Can Machine Learning Assist Locating the Excitation of Snore Sound? A Review

Kun Qian, *Senior Member, IEEE*, Christoph Janott, Maximilian Schmitt, Zixing Zhang, *Member, IEEE*, Clemens Heiser, Werner Hemmert, *Senior Member, IEEE*, Yoshiharu Yamamoto, *Member, IEEE*, and Björn W. Schuller, *Fellow, IEEE*

***Abstract*—In the past three decades, snoring (affecting more than 30 % adults of the UK population) has been increasingly studied in the transdisciplinary research community involving medicine and engineering. Early work demonstrated that, the snore sound can carry important information about the status of the upper airway, which facilitates the development of non-invasive acoustic based approaches for diagnosing and screening of obstructive sleep apnoea and other sleep disorders. Nonetheless, there are more demands from clinical practice on finding methods to localise the snore sound's excitation rather than only detecting sleep disorders. In order to further the relevant studies and attract more attention, we provide a comprehensive review on the state-of-the-art techniques from machine learning to automatically classify snore sounds. First, we introduce the background and definition of the problem. Second, we illustrate the current work in detail and explain potential applications. Finally, we discuss the limitations and challenges in the snore sound classification task. Overall, our review provides a comprehensive guidance for researchers to contribute to this area.**

***Index Terms*—Deep learning, machine learning, obstructive sleep apnoea, physiological signals, snore sound.**

Manuscript received March 27, 2020; revised June 16, 2020; accepted July 25, 2020. Date of publication July 29, 2020; date of current version April 5, 2021. This work was supported in part by Zhejiang Lab's International Talent Fund for Young Professionals (Project HANAMI), P. R. China, in part by JSPS Postdoctoral Fellowship for Research in Japan under Grant P19081 from the Japan Society for the Promotion of Science (JSPS), Japan, in part by Grants-in-Aid for Scientific Research under Grants 19F19081 and 17H00878 from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, and in part EU's HORIZON 2020 under Grant 115902 (RADAR CNS). *(Corresponding author: Kun Qian.)*

Kun Qian and Yoshiharu Yamamoto are with Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Tokyo 113-0033, Japan (e-mail: qian@p.u-tokyo.ac.jp; yamamoto@p.u-tokyo.ac.jp).

Christoph Janott and Werner Hemmert are with the Munich School of Bioengineering, Technische Universität München, 85748 Garching, Germany (e-mail: c.janott@gmx.net; werner.hemmert@tum.de).

Maximilian Schmitt is with the Chair of Embedded Intelligence for Health Care & Wellbeing, Universität Augsburg, 86159 Augsburg, Germany (e-mail: maximilian.schmitt@informatik.uni-augsburg.de).

Zixing Zhang is with GLAM – Group on Language, Audio & Music, Imperial College London, London SW7 2AZ, U.K. (e-mail: zixing.zhang@imperial.ac.uk).

Clemens Heiser is with the Department of Otorhinolaryngology/Head and Neck Surgery, Technische Universität München, 81675 Munich, Germany (e-mail: clemens.heiser@tum.de).

Björn W. Schuller is with GLAM – Group on Language, Audio & Music, Imperial College London, London SW7 2AZ, U.K., and also with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany (e-mail: schuller@ieee.org).

Digital Object Identifier 10.1109/JBHI.2020.3012666

Abbreviations

AI	Artificial Intelligence
BoAW	Bad-of-Audio-Words
CNN	Convolutional Neural Network
ComParE	Computational Paralinguistics Challenge
CSO	Competitive Swarm Optimisation
DISE	Drug-Induced Sleep Endoscopy
DL	Deep Learning
ELM	Extreme Learning Machine
EM	Expectation-Maximization
EMDF	Empirical Mode Decomposition based Features
ENT	Ear, Nose, and Throat
ES	Excitation Source
FNN	Feedforward Neural Network
FV	Fisher Vector
GAN	Generative Adversarial Network
GMM	Gausian Mixture Model
GRU	Gated Recurrent Unit
HMMs	Hidden Markov Models
HNR	Harmonics to Noise Ratio
HOG	Histogram of Oriented Gradients
KELM	Kernel based Extreme Learning Machine
KL	Kullback-Leibler
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
LLDs	Low-Level Descriptors
LPC	Linear Predictive Coding
LSTM	Long Short-Term Memory
MAP	Maximum A Posteriori
MFCCs	Mel-frequency Cepstral Coefficients
ML	Machine Learning
MLP	Multi-Layer Perceptron
MPSSC	Munich-Passau Snore Sound Corpus
MSV	Margin Sampling Voting
MV	Majority Voting
NB	Naïve Bayes
OSA	Obstructive Sleep Apnoea
PR_{800}	Power Ratio at 800 Hz
RASTA	Relative Spectral Transform

RASTA-PLP	Representations Relative Spectra Perceptual Linear Prediction
RF	Random Forest
RMSE	Root Mean Square Energy
RNN	Recurrent Neural Network
SCAT	Deep Scattering Spectrum
SERs	Subband Energy Ratios
SF	Source Flow
SFD	Source Flow Derivative
SFFS	Spectral Frequency Features
SP	Signal Processing
SVM	Support Vector Machine
SnS	Snore Sound
TL	Transfer Learning
UA	Upper Airway
UAR	Unweighted Average Recall
UBM	Universal Background Model
VOTE	Velum, Oropharyngeal lateral walls, Tongue, and Epiglottis
VQ	Vector Quantisation
WEF	Wavelet Energy Features
WPTE	Wavelet Packet Transform Energy
WTE	Wavelet Transform Energy
XAI	Explainable Artificial Intelligence
e2e	end-to-end
k-NN	k-Nearest Neighbour
scGANs	semi-supervised conditional Generative Adversarial Networks

I. Introduction

SNORING is a prevalent disorder that affects more than 30 % adults of the British population [1]. Due to the fast development in methodologies and applications of signal processing (SP) and machine learning (ML) during the past decades, snore sound (SnS) has been increasingly studied within a wide community which includes but is not limited to acoustic/audio SP, otorhinolaryngology, ML, and biomedical engineering [2]–[4]. It was found that, as a common symptom [5], SnS can be used to develop a non-invasive approach for automatically screening obstructive sleep apnoea (OSA) [6], which is a serious chronic sleep disorder affecting the general adult population ranging from 6 % to 17 % [7]. When untreated, OSA cannot only result in morning headache and daytime sleepiness [8], but also be an independent risk factor for stroke, hypertension, myocardial infarction, cardiovascular diseases, and even lead to diabetes, and cause accidents [9], [10].

As indicated in a comprehensive review article by Roebuck *et al.* [3], an audio recording based method (mainly focused on SnS analysis) can be useful as an inexpensive method for monitoring sleep. However, most existing literature aimed to use SnS for detection of OSA rather than localising the snore site. On the one hand, there are more demands from clinical practice to determine the accurate snore sound's excitation location due to the surgical options, which can be varied among different snore sites [11], [12] and facilitate a targeted surgical plan for both of the OSA suffers and the primary snorers [13]. On the other hand, there is a demand for a low-cost, convenient and non-invasive substitute for the increasingly used golden standard, drug-induced sleep endoscopy (DISE) [14]. Multichannel pressure measurement [15]–[17] is a pioneering method, which could be efficient and applicable for monitoring natural sleep, whereas it is still an invasive method that is not well tolerated by every subject. It is reasonable to leverage ML technologies to develop an approach for automatic localisation of the snore site using only SnS. Relevant studies are extremely limited but are increasingly developing given the recent advances of artificial intelligence (AI) technologies. During the past three decades, SnS analysis has witnessed three main trends: First (from 1990 to 2012), simple acoustic features were calculated and analysed with statistical methods; second (from 2013 to 2016), human hand-crafted features were used for training conventional ML models; third (from 2017 to present), state-of-the-art deep learning (DL) techniques were applied to contribute to extracting higher level representations from SnS or even leading to end-to-end learning from SnS raw data without any human expert knowledge.

In this work, we aim to provide a thorough and comprehensive review on ML methods applied to the SnS classification task. The main contributions of this review can be summarised as: First, to the best of our knowledge, this is the first review on ML based methods for localising the snore site. Second, we introduce the reader to the background (including history and definitions) of the relevant studies. In particular, we will indicate the motivation of this study and highlight its significance in clinical practice. Third, we introduce both the conventional ML methods and the advanced deep learning approaches that were successfully applied to overcome the challenges of the SnS classification task. Last but not least, we discuss the current limitations and provide perspectives on future work. We hope this review article can be a good guidance for researchers who share the common interest to improve the understanding about cutting-edge technologies for other audiences in biomedical and health informatics.

The remainder of this review article will be organised as follows: First, we give the definition of the problem we are focusing on in Section II. In Section III, the background and related work will be introduced. Then, we present methods and challenges in a comprehensive review of the existing literature in Section IV. Finally, we discuss the current work and provide an outlook in Section V before a conclusion is made in Section VI.

II. Definition of the Problem

In this section, we provide a brief introduction of the anatomy of the upper airways. Then, we explain and compare the different categories of the snore site.

A. Anatomy

The upper airways are defined as the area from the nostrils and the lips to the vocal chords. They consist of the nasal and oral cavities, the pharynx and the upper section of the larynx. The pharynx is defined as the posterior section of the head and contains several anatomical landmarks, such as the soft palate (the velum), the palatine tonsils, the posterior part of the tongue

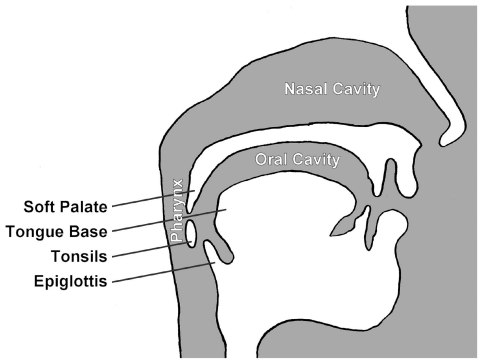

Fig. 1. The anatomy of the upper airways.

