Snoring classified: The Munich-Passau Snore Sound Corpus

Christoph Janott,*, Maximilian Schmitt, Yue Zhang, Kun Qian, Vedhas Pandit, Zixing Zhang, Clemens Heiser, Winfried Hohenhorst, Michael Herzog, Werner Hemmert, Björn Schuller

ARTICLE INFO

Keywords:
Obstructive Sleep Apnea
Primary snoring
Snore sound classification
Machine learning
Drug-Induced Sleep Endoscopy

ABSTRACT

Objective: Snoring can be excited in different locations within the upper airways during sleep. It was hypothesised that the excitation locations are correlated with distinct acoustic characteristics of the snoring noise. To verify this hypothesis, a database of snore sounds is developed, labelled with the location of sound excitation.

Methods: Video and audio recordings taken during drug induced sleep endoscopy (DISE) examinations from three medical centres have been semi-automatically screened for snore events, which subsequently have been classified by ENT experts into four classes based on the VOTE classification. The resulting dataset containing 828 snore events from 219 subjects has been split into Train, Development, and Test sets. An SVM classifier has been trained using low level descriptors (LLDs) related to energy, spectral features, mel frequency cepstral coefficients (MFCC), formants, voicing, harmonic-to-noise ratio (HNR), spectral harmonicity, pitch, and microprosodic features.

Results: An unweighted average recall (UAR) of 55.8% could be achieved using the full set of LLDs including formants. Best performing subset is the MFCC-related set of LLDs. A strong difference in performance could be observed between the permutations of train, development, and test partition, which may be caused by the relatively low number of subjects included in the smaller classes of the strongly unbalanced data set.

Conclusion: A database of snoring sounds is presented which are classified according to their sound excitation location based on objective criteria and verifiable video material. With the database, it could be demonstrated that machine classifiers can distinguish different excitation location of snoring sounds in the upper airway based on acoustic parameters.

1. Introduction

1.1. Background

Approximately one out of three adults in the western world snores [1], [2]. Snoring is excited by the inspiratory airflow causing soft tissue structures in the upper airways (UA) to vibrate [3]. Primary snoring (simple snoring) is characterized by the absence of apnoeic or hypopnoeic episodes. In contrast, Obstructive Sleep Apnea (OSA) is characterized by repeated episodes of decreased (hypopnea) or completely halted (apnea) airflow despite an ongoing effort to breathe. The averaged number of apnoeas and hypopnoeas occurring per hour of sleep is measured by the Apnea-Hypopnea-Index (AHI), which is a measure for the severity of the OSA syndrome. OSA is a serious health condition affecting 13% of men and 6% of women in the US population [4]. Symptoms associated with OSA include daytime sleepiness, excessive fatigue, and morning headache. It is an independent risk factor for cardiovascular diseases such as hypertension and myocardial infarction [1]. Loud snoring is a typical symptom associated with OSA in more than 80% of patients [5,6].

* Corresponding author.
E-mail address: c.janott@gmx.net (C. Janott).

https://doi.org/10.1016/j.compbiomed.2018.01.007
Received 12 November 2017; Received in revised form 19 January 2018; Accepted 19 January 2018

0010-4825/© 2018 Elsevier Ltd. All rights reserved.
While primary snoring 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 [7]. Further, snoring can be a reason for social disturbance, e.g., when sleeping in dormitories or camping sites, and it can affect partnerships. It is frequently mentioned as a reason for sleeping in different bedrooms and even cited as unreasonable behaviour in divorce proceedings, although reliable statistics on these aspects are not available.

Standard for the treatment of OSA is continuous positive airway pressure (CPAP), applied through a mask worn during sleep. While this treatment is highly effective, the long term compliance is only moderate.

Numerous other methods for the treatment of snoring and OSA have been developed, ranging from established conservative measures such as oral appliances to advance the mandible during sleep [8], to rather unusual methods such as didgeridoo playing as a means of oral musculature training [9]. Weight loss effectively reduces the severity of snoring and OSA in the majority of overweight patients [10].

