Annals of Biomedical Engineering, Vol. 47, No. 4, April 2019 *(© 2019)* pp. 1000–1011
https://doi.org/10.1007/s10439-019-02217-0

A Bag of Wavelet Features for Snore Sound Classification

KUN QIAN,[1,2] MAXIMILIAN SCHMITT,[2] CHRISTOPH JANOTT,[3,5] ZIXING ZHANG,[4,5] CLEMENS HEISER,[6] WINFRIED HOHENHORST,[7] MICHAEL HERZOG,[8] WERNER HEMMERT,[3] and BJÖRN SCHULLER[2,4,5]

[1]Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Arcisstr. 21, 80333 Munich, Germany; [2]ZD.B Chair of Embedded Intelligence for Health Care & Wellbeing, Universität Augsburg, Eichleitnerstr. 30, 86159 Augsburg, Germany; [3]Munich School of Bioengineering, Technische Universität München, Boltzmannstr. 11, 85748 Garching, Germany; [4]GLAM – Group on Language, Audio & Music, Department of Computing, Imperial College London, 180 Queens' Gate, Huxley Bldg., London SW7 2AZ, UK; [5]audEERING GmbH, 82206 Gilching, Germany; [6]Department of Otorhinolaryngology/Head and Neck Surgery, Klinikum rechts der Isar, Technische Universität München, Ismaningerstr. 22, 81675 Munich, Germany; [7]Department of Otorhinolaryngology/Head and Neck Surgery, Alfried Krupp Krankenhaus, Alfried-Krupp-Str. 21, 45131 Essen, Germany; and [8]Department of Otorhinolaryngology/Head and Neck Surgery, Carl-Thiem-Klinikum Cottbus, Thiemstr. 111, 03048 Cottbus, Germany

(Received 18 June 2018; accepted 21 January 2019; published online 30 January 2019)

Associate Editor Ka-Wai Kwok oversaw the review of this article.

Abstract—Snore sound (SnS) classification can support a targeted surgical approach to sleep related breathing disorders. Using machine listening methods, we aim to find the location of obstruction and vibration within a subject's upper airway. Wavelet features have been demonstrated to be efficient in the recognition of SnSs in previous studies. In this work, we use a bag-of-audio-words approach to enhance the low-level wavelet features extracted from SnS data. A Naïve Bayes model was selected as the classifier based on its superiority in initial experiments. We use SnS data collected from 219 independent subjects under drug-induced sleep endoscopy performed at three medical centres. The unweighted average recall achieved by our proposed method is 69.4%, which significantly ($p < 0.005$, one-tailed z-test) outperforms the official baseline (58.5%), and beats the winner (64.2%) of the INTERSPEECH COMPARE Challenge 2017 Snoring sub-challenge. In addition, the conventionally used features like formants, mel-scale frequency cepstral coefficients, subband energy ratios, spectral frequency features, and the features extracted by the OPENSMILE toolkit are compared with our proposed feature set. The experimental results demonstrate the effectiveness of the proposed method in SnS classification.

Keywords—Snore sound, Obstructive sleep apnea, Drug-induced sleep endoscopy, Wavelets, Bag-of-audio-words.

INTRODUCTION

Snore sound (SnS) excitation localisation can be a helpful method to support targeted surgical planning for the treatment of both primary snorers who are asymptomatic and do not have breathing interruptions during sleep, and patients suffering from obstructive sleep apnea (OSA),[52] a chronic serious sleep disorder, which is affecting 13% of men and 6% of women in the US population.[36] OSA increases the risks of stroke, hypertension, myocardial infarction, and other cardiovascular diseases, and is associated with diabetes and vulnerability to accidents.[27] The surgical options for individual subjects can be manyfold due to the multifactorial mechanisms of SnS generation and depending on the individual anatomy.[22,23] Therefore, in medical practice, it is helpful for ear, nose and throat (ENT) experts to understand the individual anatomical site of SnS generation, and the obstruction mechanism. Drug-induced sleep endoscopy (DISE)[55] can be used to identify the location of SnS. However, it requires additional time, causes cost, and means physical stress for the subjects. In addition, DISE is performed in an artificial state of sleep. Another solution, multi-channel pressure measurement,[8,44,53] requires to introduce a thin tube with multiple pressure sensors into the upper airway of the subject. This method can be used during natural sleep, however, the tube in the upper airway is not tolerated by every subject. Thus, a non-invasive method, e.g., analysis of SnS, can make diagnosis much easier both for doctors and patients compared

Address correspondence to Kun Qian, Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Arcisstr. 21, 80333 Munich, Germany. Electronic mail: andykun.qian@tum.de

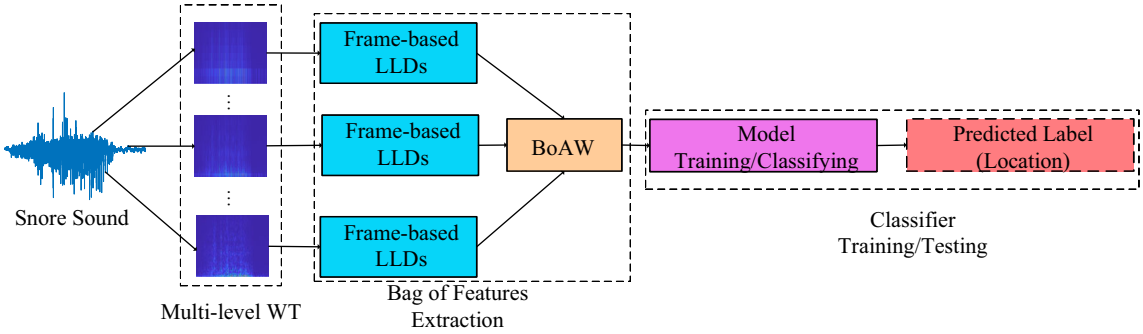

FIGURE 1. Diagram of the proposed system.

with the two methods mentioned above. Furthermore, this method can facilitate the development of portable devices for human SnS monitoring.[26]

SnS has been proven to carry information about the site and degree of obstruction in the upper airway of the subject.[37] However, studies on how to use the audio information to localise the excitation of SnS are very limited (refer to a recent literature review in Ref. 17). In the work done by Ng *et al.*,[30] wavelets had been used to form the representations to classify benign and apneic snores; the results are promising whereas the method was not performed on localising the SnS excitation. In addition, formants, spectral peak frequency, and psychoacoustic metrics were investigated to find the relationship between the human upper airway dimensions and the attributes of snores.[32] Nevertheless, the studies were still limited to the field of diagnosing benign and apneic snores. Moreover, some studies[10,28,56] showed the encouraging results on analysing tracheal respiratory sounds, and proposed another less-invasive method to screen OSA. The same as aforementioned works, those studies are not in terms of finding the locations of SnS excitation, which can be more helpful to ENT experts.