(the tongue base) and the epiglottis. The epiglottis separates the pharyngeal area from the upper gastric tract (the oesophagus) and the larynx, which contains the vocal chords. Fig. 1 shows a schematic overview of the upper airway anatomy.

Snoring is caused by vibrations of the soft tissue structures in the upper airways, especially at physiological constrictions [18]. During sleep, the muscle tone decreases and the soft tissue slackens, increasing its tendency to vibrate. The inspiratory airflow velocity increases at narrow sections within the upper airways, triggering tissue vibrations and turbulent flows, in turn causing snoring noise.

Typical areas contributing to the generation of snoring noise are the soft palate, and its very tip, the uvula, which can vibrate in anterior-posterior direction, the palatine tonsils which usually vibrate in a lateral direction, the base of the tongue which can fall back and restrict the passage between tongue and posterior pharyngeal wall causing vibrations or sounds caused by the turbulent air flow, and the epiglottis, which can collapse due to decreased structural rigidity or posterior displacement against the posterior pharyngeal wall. Furthermore, the pharyngeal walls themselves can contribute to snoring by collapsing at different levels and in different orientations.

For a targeted treatment of snoring and related sleep-related breathing disorders, it is of essence to identify the mechanisms and locations actually contributing to the airway narrowing and causing the snoring noise or the respiratory obstructions in the individual subject. Acoustic methods to distinguish between different snoring types can offer a means for tolerable and and cost-effective diagnostic measures.

B. Classification of Snoring

Numerous schemes have been suggested for classifying different types of snoring and upper airway obstructions [19]–[22]. Early classifications were limited to the distinction between palatal or non-palatal snoring, i.e., the involvement of soft palate vibrations in snoring noise generation. It was generally assumed that palatal snoring mainly occurs in primary snoring without any obstructive dispositions of the upper airways, while non-palatal snoring can be an an indicator for OSA [23].

A more accurate and widely used definition of different snoring and obstruction mechanisms is the VOTE classification

TABLE I
THE CATEGORIES OF THE VOTE CLASSIFICATION

Level	shape of constriction		
	anterior-posterior	lateral	concentric
Velum	V - ap	V - l	V - c
Oropharynx	O - ap	O - l	O - c
Tongue	T - ap	T - l	T - c
Epiglottis	E - ap	E - l	E - c

developed by Kezirian *et al.*, distinguishing between four pharyngeal levels in which snoring and airway narrowing [24] can occur. Precisely, these are

- *V-Velum:* level of soft palate, uvula, lateral pharyngeal wall tissue at velum level.
- *O-Oropharynx:* level of palatine tonsils, lateral pharyngeal wall tissues at tonsillar level.
- *T-Tongue:* level of tongue base, lingual tonsil, pharyngeal wall posterior to the tongue.
- *E-Epiglottis:* level of epiglottis.

For each of the levels, the VOTE classification describes the shape of airway constriction, using the categories anterior-posterior (*a-p*), lateral (*l*), and concentric (*c*), as well as the degree of constriction (0, no obstruction; 1, partial obstruction; 2, complete obstruction). In addition, the occurrence of snoring is noted.

Table I summarizes the resulting twelve categories of the VOTE classification. It must be noted that certain combinations of level and constriction shape are extremely rare for anatomical reasons, such as a lateral narrowing at the velum level.

A general rule in machine learning is, the bigger the number of samples in the training data set, the better the pattern generalisation and the more accurate and robust the resulting model. Furthermore, the demand for training data increases with the number of different classes that a training problem comprises. In other words, the expected recognition performance of a machine classifier on a given training set size gets better with fewer classes.

In most real-world medical ML-tasks, the amount of training data is limited, as the effort for data acquisition and preparation is considerably high. First, the raw data itself is often available in limited quantity only, and second, the effort for data preprocessing and annotation is considerable and often requires manual work by trained and experienced medical experts. In order to make best use of the MPSSC dataset, the authors have used a simplified version of the VOTE scheme for the data classification, ignoring the shape of constriction and only considering the level. Further, the degree of airway narrowing was not considered, but only the existence of audible snoring events. This resulted in a four-class scheme containing the classes *V*, *O*, *T*, and *E* [13].

The simplified VOTE scheme might present limitations in diagnostic preciseness. For example, a circular narrowing at oropharyngeal level is mainly caused by the pharyngeal walls and might lead to a different therapy decision than a vibration in lateral orientation at the same level, which indicates contribution of the tonsils.

For this reason, Janott *et. al.* developed a modified classification scheme with five classes, permitting the distinction

of selected combinations of orientation and level of vibration derived from the original VOTE classification [25]. The classes of the so-called ACLTE-scheme are defined as:

- A, *V* level, anterior-posterior vibration
- C, *V* or *O* level, concentric vibration
- L, *O* level, lateral vibration
- T, *T* level, any vibration orientation
- E, *E* level, any vibration orientation.

The resulting ACLTE-corpus contains 1 115 SnS samples from 343 subjects, and the size of classes is strongly imbalanced with the *A*-class making up for almost half of the samples, while the *T*-class is smallest with only 3 % of the samples. This reflects the frequency of occurrence of different snoring patterns in the real world, where velum snoring is relatively common, while isolated tongue-base snoring is a rare phenomenon [26].

III. Background

Early work can be traced back to Schäfer and Pirsig [27], who involved five children suffering from sleep disorders and one adult who suffered from 'simple snoring' ($n = 6$). In that study, the authors claimed that, 'simple snoring' of the adult was due in large part to vibrations of the soft palate while 'apneic snoring' of the children had a pathomechanism of enlarged adenoids and tonsils, which resulted in an impeded movement of the soft palate [27]. Their conclusions were based on observations of the frequency spectrum of the SnS. Quinn *et al.* reported differences in the waveform and frequency between palatal and tongue base snoring [28]. However, the number of subjects ($n = 6$) involved in their study was limited, therefore their conclusions cannot be easily generalised. Miyazaki *et al.* investigated the *fundamental frequency* (F0) values in four types of snoring, i.e., the soft palate, the tonsil/tongue base, the combined position, and the larynx [29]. They indicated in their findings ($n = 75$) that the average value of the fundamental frequency for the aforementioned four types of snoring was 102.8 $\pm$ 34.9 Hz (soft palate type), 331.7 $\pm$ 144.8 Hz (tonsil/tongue base type), 115.7 $\pm$ 58.9 Hz (combined type), and around 250.0 Hz (larynx type), respectively [29]. Hill *et al.* studied and made a statistical comparison ($n = 11$) of the *crest factor* (ratio of peak to root mean square value in any given epoch) between palatal and non-palatal snoring [30]. They concluded that palatal SnS can have a higher crest factor than non-palatal SnS ($p < .01$, Student-t or MannWhitney tests). In another study by Hill *et al.* [31], the values of the crest factor extracted from SnS generated by patients ($n = 5$) in natural sleep showed that the snoring mechanism may change in some individuals during the night, which means that also the snore site may change. Agrawal *et al.* calculated *peak frequency*, *centre frequency* and *power ratio* for their distinguishing capacity of palatal, tonge-based, and mixed snoring [32]. In particular, they compared the snoring sound characteristics between induced and natural sleep ($n = 11$). They claimed that induced SnS contains higher frequency components than natural SnS. Saunders *et al.* indicated that centre frequency may be efficient to distinguish pure palatal from tong base snoring ($n = 35$), but cannot be used to identify multisegmental snoring (the mixed snoring) [33]. A 2-means clustering method was used in [34] to discriminate palatal and non-palatal SnS. In their study, they used a combination of the statistical moment coefficients of *skewness* and *kurtosis* calculated from the snoring sounds from subjects ($n = 15$) performed with sleep nasendoscopy evaluation (under anaesthetic condition). Ng *et al.* continuously reported their contributions in studying *formants* extracted from SnS [35], [36], which are considered to carry important information about the status of the upper airway (UA). The first three formant frequencies, i.e., F1, F2, and F3 were indicated to be associated with the degree of constriction in the pharynx, the degree of advancement of the tongue relative to its neutral position, and the degree of lip-rounding, respectively [35], [37]–[40].

Nevertheless, the capacity of formants to localise the anatomical site of snoring was not shown in [35] ($n = 40$) or [36] ($n = 40$) while they were demonstrated to be efficient to differentiate apneic SnS from benign ones. Moreover, Ng *et al.* made their efforts to analyse and model both the source flow (SF) and its derivative (SFD) of SnS via the usage of an iterative adaptive inverse filtering approach and Gaussian probability density function [41]. In that study ($n = 40$) [41], the shapes of SF pulses are different between SnS and can be associated with the dynamic biomechanical properties (e. g., compliance and elasticity) of the SnS excitation source (ES). Particularly, the palatal (e. g., SnS from soft palate vibration) and the pharyngeal snoring (e. g., SnS from pharyngeal wall vibration) can be explained mainly by the theory of *flutter* and the concept of *static divergence*, respectively [41]–[43]. Nevertheless, Ng *et al.* clarified in [41] that clinical experiments were not conducted to warrant the accuracy of the SFD model for its relation to the occurrence and development of physiological events, e. g., closing, opening, and speed of ES vibration during snoring. Motivated by the capability of formants to represent the structure and status of the UA, Qian *et al.* and Wu *et al.* analysed the formants extracted from long duration SnS audio recordings by the *K*-means clustering method [44] and hidden Markov models (HMMs) [45], respectively. Their findings showed possible differences between the properties of formants extracted from different SnS related signals which may reflect the changes of the UA structure during the night while the accurate experts' annotation was missing. Also, the number of subjects was extremely small ($n = 1$). Additionally, Qian *et al.* found formants could also be used as an efficient marker to monitor changes of the UA by observing its tracks [46]. Xu *et al.* indicated in their study [47] ($n = 30$) that the first snoring sound after an obstructive apnoea of the upper level (above the free margin of soft palate) may have more energy components in the lower subband than its counterpart of the lower level (below the free margin of soft palate). Peng *et al.* claimed in their study [48] ($n = 74$) that, F0 and F2 were found to be lower in palatal SnS than that in non-palatal SnS.