Surgical methods to treat snoring and OSA include for example tonsillectomy, uvulopalatopharyngoplasty (UPPP), soft palate stiffening, tongue base reduction or tongue base suspension, and hypoglossal nerve stimulation. Some surgical treatment methods are highly effective, others are of limited efficacy or their evidence is limited [11].

The key to improved success rates of surgical measures is careful patient selection [12]. 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 [13,14] and has less effect when snoring is predominantly generated by the tongue base or the posterior pharyngeal walls [15]. Vice versa, procedures primarily targeting the hypopharyngeal area might not be the first choice of treatment in purely palatal snorers [16,17]. Therefore, knowledge of the mechanism and location of obstruction and snore sound excitation within the UA can be helpful for targeted interventions.

1.2. Earlier research

Extensive research has been carried out on the acoustics of sleep related breathing disorders since the 1980s. The acoustic properties of snoring sounds are well described, a comprehensive literature analysis has been published in Ref. [18]. Acoustic parameters of snoring sounds have been used in clinical trials to objectify the success of surgical snoring interventions and to assess the effectiveness of other diagnostic methods for the prediction of surgical outcomes [19,20].

Snoring sounds have been assessed for their suitability as diagnostic tools. The majority of the work pursued the goal to distinguish between primary snoring and OSA of different levels of severity, as well as the detection of apnoeic events, in order to make suitable screening systems available that are based purely or mainly on acoustic information. A literature review of publications on acoustic identification of presence and severity of OSA can be found in Ref. [21].

Less has been published on the identification of the location of the sound generation. In a literature research by the authors, eight papers have been identified on this subject, for details refer to [21].

Our group has investigated the classification of snore sound excitation locations using machine learning methods. The work was based on a predecessor of the Munich Passau Snore Sound Corpus, comprising snore sounds from 24 subjects labelled according to the simplified VOTE classification. Using a feature set based on wavelet transform with a support vector machine classifier, an unweighted average recall (UAR) of more than 70% could be achieved in this four class problem [22,23].

Applying an unsupervised feature learning approach clustering feature values within a given time-segment into acoustic words (bags-of-audio-words) based on wavelet features, formants, and MFCC, we could achieve a UAR of almost 80% [24].

The Munich Passau Snore Sound Corpus was first introduced as a sub-challenge in the INTERSPEECH 2017 Computational Paralinguistics Challenge [25]. It is freely available to researchers for scientific purposes.

1.3. Diagnostic standards

The gold standard for the diagnosis of OSA is polysomnography (PSG), a multichannel recording of physiological parameters during natural sleep [11]. In most cases, PSG is recorded in a sleep laboratory. Cardiorespiratory screening using portable devices with fewer recording channels are often used alternatively or as additional measures. In the past years, an increasing number of methods and applications have been researched and developed for sleep monitoring, e.g., [26,27]. The diagnostic accuracy of these methods, however, has not yet been validated or proven in clinical trials.

PSG and cardiorespiratory screening provide a reliable diagnosis as towards the type and severity of the sleep related breathing disorder. However, it is of very limited use to identify its underlying mechanisms.

A diagnostic procedure that has been established for the evaluation of obstruction and vibration locations and mechanisms in the UA is Drug Induced Sleep Endoscopy (DISE). It was developed in the late 1980s and first described by Croft and Pringle in 1991 [28]. In DISE, the patient is put into artificial sleep by means of titrated application of narcotics. When the patient is in an unconscious stage, the UA are intranasally inspected by means of a flexible nasopharyngoscope. Video and audio signals are often recorded for documentation or later investigation.

DISE is increasingly used by sleep surgeons and appreciated as a useful tool to identify the location of vibration and obstruction. However, it has a number of disadvantages: DISE is cost intensive, as it requires the attendance of personnel and appropriate equipment for safe administration and monitoring of sedation, as well as endoscopic equipment. Further, it is time consuming, a DISE investigation typically requires more than 20 min overall. Also, it cannot be performed during natural sleep, as the introduction of the endoscope would cause the patient to wake up.

