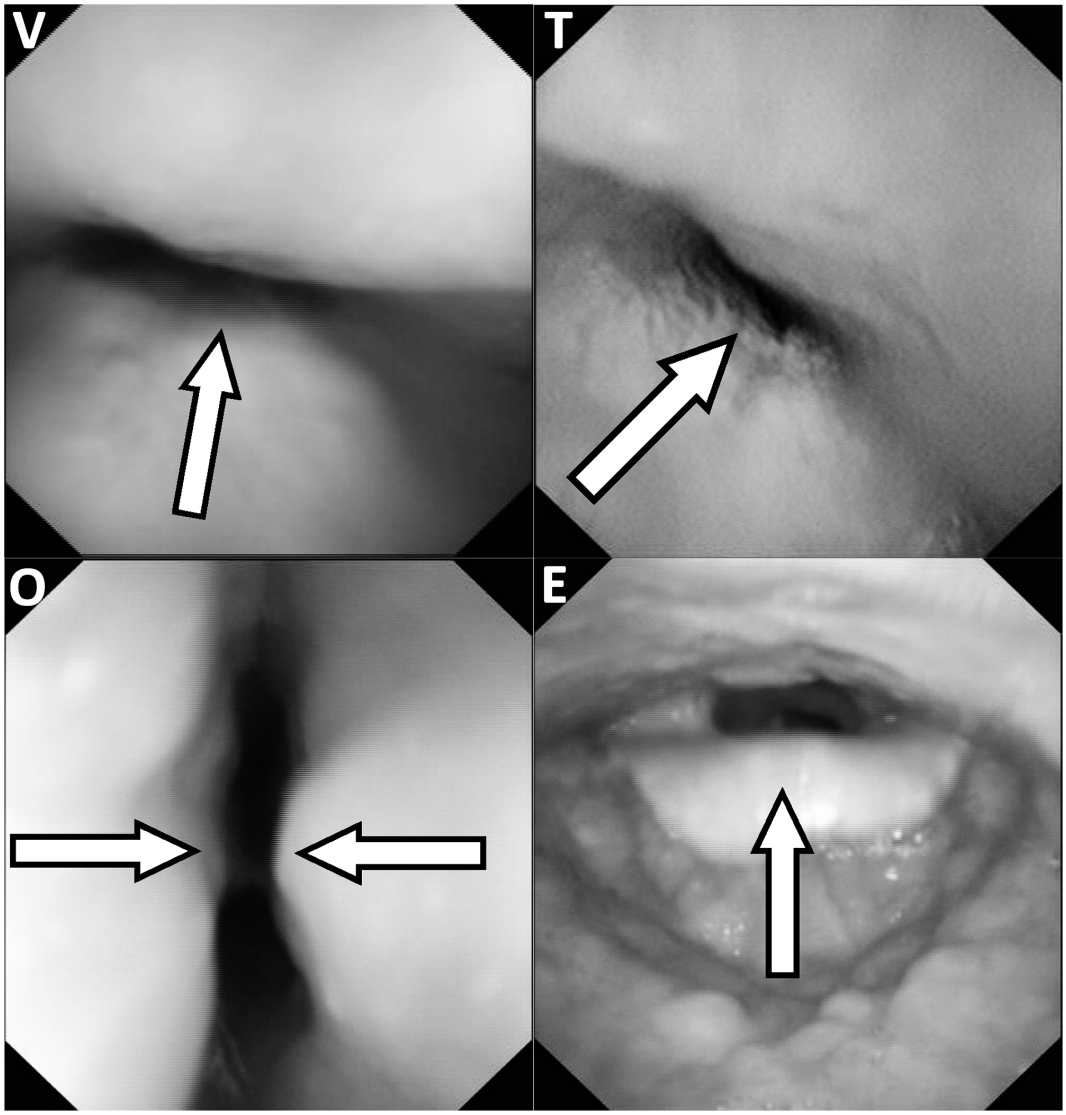

Figure 3.6: *Screenshots taken from DISE video recordings showing palatal snoring (V), oropharyngeal snoring (O), tongue base snoring (T), epiglottal snoring (E). All screenshots are taken from videos of the Essen centre.*

were disturbed by non-static background noise, such as speech or acoustic alarm signals from medical equipment, were excluded from the snore group. The same applies for snoring events that were overdriven or distorted by disturbances in the recording chain.

The criteria to include a sound event in the snoring group were therefore based on the subjective judgement of the author of this thesis. A rigid standard was

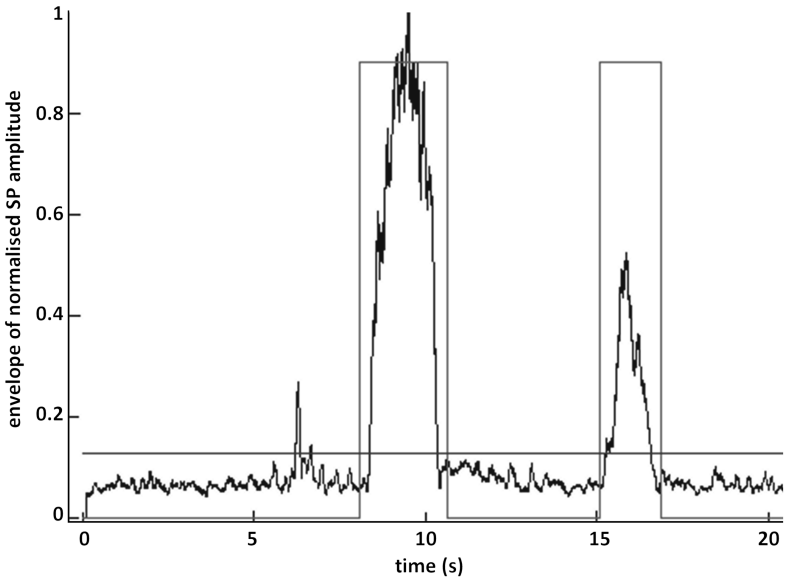

Figure 3.7: *Illustration of the segmentation procedure based on an example of a 20 s audio signal from a DISE recording.*
Based on the amplitude envelope of the snore signal, the horizontal line is the threshold amplitude of two times the background noise level. The vertical lines show onset and end of the selected audio segments identified as events of min. 300 ms length.

applied to pass as snoring sound. When in doubt, a sound event was rather excluded from the snoring group.

A subject's recording was discarded altogether if

- no acoustic event could be extracted from the original recording,
- none of the extracted acoustic events qualified as snoring signal,
- all of the snoring events were polluted by non-static background sounds, overdriven or distorted.

While the material from Halle/Saale and from Munich was already pre-selected for videos containing snore episodes, the material from Essen had not been pre-screened. Therefore, the yield of subjects with snoring events from the Essen material was distinctly lower compared to the other two centres.

In total, snoring events from 331 subjects were selected for subsequent annotation (Essen, 266 subjects; Munich, 31 subjects; Halle/Saale, 34 subjects). The total number of snoring events was 2 261, the number of snoring events per subject ranged from two to 30.

Classification

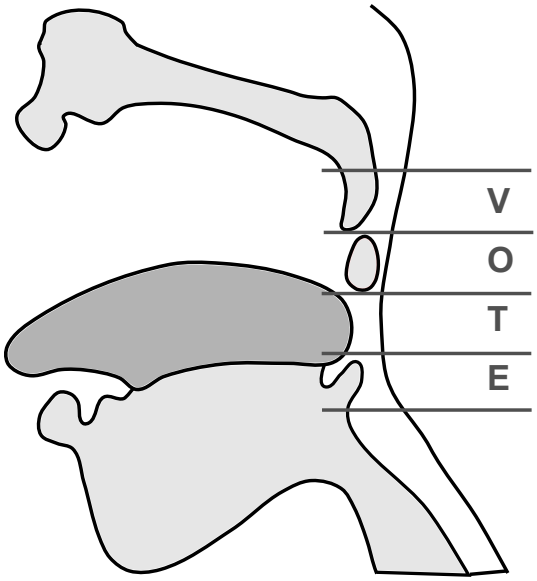

Figure 3.8: *Vibration areas in the upper airways according to the VOTE classification*
Sagittal section through the head; V = velum; O = oropharynx; T = tongue base; E = epiglottis.

Several schemes have been suggested for the classification of the location of upper airway obstructions and snoring noise generation (170; 171; 46; 172). A widely used scheme is the VOTE classification, introduced by Kezirian et. al. in 2011 (173). The VOTE classification allows a standardised description of obstruction or vibration sites with regards to their location. According to the VOTE scheme, four structures that can be involved in airway narrowing and obstruction are distinguished (174):

- *V*, Velum (palate), including the soft palate, uvula, and lateral pharyngeal wall tissue at the level of the velopharynx.

- *O*, Oropharyngeal lateral walls, including the palatine tonsils and the lateral pharyngeal wall tissues that include muscles and the adjacent parapharyngeal fat pads.

- *T*, Tongue, including the tongue base, the lingual tonsil, and the airway posterior to the tongue base.

- *E*, Epiglottis, describing folding of the epiglottis due to decreased structural rigidity or due to posterior displacement against the posterior pharyngeal wall.

Figure 3.8 illustrates the corresponding levels within the upper airways.

In addition, the VOTE classification contains a description of the shape of obstruction, using the categories anterior-posterior (*a-p*), lateral (*l*), and concentric (*c*).

Further, the degree of airway narrowing can be qualitative assessed on a scale from 0 to 2 (0, no obstruction; 1, partial obstruction; 2, complete obstruction), and the occurrence of snoring can be noted.

The VOTE classification as introduced by Kezirian et. al. is primarily used to describe airway narrowing and obstruction in OSA patients. For the research described in this chapter, a simplified version of the VOTE classification is introduced in order to describe the location of vibration of the soft tissue generating snoring noise, while no distinction is made between different degrees of airway narrowing, as only events that create vibration of the airway structures are of interest. Further, the shape of obstruction is not considered. This leads to a four-class classification described by the labels *V*, *O*, *T*, and *E*.

Annotation

For all selected sound events, the respective video files were watched by two experienced experts. Based on the video findings, each snoring event was assigned to one of the four classes. Segments where both experts were not in agreement as to the correct class were excluded.

Vibration and obstruction in the upper airway is not always limited to a single level. For this database, events were excluded if the vibration was not clearly limited to one of the four defined levels. However, during one and the same DISE examination session, the same subject might show vibration patterns at different levels in different snoring events, but limited to one vibration level per event. In this case, snoring events were included and labelled accordingly. For example, one subject showed distinct velum-level snoring when the mandible was manually advanced by the attending surgeon during the procedure using an Esmarch-maneuvre. Without this maneuvre, snoring originated from the epiglottis-level. Consequently, the database contains both *V*-type snoring and *E*-type snoring events from this very subject.

Further, only those events were included where the vibration mechanism could be clearly seen in the DISE video recording. Samples with compromised visibility

(for example due to saliva on the endoscope tip) were excluded, as were samples in which the video recording showed a different level of the upper airway than the location of excitation at the same point of time (e.g., observing the epiglottis during a suspected velum snore) and therefore the vibration mechanism could not be visually confirmed.

What's more, snoring events that coincided with an obstructive event were excluded.

For the remaining audio events, the corresponding video sections of the DISE video were reviewed, classifying the vibration location according to the simplified VOTE scale.

From the 331 subjects included in the annotation step, a total of 112 had to be excluded altogether for the following reasons:

- none of the snoring events was limited to one of our four defined levels,
- disagreement on the level of vibration between the annotators for all events,
- impaired visibility of vibration level for all events,
- obstruction occurred during all snoring events.

Of the remaining 219 subjects, a maximum of six snoring events per subject and class were included in the database. If more than six events of the same class were available in one subject, only the first six events were used. Figure 3.9 shows a summary of the selection steps taken and the number of subjects per centre included in the database after identification of snoring events and after annotation.

In order to verify the reliability of annotation, a subset of videos from 40 subjects was evaluated independently by an additional annotator. The subset included all 10 subjects that were annotated to the T class, plus 30 randomly selected subjects. There was agreement for all subjects except for one (annotator 1: O-type snoring; annotator 2: probably O-type, but not certain). Based on this sample of 18% of subjects from the total set, the inter-rater-reliability according to Cohen's Kappa is $\kappa = 0.96$. [1]

Inter-observer agreement for evaluation of DISE videos was studied by Vroegop et. al. in 2013 (175). For the level of collapse, inter-rater reliability values between $\kappa = 0.48$ for the oropharyngeal level and $\kappa = 0.71$ for the tongue base level were found for a group of seven experienced ENT surgeons. Although these results are only comparable to a limited extent (Vroegop et. al. evaluated collapse instead of

[1] Cohen's Kappa was calculated using ReCal2 0.1, dfreelon.org

vibration, and they used a classification additionally comprising the hypopharynx as a fifth level), it is safe to conclude that the inter-rater-agreement in this study offers a very high level of confidence in the annotation. Reasons for this comparatively good agreement can be that all annotators are highly experienced in the evaluation of DISE recordings, and that events with unclear level of vibration had already been excluded in a previous step.

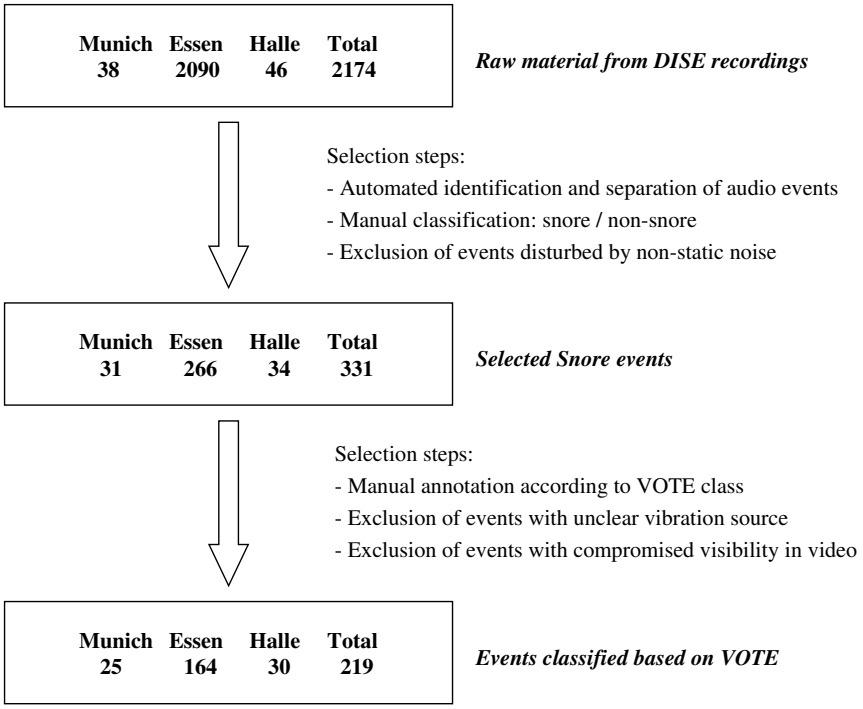

Figure 3.9: *Number of subjects per centre included in the database after each data selection step*

Partitioning

In order to prepare the corpus for machine learning experiments, the data was stratified into a train, a development (devel), and a test partition. In order to create subject disjunctive partitions, the assignment was made based on subject, not event (i. e. all snoring events from a subject are assigned to the same partition). To obtain this, the subjects were first sorted by class. Within each class, subjects were sorted by centre, then by gender, and then by age. Using this order, subjects were successively, one by one, assigned to the train, development, and test partitions. Figure 3.10 illustrates this process. A two-tailed, unpaired t-test

confirmed no significant differences between the partitions for age, gender, centre or class (p>0.05). Table 3.7 shows the resulting number of events per class and partition. Since the number of snoring events per subject differs, the partitions contain different numbers of snoring events, but equal number of subjects.

