www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

PAN ARCHIVES OF ACOUSTICS
Vol. **43**, No. 3, pp. 465–475 (2018)

DOI: 10.24425/123918

Teaching Machines on Snoring: A Benchmark on Computer Audition for Snore Sound Excitation Localisation

Kun QIAN[(1),(5)], Christoph JANOTT[(2)], Zixing ZHANG[(3)], Jun DENG[(4)], Alice BAIRD[(5)]
Clemens HEISER[(6)], Winfried HOHENHORST[(7)], Michael HERZOG[(8)]
Werner HEMMERT[(2)], Björn SCHULLER[(3),(4),(5)]

[(1)] *Machine Intelligence & Signal Processing Group, Chair of Human-Machine Communication*
Technische Universität München
Munich, Germany; e-mail: andykun.qian@tum.de

[(2)] *Munich School of Bioengineering, Technische Universität München*
Garching, Germany

[(3)] *Group on Language, Audio & Music, Department of Computing, Imperial College London*
London, UK

[(4)] *audEERING GmbH*
Gilching, Germany

[(5)] *ZD.B Chair of Embedded Intelligence for Health Care & Wellbeing, Universität Augsburg*
Augsburg, Germany

[(6)] *Department of Otorhinolaryngology, Head and Neck Surgery, Klinikum rechts der Isar*
Technische Universität München
Munich, Germany

[(7)] *Department of Otorhinolaryngology, Head and Neck Surgery, Alfried Krupp Krankenhaus*
Essen, Germany

[(8)] *Department of Otorhinolaryngology, Head and Neck Surgery, Carl-Thiem-Klinikum Cottbus*
Cottbus, Germany

(*received August 7, 2017; accepted February 28, 2018*)

This paper proposes a comprehensive study on machine listening for localisation of snore sound excitation. Here we investigate the effects of varied frame sizes, and overlap of the analysed audio chunk for extracting low-level descriptors. In addition, we explore the performance of each kind of feature when it is fed into varied classifier models, including support vector machines, k-nearest neighbours, linear discriminant analysis, random forests, extreme learning machines, kernel-based extreme learning machines, multilayer perceptrons, and deep neural networks. Experimental results demonstrate that, wavelet packet transform energy can outperform most other features. A deep neural network trained with subband energy ratios reaches the highest performance achieving an unweighted average recall of 72.8% from four types for snoring.

Keywords: snore sound; obstructive sleep apnea; acoustic features; machine learning.

1. Introduction

Snoring is a typical symptom of Obstructive Sleep Apnea (OSA), a chronic sleep disorder, which affects approximately 13% of men and 6% of women in the US alone (Peppard *et al.*, 2013). OSA is defined as a syndrome of cessation or reduction of airflow during sleep, caused by complete (apnea) or partial (hy-

popnea) collapse of the upper airway for more than ten seconds, and with five or more episodes per hour (Strollo Jr, Rogers, 1996). When untreated, OSA increases the risk of cardiovascular diseases, stroke, hypertension, myocardial infarction, diabetes as well as vulnerability to being accident prone (Young *et al.*, 1993; Peppard *et al.*, 2000; Yaggi *et al.*, 2005; Marin *et al.*, 2005; Mokhlesi *et al.*, 2016). Snoring is a by – product of OSA with more than 80% of sufferers reporting to experience it (Aldrich, 1999). In the past two decades, a number of studies have been published focusing on the combination of acoustic information with machine learning to support diagnosis of OSA by analysing acoustic events during the subjects' sleep (Pevernagie *et al.*, 2010; Roebuck *et al.*, 2014). Aiming to facilitate diagnosis and to complement the current diagnostic gold standard, i.e., polysomnography (PSG).

Despite this, there are limited publications on the use of machine learning to localise snore sound (SnS) excitation. Such studies may facilitate a more targeted and less invasive surgical approach. Among those studies on discrimination of snore excitation sites by acoustic methods, frequency features (Miyazaki *et al.*, 1998; Agrawal *et al.*, 2002), amplitude features (Hill *et al.*, 1999), statistical time series features (Beeton *et al.*, 2007), and psychoacoustic features (Herzog *et al.*, 2014) were evaluated for their suitability for SnS classification. However, no machine learning methods were used. Qian *et al.* (2013; 2014; 2015) published pilot work on classification of different SnS events from overnight audio recordings. Nevertheless, their proposed methods did not prove to be efficient for localisation of excitation of SnS. Recently, novel acoustic features like wavelet features (Qian *et al.*, 2016), and learning methods like bag-of-audio-words (Schmitt *et al.*, 2016) were proposed aiming at classification of SnS generated in different locations of the upper airway. However, the studies were based on a limited number of only 24 subjects, and did not use a development set to tune robust model learning.