1.4. Aim

It is therefore 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 snore sounds.

It was hypothesised that different excitation locations of snore sounds are correlated with distinct acoustic characteristics. The snore signal is shaped by a transfer function which depends on the cross-sectional profile of the UA from the excitation location to the nose and mouth opening [29]. The resulting sound is therefore a function of the excited wave and the shape of the upper airway. Different snoring generation mechanisms and related excitation locations go along with typical lengths of the acoustically effective part of the UA, therefore carrying characteristic acoustic properties which allow a classification of defined classes of snoring [3,30].

In order to test this hypothesis, the Munich Passau Snore Sound Corpus (MPSSC) has been developed.

For the first time, we present a database of snore sounds labelled by their class of excitation location. Annotation of the snore events has been carried out based on simultaneous endoscopic video recordings of the upper airways and is therefore objective and independently verifiable. To our knowledge, no such database is publicly available to date. On this basis, machine learning strategies can be applied to train classifiers to distinguish snore sounds according to their source of excitation. Perspectively, these methods have the potential to complement DISE investigations or even replace them by acoustic analysis of snore sounds in selected patients, and thus to decrease the physical strain for the patients undergoing snoring diagnosis and to reduce healthcare cost.

In contrast to earlier work, we do not aim 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.
1.5. Structure of this paper

This paper is structured as follows: the process of data collection, audio pre-processing, event selection, classification and labelling of the data is outlined in chapter 2. In chapters 3 and 4, the resulting properties of the database and our classification experiments are described. Results are summarised in chapter 5, discussion and a conclusion follow in chapters 6 and 7.

2. Materials and methods

2.1. Definitions

According to the International Classification of Sleep Disorders (ICSD-3), 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 [31].

A definition of snoring based on concrete acoustic parameters, and its delimitation to other nocturnal breathing sounds, does not yet exist [18]. The distinction has so far been based exclusively on the subjective assessment of human listeners. In a study by Rohrmeier et al., in which 25 human listeners were tasked to classify acoustic sequences in respiratory sounds and snoring, 16% of the events could not be assigned clearly [32].

In this work, snoring shall be distinguished from breathing sounds by the existence of predominant tonal components in the resulting sound [33,34]. Fig. 1 shows typical examples of the time signal of a breathing sound and a snoring sound.

For the sake of consistent nomenclature, in this paper, the individual sound which is produced within one breath is called a snoring event, while a sequence of snoring events (a period of continuous snoring) is referred to as a snoring episode.

2.2. Data collection

The database is derived from original endoscopic recordings of DISE examinations. 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 polysomnography (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 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.
- Alfred 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 1 shows the equipment used for recording of the DISE videos. As an example, Fig. 2 displays screenshots taken from DISE recordings of typical snoring events. The upper left image (V) shows a vibrating velum at the palatal level. In the upper right image (O), the oropharyngeal level can be seen with vibrating palatine tonsils. In the lower left image (V), the tongue base vibrates against the posterior pharyngeal wall. And the lower right image (E) shows a vibrating epiglottis. The white arrows in the images mark the respective vibrating structures.

2.3. 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. Octave 3.6.1 with GCC 4.6.2 was used as programming platform. The absolute value of the signal amplitude was averaged in 10 ms segments with no overlap and the background noise level was determined by means of a 1024-step histogram averaging 10 s segments. 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). Fig. 3 illustrates the segmentation procedure. All described values were experimentally optimised during the algorithm development based on a subset of the DISE audio recordings.

2.4. Pre-selection: snore and non-snore sound events

In a next step, an experienced human listener (the first author) listened to all selected events and classified them manually as either pure snoring (snore) or other sounds (non snore). Also, those events that contained a snore event but 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 snore events that were overdriven or distorted by disturbances in the recording chain such as slack joints.

The criteria to include a sound event in the snore group were therefore subjective. A quite rigid standard was applied to pass as snore sound.