In particular, an even distribution of the data by centre reduces the risk of learning ambient acoustic characteristics instead of snoring sound properties. However, of the T-type subjects, seven are from Essen, but only two from Munich, and one from Halle. For this reason, the instances from this class could not be balanced completely evenly by centre between the set splits. This should be considered when interpreting the classification results.

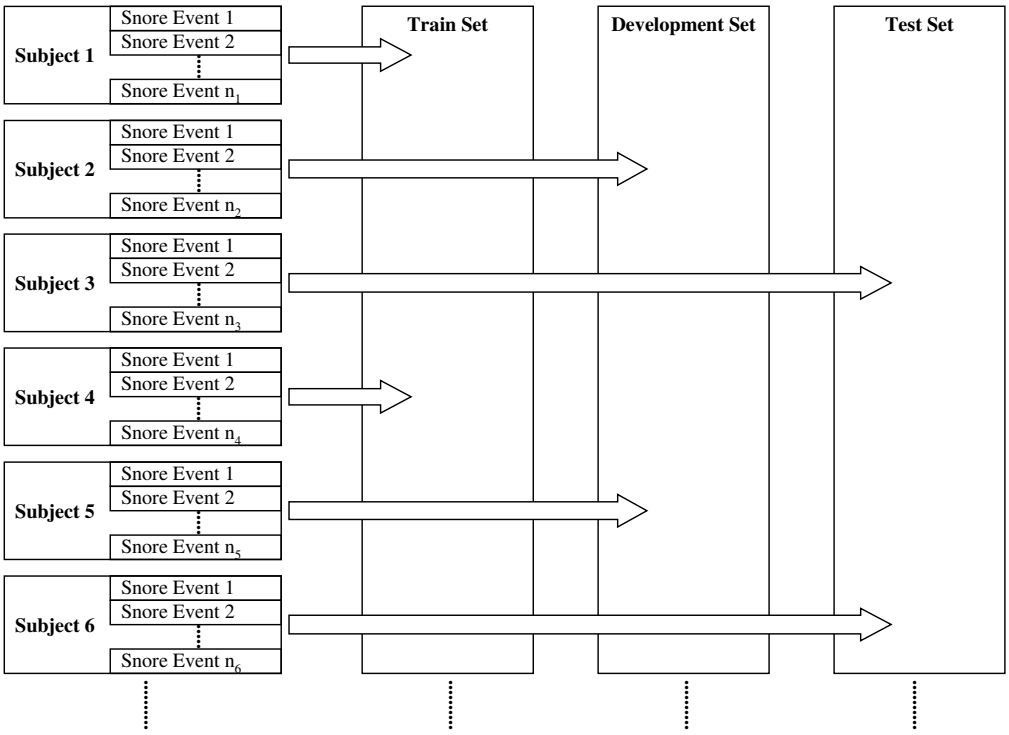

Figure 3.10: *Process of subject-disjunctive stratification*
All snoring events from a subject are successively assigned to the respective train, development or test set.

	Train	Devel	Test	Σ
V	168	161	155	484
O	76	75	65	216
T	8	15	16	39
E	30	32	27	89
Σ	282	283	263	828

Table 3.7: *Number of snoring events per class in the set splits*

Database Properties

The resulting database contains audio samples of 828 snoring events from 219 subjects. All samples in the database are available with a sampling rate of 16 000 Hz and a resolution of 16 bit.

The average sample duration is 1.46 s (range 0.73 ... 2.75 s). Samples from the T-class are significantly shorter than those from the three other classes (p<0.001, see Figure 3.11 B). [2]

Since the sample duration itself might be a descriptor for the respective class, the differences in sample length are not a sign of inhomogeneity of the database, but rather a noteworthy fact.

Average age of the subjects is 49.8 (range 24 ... 78) years, with no significant difference between classes (p>0.10), see Figure 3.11 A.

Further, notably, 93.6% of all subjects are male.

Table 3.8 contains the number of subjects per class and centre, which are included in the database. Note that the total number for all classes in Table 3.8 is 223, whereas the total number of actually included subjects is 219. Reason for this discrepancy is that one of the subjects showed both V and E type snoring, another subject showed both V and O type snoring, and again two other subjects showed both V and T type snoring during the DISE investigation. Thus, these four subjects are counted twice.

The number of events and subjects per class in the database is strongly unbalanced, with the majority of samples belonging to the V and O class (total 84.5%), whereas T and E type snoring samples only account for 4.7%, and 10.8%, respectively, of the total number of events. This was to be expected and is in line with earlier findings from DISE evaluations. Hessel et. al. described in 2003 based on DISE examinations of 380 patients that single level obstructive events at the hypopharyngeal level (thus, T, and E type according to this classification)

[2] All probability values calculated with two-tailed, unpaired t-test

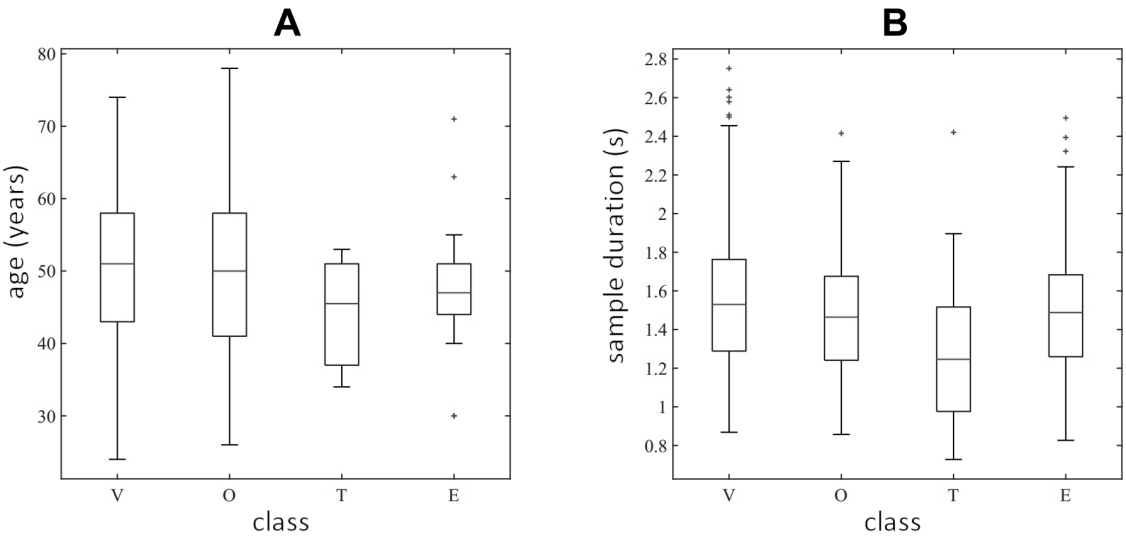

Figure 3.11: *Subject's metadata per class*
A = age per class, (in years at the time of DISE investigation). B = sample duration per class (in seconds per event).

Centre	V	O	T	E
Munich	14	4	2	5
Essen	100	46	7	15
Halle	19	6	1	4
Total	133	56	10	24

Table 3.8: *Number of subjects per centre and class*

occurred in only 2% of patients, whereas single level V and O type events occurred in 22% of patients, thus 10 times as often (48). Other researchers come to similar results (176).

It is important to note that certain acoustic properties of the sound samples from the three centres are distinctly different. Firstly, the acoustic characteristics of the room (ambient noise, room acoustics) differ between the three centres. Secondly, different types and models of microphones were used, resulting in differences in the frequency response of the microphone itself, as well as the position and distance of the microphone relative to the snorer, which again can have a considerable influence on the signal to noise ratio. In Munich, a headset microphone was used, in Halle, a stand-mounted microphone was deployed. In Essen, a handheld microphone, a headset microphone, and a microphone to be fixed on the forehead were

available, and the type of microphone used for the audio recordings was chosen according to the preference of the surgeon performing the DISE investigation.

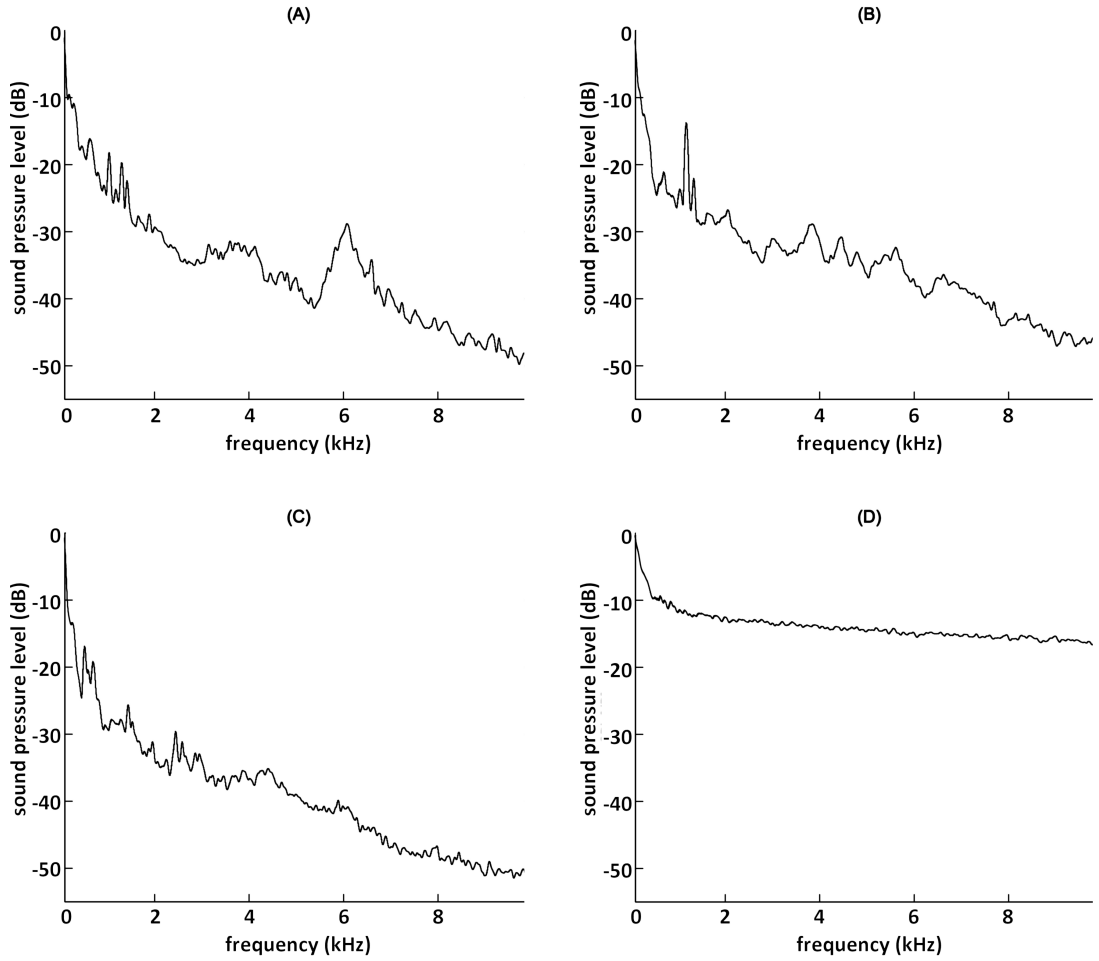

Figure 3.12: *Background noise frequency spectra for different recording settings A = Essen, using a handheld microphone. B = Essen, using a headset microphone. C = Halle, using a stand-mounted microphone. D = Munich, using a headset microphone.*

Figure 3.12 shows spectrograms of the background noise in different recording settings, taken from sections of the DISE recordings in the three centres. The spectrograms show that the background noise characteristics are distinctly different. To evaluate this effect, an additional machine learning experiment was performed in the same setup as described in the following Section 3.2.2, but using the centres as classes instead of the snoring noise type. The results show that centres can be clearly distinguished with a UAR of 88.0% (mean UAR of all partition permutations and using the INTERSPEECH ComParE baseline feature set plus

the formants subset), proving that the snoring sounds indeed carry centre-specific information. In order to evaluate the impact on the performance of our classifier setup to distinguish snoring noise, a machine learning experiment was performed exactly as described in the following section, but using samples only from Essen, resulting in a slightly worse performance compared to the results including all three centres (53.4% UAR for Essen only versus 55.8% for all centres, rated by mean UAR of all permutations and using the full ComParE feature set plus the formants subset).

These experiments show that centre-specific acoustic properties are not necessarily a weakness of the database, but can be desired for machine learning experiments. Since the task is not about distinguishing between centres, and each of the snore classes contains a balanced number of samples from all three centres, the difference in ambient sound characteristics might actually prevent the machine from learning these features, and to focus on the differences in actual snoring noise, resulting in an even more robust classifier model.

Nevertheless, care should be and has been exercised to carefully balance the number of samples from the different centres per class and per partition.

Machine Learning Experiments

Baseline experiments were performed in the framework of the *Snore* Sub-Challenge in the INTERSPEECH 2017 Computational Paralinguistics ChallengE (ComParE), yielding a UAR of 58.5% using the ComParE functionals in combination with a linear support vector machine. Details can be found in section 2.2.2 and in (11).

To obtain a more detailed insight into the suitability of different acoustic features for the task at hand, the performance of the different subsets of the ComParE feature set was evaluated. A description of the subsets is given in Table 3.1 on page 57. For the sake of a convenient overview, the feature subsets, number of low level descriptors and resulting number of features after calculating deltas and functionals are repeated in a brief form in Table 3.9. In addition to the ComParE features (lines 1 through 13), the frequency and bandwidth of the formants F1, F2, and F3 were extracted for the following experiments (lines 14 through 19).

The feature sets were extracted by the openSMILE feature extraction and audio analysis tool. All experiments were conducted using the LIBLINEAR SVM toolbox. As solver type, a linear kernel (L2-regularised L2-loss support vector classification, dual) was chosen with a bias of *1*.

Figures 3.13 and 3.14 show the principle of the machine learning setup used for the training and the test phase.

Line	Feature type	#LLDs	#Features	Description
1	audspec	1	100	Sum of audSpec
2	audspecRasta	1	100	audspec incl. RASTA
3	pcm_RMSenergy	1	100	RMS energy
4	pcm_zcr	1	100	Zero crossing rate
5	audSpec	26	2600	Mel frequency PLP cepstral coefficients
6	pcm_fftMag	15	1500	Fast fourier transform magnitudes
7	mfcc	14	1400	Mel frequency cepstral coefficients
8	F0final	1	83	Fundamental frequency
9	voicingFinalUncl.	1	78	Voicing probability
10	jitterLocal	1	78	Period length differences
11	jitterDDP	1	78	Difference of difference of period lengths
12	shimmerLocal	1	78	Amplitude variations
13	logHNR	1	78	Log. HNR ratio
14	F1frequency	1	78	First formant frequency
15	F1bandwidth	1	78	First formant bandwidth
16	F2frequency	1	78	Second formant frequency
17	F2bandwidth	1	78	Second formant bandwidth
18	F3frequency	1	78	Third formant frequency
19	F3bandwidth	1	78	Third formant bandwidth
20	F1-F3	6	468	Frequency and bandwidth of F1-F3
21	ALL wo. F1-F3	65	6373	Features Line 1 through 13
22	ALL	71	6841	Features Line 1 through 19

Table 3.9: *Feature subsets*
#LLDs = Number of low-level descriptors; #Features = Number of features including functionals and deltas.