Psychoacoustical properties combined with other acoustical features, i.e., sound pressure level ([dB], A-weighted), loudness (sone), sharpness (acum), roughness (cAsper), fluctuation strength (cVacil) and centre frequency (Hz) (mean values for each parameter), have been applied to SnS analysis in [49]. In that study, the authors summarised the statistical analysis

of the aforementioned features extracted from SnS collected within drug induced patients ($n = 41$) that, obstructive SnS had a higher loudness than non-obstructive SnS (>25 sone); velar SnS showed a higher roughness (>150 cAsper) than tonsillar and post-apnoeic SnS, and had the lowest centre frequency (<3 000 Hz); post-apnoeic SnS had the largest fluctuation strength (>50 cVacil) whereas tonsillar SnS showed the highest sharpness values (>1.6 acum).

In summary, the studies reviewed above were mainly based on statistical analysis of acoustical features extracted from SnS rather than using ML methods to localise the snore site automatically. Besides, the involved subject numbers were limited (less than 100). Early work using ML for classifying different SnS data was done by Qian *et al.* [50]–[52]. The acoustic features (e. g., *crest factor*, *power ratio*, *formants*, etc.) were investigated, and a simple machine learning model, i.e., *k*-nearest neighbour (*k*-NN) [53], [54] was used as classifier. Furthermore, the phase of *feature selection* was involved in [51], [52], in which the finally selected features can be superior to the original larger dimension of features in recognising SnS. Qian *et al.* found that frequency-based features (e. g., spectral features, Mel-frequency cepstral coefficients (MFCCs), or subband energy ratios (SERs)) performed better than amplitude-based features (e. g., crest factor). Nevertheless, their study was based on SnS data without accurate annotation in a low number of subjects ($n = 2, 1, 20$ in [50], [52], [51], respectively). Wavelet features were first introduced to the task of SnS classification in [55] ($n = 24$), which was also a first time where a machine learning based method was proposed for classifying four types of SnS annotated by ENT (ear, nose, and throat) experts, VOTE, i.e., V (Velum), O (Oropharyngeal lateral walls), T (Tongue), and E (Epiglottis). Qian *et al.* claimed that their proposed wavelet features outperformed other frequently-used features (e. g., formants, power ratio, MFCCs) by achieving a highest unweighted average recall (UAR, which is thought to be more suitable than accuracy for imbalanced data) [56] at 71.2 % by two-fold cross validation in twenty-four subjects [55]. This record was soon beaten by a bag-of-audio-words (BoAW) approach (reaching an UAR of 79.5 % using the same database as in [55]) proposed by Schmitt *et al.* [57].

A comprehensive study on the comparison of features and classifiers for recognising VOTE SnS was conducted by Qian *et al.* in [58]. In their study ($n = 40$), nine kinds of features, i.e., crest factor, F0, formants, spectral frequency features (SFFs), power ratio at 800 Hz (PR_{800}), SERs, MFCCs (0–12), empirical mode decomposition [59] based features (EMDF), and wavelet energy features (WEF) were investigated and compared. As classifiers, seven models were selected, i.e., *k*-NN [53], [54], linear discriminant analysis (LDA) [60], support vector machine (SVM) [61], random forest (RF) [62], feedforward neural network (FNN) [63], extreme learning machine (ELM) [64]–[66], and kernel based extreme learning machine (KELM) [66]. Finally, an early fusion (direct concatenation) of the overall features selected by the ReliefF algorithm [67], [68] built on a RF classifier reached the highest UAR at 78.0 % in a rigorous subject-independent case [58].

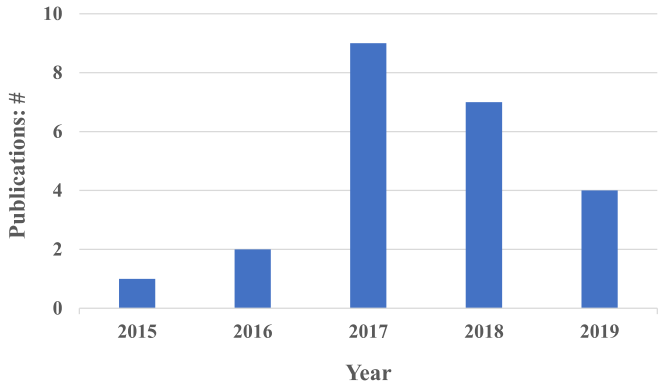

Fig. 2. The number of publications on ML for SnS classification over the recent five years (one paper officially published in January 2020 was calculated into the year of 2019 for its first online publishing time). Literature searching was under a strict manually selection processing based on Google Scholar, IEEE Xplore, and PubMed with the keywords 'snore sound,' 'snore site classification,' 'machine learning,' and 'deep learning' from the years 2015 to 2019.

The results of the aforementioned published work are encouraging and promising. However, one challenge is still unresolved: We are still lacking a standard publicly accessible annotated SnS database, which makes it difficult to develop and compare relevant algorithms and approaches for the SnS classification task. One milestone was reached by Janott *et al.* in [13], who introduced the first accurately annotated and publicly accessible SnS database, the Munich-Passau snore sound corpus (MPSSC). MPSSC was first released in the INTERSPEECH 2017 Computational Paralinguistics Challenge (ComParE) [69], which dramatically promoted the relevant studies in recent years (see Fig. 2). In Section IV, we will introduce and summarise the published literature based on MPSSC, which includes both conventional ML methods and the state-of-the-art DL approaches. On the one hand, MPSSC makes the study on SnS classification sustainable and comparable in terms of establishing the standard (subject-independent data partitioning), defining the task (V, O, T, and E types of SnS by performing DISE), and benchmarking the fundamental studies (official baseline). On the other hand, there are still several challenges to be addressed in MPSSC and among the relevant studies: First, the number of participating subjects might be sufficient ($n = 219$) whereas the number of available SnS audio instances is quite low (only 828), which constrains the capacity to learn robust higher representations by deeper models. Second, DL based methods may achieve comparable or even better performance than conventional ML methods in SnS classification, but they are not perfectly explainable. Third, the fundamental mechanism of different SnS generated from a variety of locations in the UA is not well modelled or explained. In particular, as Hill *et al.* [31] indicate, the snore site may change during night, which makes localisation of SnS using non-invasive audio based methods more complicated and difficult. Fourth, early studies [32], [70] raised an issue that the SnS collected during induced sleep may not share the same characteristics as the SnS generated under natural sleep. Nevertheless, most of the current work were based on

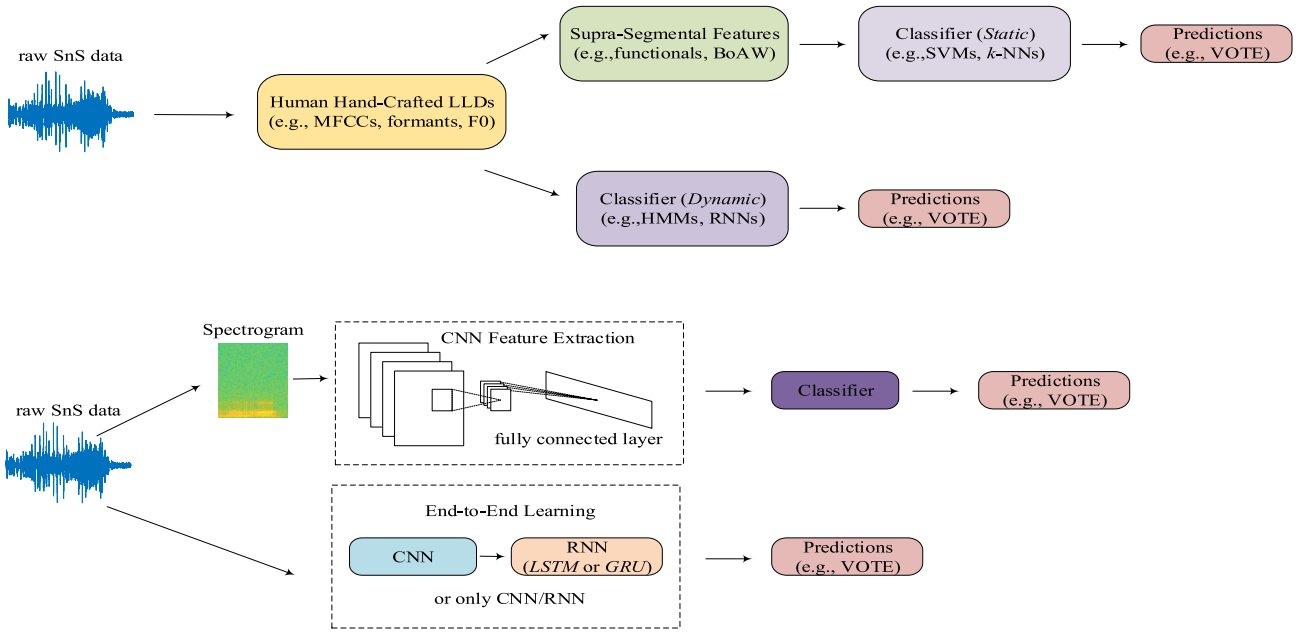

Fig. 3. Overview of conventional ML (top) and DL (bottom) based paradimgs for the SnS classification task. In conventional ML paradigms, human hand-crafted features (low-level descriptor (LLD) or higher representations) are extracted from the SnS audio signal via human expert domain knowledge. Then, a classifier will make predictions using its prior knowledge acquired via the training phase. In DL paradigms (except the DNN models trained on human hand-crafted features), DL models learn features by themselves without any human expert domain knowledge. Then, a classifier (or a fully connected layer combined with a softmax layer) will make the final predictions based on the outputs of the trained DNN models.