Table 1

<table>
<thead>
<tr>
<th>Centre</th>
<th>Recording equipment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Munich</td>
<td>Storz flexible nasopharyngoscope</td>
</tr>
<tr>
<td></td>
<td>Storz Telepack X recording system</td>
</tr>
<tr>
<td></td>
<td>Headset microphone</td>
</tr>
<tr>
<td>Essen</td>
<td>Olympus flexible nasopharyngoscope</td>
</tr>
<tr>
<td></td>
<td>Rehder/Partner rp90ene recording system</td>
</tr>
<tr>
<td></td>
<td>Handheld or headset or forehead-mounted microphone</td>
</tr>
<tr>
<td>Halle</td>
<td>Storz flexible nasopharyngoscope</td>
</tr>
<tr>
<td></td>
<td>Storz AIDA recording system</td>
</tr>
<tr>
<td></td>
<td>Stand-mounted microphone</td>
</tr>
</tbody>
</table>

Fig. 1. 300 ms section of the time domain signal of a breathing event (upper diagram) and a velar snoring event (lower diagram). The periodical waves of the tonal components representing the fundamental frequency of the snoring sound can be clearly seen in the lower diagram, whereas the breathing sound in the upper diagram has a predominantly noisy character.
When in doubt, a sound event was rather excluded from the snore 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 snore signal,
- all of the snore events were polluted by non-static background sounds, overdriven or distorted.

While the material from Halle/Saale and from Munich was already pre-selected (by MH and CH) for videos containing snore episodes, the material from Essen had not been pre-screened. Therefore, the yield of subjects with snore events from the Essen material was significantly lower than from the other two centres.

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

Several algorithms have been described and tested for automated classification of snore and non-snore sounds [35–37]. Although some prove high sensitivity and specificity of more than 97% [38], for our database, a ‘human’ classification process was preferred over automated algorithms for the following reasons: 1. The criteria to define a sound as a snore sound are not standardised by objective acoustic parameters, and none of the snore/non-snore classification algorithms that are known to us refer to commonly accepted selection schemes. From our raw material, snore sounds with overlaid non-stationary disturbing noise needed to be excluded. None of the existing algorithms are described as to their sensitivity and specificity for excluding snore sounds with such artefacts.

### 2.5. Classification

Several schemes have been suggested for the classification of the location of snoring noise and obstructions [13,39–41]. A widely used scheme is the VOTE classification, introduced by Kezirian et al., in 2011 [42]. The VOTE classification distinguishes four structures that can be involved in airway narrowing and obstruction [43]:

- 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 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.

Fig. 4 illustrates the corresponding locations within the upper airways.

In addition, the VOTE classification contains a description of the shape of obstruction (anteroposterior, lateral, and concentric), and a qualitative assessment of the degree of airway narrowing (no, partial or complete obstruction). The VOTE classification as introduced by Kezirian et al. is primarily used to describe airway narrowing and obstruction in OSA patients.

![Fig. 4. Areas in the upper airways according to the VOTE classification.](image)
For our research, we introduce a simplified version of the VOTE classification in order to describe the location of vibration of the soft tissue generating snoring noise. We do not distinguish between different levels of airway narrowing. Furthermore, only events that create vibration of the airway structures are of interest. Therefore, our selection of samples is limited to partial narrowing according to the VOTE classification. Further, we do not distinguish different obstructive patterns. This leads us to a four-class classification described by the labels V, O, T, and E.

### 2.6. Annotation to VOTE classes

For all selected sound events, the respective video files were watched by two experienced experts (CH and CJ). Based on the video findings, each snore event was assigned 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 UA is not always limited to a single level according to our classification. For this database, we excluded events in which vibration was not clearly limited to one of our four defined levels. However, during one DISE examination session, the same subject might show vibration patterns at different levels in different snore events, but limited to one vibration level per event. In this case, snore events were included and labelled accordingly. For example, one subject showed distinct velum-level snoring when the mandible was advanced using an Esmarch-maneuver. Without this maneuver, snoring originated from the epiglottis-level. Consequently, the database contains both V-type snoring and E-type snoring events from this very subject.