For all experiments, the complexity parameter of the SVM was optimised on the development set in the range of $2^{-30}, 2^{-29}, \ldots, 2^{0}$. The complexity providing the maximum UAR was selected and divided by 2 for the training of the final model, fusing train and dev set. As both sets have approximately the same size,

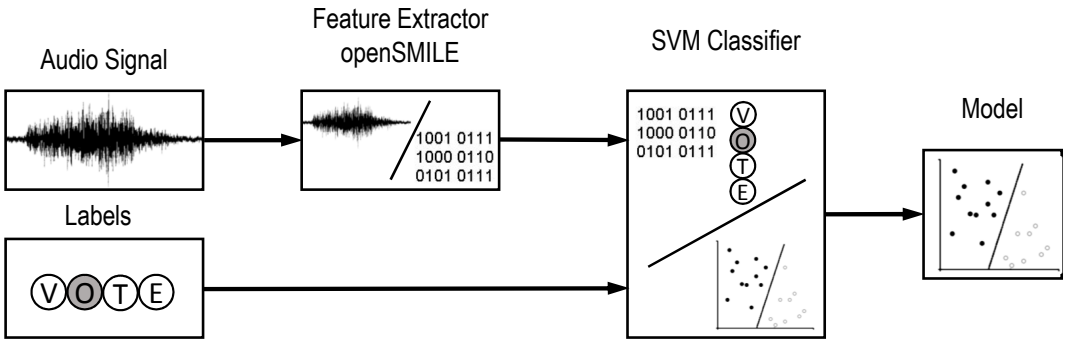

Figure 3.13: *Structure of the machine learning system used (training phase).*

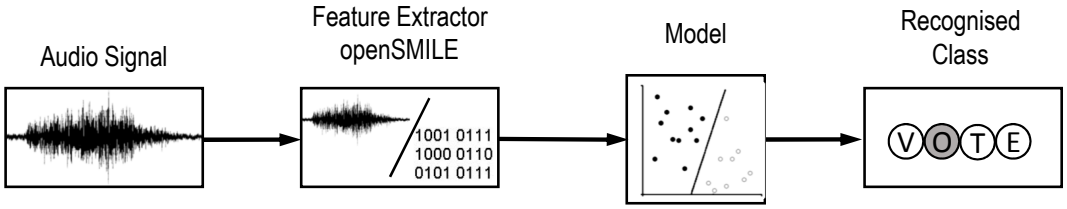

Figure 3.14: *Structure of the machine learning system used (test phase).*

this bisection of the complexity parameter has proven to be most suitable. All features were standardised to zero mean and unit standard deviation using an online approach. This means that the parameters mean and standard deviation were derived from the train set (or the fusion of train and dev set, respectively) only and then applied to the dev set (the test set, respectively). No resampling of the data was done in any of the experiments.

In order to average out potential differences in the characteristics of train, development, and test partition, the experiments were carried out six times in all possible permutations of the three partitions. In other words, the three partitions were swapped as follows.

1st permutation: partition 1 = train; partition 2 = devel; partition 3 = test
2nd permutation: partition 1 = train; partition 3 = devel; partition 2 = test
3rd permutation: partition 2 = train; partition 1 = devel; partition 3 = test
4th permutation: partition 2 = train; partition 3 = devel; partition 1 = test
5th permutation: partition 3 = train; partition 1 = devel; partition 2 = test
6th permutation: partition 3 = train; partition 2 = devel; partition 1 = test

Line	Feature type	f/d	#LLDs	#Features	mn. UAR	Range
1	mfcc	f+d	28	1400	49.9 %	7.0 %
2	mfcc	only f	14	756	**52.9 %**	10.6 %
3	mfcc	only d	14	644	33.4 %	11.4 %
4	F1-F3	f+d	12	468	30.6 %	8.8 %
5	F1-F3	only f	6	234	30.5 %	8.0 %
6	F1-F3	only d	6	234	29.7 %	8.0 %
7	ALL w/o F1-F3	f+d	130	6373	55.4 %	12.5 %
8	ALL w/o F1-F3	only f	65	3425	54.7 %	10.5 %
9	ALL w/o F1-F3	only d	65	2948	47.1 %	9.5 %
10	ALL	f+d	142	6841	**55.8 %**	7.1 %
11	ALL	only f	71	3659	54.7 %	10.1 %
12	ALL	only d	71	3182	48.0 %	4.7 %

Table 3.10: *Classification results*
mn. UAR = unweighted average recall, mean performance of all partition permutations; Range = Range of results between partition permutations; f/d = functionals (f) and/or deltas (d) used; LLDs = Number of low-level descriptors; Features = Number of features including functionals and deltas.

3.2.3 Results

Classification results are shown in Table 3.10 for the best-performing feature sets together with the corresponding number of LLDs and the final number of features after computing the functionals. The mean unweighted average recall (UAR) over all permutations of the partitions as well as the range of UAR results between the permutations are listed. *ALL w/o F1-F3* with coefficients and deltas (line 7 in Table 3.10) is the full ComParE feature set. *ALL* shows the results when applying the ComParE feature set plus *F1-F3*. For comparison, the *F1-F3* subset is also listed.

Detailed results can be found in Tables 3.13, 3.14, and 3.15. In particular, Table 3.13 lists the number of LLDs and the number of overall features for each of the feature subsets. Table 3.14 contains the obtained UAR for all feature subsets and all permutations of the train, development, and test partition. Best results for each permutation and mean performance are highlighted. Table 3.15 shows the class-specific recall of all feature subsets for the mean of all permutations of the train, development, and test partition.

Rated by UAR, the best classification performance could be achieved with the full feature set consisting of the ComParE features plus the formant set *F1-F3* including functionals and deltas. Best performing single subset is *mfcc only coef*, consisting of MFCC-related LLDs, using functionals, but not deltas. Using only formant-related features (*F1-F3*) yielded inferior classification results. Removing the formants subset from the full feature set results in only a minor deterioration of 0.4% UAR, suggesting that formant frequencies and bandwidths do not provide meaningful additional information in these experiments.

It is remarkable that the results differ considerably between the permutations. Range of performance between the best and the worst performing permutation is up to 12.5% for the ComParE feature set, and still 7.1% for the full feature set. A comparison of the confusion matrices reveals that the largest differences occur in the two small classes T and E, with a range of 18%, and 28% in class-specific recall, respectively, between the permutations. Performance differences for the large classes V and O are smaller by comparison. Table 3.12 shows confusion matrices for all permutations, and Table 3.11 summarises mean and range of all permutations of the class-specific recalls. All results are for the best-performing *ALL* feature set with functionals and deltas (line 10 in Table 3.10.

Class	Mean	Min	Max	Range
V	66.6 %	59.4 %	73.9 %	14.6 %
O	62.1 %	56.6 %	67.7 %	11.1 %
T	24.4 %	13.3 %	31.3 %	17.9 %
E	70.3 %	53.1 %	81.5 %	28.4 %

Table 3.11: *Mean, minimum, maximum, and range of class specific recall of all partition permutations.*

It can be suspected that these discrepancies are a result of chance, since the number of subjects in both classes is fairly small for a machine learning task. It is well possible that it requires a larger number of samples for a machine classifier to deduct the characteristic acoustic features of T and E snoring, which can be subject of future experiments based on larger datasets.

Applying the weighted average recall as a performance measure overweights the contribution of the larger classes V and O, thereby reducing the influence of the questioned small classes. With a WAR of 65.4%, the combination of all employed features (*ALL*) with functionals but without deltas shows the best results over all permutations.

Tr+De>Te

pred ->	V	O	T	E	Recall
V	92	37	8	18	59.4 %
O	13	44	2	7	67.7 %
T	0	7	4	5	25.0 %
E	0	4	1	22	81.5 %

De+Tr>Te

pred ->	V	O	T	E	Recall
V	94	33	9	19	60.6 %
O	15	39	4	7	60.0 %
T	0	6	5	5	31.3 %
E	0	3	2	22	81.5 %

Tr+Te>De

pred ->	V	O	T	E	Recall
V	119	32	2	8	73.9 %
O	17	49	0	9	65.3 %
T	0	12	2	1	13.3 %
E	10	4	1	17	53.1 %

Te+Tr>De

pred ->	V	O	T	E	Recall
V	118	33	1	9	73.3 %
O	15	49	3	8	65.3 %
T	0	10	4	1	26.7 %
E	7	3	0	22	68.8 %

De+Te>Tr

pred ->	V	O	T	E	Recall
V	107	30	12	19	63.7 %
O	20	43	6	7	56.6 %
T	2	2	2	2	25.0 %
E	1	5	4	20	66.7 %

Te+De>Tr

pred ->	V	O	T	E	Recall
V	115	29	13	11	68.5 %
O	18	44	6	8	57.9 %
T	2	2	2	2	25.0 %
E	2	5	2	21	70.0 %

Table 3.12: *Confusion matrices (all permutations), best-performing feature set Tr = train partition; De = development partition; Te = test partition.*

3.2.4 Discussion

Marked differences in classification performance can be found between the permutations of train, development, and test partition, mainly caused by the classes T and E. Due to the low number of subjects in these classes, misclassification of only few events can result in a considerable performance difference measured by unweighted average recall (UAR). Still, UAR should be the ultimate measure for performance in this task, as the WAR underrates the performance in the small classes. No matter how small, each of the four classes have equal importance, as a therapy decision for T or E type snorers is distinctively different than for V or O type snoring, which occurs much more frequently. More stable results could be expected with data from a higher number of subjects in the smaller classes.

Snoring and speech have a lot of acoustic similarities: both are generated in the upper airway through vibrations caused by airflow, acoustically shaped by the frequency transfer function of the upper airway and emitted through mouth and nose. The position of the tongue is of significance for shaping the different phonemes in speech and in the generation of different types of snoring, thus shaping the resulting sound in a characteristic way. Acoustic descriptors that have proven effective in speech-related machine learning tasks can therefore be expected to be well suited also for the classification of snoring noise. The findings from these experiments as well as the results from the ComParE Snore Sub-Challenge contributions underpin this assumption. The presented acoustic tube model of the upper airways (132) has yielded results that are consistent with the underlying anatomy it aims to resemble. MFCC-based features haven proven most successful in classification performance in (130), and those models using feature sets based on MFCCs and PLP cepstrum showed the best results of the challenge (133; 134). This is confirmed when investigating the performance of the INTERSPEECH ComParE feature subsets: the MFCC subset has shown a superior classification performance compared to all other single subsets. Hence, the descriptors that prove sensitive in the classification task at hand are those representing the spectral properties of the signal, which can be seen as a confirmation for the hypothesis that the upper airway transfer function is characteristic for different excitation locations of snoring sounds.

Formant characteristics have been investigated for their suitability to describe snoring sounds in earlier works. Peng et. al. have found a statistically significant difference in frequency of F2 between snoring generated by the velum versus the lateral pharyngeal walls (177). Koo et. al. looked at obstruction levels in OSA patients and found significantly higher frequencies for F1 and F2 in snorers with retrolingual obstruction compared to those with retropalatal obstruction (178). In the experiments described in this chapter, MFCCs have clearly outperformed the

subset that is based on formant characteristics alone, suggesting that formants are indeed descriptive for the excitation location of snoring sounds, but inferior to MFCCs.

There are also a number of differences between speech and snoring. In speech generation, except for plosive and fricative consonants, the sound is excited in a fixed location, the voice box in the glottis. Vowels are formed by the position of tongue, palate, mandible and lips, altering the cross-sectional profile of the upper airway. At the same time, the total length of the acoustically effective tubes change only marginally. Snoring, in contrast, can be generated in different locations within the UA, resulting in a variable length of the acoustically effective system for spectral shaping.

While the glottis wave in speech can be altered in pitch and loudness, in healthy speakers it has a characteristic shape. Also, the fundamental frequency range is defined for different speakers (male female, children), the melody of speech (so-called pitch) is mainly characterised by the prosodic content (speech melody). The excitation waveform of snoring sounds, in contrast, can vary widely, and the fundamental frequency can range from as low as 10 Hz to more than 500 Hz. Also the pitch of a snoring event can vary in a lot of forms. Novel descriptors derived from those used in speech classification tasks might help to further improve classification outcomes in future snoring sound classification experiments.

Line	Feature type	Delta	#LLDs	#Feat
1	audspec	coef+delta	2	100
2	audspecRasta	coef+delta	2	100
3	pcm_RMSenergy	coef+delta	2	100
4	pcm_zcr	coef+delta	2	100
5	audSpec	coef+delta	52	2600
6	pcm_fftMag	coef+delta	30	1500
7	mfcc	coef+delta	28	1400
8	F0final	coef+delta	2	83
9	voicingFinalUnclipped	coef+delta	2	78
10	jitterLocal	coef+delta	2	78
11	jitterDDP	coef+delta	2	78
12	shimmerLocal	coef+delta	2	78
13	logHNR	coef+delta	2	78
14	F1frequency	coef+delta	2	78
15	F1bandwidth	coef+delta	2	78
16	F2frequency	coef+delta	2	78
17	F2bandwidth	coef+delta	2	78
18	F3frequency	coef+delta	2	78
19	F3bandwidth	coef+delta	2	78
20	audspec	only coef	1	54
21	audspecRasta	only coef	1	54
22	pcm_RMSenergy	only coef	1	54
23	pcm_zcr	only coef	1	54
24	audSpec	only coef	26	1404
25	pcm_fftMag	only coef	15	810
26	mfcc	only coef	14	756
27	F0final	only coef	1	44
28	voicingFinalUnclipped	only coef	1	39
29	jitterLocal	only coef	1	39
30	jitterDDP	only coef	1	39
31	shimmerLocal	only coef	1	39
32	logHNR	only coef	1	39
33	F1frequency	only coef	1	39
34	F1bandwidth	only coef	1	39
35	F2frequency	only coef	1	39
36	F2bandwidth	only coef	1	39
37	F3frequency	only coef	1	39
38	F3bandwidth	only coef	1	39
39	audspec	only delta	1	46
40	audspecRasta	only delta	1	46
41	pcm_RMSenergy	only delta	1	46
42	pcm_zcr	only delta	1	46
43	audSpec	only delta	26	1196
44	pcm_fftMag	only delta	15	690
45	mfcc	only delta	14	644
46	F0final	only delta	1	39
47	voicingFinalUnclipped	only delta	1	39
48	jitterLocal	only delta	1	39
49	jitterDDP	only delta	1	39
50	shimmerLocal	only delta	1	39
51	logHNR	only delta	1	39
52	F1frequency	only delta	1	39
53	F1bandwidth	only delta	1	39
54	F2frequency	only delta	1	39
55	F2bandwidth	only delta	1	39
56	F3frequency	only delta	1	39
57	F3bandwidth	only delta	1	39
58	F1-F3 only	coef+delta	12	468
59	F1-F3 only	only coef	6	234
60	F1-F3 only	only delta	6	234
61	ALL (wo. F1-F3)	coef+delta	130	6373
62	ALL (wo. F1-F3)	only coef	65	3425
63	ALL (wo. F1-F3)	only delta	65	2948
64	ALL	coef+delta	142	6841
65	ALL	only coef	71	3659
66	ALL	only delta	71	3182

Table 3.13: *Feature subsets and number of features*

Delta = functionals (coef) and/or deltas used; #LLDs = Number of low-level descriptors; #Feat = Number of features including functionals and deltas.