Qian *et al.* published pilot work on using acoustic signal processing combined with machine learning for the recognition of SnS.[39–41,45] The results of their work were promising and encouraging. Nevertheless, the number of independent subjects they used are of limited size, i.e., 24 in Refs. 41,45 and 40 in Refs. 39,40 The main aims of this study are as follows: firstly, we extend the number of independent subjects to 219, which will give more persuasive results. In addition, this newly released database is publicly accessible, which makes the relevant studies reproducible and comparable. Secondly, a novel method based on multi-resolution wavelet transformation (WT)[41] and bag-of-audio-word (BoAW) approach[45] is proposed to improve the current baselines achieved by standard features and classifiers. Finally, a brief comparison between our proposed method and the state-of-the-art methods used in most recent submissions of the INTERSPEECH ComParE Challenge 2017 Snoring sub-challenge[49] will be given, which can hopefully attract more researchers' interests in this topic.

MATERIALS AND METHODS

In this section, the database we used will be firstly introduced. Then the methodology part will give a description of the proposed system. The diagram of the proposed system is shown in Fig. 1. Firstly, the frame-based low-level descriptors (LLDs) will be extracted from the SnS *via* the multi-level WT. Then, BoAW approach can transfer the LLDs into higher representations for the classifier's training and testing.

MPSSC Database

The Munich Passau Snore Sound Corpus (MPSSC)[16] was first released for a sub-challenge in the INTERSPEECH 2017 Computational Paralinguistics Challenge.[49] The MPSSC contains audio from selected audio–video recordings taken during DISE[55] at three medical centres in Germany, i.e., Klinikum rechts der Isar, Technische Universität München, Munich, Alfried Krupp Hospital, Essen, and University Hospital, Halle.

Detailed information about the SnS data acquisition system, data selection and labeling can be found in Ref. 16 In the audio-track of the DISE videos, snore events were separated using a combination of automated and manual selection steps. The selected snore events were then labelled by an ENT expert by watching the DISE videos and based on the VOTE classification[19] ('V' represents the level of the velum, 'O' represents the oropharyngeal area, 'T' represents the tongue base, 'E' represents the level of the epiglottis). Only the snore events which showed one clear vibration source were included in the corpus, the ones with mixed or unclear source of vibration were

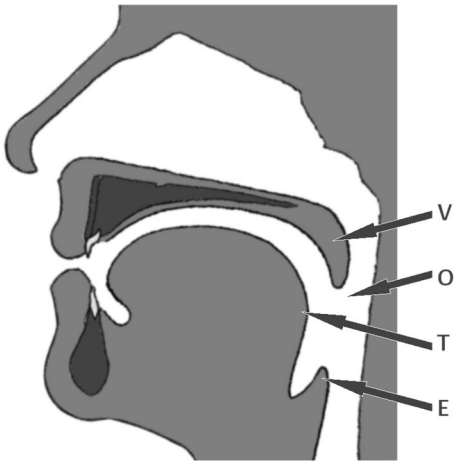

FIGURE 2. Locations of snoring vibrations or obstruction according to the VOTE scale.

TABLE 1. The number of snore events per class in each partitioned splits, as originally used in Ref. 49.

	Train	Dev	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

excluded. Figure 2 shows the corresponding vibration locations in the upper airways.

The final MPSSC database contains 828 snore events from 219 independent subjects. The overall time duration of the MPSSC is 1250.11 s, and the average length of the events is 1.51 s (ranging from 0.73 to 2.75 s). Balanced by class, centre, gender, and age, the whole database was partitioned into train, development (dev), and test set. Table 1 illustrates the number of snore events per class in each partitioned split. For the sake of comparability, we used the same splits as in the INTERSPEECH 2017 Computational Paralinguistics Snoring sub-challenge.[49] Figure 3 illustrates the waveforms and spectrograms of typical four types of SnS events.

Wavelet Features

The wavelet features were first introduced to the SnS classification task in Ref. 41 which proposed its superiority to other conventional acoustic features like formants, mel-scale frequency cepstral coefficients (MFCCs), fundamental frequency, *etc*. In previous studies,[39,41] we used an early fusion (multiple kinds of feature sets are directly fused into one feature set before fed into the classifier) of two kinds of wavelet energy feature (WEF) sets, i.e., wavelet transform energy (WTE), and wavelet packet transform energy (WPTE) as the feature representations. In this work, we separately investigate and compare WTE and WPTE. In addition, the early fusion of the two aforementioned features are studied. To be consistent with the previous study,[39] the early fusion of WTE and WPTE features are named as WEFs.

Wavelet Transform Energy

The WTE features are extracted by discrete WT (DWT),[24] which operates only on the outputs of the lowpass filter at the subsequent levels of the decomposed signal (see Fig. 4a). We calculated the vector of percentage of WTE at the jth level as:

$$\mathbf{E}_{\mathbf{v}_j} = \frac{\mathbf{w}_j^2}{\sum_{j=1}^{J_{\max}} \mathbf{w}_j^2} \times 100, \tag{1}$$

where $\mathbf{w}_j$ are the coefficients generated by DWT at the jth decomposition level. Then the *mean*, *variance*, *waveform length* (the sum of the absolute differences), and *entropy* are calculated from the vector (see Eq. 1) as basic WTE representations. In summary, for a jth decomposition of WT, we will generate $4 \times (J_{\max} + 1)$ WTE LLDs ($J_{\max}$ family of approximation coefficients plus one family of detail coefficients in the first level). $J_{\max}$ is the maximum level for wavelet decomposition by a certain *wavelet type*.

Wavelet Packet Transform Energy

The WPTE features are extracted by the wavelet packet transformation,[4,5] which not only decomposes the components of the 'approximation' (by lowpass filter), but also the components of the 'detail' (by highpass filter) (see Fig. 4b). In Ref. 20 the normalised bank filter energy was defined as:

$$\widetilde{E}_{\mathbf{v}_{j,k}} = \log \sqrt{\frac{\sum_{n=1}^{N_{j,k}} (\mathbf{w}_{j,k,n})^2}{N_{j,k}}}, \tag{2}$$

where $\mathbf{w}_{j,k}$ represents the coefficients calculated by WPT from the signal at the subspace $\mathbf{V}_{j,k}$. $N_{j,k}$ is the total number of wavelet coefficients in the kth subband at the jth level. The scale of k is $0, 1, 2, \ldots, 2^j - 1$. Totally, $2^{J_{\max}+1} - 1$ WPTE based LLDs are generated. Figure 5 shows the multi-resolution time–frequency analysis by WPT for typical four types of SnS events.

Conventional Features

To evaluate the methods proposed, we compare other features demonstrated to be efficient in SnS