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

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. Only clear vibration patterns were selected. Those which where unclear, where multiple vibration sites where simultaneously involved, and snoring events with an obstructive event, were excluded.

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 snore 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 snore 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. Fig. 5 shows a summary of the selection steps taken and the number of subjects per centre included in the database after identification of snore 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 (WH). 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 CH: O-type snoring; Annotator WH: probably O-type, but not certain). Based on this sample of 18% of subjects from the total set, the intrarater-reliability according to Cohen’s Kappa is $\kappa = 0.96$.\(^1\)

Interobserver agreement for evaluation of DISE videos was studied by Vroegop et al., in 2013 [44]. For the level of collapse, intrarater 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 vibration, and they used a classification additionally comprising the hypopharynx as a fifth level), it is safe to conclude that the intrarater-agreement in our study offers a very high level of confidence in the annotation. Reasons for this comparatively good agreement can be that all annotators are highly

---

\(^1\) Cohen’s Kappa was calculated using ReCal2 0.1, dfreelon.org.
experienced in the evaluation of DISE recordings, and that events with unclear level of vibration have already been excluded in a previous step.

2.7. Partitioning

In order to prepare the corpus for machine learning experiments, we stratified the data into a train, a development (dev), and a test partition. In order to create subject disjunctive partitions, assignment is made based on subject, not event (i.e., all snore events from a subject are assigned to the same partition). To obtain this, we first sorted the subjects 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. Fig. 6 illustrates this process. A two-tailed, unpaired t-test confirmed no significant differences between the partitions for age, gender, center or class ($p > 0.05$). Table 2 shows the resulting number of events per class and partition. Since the number of snore events per subject differ, the partitions contain different numbers of snore 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 snore 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.

3. Database properties

The resulting database contains audio samples of 828 snore 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.

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 Fig. 7B).\(^2\)

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 Fig. 7A.

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

Table 3 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 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 our classification) occurred in only 2% of patients, whereas single level $V$ and $O$ type events occurred in 22% of patients, thus 10 times as often [15]. Other researchers come to similar results [45].

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 significant 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.

Fig. 8 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. We performed a machine learning experiment in the same setup as described in the following chapter, but using the

---

\(^2\) All probability values calculated with two-tailed, unpaired t-test.
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 snore sounds indeed carry center-specific information. In order to evaluate the impact on the performance of our classifier setup to distinguish snoring noise, we performed the machine learning experiments exactly as described in the following chapter, 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 vs 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 and to create an even more stable classifier model.

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

4. Machine learning experiments

4.1. COMPARE baseline and challenge contributions

The data of the MPSSC was introduced as the Snore Sub-Challenge in the INTERSPEECH 2017 Computational Paralinguistics Challenge (COMPARE). In this context, members of our group performed baseline experiments using the official INTERSPEECH COMPARE baseline feature set, which includes low-level descriptors (LLDs) related to energy, spectral features, mel frequency cepstral coefficients (MFCC), voicing, harmonic-to-noise ratio (HNR), spectral harmonicity, F0 (pitch), and microprosodic features (jitter and shimmer). In addition to these LLDs, their 1st order derivatives (delta) 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 Refs. [46] and [47]. In addition, we employed a bag-of-audio-words (BoAW) approach as well as an end-to-end learning (e2e) model. The highest unweighted average recall (UAR) of 58.5% could be achieved using the COMPARE functionals in combination with a Support Vector Machine (SVM). The e2e learning and BoAW models have yielded inferior results. Details can be found in Ref. [48].

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

Tavarez et al. [49] used i-vector representations of MFCCs, constant Q cepstral coefficients (CQCCs) 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 supraglottal segmental 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 of 54.3% UAR on the development set and 50.6% UAR on the test set.