Line	Feature type	Delta	Tr+De ->Te	De+Tr ->Te	Tr+Te ->De	Te+Tr ->De	De+Te ->Tr	Te+De ->Tr	Mean	Range
1	audspec	coef+delta	34.8%	36.2%	39.1%	37.4%	37.8%	38.3%	37.2%	4.3%
2	audspecRasta	coef+delta	28.5%	29.0%	28.8%	32.7%	37.0%	37.1%	32.2%	8.6%
3	pcm_RMSenergy	coef+delta	29.0%	27.9%	32.2%	31.9%	31.5%	35.3%	31.3%	7.3%
4	pcm_zcr	coef+delta	33.1%	29.2%	29.1%	33.2%	29.5%	21.4%	29.2%	11.8%
5	audSpec	coef+delta	53.8%	54.4%	48.9%	48.9%	48.4%	53.4%	51.3%	6.1%
6	pcm_fftMag	coef+delta	43.9%	44.8%	31.8%	44.7%	35.5%	32.8%	38.9%	13.1%
7	mfcc	coef+delta	53.2%	49.3%	47.8%	49.1%	53.6%	46.6%	49.9%	7.0%
8	F0final	coef+delta	32.8%	28.8%	32.8%	32.5%	30.7%	31.9%	31.6%	4.0%
9	voicingFinalUncl.	coef+delta	30.0%	31.4%	29.5%	28.7%	30.7%	30.7%	30.2%	2.7%
10	jitterLocal	coef+delta	32.4%	28.3%	30.9%	30.9%	31.9%	27.4%	30.3%	5.0%
11	jitterDDP	coef+delta	30.1%	30.0%	31.5%	31.0%	27.4%	27.0%	29.5%	4.5%
12	shimmerLocal	coef+delta	31.3%	30.1%	31.2%	24.7%	23.6%	23.3%	27.4%	8.0%
13	logHNR	coef+delta	32.0%	28.8%	35.9%	30.8%	26.7%	28.9%	30.5%	9.3%
14	F1frequency	coef+delta	31.8%	29.8%	28.5%	29.2%	25.7%	26.3%	28.5%	6.1%
15	F1bandwidth	coef+delta	32.4%	26.3%	27.9%	33.7%	26.0%	26.3%	28.8%	7.7%
16	F2frequency	coef+delta	31.3%	31.3%	30.4%	30.7%	27.1%	26.5%	29.5%	4.8%
17	F2bandwidth	coef+delta	33.1%	32.7%	32.2%	32.1%	26.7%	26.9%	30.6%	6.4%
18	F3frequency	coef+delta	31.2%	32.3%	27.5%	29.0%	25.8%	26.9%	28.8%	6.5%
19	F3bandwidth	coef+delta	31.6%	31.6%	32.1%	28.2%	24.6%	24.6%	28.8%	7.5%
20	audspec	only coef	31.5%	31.5%	30.1%	32.2%	36.4%	38.0%	33.3%	7.8%
21	audspecRasta	only coef	25.8%	28.5%	29.3%	29.4%	35.7%	35.6%	30.7%	9.9%
22	pcm_RMSenergy	only coef	28.9%	28.9%	30.0%	30.5%	31.4%	31.4%	30.2%	2.5%
23	pcm_zcr	only coef	30.1%	28.4%	26.2%	29.3%	23.2%	26.0%	27.2%	6.9%
24	audSpec	only coef	50.7%	51.7%	48.8%	47.9%	46.9%	46.9%	48.8%	4.8%
25	pcm_fftMag	only coef	42.9%	47.7%	37.5%	42.5%	41.1%	40.6%	42.1%	10.2%
26	mfcc	only coef	53.1%	53.3%	53.5%	57.7%	52.5%	47.2%	52.9%	10.6%
27	F0final	only coef	31.9%	28.4%	32.3%	29.2%	29.7%	30.0%	30.3%	3.9%
28	voicingFinalUncl.	only coef	33.2%	33.2%	31.9%	29.1%	31.9%	31.9%	31.9%	4.1%
29	jitterLocal	only coef	31.2%	28.8%	29.7%	30.5%	30.2%	29.2%	29.9%	2.4%
30	jitterDDP	only coef	30.4%	31.4%	31.0%	29.5%	26.1%	29.5%	29.7%	5.3%
31	shimmerLocal	only coef	31.7%	27.9%	30.8%	26.9%	24.1%	25.5%	27.8%	7.6%
32	logHNR	only coef	30.1%	30.5%	32.3%	30.9%	26.9%	29.2%	30.0%	5.4%
33	F1frequency	only coef	31.9%	36.5%	28.9%	33.9%	26.1%	28.3%	31.0%	10.4%
34	F1bandwidth	only coef	32.3%	29.4%	32.4%	33.2%	26.7%	26.7%	30.1%	6.5%
35	F2frequency	only coef	32.3%	26.3%	28.9%	29.9%	27.7%	26.7%	28.6%	6.0%
36	F2bandwidth	only coef	32.3%	31.4%	32.3%	32.3%	26.7%	26.7%	30.3%	5.7%
37	F3frequency	only coef	32.3%	32.0%	34.6%	28.1%	25.9%	26.7%	29.9%	8.7%
38	F3bandwidth	only coef	32.3%	30.5%	33.2%	33.2%	26.7%	24.4%	30.1%	8.8%
39	audspec	only delta	30.9%	31.5%	32.2%	28.8%	31.4%	32.2%	31.2%	3.4%
40	audspecRasta	only delta	30.3%	29.7%	27.1%	27.4%	30.4%	30.7%	29.3%	3.6%
41	pcm_RMSenergy	only delta	27.0%	30.0%	33.0%	31.9%	31.8%	32.7%	31.1%	6.0%
42	pcm_zcr	only delta	30.6%	28.6%	27.4%	25.4%	25.7%	26.4%	27.4%	5.3%
43	audSpec	only delta	37.5%	42.8%	37.7%	43.5%	44.8%	45.6%	42.0%	8.1%
44	pcm_fftMag	only delta	40.2%	44.4%	35.8%	40.3%	37.7%	35.6%	39.0%	8.8%
45	mfcc	only delta	33.3%	28.2%	39.6%	32.3%	32.8%	34.0%	33.4%	11.4%
46	F0final	only delta	32.9%	29.7%	30.7%	29.3%	29.8%	29.5%	30.3%	3.6%
47	voicingFinalUncl.	only delta	24.9%	24.3%	25.1%	26.1%	24.0%	24.4%	24.8%	2.1%
48	jitterLocal	only delta	31.9%	30.3%	31.2%	31.2%	28.4%	26.5%	29.9%	5.4%
49	jitterDDP	only delta	29.6%	29.9%	29.8%	32.2%	28.3%	28.2%	29.7%	4.1%
50	shimmerLocal	only delta	28.2%	27.3%	31.2%	31.2%	23.4%	24.7%	27.7%	7.8%
51	logHNR	only delta	32.7%	26.0%	32.3%	25.8%	24.5%	24.0%	27.6%	8.8%
52	F1frequency	only delta	31.8%	28.1%	33.8%	26.4%	27.5%	26.5%	29.0%	7.4%
53	F1bandwidth	only delta	30.4%	29.8%	33.2%	28.8%	26.7%	27.9%	29.5%	6.4%
54	F2frequency	only delta	30.8%	30.0%	31.8%	27.3%	26.3%	26.6%	28.8%	5.5%
55	F2bandwidth	only delta	32.6%	32.6%	33.0%	28.9%	25.6%	27.5%	30.0%	7.4%
56	F3frequency	only delta	31.4%	31.4%	27.5%	27.7%	26.9%	27.4%	28.7%	4.5%
57	F3bandwidth	only delta	30.4%	31.6%	32.8%	31.2%	26.9%	24.9%	29.6%	7.9%
58	F1-F3 only	coef+delta	30.4%	32.5%	32.5%	34.2%	25.4%	28.4%	30.6%	8.8%
59	F1-F3 only	only coef	32.3%	32.5%	29.4%	34.7%	27.6%	26.7%	30.5%	8.0%
60	F1-F3 only	only delta	31.1%	33.1%	30.5%	30.3%	28.2%	25.1%	29.7%	8.0%
61	ALL (wo. F1-F3)	coef+delta	61.3%	58.1%	53.3%	56.5%	48.8%	54.6%	55.4%	12.5%
62	ALL (wo. F1-F3)	only coef	58.7%	59.4%	54.4%	52.9%	53.9%	48.9%	54.7%	10.5%
63	ALL (wo. F1-F3)	only delta	45.2%	48.3%	48.6%	48.1%	41.6%	51.1%	47.1%	9.5%
64	ALL	coef+delta	58.4%	58.3%	51.4%	58.5%	53.0%	55.3%	55.8%	7.1%
65	ALL	only coef	59.0%	58.6%	54.9%	54.4%	52.6%	48.8%	54.7%	10.1%
66	ALL	only delta	45.1%	45.1%	48.1%	49.8%	49.7%	49.8%	48.0%	4.7%

Table 3.14: *Classification results - performance of feature subsets and permutations*

Obtained UAR for all feature subsets and all permutations of Train, Development, and Test partition. Delta = functionals (coef) and/or deltas used; Tr = train partition; De = development partition; Te = test partition; Mean = mean performance of all permutations.

Line	Feature type	Delta	V	O	T	E
1	audspec	coef+delta	74.5%	55.1%	1.0%	18.4%
2	audspecRasta	coef+delta	74.7%	35.1%	1.0%	17.9%
3	pcm_RMSenergy	coef+delta	77.4%	40.0%	2.1%	5.7%
4	pcm_zcr	coef+delta	68.4%	21.4%	5.6%	21.6%
5	audSpec	coef+delta	56.7%	57.3%	14.9%	76.4%
6	pcm_fftMag	coef+delta	63.8%	47.5%	12.6%	31.8%
7	mfcc	coef+delta	67.6%	56.7%	13.8%	61.7%
8	F0final	coef+delta	74.0%	44.9%	7.5%	0.0%
9	voicingFinalUnclipped	coef+delta	77.5%	42.5%	0.0%	0.6%
10	jitterLocal	coef+delta	66.2%	50.8%	3.1%	1.1%
11	jitterDDP	coef+delta	77.1%	39.3%	0.0%	1.6%
12	shimmerLocal	coef+delta	66.3%	41.6%	1.0%	0.5%
13	logHNR	coef+delta	77.2%	37.3%	7.6%	0.0%
14	F1frequency	coef+delta	73.8%	32.2%	0.0%	8.2%
15	F1bandwidth	coef+delta	60.8%	51.6%	1.0%	1.6%
16	F2frequency	coef+delta	64.2%	48.7%	0.0%	5.3%
17	F2bandwidth	coef+delta	74.1%	40.9%	4.3%	3.1%
18	F3frequency	coef+delta	65.5%	46.3%	0.0%	3.3%
19	F3bandwidth	coef+delta	64.4%	50.2%	0.0%	0.6%
20	audspec	only coef	75.9%	50.7%	2.1%	4.5%
21	audspecRasta	only coef	75.3%	32.3%	0.0%	15.3%
22	pcm_RMSenergy	only coef	82.0%	33.0%	2.1%	3.7%
23	pcm_zcr	only coef	75.9%	13.4%	10.7%	8.8%
24	audSpec	only coef	56.2%	47.5%	13.9%	77.7%
25	pcm_fftMag	only coef	68.0%	50.1%	21.0%	29.0%
26	mfcc	only coef	68.3%	56.6%	16.1%	70.6%
27	F0final	only coef	77.2%	43.9%	0.0%	0.0%
28	voicingFinalUnclipped	only coef	85.3%	40.1%	2.1%	0.0%
29	jitterLocal	only coef	68.9%	47.2%	2.1%	1.6%
30	jitterDDP	only coef	78.9%	38.6%	0.0%	1.1%
31	shimmerLocal	only coef	67.4%	43.9%	0.0%	0.0%
32	logHNR	only coef	71.9%	46.9%	1.1%	0.0%
33	F1frequency	only coef	69.8%	41.0%	1.0%	12.0%
34	F1bandwidth	only coef	55.6%	64.3%	0.0%	0.5%
35	F2frequency	only coef	70.2%	38.4%	0.0%	6.0%
36	F2bandwidth	only coef	54.2%	66.9%	0.0%	0.0%
37	F3frequency	only coef	69.1%	45.5%	0.0%	5.1%
38	F3bandwidth	only coef	56.5%	62.7%	0.0%	1.0%
39	audspec	only delta	77.6%	41.8%	0.0%	5.2%
40	audspecRasta	only delta	82.9%	31.3%	1.0%	1.9%
41	pcm_RMSenergy	only delta	85.2%	35.9%	0.0%	3.3%
42	pcm_zcr	only delta	79.3%	10.1%	0.0%	20.0%
43	audSpec	only delta	60.7%	48.9%	7.4%	51.1%
44	pcm_fftMag	only delta	64.7%	43.7%	15.6%	31.9%
45	mfcc	only delta	59.5%	48.6%	15.8%	9.5%
46	F0final	only delta	71.3%	41.5%	8.5%	0.0%
47	voicingFinalUnclipped	only delta	84.6%	13.5%	0.0%	1.1%
48	jitterLocal	only delta	69.7%	47.3%	2.1%	0.6%
49	jitterDDP	only delta	78.9%	38.1%	0.0%	1.8%
50	shimmerLocal	only delta	67.9%	42.8%	0.0%	0.0%
51	logHNR	only delta	66.9%	42.3%	1.0%	0.0%
52	F1frequency	only delta	71.2%	44.8%	0.0%	0.0%
53	F1bandwidth	only delta	64.2%	52.6%	0.0%	1.1%
54	F2frequency	only delta	71.6%	43.6%	0.0%	0.0%
55	F2bandwidth	only delta	66.8%	50.8%	2.1%	0.5%
56	F3frequency	only delta	69.4%	45.0%	0.0%	0.5%
57	F3bandwidth	only delta	59.4%	57.5%	0.0%	1.6%
58	F1-F3 only	coef+delta	67.6%	43.4%	1.0%	10.3%
59	F1-F3 only	only coef	67.9%	43.5%	0.0%	10.6%
60	F1-F3 only	only delta	68.1%	47.5%	1.0%	2.2%
61	ALL (wo. F1-F3) = ComParE	coef+delta	67.0%	59.0%	26.6%	69.2%
62	ALL (wo. F1-F3)	only coef	69.5%	60.5%	16.9%	71.9%
63	ALL (wo. F1-F3)	only delta	64.6%	56.5%	19.0%	48.5%
64	ALL	coef+delta	66.6%	62.1%	24.4%	70.3%
65	ALL	only coef	69.7%	61.8%	14.9%	72.4%
66	ALL	only delta	65.3%	56.5%	22.2%	47.8%

Table 3.15: *Classification results - performance of feature subsets per class*

Per-class recall for all feature subsets (mean of all permutations). Delta = functionals (coef) and/or deltas used.

3.3 Comparison of Two Snoring Noise Classifications

3.3.1 Introduction

The simplified VOTE classification as introduced in Section 3.2 has limitations for certain therapy decisions. While it allows the detection of the level of vibration, it does not reveal any information as towards the vibratory pattern or orientation, which is an information that is routinely included in the original VOTE classification and often used for therapy decision making. On the other hand, the higher the number of classes a machine learning system should differentiate, the larger the required size of the training data set to achieve an acceptable level of classification performance. In other words, the expected recognition performance of a machine classifier with a given size of the training data set is the higher the fewer classes have to be distinguished. It is therefore desired to reduce the theoretically possible number of the twelve classes in the original VOTE scheme as much as possible whilst still providing a meaningful support in the choice of therapeutic measures.

In this section, an alternative classification scheme to the simplified VOTE classification is introduced, and its performance with the simplified VOTE scheme is compared.

3.3.2 Material and Methods

The ACLTE Classification Scheme

Figure 3.16 shows all twelve theoretically possible combinations of vibration location and pattern according to the VOTE classification. By definition, a snoring event is based on soft tissue vibration, therefore, the severity of obstruction as defined in the VOTE classification is not considered here. For the sake of comparison, Figure 3.17 depicts the classes of the simplified VOTE scheme, in which only the vibration plane is considered.

The alternative scheme introduced here, in the following referred to as *ACLTE*, allows the distinction of selected combinations of vibratory level and pattern, as shown in Figure 3.18. It comprises five classes which are defined in the following.