SnS annotated by performing the DISE (e. g., MPSSC), which means the achievements might not be directly applicable to the development of natural smart home devices. Last but not least, more attention and efforts should be contributed to this research (see Fig. 2).

In the following parts of this review, we will systematically introduce the problems, methods, and challenges. Moreover, we will discuss the current findings and limitations, and point out our perspectives for future work.

IV. Methods

In this section, we present the methods applied to SnS classification. The ML techniques including conventional methods and state-of-the-art DL approaches will be illustrated and described in detail. Fig. 3 shows the general diagram of the conventional ML and the DL based paradigms for the SnS classification task.

A. Human Hand-crafted Low-Level Descriptors

In the conventional ML paradigm, features are designed by human experts with a specific domain knowledge (e. g., medicine). Due to the similar characteristics of speech and SnS, early work on SnS classification tended to process the SnS data as speech. The low-level descriptors (LLDs) were firstly extracted from frame-based SnS signals. Those LLDs may have specific physiological meanings in SnS analysis, and can be seen as the raw representations extracted from short-time frames of the analysed SnS. Table II lists the main LLDs used in published literature on SnS classification task and their corresponding findings.

Most of the studied LLDs are typical acoustical features (e. g., MFCCs, F0, formants), while some others are not originally designed for audio analysis (s. g., WEF, local binary pattern (LBP), histogram of oriented gradients (HOG)). Note that, SnS has similar characteristics as speech, whereas it also has some properties belonging to physiological signals. These human hand-crafted LLDs carry important information about the snore site and can be interpreted in the time and the frequency domain of the SnS. A large scale acoustical feature set, i.e., ComParE, and a simplified acoustical feature set, i.e., eGeMAPS, were investigated for the SnS classification task (summarised in Table II). Both the two feature sets can be extracted by our open-source toolkit, openSMILE [95], [96].

B. Higher Representations

The aforementioned LLDs can be used directly for dynamic ML models (e. g., HMMs [97] and Recurrent Neural Networks (RNNs) [98]), while higher representations (independent of the SnS audio clip length) containing the statistical information of the LLDs over a given time are needed for training static models (e. g., SVMs [61], or ELMs [65]). In this subsection, we will introduce the higher representations investigated in the literature that can be extracted from LLDs for the SnS classification task.

1) Statistical Functionals: The *statistical functionals* are calculated from the frame based LLDs from a given period of the audio signal, which include the arithmetic mean, standard deviation, extremes (minimum value, maximum value), and further more [99]. Some more advanced functionals, e. g., moments, percentiles, kurtosis, skewness, slope, and bias of the linear

TABLE II
HUMAN HAND-CRAFTED LOW-LEVEL DESCRIPTORS (LLDS) FOR SNS CLASSIFICATION IN PUBLISHED LITERATURE. LPC: LINEAR PREDICTIVE CODING. SFFS: SPECTRAL FREQUENCY FEATURES. SERS: SUBBAND ENERGY RATIOS. EMDF: EMPIRICAL DECOMPOSITION BASED FEATURES. WTE: WAVELET TRANSFORM ENERGY. WPTE: WAVELET PACKET TRANSFORM ENERGY. WEF: WAVELET ENERGY FEATURE. GMM: GAUSSIAN MIXTURE MODEL. RASTA-PLP: REPRESENTATIONS RELATIVE SPECTRA PERCEPTUAL LINEAR PREDICTION. SCAT: DEEP SCATTERING SPECTRUM. LBP: LOCAL BINARY PATTERN. HOG: HISTOGRAM OF ORIENTED GRADIENTS

Name	Definition	Findings	Literature
Crest Factor	the ratio of peak to root mean square value	not a strong marker	[50], [58], [71]
PR_{800}	the cumulative spectrum energy below 800 Hz divided by its counterpart above 800 Hz	not a strong marker	[50], [58], [71]
F0	the lowest frequency of a periodic waveform	not a strong marker	[58], [71]
Formants	the spectral peaks (usually the first three) of the sound spectrum extracted by the LPC approach [38], [72], [73]	might be useful, but very limited	[51], [52], [55] [57], [58], [71] [74], [75]
MFCCs	the Mel-scale frequency (in Mels) cepstrum coefficients, calculated by mapping the real scale frequency (in Hz) via triangular overlapping filters, widely and successfully used in speech recognition [76]	might be useful, further study needed	[13], [51], [52] [55], [57], [58] [69], [71], [74] [75]
SFFs	the peak frequency, the centre frequency, and the mean frequency of the whole spectrum, the mean frequencies in each subband spectrum	might be useful, but very limited	[51], [52], [58] [71], [74], [75]
SERs	the ratios of subband spectrum energy to the whole spectrum energy	might be useful, but inconsistent	[51], [52], [58] [71], [74], [75]
EMDF	the energy and entropy based EMD coefficients [59]	further study needed	[51], [52], [58]
WTE	the energy based WT coefficients [77]	useful	[57], [71], [74] [75], [78]
WPTE	the energy based WPT coefficients [79]	useful	[71], [74], [75] [78]
WEF	the early fusion (concatenation) of WTE and WPTE	useful	[55], [58], [71] [74], [75], [78]
Filter Coefficients	the estimated parameters from a dual source-filter model of SnS	with limited usage	[80]
RASTA-PLP	the modified PLP analysis [81] with a spectral estimate in which each frequency channel is band-pass filtered by a filter with a sharp zero at the zero frequency [83]	might be useful, further study needed	[82]
SCAT	the time-averaged coefficients extracted from deep scattering spectrum analysis [85], [86]	useful	[84]
LBP & HOG	the LBP [87] descriptors extracted from the spectrogram the HOG [89] descriptors extracted from the spectrogram	useful (LBP is better than HOG, fusion is better than individual)	[88]
COMPARE	the large scale acoustic feature set [90], cf. Table III	might be useful, but limited	[13], [71], [74] [75], [78], [82] [25], [91]–[93]
EGEMAPS	the simplified acoustic feature set [94], cf. Table IV	with limited usage	[71]

regression estimation of the LLDs can also be applied in this method [99]. For details on OPENSMILE LLDs (i.e., COMPARE and EGEMAPS), interested readers are referred to [99]. Qian *et al.* further investigated and compared nine functionals (maximum, minimum, and mean values, range, standard deviation, slope and bias of linear regression estimation, skewness, kurtosis) in [55], [58].

2) Bag-of-Audio-Words: The bag-of-audio-words (BoAW) approach originates from the Bag-of-Words (BoW, cf. [100]) approach, which had been demonstrated to be efficient in *natural language processing* [101] and *computer vision* [102], [103]. In the BoAW approach, the numerical LLDs or alternatively the higher level derived features extracted from the SnS data will first undergo a vector quantisation (VQ) step, which employs a *codebook* of template LLDs which was previously learnt from a certain number of training data [74]. For generating the codebook, Schmitt *et al.* and their followers used the initialisation step of *k-means++ clustering* [104], which is comparable to an optimised *random sampling* of LLDs [105] instead of the traditional *k-means clustering* [106], [107] method to improve the computational speed and at the same time guarantees a comparable performance. To improve the robustness of this approach, the N_a (*assignment number*) words (i.e., LLDs) with the lowest *Euclidean* distance are considered instead of assigning each LLD to only the most similar word in the codebook. Finally, the term frequency histograms (logarithm with a bias of one) are used as higher representations extracted from the SnS via the BoAW approach. The BoAW approach was first introduced by

TABLE III
THE HUMAN HAND-CRAFTED LOW-LEVEL DESCRIPTORS (LLDS) IN THE COMPARE FEATURE SET. RASTA: RELATIVE SPECTRAL TRANSFORM; HNR: HARMONICS TO NOISE RATIO; RMSE: ROOT MEAN SQUARE ENERGY

4 energy related LLDs	**Group**
RMSE, zero-crossing rate	Prosodic
Sum of auditory spectrum (loudness)	Prosodic
Sum of RASTA-filtered auditory spectrum	Prosodic
6 voicing related LLDs	**Group**
F_0 (SHS and Viterbi smoothing)	Prosodic
Prob. of voicing	Voice quality
log HNR, jitter (local and δ), shimmer (local)	Voice quality
55 spectral LLDs	**Group**
MFCCs 1–14	Cepstral
Psychoacoustic sharpness, harmonicity	Spectral
RASTA-filtered auditory spectral bands. 1–26 (0–8 kHz)	Spectral
Spectral energy 250–650 Hz, 1 k–4 kHz	Spectral
Spectral flux, centroid, entropy, slope	Spectral
Spectral roll-off point 0.25, 0.5, 0.75, 0.9	Spectral
Spectral variance, skewness, kurtosis	Spectral

TABLE IV
THE HUMAN HAND-CRAFTED LOW-LEVEL DESCRIPTORS (LLDS) IN THE EGEMAPS FEATURE SET. RASTA: RELATIVE SPECTRAL TRANSFORM; HNR: HARMONICS TO NOISE RATIO; RMSE: ROOT MEAN SQUARE ENERGY

3 energy/amplitude related LLDs	**Group**
Sum of auditory spectrum (loudness)	Prosodic
log. HNR, shimmer (local)	Voice quality
8 frequency related LLDs	**Group**
F_0 (linear and semi tone)	Prosodic
Jitter (local), formant 1 (bandwidth)	Voice quality
Formants 1, 2, 3 (frequency)	Voice quality
Formants 2, 3 (bandwidth)	Voice quality
14 spectral LLDs	**Group**
Alpha ratio (50–1000 Hz/1–5 kHz)	Spectral
Hammarberg index	Spectral
MFCCs 1–4	Cepstral
Formants 1, 2, 3 (relative energy)	voice quality
Harmonic difference H1-H2, H1-A3	voice quality
Spectral flux	Spectral
Spectral slope (0–500 Hz, 0–1 kHz)	Spectral

Schmitt *et al.* to the SnS classification task in [57]. Qian *et al.* extended this study on wavelet-based features [55], [78] and extended the findings by involving the BoAW approach into their multi-resolution analysis for SnS classification in [75].