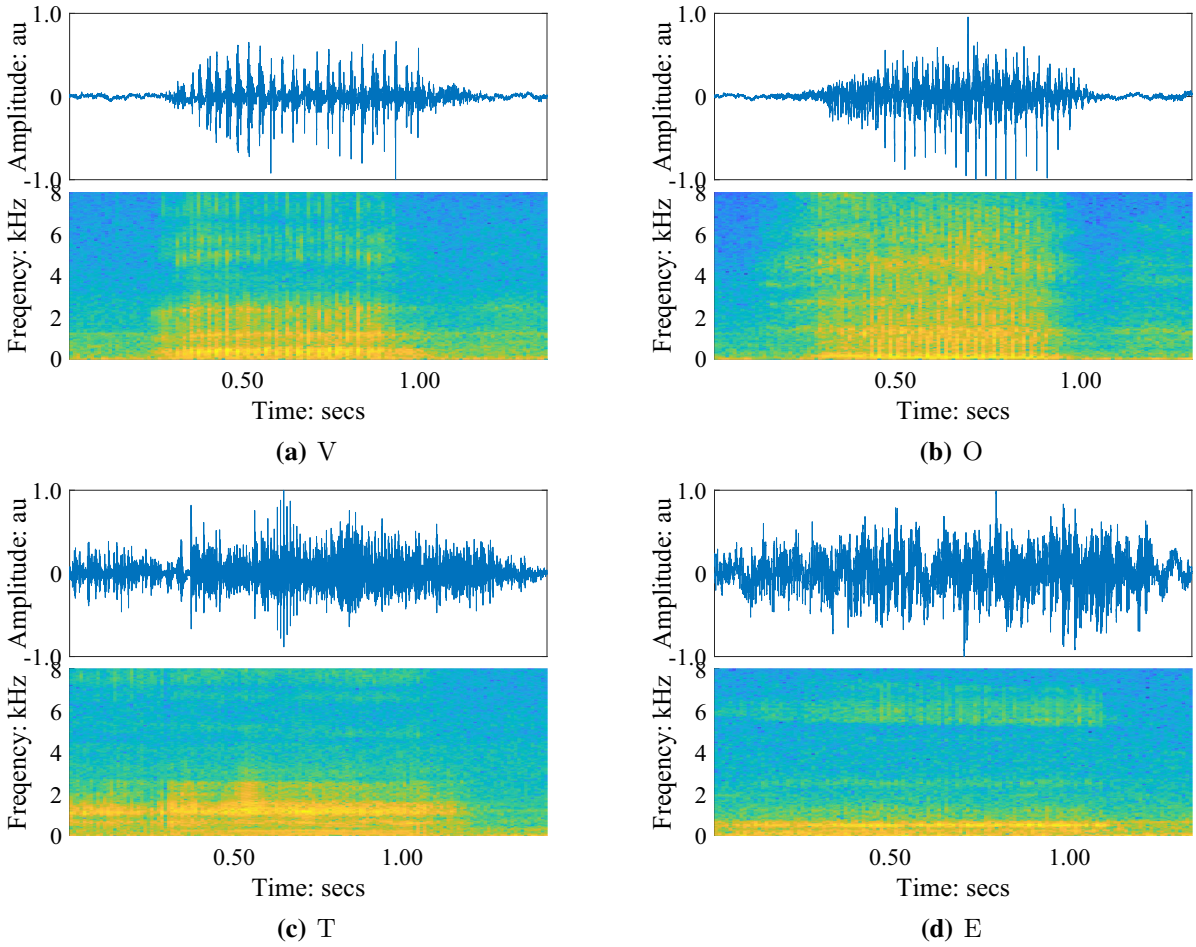

FIGURE 3. The examples of waveforms (top row) and spectrograms (bottom row) for the snore event labelled as type of V, O, T, and E. The waveforms are normalised and the amplitude has an arbitrary unit (au).

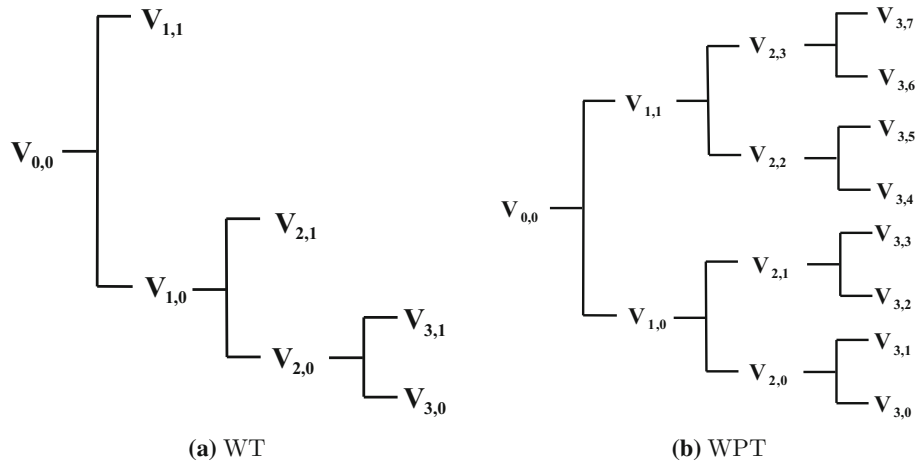

FIGURE 4. Examples of 3-level decomposition by WT (a) and WPT (b).

recognition like *Formants*[31,51] (the first three formant frequencies calculated from the *linear predictive coding coefficients*[7]), MFCCs,[34] *subband energy ratios* (SERs, the ratios of the energy in every subband to that of the whole sound spectrum),[3] and *spectral frequency features* (SFFs, the peak, centre, and mean frequency of

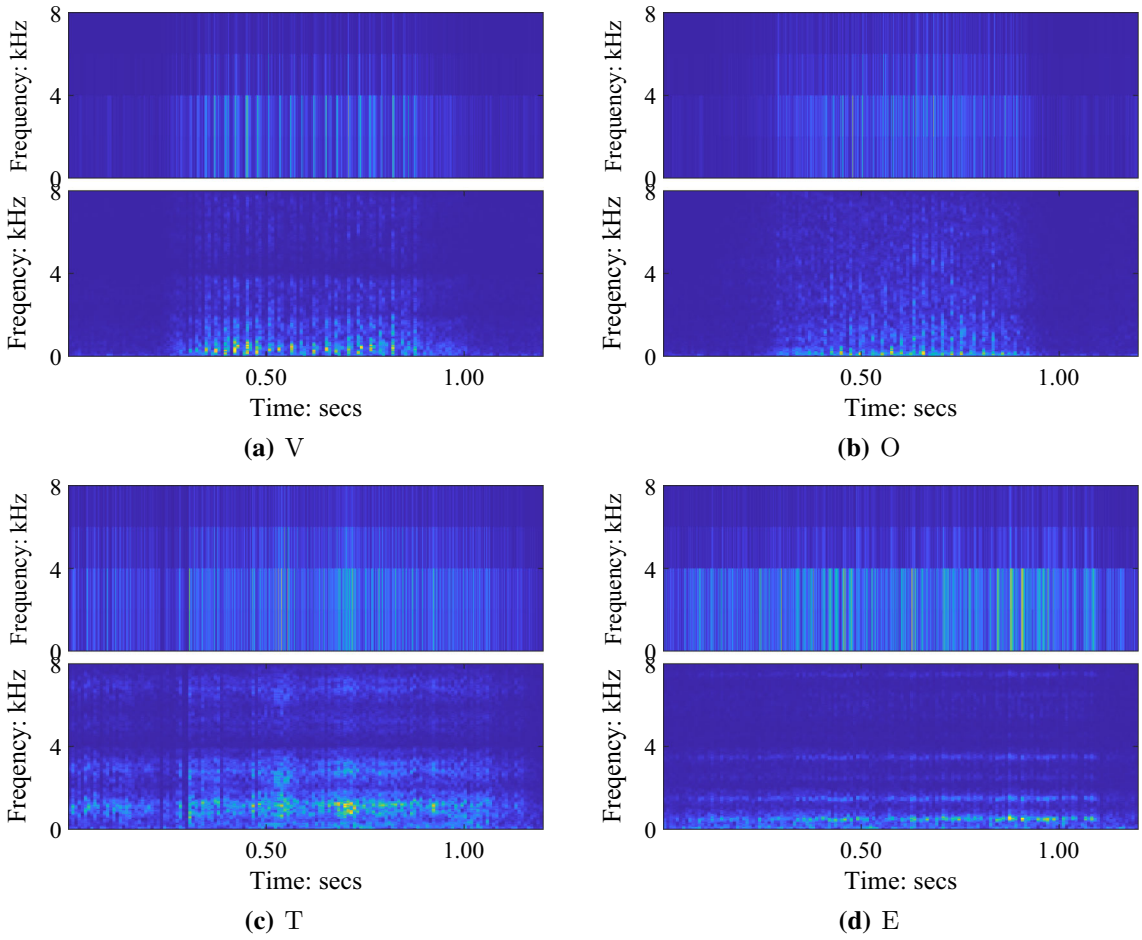

FIGURE 5. The examples of multi-resolution time–frequency analysis of four types of SnS by WPT (top row: $J = 2$, bottom row: $J = 7$). Wavelet type: 'haar'. J decomposition level. The audio examples are the same as used in Fig. 3.