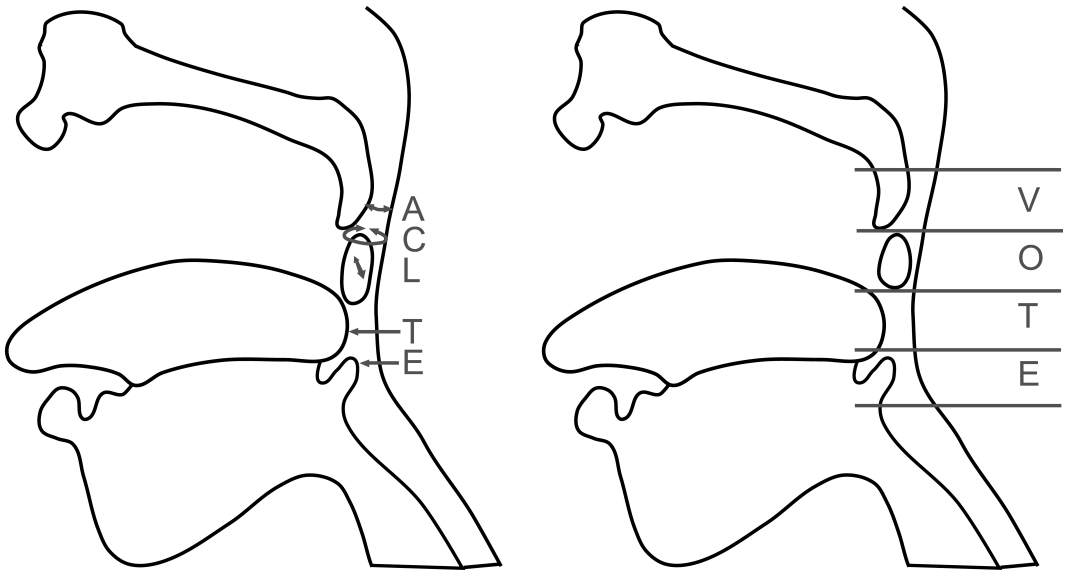

Figure 3.15: *Vibration levels and orientations, ACLTE and simplified VOTE scheme, sagittal section through the head.*
The simplified VOTE scheme (right drawing) considers only the vibration plane, the classes correspond exactly to the levels described in the original VOTE classification. The ACLTE scheme (left drawing) distinguishes different combinations of vibration location and orientation.

- A, anterior-posterior vibration of the soft palate and/or uvula,
- C, concentric vibration at the level of the soft palate, the uvula or in the area of the oropharynx,
- L, lateral vibrations at the level of the oropharynx,
- T, vibration at the level of the tongue base,
- E, vibration in the area of the epiglottis.

For classes T and E, the orientation of the vibration is not taken into account.

Figure 3.15 shows the corresponding anatomical regions in the upper airways in direct comparison to the simplified VOTE scheme using a schematic representation of a sagittal section through the head.

The ACLTE scheme reflects certain diagnostic requirements for snoring and OSA therapy planning. An anterior-posterior vibration at soft palate level can be

treated with comparably high chances of success with uvulectomy or minimally-invasive methods for the stiffening of the soft palate. A lateral vibration at oropharyngeal level is a frequent indication for tonsillotomy or tonsillectomy. In contrast, the chances of a successful surgical treatment of concentric vibration or occlusion patterns at these levels is comparatively low, it is even a contraindication for hypoglossal nerve stimulation (179). In these cases it might be recommended to revert to weight loss in snoring or continuous positive airway pressure therapy in OSA. In the hypopharyngeal area (tongue base and epiglottis level), the vibratory pattern is less important for therapy planning.

Snoring Databases

For the creation of the ACLTE snoring database, the source material from the MPSSC corpus was partly reused. Moreover, new material from the Alfried Krupp Hospital was included and further recordings from an additional centre were added. In summary, the ACLTE-corpus is created from audio and video recordings of DISE examinations carried out between 2006 and 2016 at the ENT clinics of four clinical centres:

- Klinikum rechts der Isar, Munich, Germany
- Alfried Krupp Hospital, Essen, Germany
- University Hospital Halle/Saale, Halle/Saale, Germany
- Hospital Dr Peset, Valencia, Spain

The recording equipment and setup in the three centres in Munich, Essen, and Halle is described in Table 3.6 in Section 3.2.2. In Valencia, an AverMedia C285 Capture Box (AverMedia, Taiwan) was used for recording in combination with the lavalier microphone provided with the system, mounted at collarbone height, recording distance approx. 20 cm from the patient's mouth.

In total, DISE recordings of over 2500 patients were analysed for the ACLTE database.

For pre-processing, pre-selection and annotation of the additional raw data, the same procedure was followed as described in Section 3.2.2: snoring events were identified in the audio tracks using a semi-automatic procedure, then amplitude-normalised, stored in separate audio files and the times of occurrence in the video were noted. Snoring events containing non-static background noise and distorted audio signals were discarded. Annotation was carried out by two experienced and blinded examiners based on the video findings.

Level	Pattern: anterior-posterior	lateral	concentric
Velum	V – a-p	V - l	V - c
Oropharynx	O – a-p	O - l	O - c
Tongue Base	T – a-p	T - l	T - c
Epiglottis	E – a-p	E - l	E - c

Figure 3.16: *Vibration levels and patterns according to the original VOTE classification.*

Level	Pattern: anterior-posterior	lateral	concentric
V	V		
O	O		
T	T		
E	E		

Figure 3.17: *Defined classes according to the simplified VOTE classification.*

Level	Pattern: anterior-posterior	lateral	concentric
V	A		C
O		L	
T	T		
E	E		

Figure 3.18: *Defined classes according to the ACLTE classification. Combinations of level and pattern which are anatomically not possible are depicted in grey colour.*

Those events labelled V and O from the MPSSC database were re-evaluated according to the new ACLTE scheme and newly assigned to one of the classes A, C, or L. Only events in which the new class was unequivocally recognised by both examiners were included in the ACLTE database, or otherwise discarded. The events labelled T and E from the MPSSC database were re-evaluated, the classification was confirmed in all cases.

Annotation according to the two classification schemes was performed at separate times. The simplified VOTE annotation was performed between 2015 and 2017, the ACLTE annotation between 2017 and 2018.

For an easier comparison, the size and properties of both databases are summarised in Table 3.16.

For the sake of comparability to the results described in Chapter 3.2, feature extraction and partitioning was carried out following the same method and tools as for the MPSSC database.

3.3.3 Results

As in the previous experiments, the unweighted average recall (UAR) was used as a measure of recognition accuracy, defined as the average share of correctly assigned events over all classes.

Overall, the five-class model of the ACLTE scheme achieves a slightly lower UAR than the four-class model of the simplified VOTE scheme (49.1% versus 55.4%). This is plausible due to the higher number of classes. Overall, however, the results show a similar performance.

Figure 3.19 shows the class-specific recall for both databases. It is noteworthy that in both schemes, the snoring noises occurring at epiglottis level are the ones which are best differentiated. Furthermore, a good recognition rate of the velar snoring sounds is noticeable. The V-class of the simplified VOTE scheme contains both circular and anterior-posterior vibration patterns at the velum level, so that it can be concluded from the results that especially vibrations in the anterior-posterior direction can be differentiated well. In both schemes, the confusion is highest for snoring at tongue base level. In the simplified VOTE scheme, it is approximately at a random level.

Tables 3.20 and 3.21 show the confusion matrices of both schemes. The diagonal fields contain the correctly recognised percentage for the respective class and thus the class-specific recall. Confusion values >=20% are marked.

In the simplified VOTE scheme velar and oropharyngeal snoring events are most frequently confused, while tongue-type snoring is particularly often misrecognised as oropharyngeal snoring, but also as epiglottis snoring. The ACLTE

ACLTE	**Number**	**Share**
Total number of subjects	343	100%
male	306	89%
female	37	11%
Age (years)	48.6	(Range 20-74)
Data origin (number of subjects)	Number	Share
Essen	278	81%
Munich	24	7%
Halle	22	6%
Valencia	19	6%
Snoring events	Number	Share
Total	1115	100%
A	521	47%
C	172	15%
L	263	24%
T	37	3%
E	122	11%
Events per partition	Number	Share
training	373	33.5%
development	369	33.1%
test	373	33.5%

simpified VOTE	**Number**	**Share**
Total number of subjects	219	100%
male	205	94%
female	14	6%
Age (years)	49.8	(Range 24-78)
Data origin (number of subjects)	Number	Share
Essen	164	75%
Munich	25	11%
Halle	30	14%
Snoring events	Number	Share
Total	828	100%
V	484	58%
O	216	26%
T	39	5%
E	89	11%
Events per partition	Number	Share
training	282	34.1%
development	283	34.2%
test	263	31.8%

Table 3.16: *Size and properties of the MPSSC database and the ACLTE database*

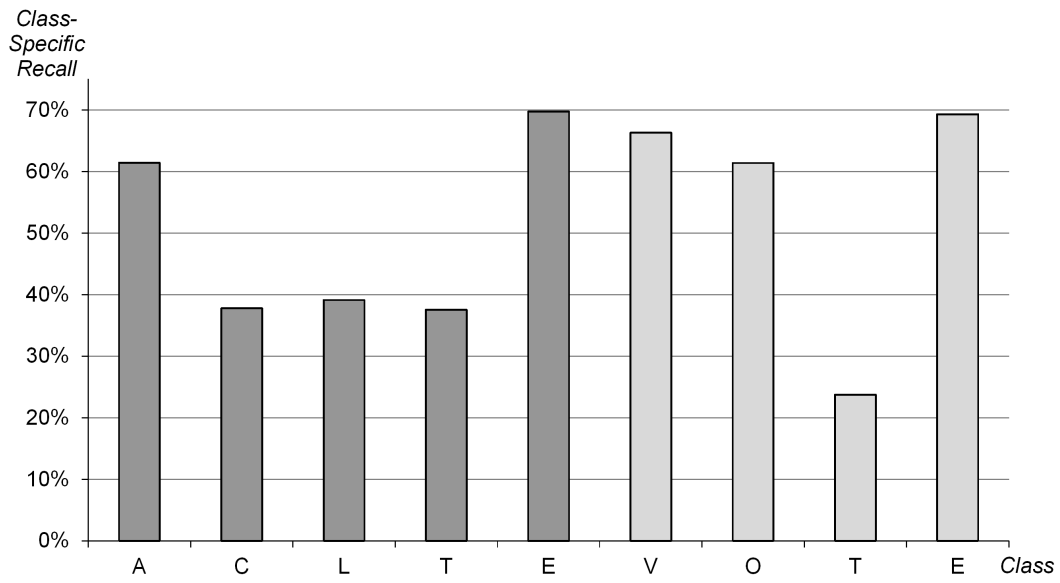

Figure 3.19: *Class-specific recall of the ACLTE versus the simplified VOTE classification, averaged over all permutations*

scheme shows a markable confusion of circular and lateral vibrations at the velum and oropharyngeal level. In addition, circular vibration patterns are frequently misidentified as anterior-posterior velum snoring. Also in this scheme, tongue snoring is confused with (lateral) oropharyngeal snoring and epiglottis snoring.

3.3.4 Discussion

Machine classifiers are 'data hungry'. The larger the available training data sets, the better the generalisation of the patterns present in the data, and consequently, the more accurate the potential classification results. On the other hand, the amount of available training data for this very specific application is limited. Audio and video recordings of DISE studies are only available in limited amounts. The effort for data selection and annotation is considerably high, requiring manual work carried out by ENT experts that cannot be automated. In order to achieve the best possible results despite the limited amount of data, two classification schemes were compared which encode relevant anatomical information in as few different classes as possible.

In addition to the level within the upper respiratory tract at which snoring noise is produced, the orientation of vibration also contains important informa-

tion which is relevant for the therapy decision. At the levels of velopharynx and oropharynx, for example, an isolated vibration of the soft palate or the uvula can be assumed in the case of anterior-posterior oscillation. Circular vibrations indicate the involvement of the pharyngeal walls, while lateral vibrations are highly likely to be caused by the palatine tonsils. This plus of information in the ACLTE scheme compared to the simplified VOTE scheme is achieved at the cost of an additional class, which places higher demands on the machine learning system.

pred. ->	V	O	T	E
V	***66%***	**20%**	5%	9%
O	**23%**	***61%***	5%	11%
T	5%	**50%**	***24%***	**21%**
E	11%	14%	6%	*69%*

Figure 3.20: *Confusion matrix, simplified VOTE, averaged over all permutations*

pred. ->	A	C	L	T	E
A	*61%*	15%	12%	3%	9%
C	**30%**	***38%***	**20%**	5%	6%
L	19%	**25%**	***39%***	9%	9%
T	7%	3%	**20%**	***38%***	**33%**
E	12%	6%	8%	4%	*70%*

Figure 3.21: *Confusion matrix, ACLTE, averaged over all permutations*

In summary, on the basis of the available results, a good differentiation of epiglottis snoring is achieved, and isolated velopharyngeal snoring in anterior-posterior orientation, which is typical for isolated, primary soft palate snoring, is well recognised. This information is valuable for clinical use because these snoring types have to be treated in a distinct way. The differentiation of different vibration patterns at the level of the oropharynx is only moderately successful. There is a need for further research, since the shape of the airway constriction (circular or lateral) is an important information for the therapy decision. Isolated tongue base snoring occurs very rarely in the investigated data and is therefore not well recognised. There is a discrepancy to the prevalence of obstructions at the tongue level, which is diagnosed in a considerably higher proportion of patients. This descripancy may be explained by the fact that the latter predominantly occur as multilevel obstructions with the participation of velo- and oropharyngeal struc-

tures. However, multi-level snoring events were deliberately excluded from the database.

The acoustic properties of the upper respiratory tract provide a possible explanation for the different class-specific recalls. The distance from the point of vibration to the sound emitted into the environment at nostrils and lips acts as an attachment tube that represents an acoustically effective filter, whose frequency response depends on its length and cross-section. The good differentiation of epiglottis snoring can therefore be explained by the fact that the tube has a longer overall length with vibrations at the epiglottis level than with velar or oropharyngeal vibrations. In addition, there is a constriction in the cross-section at the level of the base of the tongue which has a characteristic influence on the transfer function. This is not present in all other classes. Similarly, the more frequent confusion of the classes A, C and L, or V and O, can be explained by the fact that the vibration locations are close to each other. Another argument in favour of this explanation is that the best differentiating feature subgroups describe the spectral signal properties and thus the filter properties of the attachment tube (7). In comparison, the characteristics that describe the properties of the stimulating signal, i. e. the characteristics of the sound source instead of the length and shape of the upper respiratory tract, differentiate less well.

For the databases, only those sound events were used that could be unequivocally and clearly assigned to a class by two annotators. Nevertheless, the transition between different vibration locations and orientations, especially in the oropharyngeal region, is fluent in reality. This can be a further explanation for the confusion of classes C and L in particular. In borderline cases, the trained model may decide differently based on the acoustic properties than a human evaluator who can revert to video images for the decision.

The poor detection rate of the tongue base snoring samples, on the other hand, can be explained by the low number of sound events in this class. With only 3% (ACLTE) and 5% (simplified VOTE) of all events, the T-class is clearly underrepresented This is not surprising, it is known that isolated vibrations and obstructions at the base of the tongue occur comparatively rarely (48). The amount of training data is thus very small for a machine learning task, so the results can be fairly random. This is also supported by the fact that the deviations of the class specific recalls of the T-class across the different permutations are comparatively large.

Chapter 4

Summary

"The great expectation from computerisation was to ... enhance our ability to understand physiologic information in new, clinically useful ways; however, we are still waiting for these hopes to be realised."
(Max Hirshkowitz, Handbook of Clinical Neurology, 2011)

This thesis deals with the analysis of the acoustic phenomenon of snoring. It could be demonstrated that snoring and voice production have a number of similarities. In both cases, an excitation wave is generated within the upper respiratory tract, and the harmonic spectrum of this signal is shaped depending on the shape of the upper airways as the sound travels from its excitation location to the lips and nostrils, where it is emitted into the environment. Although snoring and voice production also differ in a number of aspects, it was an appropriate approach to use analytic strategies and acoustic features known from speech analysis and automatic speech recognition to approach the topic of snore analysis. Indeed, those features that have proven to be well suited for automatic speech processing also showed a high sensitivity both in the detection of snoring events and in the distinction of different snoring types. Especially, Mel-frequency cepstral coefficients (MFCC) and perceptual linear predictive cepstral coefficients (PLPCC) outperformed other acoustic descriptors in the given tasks. All of these features aim to describe the acoustic properties of the upper airways and their sound-shaping characteristics on the excited signal. In contrast, features describing the temporal characteristics of the excitation signal itself were inferior, a fact that is known also in speech signal analysis.