In this work, we further the study in (Qian *et al.*, 2017), aim to make a comprehensive investigation on the effects of frame size and overlaps of SnS chunk for extraction of low-level descriptors. In addition, we incorporate deep neural networks for SnS classification, and compare the performance of each kind of feature set extracted from the chunk within optimised frame size and overlap by varied classifiers. The article will be organised as follows: Firstly in Sec. 2, we will explain and clarify the relation to prior work. Then, the database and methods used will be described in Sec. 3. We show the experimental results and give a discussion in Sec. 4. Finally, a conclusion will be made in Sec. 5.

2. Relation to prior work

This work is a continuation based on (Qian *et al.*, 2017), which proposed an acoustic multi-feature method for SnS classification. The main contributions of this work are: Firstly, we investigate the effects of fame size and overlap for analysed chunk for extraction of low-level descriptors, which were ignored by most of the previous studies. These findings could be important prior knowledge for further study on recurrent neural networks (Sak *et al.*, 2014), convolutional neural networks (Abdel-Hamid *et al.*, 2014), and bag-of-audio-words (Pancoast, Akbacak, 2012). Secondly, a refined set of functionals are used to eliminate the feature selection process by human-involved parameter settings in (Qian *et al.*, 2017). In addition, we give a comparison between the proposed features and the state-of-the-art acoustic feature extracted by openSMILE toolkit (Eyben *et al.*, 2010; 2013). Finally, a significant level analysis will be given through different feature sets fed into varied classifiers including deep neural networks.

3. Materials and methods

In this section, we will firstly give a brief description on the database (see Subsec. 3.1) used in this work. Then, the methodology will be proposed separately as *Acoustic Features* (see Subsec. 3.2), and *Classifiers* (see Subsec. 3.3).

3.1. Database

This study was approved by the ethic committee of Klinikum rechts der Isar, Technische Universität München, Germany. The SnS data was collected from three sites: Klinikum rechts der Isar, Munich, Germany; Alfried Krupp Hospital, Essen, Germany; and, University Hospital Halle (Saale), Germany. Videos were recorded in MP4 format with an image resolution of 720×288 pixels (Munich), 720×544 pixels (Essen), or 1280×720 pixels (Halle) and 25 frames/s (all centres). Audio embedded in the MP4 file was recorded at a sample rate of 44 kHz (all centres) and a bitrate of 128 kBit/s (Munich and Halle), or 705 kBit/s (Essen). For further processing in our experiments, the audio information was extracted from the MP4 file and converted into wav format with a sample rate of 16 kHz and a resolution of 16 bit (refer to (Qian *et al.*, 2017)). Figure 1 shows a SnS data acquisition system setting in Munich, Germany. All the subjects underwent a drug induced sleep endoscopy (DISE) (El Badawey *et al.*, 2003). By watching the videos, experts on ENT (ear, nose, and throat) labelled the SnS data with 'V' (the level of the velum), 'O' (the oropharyngeal area), 'T' (the tongue base), and 'E' (the level of the epiglottis)

Fig. 1. Snore sound data acquisition system (Munich). A flexible nasopharyngoscope was used for recording the video of the upper airway (Storz, Germany at the Munich and Halle sites; Olympus, Germany, at the Essen site) connected to a video recording system (Telepack X, Storz, Germany, at the Munich site; AIDA, Storz, Germany, at the Halle site; rpSzene, Rehder/Partner, Hamburg, Germany, at the Essen site).

(refer to Fig. 2) localisation classification (Kezirian *et al.*, 2011).

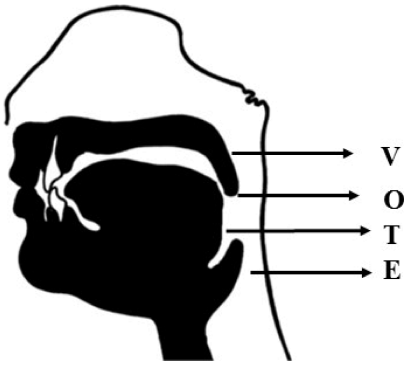

Fig. 2. The anatomical positions of 'V', 'O', 'T', and 'E' in the upper airway. 'V' represents the level of the velum. 'O' represents the oropharyngeal area. 'T' represents the tongue base. 'E' represents the level of the epiglottis.

In total, 164 snoring episodes from 40 independent male patients (with primary or OSA snoring) showing clearly identifiable, single source snoring sounds have been selected for this study. To generate sufficient data for machine learning, we segmented 164 snoring episodes into single 200 ms episodes, which share 50% overlap with neighbours. In addition, we partitioned the whole data set into train, dev (development) and test sets. The classifiers' parameters will be optimised by dev set, and applied to a test set. The train, dev, and test sets are all from independent subjects. Detailed information about data sets can be found in Table 1 and Table 2.