3) GMM Supervectors: The GMM supervecors are generated by the GMM approach [108], [109], which was successfully applied to text-independent speaker recognition tasks. In essence, the GMM supervectors are the stacked mean vectors of Gaussian mixture components [110]. In this paradigm, a universal background model (UBM) is first trained by the expectation-maximization (EM) algorithm [111] from a background dataset, which includes a wide range of corpora. Then, the GMM supervectors (usually the mean vectors) can be extracted from the models that are adapting the UBM model via the maximum a posteriori (MAP) criterion [112]. This approach was used for the SnS classification task in [84], [91]. In particular, Nwe *et al.* extracted not only the first-order statistics of mean (representing the acoustical characteristics), but also the second-order statistics of covariance (representing the shape of the distribution) [91], [110]. Specifically, in their study, the Bhattacharyya distance [113] was used instead of the widely used Kullback-Leibler (KL) divergence [114] to measure the dissimilarity between two GMM distributions.

4) Fisher Vectors: The aim of the Fisher vector (FV) method is to quantify the gradient of the parameters from a generative probability model [115]. Actually, the gradient of the log-likelihood describes the direction that the parameters should be adapted to in order to fit the data (LLDs) [115] best. Kaya and Karpov introduced the FV method into the SnS classification task in [82]. In their study, only the gradients of a K-component GMM are used as Fisher vectors.

C. Deep Learning

In the past decade, DL [116] has become a very hot and popular subject of the ML community due to its continuous breakthroughs in speech recognition [117], image classification [118], and object detection [119]. With the help of a series of nonlinear transformation of the inputs, DL models can usually learn more robustly and generalise higher representations from a big data size compared with the classical ML models (shallow architectures). Specifically, DL models can facilitate the technique development in the domain of biomedical and health informatics with the ever-increasing big data [120]–[122]. For the SnS classification task, DL was found efficient in several studies even with a limited size of data. In summary, among those DL based models for SnS classification, there are two typical paradigms: First, training the models with human hand-crafted features under a deep architecture (e. g., a more hidden layers based multi-layer perceptron (MLP) [71], [74], [78], [84], stacked autoencoders [71], [74], [78], or deep recurrent neural networks [123]); second, using a pre-trained deep convolutional neural network (CNN) [124] model to learn higher representations from the SnS data (its spectrograms), or learn the higher representations from the raw SnS data (its audio) via the CNN plus a RNN structure (end-to-end). In the first paradigm, the human hand-crafted features are still needed, which restrains the strength of DL compared with the traditional ML models. Therefore, we will emphasize successful applications for the SnS classification task via transfer learning (TL – see next Subsection) [125] and end-to-end learning (e2e) [126]. In addition, a recent study using generative adversarial network (GAN) [127] for addressing the data scarcity in SnS will be introduced.

1) Transfer Learning: This method was first introduced to SnS classification in [128], [129], by which the authors used the TL paradigm to extract deep spectrum features from spectrograms of snoring. By leveraging pre-trained CNNs (AlexNet [118], and VGG 19 [130]), high level representations of the spectrograms can be extracted from the activations of the fully connected layers of the aforementioned deep models. It was demonstrated that these CNN descriptors can achieve excellent performance on SnS classification without any human expert domain knowledge. Moreover, to reduce the redundancy

of the learnt deep spectrum features, Freitag *et al.* [129] involved a feature selection phase by applying the competitive swarm optimisation (CSO) algorithm [131] to a wrapper based paradigm [132].

2) End-to-End Learning: The e2e model was introduced in the baseline work by the INTERSPEECH COMPARE snoring sub-challenge [69]. As indicated by Schuller *et al.*, one attractive characteristic of the e2e model is that the optimal features can be learnt automatically from the data at hand [69]. In other words, feature engineering work needing much of human experts' efforts (e. g., acoustic and medical knowledge for snoring) is excluded in that paradigm. In the baseline e2e model [69], a convolutional neural network was used to extract features from raw time representations of SnS data, and a subsequent recurrent neural network (with long short-term memory (LSTM) cells [133]) was adopted to perform the final classification, which was similar to the model first applied successfully to a speech emotion recognition task [134]. A dual convolutional layer topology was proposed by Wang *et al.* in [135], by which the outputs of two separate convolutional layers (having different kernel dimension on the frequency axis, but equal dimension on the time axis) were merged via the element-wise average. Subsequently, a channel slice model (instead of fully connected layers) and two reccurrent layers (with a gated recurrent unit (GRU) cell [136]—a simpler structure compared to LSTM) were used to implement the classification capacity. Schmitt and Schuller made an in-depth investigation on different topologies of e2e for SnS classification [137]. They claimed in their findings that a convolutional layer followed by a pooling step was superior to an LSTM layer.

3) Generative Adversarial Network: Zhang *et al.* were the first group introducing GANs [127] to the SnS classification task [123], which provides a solution for addressing the data scarcity (specifically the annotated data) problem in almost all intelligent healthcare topics. The authors proposed the semi-supervised conditional generative adversarial networks (scGANs), which can automatically generate data by mapping a random noise space to the original data distribution. In doing this, one can simulate an infinite number of training data without the need of an additionally exhausting human expert annotation process due to a generation process.

Further, by integration of the semi-supervised paradigm, scGANs require only one model to synthesise different categorical SnS data. Moreover, an ensemble of scGANs are employed to overcome the model collapse issue when generating the data.

V. DISCUSSION

In this section, we discuss the findings showing interesting scientific significance of the current studies. Also limitations in the work covered by this literature review will be given. In addition, we indicate some possible future directions, which may help facilitate attracting more work to this topic.

A. Current Findings

Generally speaking, in the conventional ML paradigm, there is no huge gap between the performance among different ML models, while the features matter indeed [74]. As demonstrated in [75], very well designed features can be excellent representations for recognising SnS even with a simple classifier like Naïve Bayes (NB) [54]. Among the features, spectrum based descriptors (e. g., MFCCs) outperformed amplitude based representations (e. g., crest factor). Qian *et al.* investigated the effects of frame sizes and overlap lengths of the analysed audio chunk for extracting LLDs from SnS [71]. They indicate that WEF may need a longer frame size (64 ms) than other feature sets (16 ms or 32 ms). In addition, the higher representation extraction methods (cf. Section IV-B) are essential for final performance. But a direct comparison between methods (e. g., BoAW vs FV) is still missing.

For the DL paradigm, the main limitation is data size, which constrains the power of deep models to learn robust and generalise representations from the SnS data. Encouragingly, DL has demonstrated that some efficient high level representation can be extracted automatically from the SnS without any human expert knowledge [128], [129], [135], [137]. In particular, CNN layers were found superior to RNN layers in extracting features for SnS classification [137]. In fact, directly using a CNN+LSTM architecture did not reach an excellent performance in an early study [69]. The RNN based models were found to be efficient when a data augmentation phase was involved [123]; they reached a UAR at 67.4 % on the development set, while the performance decreased on the test set (UAR at 54.4 %). But likely their main contribution to the SnS literature were their proposed scGANs, which were successfully validated in both the static acoustic data and the sequential acoustic data [123], which was demonstrated to outperform other conventional data augmentation methods (e. g., the synthetic minority oversampling technique (SMOTE) [138], and a transformation method [139]).

One significant finding is that the multi-resolution method (e. .g, wavelets) is very efficient for SnS classification. Qian *et al.* extensively validated their wavelet based approaches for SnS classification in [55], [58], [71], [74], [75], [78]. This finding was also supported by the work in [135], in which Wang *et al.* found that fusing the global and local frequency information by using different kernel sizes of CNN models can facilitate the extraction of deep representations from snoring.

Fig. 4 shows the UARs achieved by different models which achieved better results than the MPSSC baseline in recent years. The current best result on the test set ($p < .001$ by one-tailed z-test, compared to the baseline) was achieved by Demir *et al.* in [88]. They used the LLDs extracted from spectrograms of SnS via image processing methods. However, we should note that there was a big gap between the performance on the development and test sets (37.8 % vs 72.6 % of UAR) in their study. We can find this phenomenon in almost any other studies based on the MPSSC database. We think this could be due to the factor that MPSSC has different data collection environment conditions and acoustic property distributions among the partitions. One exception is the work done by Vesperini *et al.* [84], in which their model had an excellent performance on both the development and the test sets (67.1 % vs 67.7 % UAR). In their proposed method, a well designed MLP based deep model (with specifically tuned hyper-parameters) was used, which might need large amount of efforts from experienced AI experts.