the whole sound spectrum and the mean frequency in every subband).[39] The detailed definitions of these LLDs can be found in Ref. 39 Based on the study in Ref. 39 we further applied *functionals* to the extracted frame-level LLDs. The functionals refer to the statistical information of each LLD contour comprised of continuous LLDs, including *maximum*, *mean*, *minimum*, and *bias* of the linear regression estimation of LLDs in one instance. In this study, *functionals* are the counterpart which will be compared with BoAW approach.

In addition, we involve the large scale acoustic feature set ComParE[50] (extracted by the open-source toolkit, openSMILE[12]), which was selected as the official baseline feature set in snore sub-task in the INTERSPEECH 2017 Computational Paralinguistics Challenge.[49] The LLDs used in ComParE are listed in Table 2, and the *functionals* can be found in Ref. 11 In total, 65 LLDs, and 6373 statistic features are used in ComParE.

Bag of Audio Words

In the BoAW approach (see Fig. 6), the acoustic LLDs extracted from each audio instance are summarised and represented as a *term-frequency histogram*. Compared to the *bag-of-words* approach known from *natural language processing*, where text documents are represented as word histograms, the numerical LLDs extracted from the speech signal need to undergo a vector quantisation (VQ) step first. The VQ is done employing a *codebook* of template LLDs ('audio words') which is previously learnt from a certain amount of training data. Although the codebook generation usually employs *K-means clustering*,[35] similar results can be achieved using a *random sampling* of the LLDs[43] (using the default random seed in openXBOW), where the sampling follows the initialisation step of *K-means++ clustering*,[2] i.e., far-off LLDs are prioritised. Instead of assigning each LLD to only the most similar word in the codebook, the N_a words with the lowest Euclidean distance are considered, which

usually results in an improved robustness of the approach.[46] Counting the term-frequencies, i.e., the number, each audio word has been chosen as the nearest neighbour for the LLDs in one audio instance, the *term-frequency histogram* is generated. In the resulting histogram, the logarithm (with a bias of 1) is then taken from the word frequencies, in order to compress the range of values.

In this study, the open-source toolkit openXBOW[47] is used. In order to reduce the effect of different magnitudes between the LLDs, they are subject to standardisation. Accordingly, also the resulting term-frequency histograms are standardised before they are fed into a classifier. The *codebook size* and the *number of assignments* are empirically fixed to 5000 and 10 in initial experiments, respectively.

TABLE 2. ComParE acoustic feature set: 65 LLDs.

Four energy related LLDs	Group
RMSE, zero-crossing rate	Prosodic
Sum of auditory spectrum (loudness)	Prosodic
Sum of RASTA-filtered auditory spectrum	Prosodic
55 Spectral LLDs	**Group**
MFCC 1–14	Cepstral
Psychoacoustic sharpness, harmonicity	Spectral
RASTA-filtered auditory spectrum bds. 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 pt. 0.25, 0.5, 0.75, 0.9	Spectral
Spectral variance, skewness, kurtosis	Spectral
Six voicing related LLDs	**Group**
F_0 (SHS and Viterbi smoothing)	Prosodic
Probability of voicing	Voice quality
Log. HNR, jitter (local and δ), shimmer (local)	Voice quality

Refer to Ref. 11 for more details.
RMSE root mean square energy, *RASTA* relative spectral transform, *HNR* harmonics to noise ratio.

Naïve Bayes Classifier

In this study, a Naïve Bayes classifier has been chosen due to its efficient performance both in terms of classification accuracy and implementation speed in our initial experiments. Further, Naïve Bayes classifiers are known to be less prone to overfitting than other classifier types especially on small datasets, which is an important aspect in this case.

The Naïve Bayes classifier is based on a conditional probability model, which assigns given instance probabilities $\boldsymbol{p}(\boldsymbol{C}_\lambda|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_\sigma)$ for each of λ possible *classes* $\boldsymbol{C}_\lambda$[29] $(\boldsymbol{C}=(C_1,\ldots,C_\Lambda))$, where the vector $\boldsymbol{x}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_\sigma)$ represents σ features. For Naïve Bayes classifiers, the assumption is made that each feature is independent of the value of the other features when given the *class* variable. Therefore, based on the Bayes Theorem,[29] and the chain rule for repeated applications of the definition of conditional probability, the

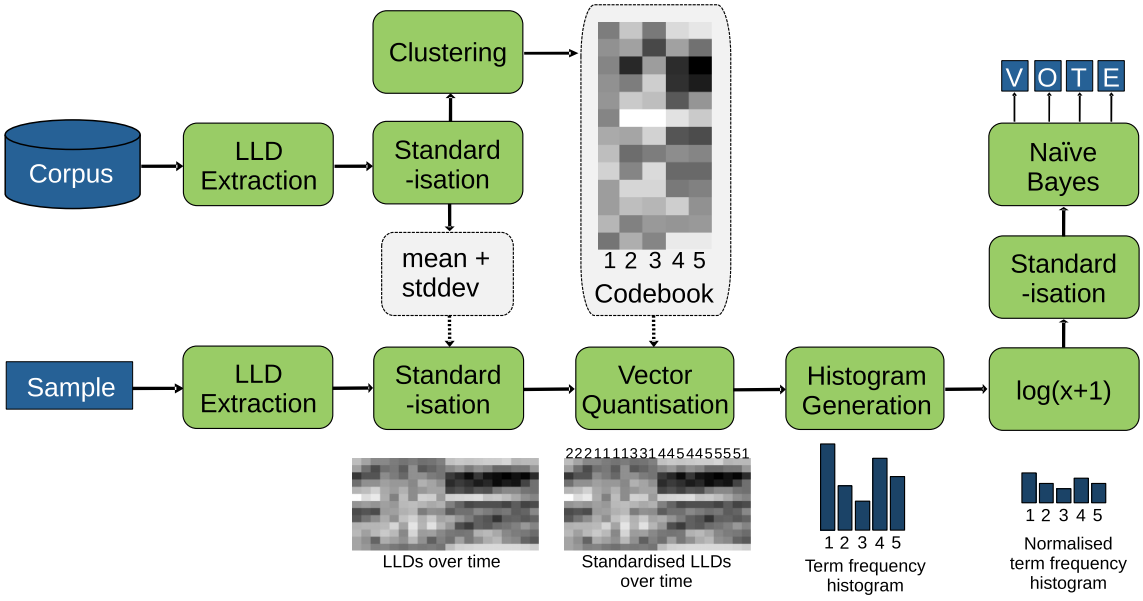

FIGURE 6. An example of the whole bag-of-audio-words approach process. The *codebook generation* is only performed in the training phase. In this example, there are 13 frame-level LLDs extracted for each frame. The codebook size is 5.