Nwe et al. [50] approached the snore 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 6373 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 snore 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

| Table 2 | Number of snoring events per class in the set splits. |
|-----------------|---------------------|---------------------|------------------|
|                | 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 | Number of subjects per centre and class. |
|-----------------|---------------------|---------------------|------------------|
| 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 |

![Fig. 7. Subject’s metadata per class. A: age per class, (age in years at the time of DISE investigation). B: sample duration per class (in seconds per event).]
the airways was proposed by Rao et al. for feature extraction [51]. The model consists of two all-pole filters resembling the acoustic properties of two consecutive tubes: the first one 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, achieving a UAR of up to 52.8% on the test partition. Interestingly, comparison of the confusion matrices reveals that the classification error between V and E class samples is reduced compared to the COMPARE baseline approach. Anatomically, Velum and Epiglottis, representing the excitation locations in the upper airways for these two classes, are farthest apart, which might result in significantly different filter coefficient estimates. On the other hand, V and O 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. [52] extracted features at frame level using a feature set first proposed in the INTERSPEECH 2013 COMPARE challenge [53], consisting of 39-dimensional MFCCs, voicing probability, harmonics-to-noise-ratio, F0 and zero-crossing rate and their respective 1st and 2nd derivatives, as well as mean and standard deviation over nine neighbouring frames. Further, each instance is divided into 10 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. [54] 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 1st and 2nd 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' (WKELM) and a 'Weighted Kernel Partial Least Squares (PLS)' classifier are introduced and their performance compared to the
unweighted models by applying a 2-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. [55] generated feature vectors using deep image CNNs trained with spectrogram plots of the snoring audio data. The feature vectors, with a dimension of 4096 features, were extracted from the first and second fully connected neuronal layers, respectively, and used to train linear kernel SVMs, achieving UARs of 44.8% on the development set and 67.0% on the test set. Notably, the choice of colour map for the spectrogram plots had a significant 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. [56] 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 dev 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.

### 4.2. CompPare baseline feature subsets

To obtain a more detailed insight into the suitability of acoustic features for the task at hand, we evaluated the performance of the different subsets of the CompPare feature set.

Please refer to Table 4 for a description of the features deployed. In addition to the CompPare features (lines 1 through 13), we extracted frequency and bandwidth of the formants F1-F3 (lines 14 through 19). The number of low level descriptors for each feature subset is given, as well as the resulting number of low level descriptors for each feature subset after calculating deltas and functionals.

The feature sets were extracted by the openSMILE feature extraction and audio analysis tool [57,58]. All experiments were conducted using an SVM with linear kernel. We used the open-source toolkit LIBLINEAR [59]. As solver type, the default configuration (L2-regularised L2-loss support vector classification, dual) was chosen with a bias of 1. For all experiments, the complexity parameter of the SVM was optimised on the vector classification times in all possible permutations of the three partitions.

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. [55] generated feature vectors using deep image CNNs trained with spectrogram plots of the snoring audio data. The feature vectors, with a dimension of 4096 features, were extracted from the first and second fully connected neuronal layers, respectively, and used to train linear kernel SVMs, achieving UARs of 44.8% on the development set and 67.0% on the test set. Notably, the choice of colour map for the spectrogram plots had a significant 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. [56] 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 dev 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.

#### 4.2.1. CompPare baseline feature subsets

To obtain a more detailed insight into the suitability of acoustic features for the task at hand, we evaluated the performance of the different subsets of the CompPare feature set.

Please refer to Table 4 for a description of the features deployed. In addition to the CompPare features (lines 1 through 13), we extracted frequency and bandwidth of the formants F1-F3 (lines 14 through 19). The number of low level descriptors for each feature subset is given, as well as the resulting number of low level descriptors for each feature subset after calculating deltas and functionals.

The feature sets were extracted by the openSMILE feature extraction and audio analysis tool [57,58]. All experiments were conducted using an SVM with linear kernel. We used the open-source toolkit LIBLINEAR [59]. As solver type, the default configuration (L2-regularised L2-loss support vector classification, dual) was chosen with a bias of 1. For all experiments, the complexity parameter of the SVM was optimised on the development set in the range of $2^{-30}$ - $2^{-20}$ ... $2^0$. 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, this bisection of the complexity parameter has proven to be more suitable. Standardisation of all features (to zero mean and unit standard deviation) was employed in an on-line approach, i.e., the parameters (mean and standard deviation) were derived from the train set (the fusion of train and dev set) only and then applied to the dev set (the test set). No resampling of the data was employed in any of the experiments.