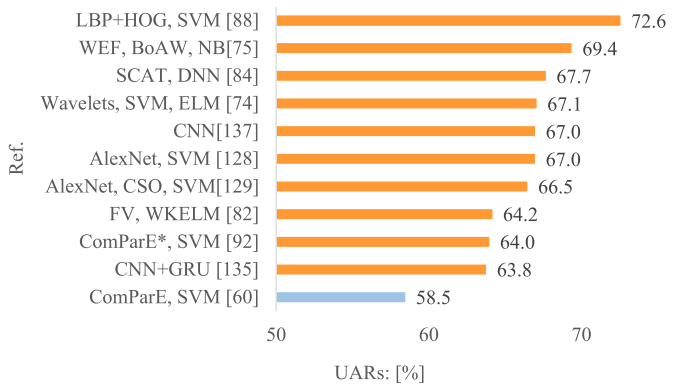

Fig. 4. The UARs (in [%]) achieved by models (on the test set) in published work based on the MPSSC Database. Only the work which achieved better results than the MPSSC baseline [69] (in blue bar) are shown. Wavelets here means the late fusion of WTE, WPTE, and WEF. The work in [92] used a slightly different ComParE feature set as used in [69].

Another point which should be noted is that feature selection may help improve or at least keep a comparable performance when using a much lower dimension than the original feature space [58], [91], [129]. Nevertheless, this step may dramatically increase the computational complexity of the whole paradigm, and may result in inconsistent feature selection results on development and test sets.

B. Limitations and Outlook

Even though the existing studies have shown encouraging and promising results, there are still some directions that require in-depth research. Based on the limitations of the existing work, we give a brief summary on the future outlook as follows:

1) Fundamental Studies: To the best of our knowledge, there is not a thorough and solid conclusion revealing the relationship between acoustical properties of SnS and the anatomical positions of snore sites. We still lack a fundamental understanding about the characteristics of SnS, particularly, for discriminating snore sites. Pevernagie *et al.* presented a comprehensive review on the acoustics of snoring [2] while the main attention was given to OSA diagnosis. Similarly, mechanism modelling of SnS was built on the target of OSA detection rather than the localisation of the snore site [36], [41]. Therefore, the acoustical analysis of SnS based on large scale investigations should be given more attention in future investigations. This will not only enrich the knowledge of experts in acoustics and medicine, but also help the ML community to design more efficient and robust features specifically for snore site localisation.

It is a matter of ongoing discussion in the medical community to what extent SnS collected during drug-induced sleep resemble those generated in natural sleep [2], [32], [70]. There are indications that the type of snoring changes during the course of the night in normal sleep [31], and it is known that the type of snoring and the mechanisms of obstructions observed during DISE do to a certain extent change with the depth of sedation. On the other hand, it can be argued that the actual sound of different types of snoring sound should remain very similar independent of the type of sleep, since the underlying pathomechanical properties are no different. Comparing snoring during artificial and natural sleep in the same subjects using an ML model of sufficient accuracy might even help contributing to this problem. Also, making use of multichannel pressure measurement in combination with DISE for the annotation of SnS types in artificial and natural sleep might help to shed more light to this unanswered question.

Last but not least, it is known that snoring properties depend on anthropometric parameters [140], but little is known on the differences of SnS properties between different ethnicities. The MPSSC is assembled using recordings from three German hospitals with patients predominantly coming from central Europe. Comparing acoustic properties from different SnS databases using raw data form different parts of the world can be an interesting aspect of future snoring research.

2) Explainable Models: Explainable AI (*aka* XAI [141]) aims to improve the trust and the transparency of AI-based systems by making the ML algorithms interpretable. As highlighted in [141], knowing the reasons behind a critical decision is important in disease diagnosis. Recently, scholars in the community of biomedical engineering are making efforts to improve the interpretability in both conventional ML models [142] and DL models [143]. Looking back at the SnS classification task, the lack of interpretability in existing successful methods limit the power of AI in clinical practice. In particular, compared with conventional ML models, DL models have their own black-box-like characteristics, which makes the explanation dramatically difficult once the model is sufficiently complex. Moreover, there is a trade-off between interpretability and accuracy [144]. Current studies in SnS focused more on the interpretability of features (both the conventional ML and DL methods) rather than the models. Adadi *et al.* systematically summarised the emerging techniques used in XAI [141]. We think visualisation appears as a promising method for understanding the higher representations extracted from SnS by DL models, which has already been successfully applied to the field of acoustic scene classification [145].

3) Fusion Strategies: As summarised by Han *et al.*, the main fusion strategies can be categorised into three groups, i.e., *feature-level* fusion, *decision-level* fusion, and *model-level* fusion [146]. Among these strategies, feature-level fusion (*a. k. a.* early fusion) and decision-level fusion (*a. k. a.* late fusion) have already been applied to the SnS classification task. Model-level fusion means fusing the intermediate representations of different modalities (e. g., audio, and video) [147]. The authors believe that in the future, other non/less-invasive modalities (e. .g, audio, scalp electroencephalography, respiratory, heart rate, and blood pressure) can be fused together for a better localisation of the snore site. In particular, with the fast development of wearables and technologies of distributed/edge computing, we can easily collect and get more useful modalities for the SnS classification task. For this review article, we mainly focused on using audio based methods. Adding other features requires work in early and late fusion. For early fusion, we should learn from previous work that 'more' does not always mean 'better' when concatenating features. We should take both the final prediction performance

and the dimensionality of the feature space into account. To this end, selecting efficient and robust features, or feature reduction methods, can be a good direction in future SnS classification tasks. Qian *et al.* systematically evaluated the contributions of each feature set to SnS classification, but their method involved human experts' effects [58]. In future work, automatic feature selection approaches will be more telling. For late fusion, finding an efficient voting strategy is the key to a successful implementation. In a recent doctoral thesis [74], two popular voting strategies were compared, i.e., majority voting (MV) and margin sampling voting (MSV). The former one is based on the major prediction made by an ensemble of ML models while the latter one is based on the prediction made by the ML model which achieved the highest *margin sampling value* [148] (the difference between the first and the second highest posterior probability). In that study [74], MV outperformed MSV in late fusion of multiple ML models for SnS classification. Future work could explore more generalised late fusion strategies, specifically, for evaluating the *confidence level* of the trained ML models.

4) Data Enrichment: We need to face and address one serious challenge almost for all AI applications in medicine: *data scarcity*. It is relatively easy to collect a large amount of SnS, whereas the annotation work is expensive, time-consuming, and even not sufficiently accurate. In particular, for SnS, its natural imbalanced characteristic [26] cannot be ignored. Take the VOTE-category as an example, SnS belonging to the V and the O class occupy 84.5 % in MPSSC while T and E type snoring samples only account for 4.7 %, and 10.8 %, respectively [13]. To overcome this issue, Zhang *et al.* proposed the scGANs based system, which was demonstrated to be more efficient than other classical data augmentation methods. In future work, some other state-of-the-art methods like unsupervised learning [149], semi-supervised learning [150], active learning [151], and cooperative learning [152] are worth being explored for the SnS classification task.

5) Open Resources: Reproducibility is crucial for a sustainable research. We encourage more researchers who share the same interests in SnS classification to contribute to open resources (e. g., databases, toolkits). Before MPSSC, there was no significant public SnS database available. We also released our toolkits like openSMIE [95], [96], openXBOW [153], auDeep [154], and End2You [155], which include both the state-of-the-art conventional ML and DL paradigms. It will be very helpful to make a fair and efficient comparison on algorithms and systems for automatically localising SnS. Specifically, we hope SnS collected in natural sleep can be added into this field, which will significantly facilitate a real application in clinical or home based situations.

VI. Conclusion

This article provided a comprehensive review of the research using audio data to localise snore sites. While the mechanism of snoring is clear, there are various definitions of the categories of the snore site. We also compared both traditional machine learning and state-of-the-art deep learning technologies and gave a detailed analysis how they can be used, and to what extent, for overcoming the challenges posed by SnS localisation. Compared to other applications in AI for healthcare, acoustical analysis of SnS is a younger field, which means that we still have insufficient fundamental knowledge about the acoustical properties of SnS. Moreover, the availability of publicly accessible databases is also extremely limited, which constrains the relevant studies. Deep learning methods are promising, but there is a far way to go to build a robust and explainable system for SnS analysis. In the discussion, we shared final insights and perspectives. We think that the combination of the conventional solid knowledge in signal processing and machine learning together with the increasingy advanced deep learning methods can leverage the power of AI to finally provide a robust and accurate system for the non-invasive localisation of the snoring site via an audio based approach.

References

[1] M. Lechner, C. E. Breeze, M. M. Ohayon, and B. Kotecha, "Snoring and breathing pauses during sleep: Interview survey of a United Kingdom population sample reveals a significant increase in the rates of sleep apnoea and obesity over the last 20 years-data from the U.K. sleep survey," *Sleep Med.*, vol. 54, pp. 250–256, 2019.

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

[3] A. Roebuck *et al.*, "A review of signals used in sleep analysis," *Physiol. Meas.*, vol. 35, no. 1, pp. R1–R57, 2014.

[4] F. Mendonça, S. S. Mostafa, A. G. Ravelo-García, F. Morgado-Dias, and T. Penzel, "A review of obstructive sleep apnea detection approaches," *IEEE J. Biomed. Health Inf.*, vol. 23, no. 2, pp. 825–837, 2019.

[5] D. J. Eckert and A. Malhotra, "Pathophysiology of adult obstructive sleep apnea," *Proc. Amer. Thoracic Soc.*, vol. 5, no. 2, pp. 144–153, 2008.

[6] P. J. Strollo Jr and R. M. Rogers, "Obstructive sleep apnea," *N. Eng. J. Med.*, vol. 334, no. 2, pp. 99–104, 1996.