conditional distribution over the *class* $\boldsymbol{C}$ can be written as:

$$\boldsymbol{p}(\boldsymbol{C}_\lambda|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_\sigma) = \frac{1}{\boldsymbol{Z}}\boldsymbol{p}(\boldsymbol{C}_\lambda)\prod_{i=1}^{\sigma}\boldsymbol{p}(\boldsymbol{x}_i|\boldsymbol{C}_\lambda), \quad (3)$$

where $\boldsymbol{Z} = \boldsymbol{p}(\boldsymbol{x})$ is constant, i.e., independent from the class, when the feature values $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_\sigma$ are known. When constructing a classifier, the *maximum a posteriori*[38] decision rule is used:

$$\hat{y} = \arg\max_{\lambda\in\{1,\ldots,\Lambda\}}\boldsymbol{p}(\boldsymbol{C}_\lambda)\prod_{i=1}^{\sigma}\boldsymbol{p}(\boldsymbol{x}_i|\boldsymbol{C}_\lambda), \quad (4)$$

where $\hat{y}$ is the predicted label.

Evaluation Metrics

Classification Evaluation

The unweighted average recall (UAR)[48] is used as the metric to evaluate the models' performance due to the highly unbalanced distributions of instances among classes in the SnS dataset. The UAR is defined as:

$$UAR = \frac{\sum_{\lambda=1}^{\Lambda}\text{Recall}_\lambda}{\Lambda}, \quad (5)$$

where Λ is the number of classes, and Recall_λ is the class-specific *recall*, i.e., the ratio of instances of class λ that is classified correctly of the λ-th class.

Significance Tests

To compare the difference of the classification performances, i.e., the $UARs$ of two methods, a one-tailed z-test is used in this study. The standard score z can be calculated as[9]:

$$z = \frac{m_\text{A} - m_\text{B}}{\sqrt{2m(1-m)/NS}}, \quad (6)$$

where $m = (m_\text{A} + m_\text{B})/2$, m_A and m_B are the measure value (i.e., UAR) of methods A and B, respectively, NS is the total number of instances. In the one-tailed case (e.g., $m_\text{A} > m_\text{B}$), the p-value is calculated as:

$$p = 1 - \Phi(z) < \alpha, \quad (7)$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function, α is the significance level (e.g., 0.05, 0.01, 0.001). Generally, the p-value represents the probability of rejecting the null hypothesis. A smaller p-value means a more significant difference between the compared two methods. In this study, $p < 0.05$ is considered as the threshold to demonstrate the significance level.

EXPERIMENTAL RESULTS

Experimental Setup

When extracting the LLDs, the whole instance is firstly segmented into numerous chunks (usually 30–40 ms). Then the frame-level LLDs are extracted from these chunks. In our previous study,[40] we found that the frame length and the overlap of the analysed chunks for extraction of LLDs can affect the final classification performance. We transfer the empirical knowledge from massive experiments in Ref. 40 to design the frame length and overlap as Table 3 shows.

On the other hand, some other parameters are optimised *via* initial experiments on the train and development sets. The *wavelet type*, and the maximum decomposition level $\boldsymbol{J}_\text{max}$ are listed in Table 4. The names of *wavelet types* and the decomposition scripts are based on the Wavelet Toolbox[25] of Matlab by MathWorks. The LLDs of SERs and SFFs (subband mean frequency[39]) are extracted based on the subbands at 500 Hz. The original audio files are normalised, mono channel with a sampling rate of 16 kHz and 16 Bit resolution. The ComParE feature set is extracted by the openSMILE toolkit.[12] All other feature sets are extracted by the scripts of Matlab.

Before fed into the classifier, all the features (both for *functionals* and BoAW) are *standardised* to eliminate the effects by outliers. The unbalanced MPSSC data are *upsampled* when training the classifier. In this

TABLE 3. Configurations of each feature set.

	Frame length (ms)	Overlap length (ms)	Dimensions of LLDs
Formants	16	12	3
MFCCs	32	24	13
SERs	32	8	16
SFFs	32	16	19
ComParE	20	10	65
WTE	16	4	16
WPTE	32	16	1023
WEF	64	32	87

TABLE 4. Parameters for wavelet features.