4.1 Research Questions

1. What is snoring? In order to approach this quite universal question, the delimination of snoring and loud breathing using acoustic descriptors has been investigated based on a corpus of nocturnal breathing sounds that had been annotated by a group of blinded raters in an emotionally neutral environment. The events were then labelled as snoring or non-snoring according to the inter-rater agreement (164). The authors of this corpus originally compared the human ratings with psychoacoustic features, which is an obvious idea since the decision of the raters is based on subjective perception. The features used, however, consider volume and therefore the absolute sound pressure level of the signal, a fact that makes the findings difficult to be applied in real-world applications since a calibrated recording equipment and a defined distance between snorer and microphone would be required.

   Hence, in the experiments described in this thesis, a number of SPL-independent acoustic descriptors have been applied to the normalised set of samples, and a very good agreement of machine classification results with the human rating could be achieved. The obtained specificity and sensitivity of both more than 90% clearly exceeds previously published results on this task. The performance is sufficient to be applied to real-world applications and it adds a layer of objectiveness to a task that is in today's applications largely based on the subjective judgement of single raters.

   It must be noted that the proposed solution only addresses the very specific task of distinguishing between snoring and loud breathing. To be useful in universal snoring detection applications, a step of excluding non-respiratory sounds caused by other sources should be applied beforehand. Appropriate strategies for such a task have been described by several research groups and that go beyond the scope of this thesis.

   Care should be exercised when generalising these results, since the corpus used is relatively small for a machine-learning task, and the experiments may be repeated on larger datasets with the aim to confirm and refine the results.

2. Can different snoring types be distinguished acoustically? The results on the classification of different snoring sound excitation locations and orientations show that it is in principle possible to distinguish different sound generation mechanisms by their acoustic properties.

   The performance of the trained machine classification models of 55.4% in a four-class model, and 49.1% in a five-class model, respectively, are encouraging. However, they would still be of limited use in real-world medical applications. The trained models can be helpful as a pre-screening step which,

depending on the detected snoring class and with the knowledge of the class-specific recall, guides the decision which further diagnostic measures must be applied, or can be spared in certain cases to ease strain for patients and to save time and cost.

3. How relevant are the results achieved for therapy decision making? With the highest class-specific recall obtained in snoring originated from the epiglottis (E-type snoring), and the second highest recall achieved in the detection of palatal snoring in anterior-posterior orientation (A-type snoring), two snoring types are addressed that require distinct therapeutic approaches. While A-type snoring can be treated comparatively easily with minimally invasive interventions targeting the velum and/or the uvula, such as soft palate implants or thermocoagulation, E-type snoring requires rather complicated surgical measures, but can also respond well to conservative treatment such as mandibular advancement using oral appliances. In this respect, the trained models can be useful for real-world therapy decision making in a selected subgroup of patients.

   It must be noted that the databases introduced and assessed in this work contain sounds of vibration events in the upper respiratory tract without obstructive disposition. The localisation and orientation of the vibrations have been analysed and models have been trained which permit a direct conclusion with regards to the anatomical causes of snoring. However, they do not reveal any information about an obstructive disposition of the investigated subject. Measurement of OSA severity is not the aim of these models and must be carried out independently using established and well-proven diagnostic methods.

   The VOTE classification according to Kezirian et. al. defines three degrees of airway narrowing (no, partial, complete obstruction). It is their observation that snoring usually occurs during a stage of partial narrowing without complete occlusion. In the symptomatic treatment of primary snoring, information is required as to the location of the snoring sound generation in order to allow targeted therapy. It has not been investigated to what extent vibration patterns correspond with constriction or obstruction patterns occurring during OSA. In order to apply the findings from this thesis as a basis for therapy decisions in OSA patients, research needs to be done regarding the correlation of snoring and occlusion patterns during sleep.

4.2 Limitations and Areas of Future Work

It must be noted that the models and strategies used in this thesis are a simplified approximation to the real world, in particular with regards to the following aspects.

1. The databases analysed in Sections 3.2 and 3.3 are based on recordings taken during DISE examinations. It is an ongoing subject of the scientific debate as to which extent the vibrational and obstructive patterns observed under DISE are similar to those in natural sleep. Here, however, this question might not be of central relevance. The aim of the work presented here is to provide material for the automatic classification of different snoring sound excitation locations by means of machine learning methods. It can be hypothesised that the form of sleep (natural or drug induced) has no meaningful influence on the acoustic characteristics of snoring sounds from different excitation locations. In other words, a velum snoring sounds the same, no matter if it is generated in natural sleep or during drug-induced sedation, as long as it stems from the palate level. In turn, there will be characteristic acoustic properties for the different snoring sound classes, independent of the type of sleep. Given this hypothesis is valid, results based on this database material will be transferable to snoring sound examinations during natural sleep.

2. Only snoring events that could be clearly assigned to one of the defined classes were used for the experiments. In reality, there is no sharp delimination between different snore classes, but it is rather a continuum with transitions from one class to another, which sometimes may difficult to be unequivocally interpreted.

3. It is an inherent limitation of DISE examinations as a diagnostic method that only the upper obstruction level is directly visible. Potential obstruction levels caudally of the visible one might be overlooked. However, it is possible to move the endoscope through the obstruction plane and observe what happens caudally. Advancing the endoscope further than the cranial obstruction plane usually only slightly affects the vibratory pattern and a second caudal plane can be detected. When selecting the snoring events, care was taken to ensure that vibrations occurred in one plane only. Multilocular vibrations were deliberately excluded in order to obtain as clear a sample selection as possible for the acoustic evaluations and machine learning experiments. The same applies to recordings in which an acoustic impairment due to excessive accumulation of mucus and saliva was found. It can be assumed that multilocular vibration patterns have more complex acoustic characteristics than unilocular and also excessive salivation leads to acoustic alteration of the

sounds. Hessel et. al. report that single level obstructions only occur in 35% of patients (48). With a well-trained classifier, multi-level snoring events could be added to the data probing the capability of the classifier models in dealing with this new group of data.

4. Due to the strongly imbalanced nature of the database, the number of actual subjects with tongue-base and epiglottis type snoring is fairly small, leading to potentially unreliable results using machine learning strategies.

Nevertheless, the results from this thesis provide a good basis for future research, in which the trained machine learning models can be applied to independent routine clinical data. By comparing the classification results to expert diagnoses, it can be evaluated how the models perform in difficult, real-world situations. For a broad clinical application of machine-based acoustic analysis of snoring sounds, further differentiated evaluation models will be required in order to capture the phenomenon of snoring in all its complexity.

Adding more subjects to the databases, refining the snoring classes and developing novel descriptors for snoring sound characteristics are areas of future work to further improve classification performance of different types of snoring, with the perspective of complementing DISE as a diagnostic measure in the targeted treatment of sleep-related breathing disorders.

4.3 Conclusion

In the handbook of clinical neurology, issued in 2011, it was stated that "the great expectation from computerisation was to ... enhance our ability to understand physiologic information in new, clinically useful ways; however, we are still waiting for these hopes to be realised" (180).

The research results presented in this thesis show that machine learning methods for the classification of snore data can provide valuable information for the interpretation of the underlying anatomical processes of this frequently occurring acoustic phenomenon. The results are promising to support human experts in the interpretation of sleep data and support diagnosis. Perspectively, findings from this work have the potential to complement DISE investigations or even replace them in selected patients, and thus to decrease the physical strain for the patients undergoing snoring diagnosis and to reduce healthcare cost.

Sleep has been a fascinating field of research over the past centuries, and it will retain is fascination within the ever advancing field of machine learning and artificial intelligence, with the perspective of providing ever more sophisticated tools and methods to support medical diagnosis and therapy decision making.

List of Figures

List of Tables

List of Abbreviations

AASM	American Academy of Sleep Medicine
AHI	Apnoea-Hypopnoea-Index
ASDA	American Sleep Disorders Association
BoAW	Bag-of-Audio-Words
CNN	Convolutional Neural Network
ComParE	Computational Paralinguistics ChallengE
COPD	Chronic Obstructive Pulmonary Disease
CPAP	Continuous Positive Airway Pressure
CQCC	Constant Q Cepstral Coefficient
CSA	Central Sleep Apnoea
dB	Decibel
DISE	Drug Induced Sleep Endoscopy
DNN	Deep Neural Network
ECG	Electrocardiography
EEG	Electroencephalography
ELM	Extreme Learning Machine
EMG	Electromyography
ENT	Ear, Nose and Throat
EOG	Electrooculography
F0	Fundamental Frequency
GMM	Gaussian Mixture Model
HNR	Harmonic-to-Noise Ratio
Hz	Hertz
ICSD	International Classification of Sleep Disorders

kHz	Kilohertz
LLD	Low Level Descriptor
MAD	Mandibular Advancement Device
MFCC	Mel Frequency Cepstral Coefficients
MPSSC	Munich Passau Snore Sound Corpus
ms	Millisecond
openSMILE	Speech & Music Interpretation by Large-Space Extraction
OSA	Obstructive Sleep Apnoea
PLP	Perceptual Linear Prediction
PLPCC	Perceptual Linear Predictive Cepstral Coefficients
PLS	Partial Least Squares
PSG	Polysomnography
RASTA	Relative Spectral Transform
RBF	Radial Basis Function
RDI	Respiratory Disturbance Index
REM	Rapid Eye Movements
RMS	Root Mean Square
RPS	Relative Phase Shift
s	Second
scGAN	semi-supervised conditional Generative Adversarial Network
SER	Spectral Energy Ratio
SP	Sound Pressure
SPL	Sound Pressure Level
STFT	Short-Time Fourier Transformation
SVM	Support Vector Machine
UAR	Unweighted Average Recall
UPPP	Uvulopalatopharyngoplasty
VAS	Visual Analogue Scale
WAR	Weighted Average Recall
WHO	World Health Organization

Bibliography

[1] Z. Zhang, J. Han, K. Qian, C. Janott, Y. Guo, and B. Schuller, "Snoregans: Improving automatic snore sound classification with synthesized data," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 300–310, 2020.

[2] C. Janott, M. Schmitt, C. Heiser, W. Hohenhorst, M. Herzog, M. Carrasco Llatas, W. Hemmert, and B. Schuller, "Vote versus aclte: comparison of two snoring noise classifications using machine learning methods," *HNO*, vol. 67, no. 9, pp. 670–678, 2019.

[3] K. Qian, M. Schmitt, C. Janott, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller, "A bag of wavelet features for snore sound classification." *Annals of biomedical engineering*, vol. 47, pp. 1000–1011, 2019.

[4] C. Janott, M. Schmitt, Y. Zhang, K. Qian, V. Pandit, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller, "Snoring classified: The munich-passau snore sound corpus," *Computers in Biology and Medicine*, vol. 94, pp. 106–118, March 2018.

[5] K. Qian, C. Janott, Z. Zhang, J. Deng, A. Baird, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmer, and B. Schuller, "Teaching machines on snoring: A benchmark on computer auditionfor snore sound excitation localisation," *Achives of Acoustics*, vol. 43, no. 3, pp. 465–475, 2018.

[6] C. Janott, B. Schuller, and C. Heiser, "Acoustic information in snoring noises," *HNO*, vol. 65, no. 2, pp. 107–116, 2017.

[7] K. Qian, C. Janott, V. Pandit, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller, "Classification of the excitation location of snore sounds in the upper airway by acoustic multi-feature analysis," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1731–1741, 2017.

[8] C. Janott, W. Pirsig, and C. Heiser, "Acoustic analysis of snoring sounds," *Somnologie-Schlafforschung und Schlafmedizin*, vol. 18, no. 2, pp. 87–95, 2014.

[9] C. Janott, C. Rohrmeier, M. Schmitt, W. Hemmer, and B. Schuller, "Snoring - an acoustic definition," in *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Berlin, Germany, 2019, pp. 3653–3657.

[10] K. Qian, C. Janott, J. Deng, C. Heiser, W. Hohenhorst, M. Herzog, N. Cummins, and B. Schuller, "Snore sound recognition: On wavelets and classifiers from deep nets to kernels," in *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Seogwipo, South Korea, 2017, pp. 3737–3740.

[11] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 20–24.

[12] M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller, "A bag-of-audio-words approach for snore sounds' excitation localisation," in *Proceedings of ITG Speech Communication*, Paderborn, Germany, 2016, pp. 230–234.

[13] K. Qian, C. Janott, Z. Zhang, C. Heiser *et al.*, "Wavelet features for classification of vote snore sounds," in *2016 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 221–225.

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

[15] M. Ohayon, C. Guilleminault, R. Priest, and M. Caulet, "Snoring and breathing pauses during sleep: telephone interview survey of a united kingdom population sample," *BMJ: British Medical Journal*, vol. 314, pp. 860–863, 1997.

[16] J. M. Parish and P. J. Lyng, "Quality of life in bed partners of patients with obstructive sleep apnea or hypopnea after treatment with continuous positive airway pressure," *Chest*, vol. 124), pp. 942–947, 2003.

[17] E. Seiler, "The incidence of snoring as a sleep problem in parents," *J Roy Coll Gen Practit*, vol. 19, p. 247, 1970.

[18] I. G. Robin, "Snoring," *Proceedings of the Royal Society of Medicine*, vol. 61, no. 6, pp. 575–582, 1968.

[19] "Snoring (editorial comment)," *Canadian Medical Association Journal*, vol. 59, no. 4, pp. 383–384, 1948.

[20] N. Fabricant, "Snoring: a universal nuisance," *Eye, Ear, Nose and Throat Monthly*, vol. 41, p. 56, 1962.

[21] K. J. Finkel, A. C. Searleman, H. Tymkew, C. Y. Tanaka, L. Saager, E. Safer-Zadeh, M. Bottros, J. A. Selvidge, E. Jacobsohn, D. Pulley, S. Duntley, C. Becker, and M. S. Avidan, "Prevalence of undiagnosed obstructive sleep apnea among adult surgical patients in an academic medical center." *Sleep medicine*, vol. 10, pp. 753–758, 2009.

[22] F. Dalmasso and R. Prota, "Snoring: analysis, measurement, clinical implications and applications," *European Respiratory Journal*, vol. 9, no. 1, pp. 146–159, 1996.

[23] E. Kohlschuetter, "Messungen zur festigkeit des schlafes," *Z Rat Med*, vol. 17, pp. 209–253, 1863.

[24] R. MacNish, *The Philosophy of Sleep.* WR M'Phun, 1830.

[25] R. Wittern, *Sleep theories in the antiquity and in the Renaissance.* Fischer Verlag, 1989, pp. 11–22.

[26] A. E. Durham, "The physiology of sleep," *Guy's hospital reports*, vol. 6, pp. 149–173, 1860.

[27] J. Russell, "On sleepiness," *British Medical Journal*, vol. 2, no. 45, pp. 488–490, 1861.

[28] C. M. Morin, P. J. Hauri, and C. A. Espie, "Nonpharmacologic treatment of chronic insomnia. an american academy of sleep medicine review," *Sleep*, no. 22, pp. 1134–1156, 1999.

[29] H. Berger, "Ueber das elektrenkephalogramm des menschen," *Archiv fuer Psychiatrie und Nervenkrankheiten*, vol. 87, no. 1, pp. 527–570, 1929.

[30] A. L. Loomis, E. N. Harvey, and G. A. Hobart, "Cerebral states during sleep, as studied by human brain potentials," *Journal of Experimental Psychology*, vol. 21, no. 2, pp. 127–144, 1937.