[7] C. V. Senaratna *et al.*, "Prevalence of obstructive sleep apnea in the general population: A systematic review," *Sleep Med. Rev.*, vol. 34, pp. 70–81, 2017.

[8] P. Smith *et al.*, "Indications and standards for use of nasal continuous positive airway pressure (CPAP) in sleep-apnea syndromes," *Amer. J. Respir. Critical Care Med.*, vol. 150, no. 6, pp. 1738–1745, 1994.

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

[10] B. Mokhlesi, S. Ham, and D. Gozal, "The effect of sex and age on the comorbidity burden of OSA: An observational analysis from a large nationwide US health claims database," *The Eur. Respir. J.*, vol. 47, no. 4, pp. 1162–1169, 2016.

[11] K. K. Li, "Surgical therapy for adult obstructive sleep apnea," *Sleep Med. Rev.*, vol. 9, no. 3, pp. 201–209, 2005.

[12] 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," *Laryngoscope*, vol. 118, no. 5, pp. 902–908, 2008.

[13] C. Janott *et al.*, "Snoring classified: The Munich-Passau snore sound corpus," *Comput. Biol. Med.*, vol. 94, pp. 106–118, 2018.

[14] A. V. Vroegop *et al.*, "Drug-induced sleep endoscopy in sleep-disordered breathing: Report on 1,249 cases," *Laryngoscope*, vol. 124, no. 3, pp. 797–802, 2014.

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

[16] 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," *Laryngoscope*, vol. 112, no. 11, pp. 2081–2085, 2002.

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

[18] F. Dalmasso and R. Prota, "Snoring: Analysis, measurement, clinical implications and applications," *Eur. Respir. J.*, vol. 9, no. 1, pp. 146–159, 1996.

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

[20] K. Iwanaga *et al.*, "Endoscopic examination of obstructive sleep apnea syndrome patients during drug-induced sleep." *Acta Oto-Laryngol.*, no. 550, pp. 36–40, 2003.

[21] V. Abdullah, Y. Wing, and C. Van Hasselt, "Video sleep nasendoscopy: the hong kong experience," *Otolaryng. Clin. North Amer.*, vol. 36, no. 3, pp. 461–471, 2003.

[22] C. Vicini *et al.*, "The nose oropharynx hypopharynx and larynx (NOHL) classification: A new system of diagnostic standardized examination for osahs patients," *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 269, no. 4, pp. 1297–1300, 2012.

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

[24] E. J. Kezirian, W. Hohenhorst, and N. de Vries, "Drug-induced sleep endoscopy: The VOTE classification," *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 268, no. 8, pp. 1233–1236, 2011.

[25] C. Janott *et al.*, "VOTE versus ACLTE: Comparison of two snoring noise classifications using machine learning methods," *HNO*, vol. 67, no. 9, pp. 670–678, 2019.

[26] N. S. Hessel and N. de Vries, "Diagnostic work-up of socially unacceptable snoring," *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 259, no. 3, pp. 158–161, 2002.

[27] J. Schäfer and W. Pirsig, "Digital signal analysis of snoring sounds in children," *Int. J. Pediatr. Otorhinolaryngol.*, vol. 20, no. 3, pp. 193–202, 1990.

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

[29] S. Miyazaki, Y. Itasaka, K. Ishikawa, and K. Togawa, "Acoustic analysis of snoring and the site of airway obstruction in sleep related respiratory disorders," *Acta Oto-Laryngol.*, vol. 118, no. 537, pp. 47–51, 1998.

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

[31] 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," *Clin. Otolaryngol. Allied Sci.*, vol. 25, no. 6, pp. 507–510, 2000.

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

[33] N. Saunders, P. Tassone, G. Wood, A. Norris, M. Harries, and B. Kotecha, "Is acoustic analysis of snoring an alternative to sleep nasendoscopy?" *Clin. Otolaryngol. Allied Sci.*, vol. 29, no. 3, pp. 242–246, 2004.

[34] 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," *Physiol. Meas.*, vol. 28, no. 10, pp. 1225–1236, 2007.

[35] 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 Med.*, vol. 9, no. 8, pp. 894–898, 2008.

[36] A. K. Ng, T. San Koh, E. Baey, and K. Puvanendran, "Role of upper airway dimensions in snore production: Acoustical and perceptual findings," *Ann. Biomed. Eng.*, vol. 37, no. 9, pp. 1807–1817, 2009.

[37] T. Murry and R. C. Bone, "Acoustic characteristics of speech following uvulopalatopharyngoplasty," *Laryngoscope*, vol. 99, no. 12, pp. 1217–1219, 1989.

[38] J. R. Deller Jr, J. H. L. Hansen, and J. G. Proakis, *Discrete Time Process. Speech Signals*. New York, NY, USA: Wiley-IEEE Press, 1999.

[39] A. Behrman, M. J. Shikowitz, and S. Dailey, "The effect of upper airway surgery on voice," *Otolaryngol. Head Neck Surg.*, vol. 127, no. 1, pp. 36–42, 2002.

[40] G. Bertino, E. Matti, S. Migliazzi, F. Pagella, C. Tinelli, and M. Benazzo, "Acoustic changes in voice after surgery for snoring: Preliminary results," *Acta Otorhinolaryngol. Italica*, vol. 26, no. 2, p. 110, 2006.

[41] A. K. Ng and T. San Koh, "Analysis and modeling of snore source flow with its preliminary application to synthetic snore generation," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 3, pp. 552–560, Mar. 2010.

[42] L. Huang, S. J. Quinn, P. D. Ellis, and J. E. F. Williams, "Biomechanics of snoring," *Endeavour*, vol. 19, no. 3, pp. 96–100, 1995.

[43] L. Huang, "Mechanical modeling of palatal snoring," *The J. Acoustical Soc. Amer.*, vol. 97, no. 6, pp. 3642–3648, 1995.

[44] K. Qian, Y. Fang, Z. Xu, and H. Xu, "All night analysis of snoring signals by formant features," in *Proc. Int. Conf. Comput. Sci. Electron. Eng.*, Hangzhou, P. R. China, 2013, pp. 984–987.

[45] Y. Wu, Z. Zhao, K. Qian, Z. Xu, and H. Xu, "Analysis of long duration snore related signals based on formant features," in *Proc. ITA*, Chengdu, P. R. China, 2013, pp. 91–95.

[46] K. Qian, Y. Fang, and H. Xu, "A method for monitoring the cariations in the upper airway of individual osahs patients by observing two acoustic feature tracks," in *Proc. Appl. Mech. Mater.*, vol. 380–384, Trans Tech Publications Ltd., Stafa-Zurich, Switzerland, 2013, pp. 971–974.

[47] H. Xu, W. Huang, L. Yu, and L. Chen, "Sound spectral analysis of snoring sound and site of obstruction in obstructive sleep apnea syndrome," *Acta Oto-Laryngol.*, vol. 130, no. 10, pp. 1175–1179, 2010.

[48] H. Peng *et al.*, "Acoustic analysis of snoring sounds originating from different sources determined by drug-induced sleep endoscopy," *Acta Oto-Laryngol.*, vol. 137, no. 8, pp. 872–876, 2017.

[49] M. Herzog *et al.*, "Evaluation of acoustic characteristics of snoring sounds obtained during drug-induced sleep endoscopy," *Sleep Breathing*, vol. 3, no. 19, pp. 1011–1019, 2014.

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

[51] K. Qian, Z. Xu, H. Xu, and B. P. Ng, "Automatic detection of inspiration related snoring signals from original audio recording," in *Proc. ChinaSIP*, Xi'an, China, 2014, pp. 95–99.

[52] 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 Process.*, vol. 9, no. 1, pp. 21–29, 2015.

[53] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[54] M. N. Murty and V. S. Devi, *Pattern Recognition: An Algorithmic Approach*. Dordrecht, Netherlands: Springer Science & Business Media, 2011.

[55] K. Qian, C. Janott, Z. Zhang, C. Heiser, and B. Schuller, "Wavelet features for classification of VOTE snore sounds," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, P. R. China, 2016, pp. 221–225.

[56] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. INTERSPEECH*, Brighton, U.K., 2009, pp. 312–315.

[57] M. Schmitt *et al.*, "A bag-of-audio-words approach for snore sounds excitation localisation," in *Proc. ITG Speech Commun.*, Paderborn, Germany, 2016, pp. 230–234.

[58] K. Qian *et al.*, "Classification of the excitation location of snore sounds in the upper airway by acoustic multi-feature analysis," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1731–1741, 2017.

[59] N. E. Huang *et al.*, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," in *Proc. Roy. Soc. London A: Math., Phys. Eng. Sci.*, vol. 454, no. 1971. The Royal Society, 1998, pp. 903–995.

[60] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

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

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

[63] C. M. Bishop, *Pattern Recognit. Mach. Learn*. New York, NY, USA: Springer, 2006.

[64] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *Proc. IJCNN*, Budapest, Hungary, 2004, pp. 985–990.

[65] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[66] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, and Cybern., Part B (Cybern.)*, vol. 42, no. 2, pp. 513–529, Apr. 2012.

[67] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proc. AAAI*, vol. 2, San Jose, USA, 1992, pp. 129–134.

[68] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Appl. Intell.*, vol. 7, no. 1, pp. 39–55, 1997.

[69] B. Schuller *et al.*, "The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3442–3446.

[70] T. Jones, M. Ho, J. Earis, A. Swift, and P. Charters, "Acoustic parameters of snoring sound to compare natural snores with snores during steady-statepropofol sedation," *Clin. Otolaryngol.*, vol. 31, no. 1, pp. 46–52, 2006.