	Wavelet types	J_max
WTE	'bior2.8'	3
WPTE	'haar'	9
WEF	'coif5'	5

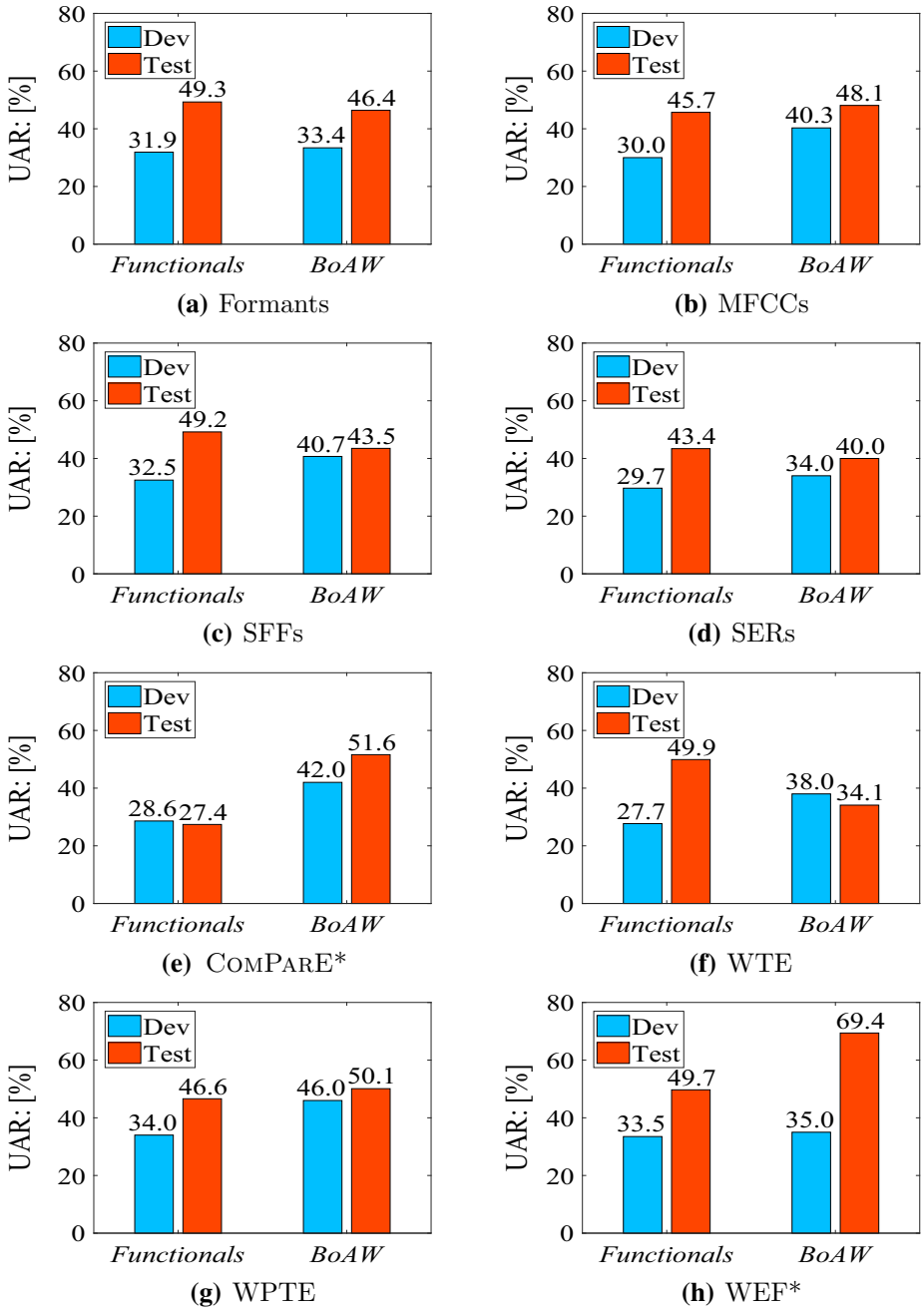

FIGURE 7. Results *UARs* (%)] achieved by each acoustic feature set within *Functionals* and *BoAW*. Feature sets showing significant improvement (p<0.001, one-tailed z-test) by *BoAW* compared with *Functionals* on test set are marked by an asterisk.

study, we manually replicate the MPSSC data to achieve an equal distribution of all snore classes.[49]

The Naïve Bayes classifier is implemented by the open-source toolkit WEKA.[15] The parameters of the classifier (using *kernel density* or *normal* estimator) are optimised on the development set and applied to the test set. The BoAW methodology is implemented using the toolkit described in Ref. 47.

Results

Figure 7 illustrates the experimental results of each feature set both on development and test set. In this study, we find that, on the test set, significant improvements are only achieved with the ComParE and WEF sets ($p < 0.001$, one-tailed z-test).

Among the *Functionals*, WTE and WEF perform best with an UAR of 49.9, and 49.7%, respectively.

TABLE 5. Confusion matrices by the ComParE feature set with *Functionals* and *BoAW*.

Pred →	V	O	T	E	Recall (%)
Functionals					
V	109	35	4	7	70.3
O	36	23	5	1	35.4
T	9	6	0	1	0.0
E	6	18	2	1	3.7
BoAW					
V	49	73	10	23	31.6
O	14	37	3	11	56.9
T	8	0	7	1	43.8
E	1	5	1	20	74.1

TABLE 6. Confusion matrices by the WEF feature set with *Functionals* and *BoAW*.

Pred →	V	O	T	E	Recall (%)
Functionals					
V	95	50	2	8	61.3
O	26	26	1	12	40.0
T	7	7	2	0	12.5
E	2	2	0	23	85.2
BoAW					
V	77	61	8	9	49.7
O	12	44	1	8	67.7
T	4	0	12	0	75.0
E	1	3	0	23	85.2

Furthermore, Formants (*UAR* of 49.3%) and SFFs (*UAR* of 49.2%) show comparable performance to the two aforementioned wavelet features when using *Functionals*. The LLDs of ComParE yield to others when applied with *Functionals*, only reaching an *UAR* of 27.4%.

On *BoAW*, WEF and ComParE achieve the best results among all feature sets, with an *UAR* of 69.4, and 51.6%, respectively. MFCCs are comparable to WPTE (*UAR* 48.1 vs. 50.1%). In particular, LLDs of ComParE considerably improve the performance from 27.4 to 51.6% ($p < 0.001$, one-tailed z-test) when using *BoAW* rather than *Functionals*. In addition, the best model, i.e., the enhanced WEF feature set, reaches an *UAR* of 69.4%, which improves 19.7% from the baseline by *Functionals* ($p < 0.001$, one-tailed z-test). However, for the feature sets of Formants, SFFs, SERs, and WTE, *BoAW* decreases the performances (on the test set) compared with *Functionals*.

Tables 5 and 6 present the confusion matrix of the two best performing models on the test set for ComParE and WEF, respectively. One common finding, both for ComParE and WEF, is that *BoAW* decrease the *Recall* on the recognition of 'V' type snores. Nevertheless, for 'T' type snores, *BoAW* can dramatically improve the *Recall* for ComParE (from 0.0 to 43.8%, $p < 0.001$, one-tailed z-test), and WEF (from 12.5 to 75.0%, $p < 0.001$, one-tailed z-test), which is the main contribution of the improvement in *UAR*. In particular, for ComParE, the *Recall* of recognising 'E' type snores has been improved from 3.7 to 74.1% ($p < 0.001$, one-tailed z-test), which results in another considerable enhanced performance for the final *UAR*. On recognition of 'O' type snores, ComParE and WEF respectively show an increase of 21.5% ($p < 0.01$, one-tailed z-test), and 27.7% ($p < 0.001$, one-tailed z-test) on *Recall* after using *BoAW* instead of *Functionals*.

DISCUSSION

Main Findings in this Study

We can see that, for analysis of SnSs, WT based features outperform the Fourier transformation based features in our experiments. A possible explanation for this excellent performance might be that, the WT contains a better balance of time and spectral information in a non-stationary signal (e.g., SnS) than the traditional Fourier transformation, which is always subject to a Heisenberg-alike time–frequency trade-off.[6]

It is noticeable that the classification results using Formants is comparable to MFCCs, SFFs, and SERs even though it has a low dimension of only three LLDs (see Table 3). Previous studies have shown that Formants have the capability to reveal the status of the human upper airway.[31,32] Future work can be done on finding more sophisticated formant-related features, and combining them with wavelets for the classification of SnS. Most features showed a superior performance to ComParE when applied with *Functionals*. ComParE needs considerable improvement for enabling it to fulfil the task on SnS classification (to some extent, the usage of *BoAW* helps ComParE achieve this target). SERs has not shown efficient performance in this study, which might be caused by the difficulties on designing a reasonable sub-band region for extraction of LLDs. Among the wavelet features, WTE performs best when using *Functionals* while it yields to WPTE and WEF when using *BoAW*. As a complimentary feature to WPTE, WTE makes the final fused feature set WEF reach the best performance with *BoAW*.

From Tables 5 and 6, we can see that, for some rare types of SnS, i.e., 'O', 'T' and 'E', *BoAW* can improve their *Recall*s compared with *Functionals*, which will be beneficial to the unbalanced distribution of MPSSC data. The types 'V' and 'O' are still the most misclassified samples (both for *Functionals* and *BoAW*). Most

TABLE 7. List of results of the INTERSPEECH ComParE Challenge 2017 Snoring sub-challenge on the final *test* set.