[31] E. Aserinsky and N. Kleitman, "Regularly occurring periods of eye motility, and concomitant phenomena, during sleep," *Science*, vol. 118, no. 3062, pp. 273–274, 1953.

[32] A. Rechtschaffen and A. Kales, *A Manual of Standardised Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects.* Los Angeles: Brain Information Service/Brain Research Institute, 1968.

[33] A. A.-Q. K. I. A. Al-Zahrawi, *La Chirurgie d'Abulcasis (Reprint).* Forgotten Books, 2018.

[34] L. Lemnius, *De miracvlis occvltis natvrae, libri III.* Coloniae Agrippinae, 1581.

[35] G. Catlin, *Shut your mouth and save your life.* Barcelona, Spain: Kegan Paul, Trench, Trü & Co LTD, 1891.

[36] A. Dioguardi and M. Al-Halawani, "Oral appliances in obstructive sleep apnea," *Otolaryngologic Clinics of North America*, vol. 49, no. 6, 2016.

[37] M. Puhan, A. Suarez, C. Lo Cascio, A. Zahn, M. Heitz, and O. Braendli, "Didgeridoo playing as alternative treatment for obstructive sleep apnoea syndrome: randomised controlled trial," *BMJ: British Medical Journal*, vol. 332, 2006.

[38] H. Ashrafian, T. Toma, S. Rowland, L. Harling, A. Tan, E. Efthimiou, A. Darzi, and T. Athanasiou, "Bariatric surgery or non-surgical weight loss for obstructive sleep apnoea? a systematic review and comparison of meta-analyses," *Obesity Surgery*, vol. 25, pp. 1239–1250, July 2015.

[39] T. Ikematsu, "Study of snoring, 4th report. therapy (in japanese)," *J. Jap Oto-Rhino-Laryngol*, no. 64, pp. 434–435, 1964.

[40] F. B. Simmons, C. Guilleminault, W. C. Dement, A. G. Tilkian, and M. Hill, "Surgical management of airway obstructions during sleep." *The Laryngoscope*, vol. 87, pp. 326–338, 1977.

[41] C.-H. Chouard, "Proceedings of the 1st international congress on chronic rhonchopathy, paris, 6-8 july 1987." Paris, France: Libbey, 1988, pp. 153–178.

[42] H.-C. Lin, M. Friedman, H.-W. Chang, and B. Gurpinar, "The efficacy of multilevel surgery of the upper airway in adults with obstructive sleep apnea/hypopnea syndrome," *The Laryngoscope*, vol. 118, no. 5, pp. 902–908, 2008.

[43] A. E. Sher, K. B. Schechtman, and J. F. Piccirillo, "The efficacy of surgical modifications of the upper airway in adults with obstructive sleep apnea syndrome." *Sleep*, vol. 19, pp. 156–177, 1996.

[44] R. W. Riley, N. B. Powell, and C. Guilleminault, "Obstructive sleep apnea syndrome: a surgical protocol for dynamic upper airway reconstruction." *Journal of oral and maxillofacial surgery*, vol. 51, pp. 742–727, 1993.

[45] H. J. Xu, R. F. Jia, H. Yu, Z. Gao, W. N. Huang, H. Peng, Y. Yang, and L. Zhang, "Investigation of the Source of Snoring Sound by Drug-Induced Sleep Nasendoscopy," *ORL J. Otorhinolaryngol. Relat. Spec.*, vol. 77, no. 6, pp. 359–365, 2015.

[46] K. Iwanaga, K. Hasegawa, N. Shibata, K. Kawakatsu, Y. Akita, K. Suzuki, M. Yagisawa, and T. Nishimura, "Endoscopic examination of obstructive sleep apnea syndrome patients during drug-induced sleep," *Acta Otolaryngol Suppl*, no. 550, pp. 36–40, 2003.

[47] N. S. Hessel and N. Vries, "Increase of the apnoea-hypopnoea index after uvulopalatopharyngoplasty: analysis of failure," *Clin Otolaryngol Allied Sci*, vol. 29, no. 6, pp. 682–685, 2004.

[48] N. Hessel and N. de Vries, "Diagnostic work-up of socially unacceptable snoring. ii. sleep endoscopy," *European Archives of Oto-Rhino-Laryngology*, vol. 259, pp. 158–161, March 2003.

[49] C. den Herder, D. Kox, H. van Tinteren, and N. de Vries, "Bipolar radiofrequency induced thermotherapy of the tongue base: Its complications, acceptance and effectiveness under local anesthesia," *European Archives of Oto-Rhino-Laryngology*, vol. 263, no. 11, pp. 1031–1040, 2006.

[50] D. Soares, H. Sinawe, A. J. Folbe, G. Yoo, S. Badr, J. A. Rowley, and H. S. Lin, "Lateral oropharyngeal wall and supraglottic airway collapse associated with failure in sleep apnea surgery," *Laryngoscope*, vol. 122, no. 2, pp. 473–479, 2012.

[51] M. J. Sateia, "International classification of sleep disorders-third edition," *Chest*, vol. 146, pp. 1387–1394, 2014.

[52] M. Partinen, "Epidemiology of sleep disorders," *Handbook of clinical neurology*, vol. 98, pp. 275–314, 2011.

[53] M. Partinen and C. Hublin, *Epidemiology of sleep disorders*, 3rd ed. WB Saunders Co, 2000, pp. 558–79.

[54] S. Ram, H. Seirawan, S. K. S. Kumar, and G. T. Clark, "Prevalence and impact of sleep disorders and sleep habits in the united states." *Sleep & breathing*, vol. 14, pp. 63–70, 2010.

[55] Y. Takaesu, Y. Inoue, A. Murakoshi, Y. Komada, A. Otsuka, K. Futenma, and T. Inoue, "Prevalence of circadian rhythm sleep-wake disorders and associated factors in euthymic patients with bipolar disorder." *Plos one*, vol. 11, p. e0159578, 2016.

[56] L. Zhu and P. C. Zee, "Circadian rhythm sleep disorders," *Neurologic clinics*, vol. 30, pp. 1167–1191, 2012.

[57] Y. Dauvilliers and A. Buguet, "Hypersomnia," *Dialogues in clinical neuroscience*, vol. 7, pp. 347–356, 2005.

[58] C. V. Senaratna, J. L. Perret, C. J. Lodge, A. J. Lowe, B. E. Campbell, M. C. Matheson, G. S. Hamilton, and S. C. Dharmage, "Prevalence of obstructive sleep apnea in the general population: A systematic review." *Sleep medicine reviews*, vol. 34, pp. 70–81, 2017.

[59] M. W. Calik, "Update on the treatment of narcolepsy: clinical efficacy of pitolisant," *Nature and science of sleep*, vol. 9, pp. 127–133, 2017.

[60] S. W. Black, A. Yamanaka, and T. S. Kilduff, "Challenges in the development of therapeutics for narcolepsy," *Progress in neurobiology*, vol. 152, pp. 89–113, May 2017.

[61] C. F. Reynolds and R. O'Hara, "Dsm-5 sleep-wake disorders classification: overview for use in clinical practice," *The American journal of psychiatry*, vol. 170, pp. 1099–1101, 2013.

[62] A. V. Benjafield, N. T. Ayas, P. R. Eastwood, R. Heinzer, M. S. M. Ip, M. J. Morrell, C. M. Nunez, S. R. Patel, T. Penzel, J.-L. Pepin, P. E. Peppard, S. Sinha, S. Tufik, K. Valentine, and A. Malhotra, "Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis." *The Lancet. Respiratory medicine*, vol. 7, pp. 687–698, 2019.

[63] P. E. Peppard, T. Young, J. H. Barnet, M. Palta, E. W. Hagen, and K. M. Hla, "Increased prevalence of sleep-disordered breathing in adults," *American Journal of Epidemiology*, vol. 177, no. 9, pp. 1006–1014, 2013.

[64] C. Iber, S. Ancoli-Israel, C. A. L. Jr., and S. F. Quan, "The aasm manual for the scoring of sleep and associated events: rules, terminology and technical specifications. 1st ed." *American Academy of Sleep Medicine*, 2007.

[65] B. Stuck, A. Dreher, C. Heiser, M. Herzog, T. KÃ¼hnel, J. Maurer, H. Pistner, H. Sitter, A. Steffen, and T. Verse, "Sk2 guidelines diagnosis and therapy of snoring in adults compiled by the sleep medicine working group of the german society of otorhinolaryngology, head and neck surgery," *HNO*, vol. 61, no. 11, pp. 944–957, 2013.

[66] C. Dickens, *The Posthumous Papers of the Pickwick Club.* Chapman & Hall, 1837.

[67] A. G. Bickelmann, C. S. Burwell, E. D. Robin, and R. D. Whaley, "Extreme obesity associated with alveolar hypoventilation; a pickwickian syndrome," *The American journal of medicine*, vol. 21, pp. 811–818, 1956.

[68] K. Whyte, M. Allen, A. Jeffrey, G. Gould, and N. Douglas, "Clinical features of the sleep apnoea/hypopnoea syndrome," *The Quarterly journal of medicine*, vol. 72(267), pp. 659–666, July 1989.

[69] M. S. Aldrich, *Sleep Medicine.* Transaction Publishers, 1999.

[70] C. Guilleminault, J. Van Den Hoed, and M. Mitler, *Clinical overview of the sleep apnea syndrome.* Alan R. Liss, Inc., 1978.

[71] G. M. Barthlen, *Schlafdiagnostik (Polysomnographie).* Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 103–111.

[72] G. Mayer, I. Fietze, J. Fischer, T. Penzel, D. Riemann, A. Rodenbeck, H. Sitter, and H. Teschler, *S 3-Leitlinie. Nicht erholsamer Schlaf - Schlafstoerungen.* Springer, 2011.

[73] E. Saifutdinova, D. U. Dudysova, L. Lhotska, V. Gerla, and M. Macas, "Artifact detection in multichannel sleep eeg using random forest classifier," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec 2018, pp. 2803–2805.

[74] M. T. Bianchi, "Sleep devices: wearables and nearables, informational and interventional, consumer and clinical." *Metabolism: clinical and experimental*, vol. 84, pp. 99–108, Jul 2018.

[75] C. Kalkbrenner, R. Brucher, T. Kesztyus, M. Eichenlaub, W. Rottbauer, and D. Scharnbeck, "Automated sleep stage classification based on tracheal body sound and actigraphy." *German medical science : GMS e-journal*, vol. 17, p. Doc02, 2019.

[76] H. Demin, Y. Jingying, W. J. Y. Qingwen, L. Yuhua, and W. Jiangyong, "Determining the site of airway obstruction in obstructive sleep apnea with airway pressure measurements during sleep," *The Laryngoscope*, vol. 112, no. 11, pp. 2081–2085, 2002.

[77] M. Reda, G. J. Gibson, and J. A. Wilson, "Pharyngoesophageal pressure monitoring in sleep apnea syndrome," *Otolaryngology–Head and Neck Surgery*, vol. 125, no. 4, pp. 324–331, 2001.

[78] B. A. Stuck and J. T. Maurer, "Airway evaluation in obstructive sleep apnea," *Sleep Medicine Reviews*, vol. 12, no. 6, pp. 411–436, 2008.

[79] B. Borowiecki, C. P. Pollak, E. D. Weitzman, S. Rakoff, and J. Imperato, "Fibro-optic study of pharyngeal airway during sleep in patients with hypersomnia obstructive sleep-apnea syndrome," *Laryngoscope*, vol. 88, no. 8, pp. 1310–1313, 1978.

[80] C. B. Croft and M. Pringle, "Sleep nasendoscopy: a technique of assessment in snoring and obstructive sleep apnoea," *Clin Otolaryngol Allied Sci*, vol. 16, no. 5, pp. 504–509, 1991.

[81] M. R. El Badawey, G. McKee, H. Marshall, N. Heggie, and J. A. Wilson, "Predictive value of sleep nasendoscopy in the management of habitual snorers," *Annals of Otology, Rhinology & Laryngology*, vol. 112, no. 1, pp. 40–44, 2003.

[82] H. W, *Pitch Determination of Speech Signals*. Springer, 1983.

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

[84] R. A, *Principes de PhonÃ©tique expÃ©rimentale*. Didier, 1904.

[85] K. Wilson, R. A. Stoohs, T. F. Mulrooney, L. J. Johnson, C. Guilleminault, and Z. Huang, "The snoring spectrum: acoustic assessment of snoring sound intensity in 1,139 individuals undergoing polysomnography." *Chest*, vol. 115, pp. 762–70, 1999.

[86] J. Sola-Soler, R. Jane, J. A. Fiz, and J. Morera, "Towards automatic pitch detection in snoring signals," in *Proceedings of the IEEE EMBS 22nd Annual International Conference*, vol. 4. IEEE, 2000, pp. 2974–2976.

[87] R. Jane, J. Sola-Soler, J. A. Fiz, and J. Morera, "Automatic detection of snoring signals: validation with simple snorers and osas patients," in *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No.00CH37143)*, vol. 4, 2000, pp. 3129–3131 vol.4.

[88] J. Sola-Soler, R. Jane, J. A. Fiz, and J. Morera, "Spectral envelope analysis in snoring signals from simple snorers and patients with obstructive sleep apnea," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439)*, vol. 3, 2003, pp. 2527–2530 Vol.3.

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

[90] J. Sola-Soler, R. Jane, J. A. Fiz, and J. Morera, "Variability of snore parameters in time and frequency domains in snoring subjects with and without obstructive sleep apnea," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, Jan 2005, pp. 2583–2586.

[91] A. K. Ng, T. S. Koh, E. Baey, and K. Puvanendran, "Speech-like analysis of snore signals for the detection of obstructive sleep apnea," in *2006 International Conference on Biomedical and Pharmaceutical Engineering*, Dec 2006, pp. 99–103.

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

[93] J. Sola-Soler, J. Morera, R. Jane, and J. Fiz, "Automatic classification of subjects with and without sleep apnea through snoring analysis," in *Proc IEEE 29th Annu Intl Conf Eng Med Bio Soc*, 2007, pp. 6093–6096.

[94] A. S. Karunajeewa, U. R. Abeyratne, and C. Hukins, "Multi-feature snore sound analysis in obstructive sleep apnea–hypopnea syndrome," *Physiological Measurement*, vol. 32, no. 1, pp. 83–97, 2011.

[95] E. Karci, Y. S. Dogrusoz, and T. Ciloglu, "Detection of post apnea sounds and apnea periods from sleep sounds," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2011, pp. 6075–6078.

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

[97] M. E. Tagluk, M. Akin, and N. Sezgin, "Time-frequency analysis of snoring sounds in patients with simple snoring and osas," in *2009 IEEE 17th Signal Processing and Communications Applications Conference*, April 2009, pp. 293–296.

[98] H. Alshaer, F. Rudzicz, T. H. Falk, W. Tseng, and T. D. Bradley, "Classification of vibratory patterns of the upper airway during sleep," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2013, pp. 2080–2083.

[99] M. Kizilkaya, F. Ari, and D. D. Demir gunes, "Detection of sleep apnea with chaotic sound features," in *2013 21st Signal Processing and Communications Applications Conference (SIU)*, April 2013, pp. 1–4.

[100] H. Nakano, K. Hirayama, Y. Sadamitsu, A. Toshimitsu, H. Fujita, S. Shin, and T. Tanigawa, "Monitoring sound to quantify snoring and sleep apnea severity using a smartphone: proof of concept." *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine*, vol. 10, pp. 73–8, 2014.

[101] A. K. Ng and T. S. Koh, "Using psychoacoustics of snoring sounds to screen for obstructive sleep apnea," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2008, pp. 1647–1650.

[102] N. Ben-Israel, A. Tarasiuk, and Y. Zigel, "Nocturnal sound analysis for the diagnosis of obstructive sleep apnea." *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, vol. 2010, pp. 6146–6149, 2010.