In order to average out potential differences in the characteristics of train, development, and test partition, we carried out the experiments six times in all possible permutations of the three partitions.

#### 5. Results

Classification results are shown in Table 5 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 standard deviation between the permutations are listed. Detailed results can be found in the appendix. ALL w/o F1-F3 with co-efficients and deltas (line 7 in Table 5) is the full CompPare feature set. ALL shows the results when applying the CompPare feature set plus F1-F3. For comparison, we also list the F1-F3 subset, applying the same coefficients as for jitter and shimmer.

Rated by UAR, the best classification performance could be achieved with the full feature set consisting of the CompPare 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 significant additional information in our 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 CosPaE 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 6 shows confusion matrices for all permutations, and Table 7 summarizes mean and range of all permutations of the class-specific recalls. All results are for the best-performing ALL feature set with coef and delta. Class-specific recall results of the four classes for all feature subsets can be found in the appendix.

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. The first author has listened to all T and E events with a ‘trained human ear’ and found that the snoring of the included subjects does indeed sound distinctively different. At the same time, the ‘typical sound’ of a tongue base or an epiglottis snorer could be described in all samples. This subjective judgment is based on an extensive experience of the first author in the assessment of snore sounds in several projects for a machine learning task. The first author has listened to all samples and found that the snoring of the included subjects demonstrates different excitation locations of snore sounds.

Applying the weighted average recall (WAR) 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 coefs but without deltas shows the best results over all permutations.

6. Discussion

6.1. Classification performance by class

Marked differences in classification performance can be found between the permutations of train, dev, 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 significant 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 has equal importance, as a therapy decision for T or E type snorers is distinctly different than for V or O type snoring, which occurs much more frequently. We expect more stable results with data from a higher number of subjects in the smaller classes, which will be available when adding data to the corpus over time.

6.2. Performance of feature subsets

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 are therefore likely to be well suited also for the classification of snoring noise. Our findings as well as the results from the CosPaE Snore Sub-Challenge contributions underpin this assumption. The presented acoustic tube model of the upper airways [51] has yielded results that are consistent with the underlying anatomy it aims to resemble. MFCC-based features have proven most successful in classification performance in Ref. [49], and those models using feature sets based on MFCCs and PLP cepstrum showed the best results of the challenge [52,54]. Our own findings when investigating the performance of the INTERSPEECH CosPaE feature subsets confirm this: 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 [60]. Koo et al. looked at obstruction levels in OSAS patients and found significantly higher frequencies for F1 and F2 in snorers with retrolingual obstruction compared to those with retropalatal obstruction [61]. In our experiments, MFCCs have clearly outperformed the subset that is based on formant characteristics alone, suggesting that formants are indeed descriptive for the excitation location of snore sounds, but inferior to MFCCs.

There are also a number of differences between speech and snoring. In speech generation, the sound is excited in a fixed location, the voice box in the glottis, whereas 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 characterized 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 as high as more than 500 Hz. 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 snore sound classification experiments.

6.3. Snoring and OSA

The majority of research in diagnosis and treatment of sleep related breathing disorders is undertaken with a focus on OSA and obstructive events. In contrast, this database includes sounds of vibration events in the UA without obstructive disposition. The VOTE classification
according to Kezirian et al. defines three levels 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. To what extent the excitation location of snore sounds and the obstruction sites in OSA correlate is not known and should be subject of future studies.

6.4. Drug induced and natural sleep

Our database is 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. In our case, however, this question might not be of relevance. The aim of our database is to provide material for the automatic classification of different snore sound excitation locations by means of machine learning methods. We hypothesise that the form of sleep (natural or drug induced) has no significant influence on the acoustic characteristics of snore sounds from different excitation locations.