	UAR (%)	Main methods
Official baseline[49]	58.5	ComParE features SVM (*linear* kernel)
Amiriparian *et al.*[1†]	67.0	CNN-based spectrum features SVM (*linear* kernel)
Freitag *et al.*[13†]	66.5	CNN-based spectrum features Evolutionary feature selection SVM (*linear* kernel)
Gosztolya *et al.*[14]	64.0	ComParE features MFCC, HNR, F0, ZCR SVM (*linear* kernel)
Kaya and Alexey[18‡]	64.2	ComParE features, MFCC RASTA-PLP, Fisher vector WKPLS, WKELM
Rao *et al.*[42]	52.8	Dual source-filter model SVM (*radial basis function* kernel)
Nwe *et al.*[33]	52.4	MFCC, LPCC, PLPC, spectrogram ComParE features Correlation feature selection SVM, RF, CNN
Tavarez *et al.*[54]	50.6	MFCC, RPS, SC Cosine distance
Our method	69.4	Wavelet features BoAW, Naïve Bayes

CNN convolutional neural network, *SVM* support vector machine, *RASTA-PLP* representations relative spectra perceptual linear prediction, *WKPLS* weighted kernel partial least squares, *WELM* weighted kernel extreme learning machine, *LPCC* linear predictive cepstral coefficient, *PLPC* perceptual linear prediction coefficients, *RF* random forest, *RPS* relative phase shift, *SC* spectral contrast. † Marks the two submissions without participation in the challenge; ‡ marks the submission of the winner in the challenge.

probably, this is due to the small sample size in our dataset, a limitation that should be targeted in future work.

Comparison with the Results of the INTERSPEECH ComParE Challenge 2017 Snoring Sub-challenge

Table 7 shows the main results of submissions to the INTERSPEECH ComParE Challenge 2017 Snoring sub-challenge. Our proposed method outperforms the winner system in this challenge, which achieves the *UAR* of 64.2% on the final test set. Among the excellent results (the ones with above 60.0% *UAR*), we find that, sophisticated features are essential for the final performance of the model. It is noticeable that, some deep learning based techniques[1,13] have been proven to be feasible to extract efficient features for SnS classification. However, the extremely limited number of SnS instances restrained the capacity of models directly using deep neural networks, which was found in our initial experiments and noted in Ref. 14 Since numerous hyper parameters need to be tuned when training deep neural networks, limited data size can easily make the model overfitting.

TABLE 8. Confusion matrix of the best model achieved from the work by Amiriparian et al. The source is from Ref. 1.

Pred →	V	O	T	E	Recall (%)
V	96	27	17	15	61.9
O	19	38	3	5	58.5
T	1	2	10	3	62.5
E	0	2	2	23	85.2

From Tables 5 and 6 we can find that, for this highly unbalanced SnS dataset, a minor change in the recall of the smallest class (i.e., the 'T' class), will dramatically change the resulting *UAR*. Table 8 shows the confusion matrix of the best model from the work by Amiriparian *et al.*[1] Compared with the proposed method in this paper, the best model in Ref. 1 has a better recall performance on the class of 'V' (61.9 vs. 49.7%). For the 'O' and the 'T', the proposed method outperforms the model (67.7 vs. 58.5 and 75.0 vs. 65.2%, respectively) presented by Amiriparian *et al.* The two models show an equal recall on the 'E' class, i.e., 85.2%. We also need to note that, for this database, there is a huge gap between the performance on the *dev* and the *test* set, for both the proposed method and other compared methods. One reasonable explanation for this phenomenon is that, MPSSC is a database collected from three different medical centres, which makes the corpus more complicated than only using one single source data set. Besides, tuning parameters for models is largely dependent on human experiences, which needs to be overcome in future. One possible direction of future work could the combination of deep learnt spectrum features[1,13] with our proposed wavelet features *via* a BoAWs approach.

CONCLUSIONS

In conclusion, we applied an enhanced wavelet features *via* BoAWs approach for the task of classification of snore sounds by excitation localisation in the upper airway. The snore sound data were collected from 219 independent subjects from three medical centres. Experimental results showed that our proposed method achieves 69.4% *UAR*, which significantly outperformed the official baseline of the INTERSPEECH ComParE Challenge 2017 Snoring sub-challenge of 58.5% ($p < 0.005$, one-tailed z-test). In future work, we will continue to collect more SnS data, and apply further state-of-the-art machine learning methods like deep neural networks[21] to study SnS.

ACKNOWLEDGMENTS

This work was partially supported by the China Scholarship Council (CSC), and the European Union's Seventh Framework under Grant Agreements No. 338164 (ERC StG iHEARu).

REFERENCES

[1]Amiriparian, S., M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller. Snore sound classification using image-based deep spectrum features. In: *Proceedings of INTERSPEECH*, 2017, Stockholm, Sweden, pp. 3512–3516.

[2]Arthur, D. and S. Vassilvitskii. K-means + +: the advantages of careful seeding. In: *Proceedings of ACM–SIAM SODA*, 2007, New Orleans, LA, USA, pp. 1027–1035.

[3]Azarbarzin, A. and Moussavi, Z. Automatic and unsupervised snore sound extraction from respiratory sound signals. IEEE Trans. Biomed. Eng. 58(5):1156–1162, 2011.

[4]Coifman, R. R., Y. Meyer, S. Quake, and V. Wickerhauser. Signal processing and compression with wavelet packets. In: *Wavelets and Their Applications*, edited by J. S. Byrnes, J. L. Byrnes, K. A. Hargreaves, and K. Berry. Dordrecht: Springer, 1994, pp. 363–379.

[5]Coifman, R. R. and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. IEEE Trans. Inf. Theory 38(2):713–718, 1992.

[6]De Bruijn, N. Uncertainty principles in Fourier analysis. In: *Inequalities (Proceedings of Symposium of Wright-Patterson Air Force Base, Ohio, 1965)*. New York: Academic , 1967, pp. 57–71.

[7]Deller Jr., J. R., J. H. L. Hansen, and J. G. Proakis. Discrete Time Processing of Speech Signals. New York: Wiley-IEEE Press, 1999.

[8]Demin, H., 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* 112(11):2081–2085, 2002.

[9]Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput*. 10(7):1895–1923, 1998.

[10]Elwali, A. and Z. Moussavi. Obstructive sleep apnea screening and airway structure characterization during wakefulness using tracheal breathing sounds. *Ann. Biomed. Eng*., 45(3):839–850, 2017.

[11]Eyben, F. Real-time Speech and Music Classification by Large Audio Feature Space Extraction. Doctoral Thesis, Springer, Cham, 2015.

[12]Eyben, F., F. Weninger, F. Groß, and B. Schuller. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: *Proceedings of ACM MM*, Barcelona, Catalunya, Spain. ACM, 2013, pp. 835–838.

[13]Freitag, M., S. Amiriparian, N. Cummins, M. Gerczuk, and B. Schuller. An end-to-evolution hybrid approach for snore sound classification. In: *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3507–3511.

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

[15]Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explor. Newsl.* 11(1):10–18, 2009.

[16]Janott, C., M. Schmitt, Y. Zhang, K. Qian, V. Pandit, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller. Snoring classified: the Munich Passau Snore Sound Corpus. *Comput. Biol. Med.* 94:106–118, 2018.

[17]Janott, C., B. Schuller, and C. Heiser. Acoustic information in snoring noise. *HNO* 65(2):107–116, 2017.

[18]Kaya, H. and K. A. Alexey. Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: snoring, addressee and cold. In: *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3527–3531.