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

[104] A. S. Wakwella, U. R. Abeyratne, and C. Hukins, "Snore based systems for the diagnosis of apnoea: a novel feature and its receiver operating characteristics for a full-night clinical database," in *IEEE International Workshop on Biomedical Circuits and Systems, 2004.*, Dec 2004, pp. S2/3–S5.

[105] J. Sola-Soler, R. Jane, J. A. Fiz, and J. Morera, "Formant frequencies of normal breath sounds of snorers may indicate the risk of obstructive sleep apnea syndrome," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2008, pp. 3500–3503.

[106] D. L. Herath, U. R. Abeyratne, and C. Hukins, "Hidden markov modelling of intra-snore episode behavior of acoustic characteristics of obstructive sleep apnea patients." *Physiological measurement*, vol. 36, pp. 2379–404, 2015.

[107] N. Ben-Israel, A. Tarasiuk, and Y. Zigel, "Obstructive apnea hypopnea index estimation by analysis of nocturnal snoring signals in adults." *Sleep*, vol. 35, pp. 1299–305C, 2012.

[108] J. A. Fiz, R. Jane, J. Sola-Soler, J. Abad, M. A. Garcia, and J. Morera, "Continuous analysis and monitoring of snores and their relationship to the apnea-hypopnea index." *The Laryngoscope*, vol. 120, pp. 854–62, 2010.

[109] S. de Silva, U. R. Abeyratne, and C. Hukins, "Impact of gender on snore-based obstructive sleep apnea screening." *Physiological measurement*, vol. 33, pp. 587–601, 2012.

[110] S. de Silva, U. Abeyratne, and C. Hukins, "Gender dependant snore sound based multi feature obstructive sleep apnea screening method." *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, vol. 2012, pp. 6353–6, 2012.

[111] U. R. Abeyratne, S. de Silva, C. Hukins, and B. Duce, "Obstructive sleep apnea screening by integrating snore feature classes." *Physiological measurement*, vol. 34, pp. 99–121, 2013.

[112] R. Plywaczewski, P. Bielen, M. Bednarek, L. Jonczak, D. Gorecka, and P. Sliwinski, "Influence of neck circumference and body mass index on obstructive sleep apnoea severity in males," *Pneumonologia i alergologia polska*, vol. 76, pp. 313–320, 2008.

[113] Y. Kawaguchi, S. Fukumoto, M. Inaba, H. Koyama, T. Shoji, S. Shoji, and Y. Nishizawa, "Different impacts of neck circumference and visceral obesity

on the severity of obstructive sleep apnea syndrome." *Obesity*, vol. 19, pp. 276–282, 2011.

[114] J. R. Perez-Padilla, E. Slawinski, L. M. Difrancesco, R. R. Feige, J. E. Remmers, and W. A. Whitelaw, "Characteristics of the snoring noise in patients with and without occlusive sleep apnea." *The American review of respiratory disease*, vol. 147, pp. 635–644, 1993.

[115] Y. Yang, Y. Qin, W. Haung, H. Peng, and H. Xu, "Acoustic characteristics of snoring sound in patients with obstructive sleep apnea hypopnea syndrome." *Lin Chung Er Bi Yan Hou Tou Jing Wai Ke Za Zhi*, vol. 26, no. 8, pp. 360–363, 2012.

[116] A. Yadollahi, E. Giannouli, and Z. Moussavi, "Sleep apnea monitoring and diagnosis based on pulse oximetry and tracheal sound signals." *Medical & biological engineering & computing*, vol. 48, pp. 1087–97, 2010.

[117] V. Hoffstein, S. Mateika, and D. Anderson, "Snoring: is it in the ear of the beholder?" *Sleep*, vol. 17, pp. 522–6, 1994.

[118] P. P. Caffier, J. C. Berl, A. Muggli, A. Reinhardt, A. Jakob, M. Moser, I. Fietze, H. Scherer, and M. Holzl, "Snoring noise pollution–the need for objective quantification of annoyance, regulatory guidelines and mandatory therapy for snoring." *Physiological measurement*, vol. 28, pp. 25–40, 2007.

[119] C. Rohrmeier, M. Herzog, F. Haubner, and T. S. Kuehnel, "The annoyance of snoring and psychoacoustic parameters: a step towards an objective measurement." *European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery*, vol. 269, pp. 1537–43, 2012.

[120] R. Fischer, T. S. Kuehnel, A.-K. Merz, T. Ettl, M. Herzog, and C. Rohrmeier, "Calculating annoyance: an option to proof efficacy in ent treatment of snoring?" *European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery*, vol. 273, pp. 4607–4613, 2016.

[121] J. Schafer, "How can one recognize a velum snorer?" *Laryngorhinootologie*, vol. 68, no. 5, pp. 290–294, 1989.

[122] S. J. Quinn, L. Huang, P. Ellis, and J. Williams, "The differentiation of snoring mechanisms using sound analysis," *Clin Otolaryngol Allied Sci*, vol. 21, no. 2, pp. 119–123, 1996.

[123] P. Hill, B. Lee, J. Osborne, and E. Osman, "Palatal snoring identified by acoustic crest factor analysis," *Physiological Measurement*, vol. 20, no. 2, pp. 167–174, 1999.

[124] P. Hill, E. Osman, J. Osborne, and B. Lee, "Changes in snoring during natural sleep identified by acoustic crest factor analysis at different times of night," *Clinical Otolaryngology & Allied Sciences*, vol. 25, no. 6, pp. 507–510, 2000.

[125] R. J. Beeton, I. Wells, P. Ebden, H. Whittet, and J. Clarke, "Snore site discrimination using statistical moments of free field snoring sounds recorded during sleep nasendoscopy," *Physiological Measurement*, vol. 28, no. 10, pp. 1225–1236, 2007.

[126] S. Agrawal, P. Stone, K. McGuinness, J. Morris, and A. Camilleri, "Sound frequency analysis and the site of snoring in natural and induced sleep," *Clinical Otolaryngology & Allied Sciences*, vol. 27, no. 3, pp. 162–166, 2002.

[127] M. Herzog, S. Plossl, A. Glien, B. Herzog, C. Rohrmeier, T. Kuhnel, S. Plontke, and P. Kellner, "Evaluation of acoustic characteristics of snoring sounds obtained during drug-induced sleep endoscopy," *Sleep Breath*, vol. 19, no. 3, pp. 1011–1019, 2015.

[128] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proceedings of INTERSPEECH*, Lyon, France, 2013, pp. 148–152.

[129] F. Eyben, "Real-time speech and music classification by large audio feature space extraction," 2015.

[130] D. Tavarez, X. Sarasola, A. Alonso, J. Sanchez, L. Serrano, E. Navas, and I. HernÃ¡ez, "Exploring fusion methods and feature space for the classification of paralinguistic information," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3517–3521.

[131] T. L. Nwe, H. D. Tran, W. Z. T. Ng, and B. Ma, "An integrated solution for snoring sound classification using bhattacharyya distance based gmm

supervectors with svm, feature selection with random forest and spectrogram with cnn," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3467–3471.

[132] M. V. Achuth Rao, S. Yadav, and P. K. Ghosh, "A dual source-filter model of snore audio for snorer group classification," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3502–3506.

[133] G. Gosztolya, R. Busa-Fekete, T. Grosz, and L. Toth, "Dnn-based feature extraction and classifier combination for child-directed speech, cold and snoring identification," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3522–3526.

[134] H. Kaya and A. A. Karpov, "Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: Snoring, addressee and cold," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3527–3531.

[135] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3512–3516.

[136] M. Freitag, S. Amiriparian, N. Cummins, M. Gerczuk, and B. Schuller, "An end-to-evolution hybrid approach for snore sound classification," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3507–3511.

[137] M. Herzog, A. Schmidt, T. Bremert, B. Herzog, W. Hosemann, and H. Kaftan, "Analysed snoring sounds correlate to obstructive sleep disordered breathing." *European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery*, vol. 265, pp. 105–113, 2008.

[138] T. M. Jones, A. C. Swift, P. M. A. Calverley, M. S. Ho, and J. E. Earis, "Acoustic analysis of snoring before and after palatal surgery," *The European respiratory journal*, vol. 25, pp. 1044–1049, 2005.

[139] M. Cremer, Lothar & Hubert, *Vorlesungen Ã¼ber Technische Akustik*. Springer-Verlag, 1990.

[140] G. H. Rene Flosdorff, *Elektrische Energieverteilung*. Vieweg+Teubner, 2005.

[141] J. A. Fiz, J. Abad, R. Jane, M. Riera, M. A. Mananas, P. Caminal, D. Rodenstein, and J. Morera, "Acoustic analysis of snoring sound in patients with simple snoring and obstructive sleep apnoea." *The European respiratory journal*, vol. 9, pp. 2365–2370, 1996.

[142] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. Wiley, 1996.

[143] G. Fant, *Acoustic theory of speech production*. Mouton, 1970.

[144] A. Cohen and A. Lieberman, "Analysis and classification of snoring signals," in *Proceedings of the IEEE ICASSP 1986*. Tokyo, Japan: IEEE, 1986, pp. 693–696.

[145] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete Time Processing of Speech Signals*. Prentice Hall PTR, 1993.

[146] E. Dafna, A. Tarasiuk, and Y. Zigel, "Automatic detection of whole night snoring events using non-contact microphone." *PloS one*, vol. 8, p. e84139, 2013.

[147] L. L. Beranek, *Acoustic measurements*. J. Wiley, 1949.

[148] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[149] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[150] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[151] R. R. Coifman, Y. Meyer, and V. Wickerhauser, "Wavelet analysis and signal processing," in *Wavelets and their Applications*. Sudbury: MA: Jones and Barlett, 1992, pp. 153–178.

[152] W. Pschyrembel, *Pschyrembel Klinisches Wörterbuch*. De Gruyter, Berlin, 2007.

[153] M. Norman, S. Middleton, O. Erskine, P. Middleton, J. Wheatley, and C. Sullivan, "Validation of the sonomat: a contactless monitoring system used for the diagnosis of sleep disordered breathing," *Sleep*, vol. 37, no. 9, pp. 1477–1487, 2014.

[154] M. Thorpy, *The international classification of sleep disorders: diagnostic and coding manual*. Lawrence KS ed Allen Press, USA, 1990.

[155] V. Swarnkar, U. Abeyratne, and R. Sharan, "Automatic picking of snore events from overnight breath sound recordings," in *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Seogwipo, South Korea: IEEE, July 2017, pp. 2822–2825.

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

[157] A. K. Ng, T. San Koh, K. Puvanendran, and U. Ranjith Abeyratne, "Snore signal enhancement and activity detection via translation-invariant wavelet transform." *IEEE transactions on bio-medical engineering*, vol. 55, pp. 2332–42, 2008.

[158] A. S. Karunajeewa, U. R. Abeyratne, and C. Hukins, "Silence-breathing-snore classification from snore-related sounds," *Physiological Measurement*, vol. 29, no. 2, pp. 227–243, 2008.

[159] A. Yadollahi and Z. Moussavi, "Automatic breath and snore sounds classification from tracheal and ambient sounds recordings," *Medical Engineering & Physics*, vol. 32, no. 9, pp. 985–990, 2010.

[160] A. Azarbarzin and Z. Moussavi, "Automatic and unsupervised snore sound extraction from respiratory sound signals," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 5, pp. 1156–1162, 2011.

[161] K. Qian, Y. Fang, Z. Xu, and H. Xu, "Comparison of two acoustic features for classification of different snore signals," *Chinese Journal of Electron Devices*, vol. 36, no. 4, pp. 455–459, 2013.

[162] K. Qian, Z. Xu, H. Xu, and B. P. Ng, "Automatic detection of inspiration related snoring signals from original audio recording," in *Proceedings of the IEEE ChinaSIP 2014*. IEEE, 2014, pp. 95–99.

[163] K. Qian, Z. Xu, H. Xu, Y. Wu, and Z. Zhao, "Automatic detection, segmentation and classification of snore related signals from overnight audio recording," *IET Signal Processing*, vol. 9, no. 1, pp. 21–29, 2015.

[164] C. Rohrmeier, M. Herzog, T. Ettl, and T. Kuehnel, "Distinguishing snoring sounds from breath sounds: a straightforward matter?" *Sleep and Breathing*, vol. 18, no. 1, pp. 169–176, 2014.

[165] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*. Florence, Italy: ACM, 2010, pp. 1459–1462.

[166] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in

Proceedings of the 21st ACM International Conference on Multimedia. Barcelona, Spain: ACM, 2013, pp. 835–838.

[167] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. Scherer, "On the acoustics of emotion in audio: What speech, music and sound have in common," *Frontiers in Emotion Science*, vol. 4, pp. 1–12, 2013.

[168] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. August, pp. 1871–1874, 2008.

[169] H. Fastl and E. Zwicker, *Psychoacoustics, Facts and Models*. Springer-Verlag, 2007.

[170] V. Abdullah, Y. Wing, and C. van Hasselt, "Video sleep nasendoscopy: the hong kong experience," *Otolaryngologic Clinics of North America*, vol. 36, no. 3, pp. 461–471, 2003.

[171] M. Friedman, H. Ibrahim, and L. Bass, "Clinical staging for sleep-disordered breathing," *Otolaryngology and Head and Neck Surgery*, vol. 127, no. 1, pp. 13–21, 2002.

[172] C. Vicini, A. De Vito, M. Benazzo, S. Frassineti, A. Campanini, P. Frasconi, and E. Mira, "The nose oropharynx hypopharynx and larynx (NOHL) classification: a new system of diagnostic standardized examination for OSAHS patients," *European Archives of Oto-Rhino-Laryngology*, vol. 269, no. 4, pp. 1297–1300, 2012.

[173] E. J. Kezirian, W. Hohenhorst, and N. de Vries, "Drug-induced sleep endoscopy: the vote classification," *European Archives of Oto-Rhino-Laryngology*, vol. 268, no. 8, pp. 1233–1236, 2011.

[174] N. Charakorn and E. J. Kezirian, "Drug-Induced Sleep Endoscopy," *Otolaryngologic Clinics of North America*, vol. 49, no. 6, pp. 1359–1372, 2016.

[175] A. V. Vroegop, O. M. Vanderveken, K. Wouters, E. Hamans, M. Dieltjens, N. R. Michels, W. Hohenhorst, E. J. Kezirian, B. T. Kotecha, N. de Vries *et al.*, "Observer variation in drug-induced sleep endoscopy: experienced versus nonexperienced ear, nose, and throat surgeons," *Sleep*, vol. 36, no. 6, p. 947, 2013.

[176] J. Fiz and R. Jane, "Snoring analysis. a complex question." *Journal of Sleep Disorders: Treatment and Care*, no. 1, 2012.

[177] H. Peng, H. Xu, Z. Xu, W. Huang, R. Jia, H. Yu, Z. Zhao, J. Wang, Z. Gao, Q. Zhang, and W. Huang, "Acoustic analysis of snoring sounds originating from different sources determined by drug-induced sleep endoscopy," *Acta Otolaryngol.*, pp. 1–5, Mar 2017.

[178] S. K. Koo, S. B. Kwon, Y. J. Kim, J. I. S. Moon, Y. J. Kim, and S. H. Jung, "Acoustic analysis of snoring sounds recorded with a smartphone according to obstruction site in OSAS patients," *European Archives of Oto-Rhino-Laryngology*, vol. 274, no. 3, pp. 1735–1740, 2017.

[179] A. Steffen, H. Frenzel, B. Wollenberg, and I. R. Konig, "Patient selection for upper airway stimulation: is concentric collapse in sleep endoscopy predictable?" *Sleep & breathing = Schlaf & Atmung*, vol. 19, pp. 1373–6, 2015.

[180] M. Hirshkowitz and A. Sharafkhaneh, "Normal sleep-recording and scoring techniques." *Handbook of clinical neurology*, vol. 98, pp. 29–43, 2011.