In other words: a velum snore sounds the same, no matter if generated in natural sleep or during DISE, as long as it stems from the palate level. In turn, there will be characteristic acoustic properties for the different snore sound classes, independent of the type of sleep. Given this hypothesis is valid, results based on our database material will be transferable to snore sound examinations during natural sleep.

6.5. Weaknesses and future work

Due to the strongly imbalanced nature of the database, the number of actual subjects with V and E type snoring is fairly small, leading to unreliable results using our machine learning models. In order to overcome this weakness, additional subjects with these rare kinds of snoring should be added to the database, which will happen gradually over time.

Only clearly identifiable, single-level snoring events have been chosen to be included in the database. Hessel et al. report that single level obstructions only occur in 35% of patients [15]. 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.

7. Conclusion and outlook

For the first time, we present a database of snoring events that have been classified by the sound excitation location in the upper airways based on objective criteria and verifiable video material from several medical centres. Baseline experiments show that automatic classification models based on the acoustic properties are able to distinguish between snoring excited at the different levels of the upper airway. Adding more subjects to the database, refining the snoring classes and developing

Table 6
Confusion matrices of all permutations for the best-performing feature set.
Tr: Train partition; De: Development partition; Te: Test partition.

<table>
<thead>
<tr>
<th></th>
<th>V</th>
<th>O</th>
<th>T</th>
<th>E</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tr+De&gt;Te</td>
<td>92</td>
<td>37</td>
<td>8</td>
<td>18</td>
<td>59.4%</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>44</td>
<td>2</td>
<td>7</td>
<td>67.7%</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>25.0%</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>22</td>
<td>81.5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>V</th>
<th>O</th>
<th>T</th>
<th>E</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tr+Te&gt;De</td>
<td>119</td>
<td>32</td>
<td>2</td>
<td>8</td>
<td>73.9%</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>49</td>
<td>0</td>
<td>9</td>
<td>65.3%</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>12</td>
<td>2</td>
<td>1</td>
<td>13.3%</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>4</td>
<td>1</td>
<td>17</td>
<td>53.1%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>V</th>
<th>O</th>
<th>T</th>
<th>E</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>De+Tr&gt;Te</td>
<td>107</td>
<td>30</td>
<td>12</td>
<td>19</td>
<td>63.7%</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>43</td>
<td>6</td>
<td>7</td>
<td>56.6%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>25.0%</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>20</td>
<td>66.7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>V</th>
<th>O</th>
<th>T</th>
<th>E</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>De+Te&gt;Tr</td>
<td>115</td>
<td>29</td>
<td>13</td>
<td>11</td>
<td>68.5%</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>44</td>
<td>6</td>
<td>8</td>
<td>57.9%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>25.0%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>21</td>
<td>70.0%</td>
</tr>
</tbody>
</table>

Table 7
Mean, minimum, maximum, and range of class specific recall of all partition permutations.

<table>
<thead>
<tr>
<th>Class</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>66.6%</td>
<td>59.4%</td>
<td>73.9%</td>
<td>14.6%</td>
</tr>
<tr>
<td>O</td>
<td>62.1%</td>
<td>56.6%</td>
<td>67.7%</td>
<td>11.1%</td>
</tr>
<tr>
<td>T</td>
<td>24.4%</td>
<td>13.3%</td>
<td>31.3%</td>
<td>17.9%</td>
</tr>
<tr>
<td>E</td>
<td>70.3%</td>
<td>53.1%</td>
<td>81.5%</td>
<td>28.4%</td>
</tr>
</tbody>
</table>

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 disordered breathing.

Conflicts of interest

The corresponding author holds a German patent on a method and system for the determination of anatomical causes of snoring noise (DE102012219128B4). All other authors declare no conflict of interest.

Acknowledgement

The authors would like to thank all the colleagues involved in the collection of the labeled VOTE snoring sound data. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.compbioimed.2018.01.007.

References

[29] F. Eyben, Real-time Speech and Music Classification by Large Audio Feature Space Extraction, 2015.