[71] K. Qian *et al.*, "Teaching machines on snoring: A benchmark on computer audition for snore sound excitation localisation," *Archives Acoust.*, vol. 43, no. 3, pp. 465–475, 2018.

[72] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 22, no. 2, pp. 135–141, Apr. 1974.

[73] R. C. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 129–134, Apr. 1993.

[74] K. Qian, *Autom. Gen. Audio Signal Classification*. Munich, Germany: Technical University of Munich, 2018, Doctoral Thesis.

[75] K. Qian *et al.*, "A bag of wavelet features for snore sound classification," *Ann. Biomed. Eng.*, vol. 47, no. 4, pp. 1000–1011, 2019.

[76] D. O'Shaughnessy, *Speech Commun.: Human Mach*. New York, NY, USA: Addison-Wesley, 1987.

[77] R. N. Khushaba, *Appl. Biosignal-driven Intell. Syst. for Multifunction Prosthesis Control*. Sydney, Australia: University of Technology Sydney, 2010, Doctoral Thesis.

[78] K. Qian *et al.*, "Snore sound recognition: On wavelets and classifiers from deep nets to kernels," in *Proc. EMBC*, Jeju Island, Korea, 2017, pp. 3737–3740.

[79] R. N. Khushaba, S. Kodagoda, S. Lal, and G. Dissanayake, "Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 121–131, 2011.

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

[81] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

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

[83] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[84] F. Vesperini, A. Galli, L. Gabrielli, E. Principi, and S. Squartini, "Snore sounds excitation localization by using scattering transform and deep neural networks," in *Proc. IJCNN*, Rio de Janeiro, Brazil, 2018, pp. 1–8.

[85] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, 2014.

[86] S. Mallat, "Group invariant scattering," *Commun. Pure Appl. Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[87] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[88] F. Demir, A. Sengur, N. Cummins, S. Amiriparian, and B. Schuller, "Low level texture features for snore sound discrimination," in *Proc. EMBC*, Honolulu, HI, USA, 2018, pp. 413–416.

[89] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, vol. 1, San Diego, CA, USA, 2005, pp. 886–893.

[90] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 148–152.

[91] L. T. Nwe, D. H. Tran, T. Z. W. 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.

[92] G. Gosztolya, R. Busa-Fekete, T. Grósz, and L. Tóth, "DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3522–3526.

[93] G. Gosztolya and R. Busa-Fekete, "Posterior calibration for multi-class paralinguistic classification," in *Proc. SLT*, Athens, Greece, 2018, pp. 119–125.

[94] F. Eyben *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affective Comput.*, vol. 7, no. 2, pp. 190–202, 2015.

[95] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE–the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM MM*, Florence, Italy, 2010, pp. 1459–1462.

[96] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. ACM MM*, Barcelona, Spain, 2013, pp. 835–838.

[97] L. Rabiner and B. Juang, "An introduction to hidden markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, 1986.

[98] J. L. Elman, "Finding structure in time," *Cognitive Sci.*, vol. 14, no. 2, pp. 179–211, 1990.

[99] F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Cham, Switzerland: Springer International Publishing, 2015, Doctoral Thesis.

[100] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2–3, pp. 146–162, 1954.

[101] F. Weninger, P. Staudt, and B. Schuller, "Words that fascinate the listener: Predicting affective ratings of on-line lectures," *Int. J. Distance Edu. Technol.*, vol. 11, no. 2, pp. 110–123, 2013.

[102] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 591–606, 2009.

[103] J. Wu, W.-C. Tan, and J. M. Rehg, "Efficient and effective visual codebook generation using additive kernels," *J. Mach. Learn. Res.*, vol. 12, pp. 3097–3118, Nov. 2011.

[104] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. ACM-SIAM SODA*, New Orleans, LA, USA, 2007, pp. 1027–1035.

[105] S. Rawat, P. F. Schulam, S. Burger, D. Ding, Y. Wang, and F. Metze, "Robust audio-codebooks for large-scale event detection in consumer videos," in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 2929–2933.

[106] C. M. Bishop, *Pattern Recognition and Mach. Learn*. New York, USA: Springer, 2006.

[107] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Proc. INTERSPEECH*, Portland, OR, USA, 2012, pp. 2105–2108.

[108] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, 1995.

[109] B. L. Pellom and J. H. Hansen, "An efficient scoring algorithm for Gaussian mixture model based speaker identification," *IEEE Signal Process. Lett.*, vol. 5, no. 11, pp. 281–284, 1998.

[110] C. H. You, K. A. Lee, and H. Li, "GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1300–1312, 2010.

[111] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, 1996.

[112] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[113] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.

[114] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951.

[115] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. CVPR*, Minneapolis, MN, USA, 2007, pp. 1–8.

[116] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[117] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[118] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[119] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Proc. NIPS*, Stateline, NV, USA, 2013, pp. 2553–2561.

[120] D. Ravì *et al.*, "Deep learning for health informatics," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 4–21, 2017.

[121] J. Andreu-Perez, C. C. Poon, R. D. Merrifield, S. T. Wong, and G.-Z. Yang, "Big data for health," *IEEE J. Biomed. and Health Inf.*, vol. 19, no. 4, pp. 1193–1208, 2015.

[122] M. Viceconti, P. Hunter, and R. Hose, "Big data, big knowledge: Big data for personalized healthcare," *IEEE J. Biomed. Health Inf.*, vol. 19, no. 4, pp. 1209–1215, 2015.

[123] Z. Zhang, J. Han, K. Qian, C. Janott, Y. Guo, and B. Schuller, "Snoregans: Improving automatic snore sound classification with synthesized data," *IEEE J. Biomed. Health Inf.*, vol. 24, no. 1, pp. 300–310, 2020.
[124] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Proc. NIPS*, Denver, CO, USA, 1989, pp. 396–404.
[125] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.
[126] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Proc. ICASSP*, Florence, Italy, 2014, pp. 6964–6968.
[127] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Neural Inf. Process. Syst.*, Montreal, Canada, 2014, pp. 2672–2680.
[128] S. Amiriparian *et al.*, "Snore sound classification using image-based deep spectrum features," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3512–3516.
[129] M. Freitag, S. Amiriparian, N. Cummins, M. Gerczuk, and B. Schuller, "An end-to-evolutionhybrid approach for snore sound classification," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3507–3511.
[130] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[131] R. Cheng and Y. Jin, "A competitive swarm optimizer for large scale optimization," *IEEE Trans. Cybern.*, vol. 45, no. 2, pp. 191–204, Feb. 2015.
[132] S. Gu, R. Cheng, and Y. Jin, "Feature selection for high-dimensional classification using a competitive swarm optimizer," *Soft Comput.*, vol. 22, no. 3, pp. 811–822, 2018.
[133] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[134] G. Trigeorgis *et al.*, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. ICASSP*, Shangai, P. R. China, 2016, pp. 5200–5204.
[135] J. Wang, H. Strömfeli, and B. W. Schuller, "A CNN-GRU approach to capture time-frequency pattern interdependence for snore sound classification," in *Proc. EUSIPCO*, Rome, Italy, 2018, pp. 997–1001.
[136] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS Deep Learn. Represent. Learn. Workshop*, Montreal, Canada, 2014, pp. 1–9.
[137] M. Schmitt and B. Schuller, "End-to-end audio classification with small datasets–Making it work," in *Proc. EUSIPCO*, A Coruña, Spain, 2019, pp. 1–5.
[138] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
[139] D. e. Amodei, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *Proc. ICML*, New York, NY, USA, 2016, pp. 173–182.
[140] A. Azarbarzin and Z. Moussavi, "Do anthropometric parameters change the characteristics of snoring sound?" in *Proc. EMBC*, Boston, MA, USA, 2011, pp. 1749–1752.
[141] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.
[142] S. M. Lundberg *et al.*, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 749–760, 2018.
[143] H. Lee *et al.*, "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets," *Nature Biomed. Eng.*, vol. 3, no. 3, p. 173, 2019.
[144] S. Sarkar, T. Weyde, A. Garcez, G. G. Slabaugh, S. Dragicevic, and C. Percy, "Accuracy and interpretability trade-offs in machine learning applied to safer gambling," in *Proc. CEUR Workshop*, vol. 1773, Barcelona, Spain, 2016, pp. 1–9.
[145] Z. Ren, Q. Kong, J. Han, M. Plumbley, and B. Schuller, "Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, Brighton, U.K., 2019, pp. 56–60.
[146] J. Han, Z. Zhang, Z. Ren, and B. Schuller, "Implicit fusion by joint audiovisual training for emotion recognition in mono modality," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, Brighton, U.K., 2019, pp. 5861–5865.
[147] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affective Comput.*, vol. 2, no. 2, pp. 92–105, 2011.
[148] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *Proc. IDA*, Cascais, Portugal, 2001, pp. 309–318.
[149] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," in *Proc. Eur. Conf. Comput. Vis.*, Dublin, Ireland, 2000, pp. 18–32.
[150] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synth. Lectures Artif. Intell. Mach. Learn.*, vol. 3, no. 1, pp. 1–130, 2009.
[151] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Madison, WI, USA, Computer Sciences Technical Report 1648, 2009.
[152] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 115–126, 2015.
[153] M. Schmitt and B. W. Schuller, "openXBOW-introducing the passau open-source crossmodal bag-of-words toolkit," *J. Mach. Learn. Res.*, vol. 18, no. 96, pp. 1–5, 2017.
[154] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6340–6344, 2017.
[155] P. Tzirakis, S. Zafeiriou, and B. Schuller, "End2You–The imperial toolkit for multimodal profiling by end-to-end learning," 2018, *arXiv:1802.01115*.