[19]Kezirian, E. J., W. Hohenhorst, and N. de Vries. Drug-induced sleep endoscopy: the VOTE classification. *Eur. Arch. Oto-Rhino-Laryngol.* 268(8):1233–1236, 2011.

[20]Khushaba, R. N., S. Kodagoda, S. Lal, and G. Dissanayake. Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. *IEEE Trans. Biomed. Eng*. 58(1):121–131, 2011.

[21]LeCun, Y., Y. Bengio, and G. Hinton. Deep learning. *Nature* 521(7553):436–444, 2015.

[22]Li, K. K. Surgical therapy for adult obstructive sleep apnea. *Sleep Med. Rev*. 9(3):201–209, 2005.

[23]Lin, H.-C., 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* 118(5):902–908, 2008.

[24]Mallat, S. A Wavelet Tour of Signal Processing: The Sparse Way. Burlington: Elsevier, 2009.

[25]MathWorks. *Matlab Wavelet Toolbox*. https://www.mathworks.com/products/wavelet.html, 2018.

[26]Mlynczak, M., E. Migacz, M. Migacz, and W. Kukwa. Detecting breathing and snoring episodes using a wireless tracheal sensor-a feasibility study. *IEEE J. Biomed. Health Inform*. 21(6):1504–1510, 2017.

[27]Mokhlesi, B., 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. *Eur. Respir. J*. 47(4):1162–1169, 2016.

[28]Montazeri, A., E. Giannouli, and Z. Moussavi. Assessment of obstructive sleep apnea and its severity during wakefulness. *Ann. Biomed. Eng*. 40(4):916–924, 2012.

[29]Murty, M. N. and V. S. Devi. Pattern Recognition: An Algorithmic Approach. Dordrecht: Springer, 2011.

[30]Ng, A. K., T. San Koh, U. R. Abeyratne, and K. Puvanendran. Investigation of obstructive sleep apnea using nonlinear mode interactions in nonstationary snore signals. *Ann. Biomed. Eng*. 37(9):1796–1806, 2009a.

[31]Ng, A. K., 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*. 9(8):894–898, 2008.

[32]Ng, A. K., T. San Koh, E. Baey, and K. Puvanendran. Role of upper airway dimensions in snore production: acoustical and perceptual findings. *Ann. Biomed. Eng*. 37(9):1807–1817, 2009b.

[33]Nwe, L. T., 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: *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3467–3471.
[34]O'Shaughnessy, D. Speech Communication: Human and Machine. New York: Addison-Wesley, 1987.
[35]Pancoast, S. and M. Akbacak. Bag-of-audio-words approach for multimedia event classification. In: *Proceedings of INTERSPEECH*, Portland, OR, USA, 2012, pp. 2105–2108.
[36]Peppard, P. E., T. Young, J. H. Barnet, M. Palta, E. W. Hagen, and K. M. Hla. Increased prevalence of sleep-disordered breathing in adults. *Am. J. Epidemiol.* 177(9):1006–1014, 2013.
[37]Pevernagie, D., R. M. Aarts, and M. De Meyer. The acoustics of snoring. *Sleep Med. Rev.* 14(2):131–144, 2010.
[38]Pishro-Nik, H. *Introduction to Probability, Statistics, and Random Processes*. Electrical and Computer Engineering Educational Materials, 2014. http://scholarworks.umass.edu/ece_ed_materials/1.
[39]Qian, K., C. Janott, V. Pandit, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller. Classification of the excitation location of snore sounds in the upper airway by acoustic multi-feature analysis. *IEEE Trans. Biomed. Eng.* 64(8):1731–1741, 2017.
[40]Qian, K., C. Janott, Z. Zhang, J. Deng, A. Baird, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller. Teaching machines on snoring: a benchmark on computer audition for snore sound excitation localisation. *Arch. Acoust.* 43(3):465–475, 2018.
[41]Qian, K., C. Janott, Z. Zhang, C. Heiser, and B. Schuller. Wavelet features for classification of VOTE snore sounds. In: *Proceedings of ICASSP*, Shanghai, China, 2016, pp. 221–225.
[42]Rao, M. V. A., S. Yadav, and P. Ghosh, Kumar. A dual source-filter model of snore audio for snorer group classification. In: *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3502–3506.
[43]Rawat, S., P. F. Schulam, S. Burger, D. Ding, Y. Wang, and F. Metze. Robust audio-codebooks for large-scale event detection in consumer videos. In: *Proceedings of INTERSPEECH*, Lyon, France, 2013, pp. 2929–2933.
[44]Reda, M., G. J. Gibson, and J. A. Wilson. Pharyngoesophageal pressure monitoring in sleep apnea syndrome. *Otolaryngol. Head Neck Surg.* 125(4):324–331, 2001.
[45]Schmitt, M., C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller. A bag-of-audio-words approach for snore sounds excitation localisation. In: *Proceedings of ITG Speech Communication*, Paderborn, Germany, 2016a, pp. 230–234.
[46]Schmitt, M., F. Ringeval, and B. Schuller. At the border of acoustics and linguistics: bag-of-audio-words for the recognition of emotions in speech. In: *Proceedings of INTERSPEECH*, San Francisco, CA, USA, 2016b, pp. 495–499.
[47]Schmitt, M. and B. W. Schuller. openXBOW-introducing the Passau open-source crossmodal bag-of-words toolkit. *J. Mach. Learn. Res.* 18(96):1–5, 2017.
[48]Schuller, B., S. Steidl, and A. Batliner. The INTERSPEECH 2009 emotion challenge. In: *Proceedings of INTERSPEECH*, Brighton, UK, 2009, pp. 312–315.
[49]Schuller, B., S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, S. A. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou. The INTERSPEECH 2017 computational paralinguistics challenge: addressee, cold and snoring. In: *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3442–3446.
[50]Schuller, B., S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In: *Proceedings of INTERSPEECH*, Lyon, France, 2013, pp. 148–152.
[51]Snell, R. C. and F. Milinazzo. Formant location from LPC analysis data. *IEEE Trans. Speech Audio Process.*, 1(2):129–134, 1993.
[52]Strollo Jr., P. J. and R. M. Rogers. Obstructive sleep apnea. *N. Engl. J. Med.* 334(2):99–104, 1996.
[53]Stuck, B. A. and J. T. Maurer. Airway evaluation in obstructive sleep apnea. *Sleep Med. Rev.* 12(6):411–436, 2008.
[54]Tavarez, D., X. Sarasola, A. Alonso, J. Sanchez, L. Serrano, E. Navas, and I. Hernáez. Exploring fusion methods and feature space for the classification of paralinguistic information. In: *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3517–3521.
[55]Vroegop, A. V., O. M. Vanderveken, A. N. Boudewyns, J. Scholman, V. Saldien, K. Wouters, M. J. Braem, P. H. Van de Heyning, and E. Hamans. Drug-induced sleep endoscopy in sleep-disordered breathing: report on 1,249 cases. *Laryngoscope* 124(3):797–802, 2014.
[56]Yadollahi, A., A. Montazeri, A. Azarbarzin, and Z. Moussavi. Respiratory flow-sound relationship during both wakefulness and sleep and its variation in relation to sleep apnea. *Ann. Biomed. Eng.* 41(3):537–546, 2013.