Table 1. Demographic information of the patients in train, dev (development) and test set, respectively. BMI: Body Mass Index; AHI: Apnea Hypopnea Index.

		mean	std	range
Age [years]	train	49.1	±12.03	26.0–71.0
	dev	45.0	±5.66	37.0–50.0
	test	44.3	±14.31	26.0–64.0
BMI [kg/m^2]	train	27.4	±3.31	22.8–38.4
	dev	26.6	±3.45	21.2–31.0
	test	25.4	±1.12	23.9–27.5
AHI [events/h]	train	23.5	±14.07	1.3–59.1
	dev	20.1	±7.54	9.9–28.0
	test	17.5	±12.86	6.2–44.0

Table 2. Number of segments/[independent subjects: #] for each snore type in train, dev (dev) and test set, respectively. 'V' represents the level of the velum. 'O' represents the oropharyngeal area. 'T' represents the tongue base. 'E' represents the level of the epiglottis.

#	train	dev	test	$\sum$
V	363/[7]	104/[2]	152/[2]	619/[11]
O	326/[7]	125/[2]	122/[2]	573/[11]
T	289/[4]	90/[2]	78/[2]	457/[8]
E	323/[6]	96/[2]	148/[2]	567/[10]
$\sum$	1301/[24]	415/[8]	500/[8]	2216/[40]

3.2. Acoustic features

When extracting acoustic features, we firstly calculate the low-level descriptors (LLDs) from the frame-level of the SnS data. Then statistical functionals will be applied to the time series of LLDs. The details on calculating functionals based on LLDs can be referred to (Eyben, 2015).

3.2.1. Low-level descriptors

In this work, we comprehensively investigate and compare twelve feature set types, which can be grouped as three families as:

a) ***Conventional feature sets***: *crest factor* (CF, the maximum absolute value divided by the root-mean-square of the digitised amplitude values of the audio waveform), *fundamental frequency* (F0, the lowest frequency of a periodic audio waveform), *power ratio* at 800 Hz (PR_{800}, the ratio of the spectrum energy below 800 Hz to that above 800 Hz), *formant (1–3)* (formants, the first three frequencies and their corresponding amplitudes of the spectral peaks in the sound spectrum), *spectral frequency features* (SFFs, the centre frequency, the peak frequency, the mean frequency, and the 1 kHz-subband mean frequency in the sound spectrum), *subband energy ratios* (SERs, the ratios of energy in each 1 kHz-subband to that of the whole

sound spectrum), and *Mel-frequency cepstral coefficients* (MFCCs 0-12, the coefficients of the cepstrum in the Mel scale frequency). The detailed definitions and formulas of each feature set above can be found in (Qian *et al.*, 2017).

b) ***openSMILE feature sets***: We use the toolkit, openSMILE (Eyben *et al.*, 2010; 2013), to extract the acoustic temporal and spectral features of SnS. In this study, two popular feature sets are extracted by openSMILE, i.e., ComParE feature set (proposed as the official baseline feature set of the INTERSPEECH Computational Paralinguistics Challenge (Schuller *et al.*, 2013)), and eGeMAPS feature set (proposed as a refined feature set for speech emotion recognition with massive experiments (Eyben *et al.*, 2016)). The LLDs are listed in Table 3 and Table 4 for ComParE and eGeMAPS, respectively.

Table 3. The low-level descriptors (LLDs) for ComParE feature set. MFCCs: Mel-Frequency Cepstral Coefficients; RASTA: Relative Spectral Transform; HNR: Harmonics to Noise Ratio; RMSE: Root Mean Square Energy.

55 spectral LLDs	Group
MFCCs 1–14	cepstral
Psychoacoustic sharpness, harmonicity	spectral
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	spectral
Spectral energy 250–650 Hz, 1–4 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
6 voicing related LLDs	Group
F_0 (SHS and Viterbi smoothing)	prosodic
Prob. of voicing	voice qual.
log. HNR, jitter (local and δ), shimmer (local)	voice qual.
4 energy related LLDs	Group
RMSE, zero-crossing rate	prosodic
Sum of auditory spectrum (loudness)	prosodic
Sum of RASTA-filtered auditory spectrum	prosodic

Table 4. The low-level descriptors (LLDs) for eGeMAPS feature set. MFCCs: Mel-Frequency Cepstral Coefficients; HNR: Harmonics to Noise Ratio.

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

c) ***Wavelet feature sets***: The wavelet features were introduced and proven to be successful in VOTE SnS classification task in (Qian *et al.*, 2016). In this study, we separately investigate three kinds of wavelet feature sets, wavelet transform energy (WTE), wavelet packet transform energy (WPTE), and wavelet energy feature (WEF, an early fusion of WTE and WPTE). The detailed definitions of WTE, WPTE, and WEF can be found in (Qian *et al.*, 2017).

3.2.2. Statistical functionals

For *conventional feature sets* and *wavelet feature sets*, we apply four functionals, i.e., the maximum, mean and minimum values, and the bias of the estimated linear regression of the frame-level features, which had been demonstrated to be efficient in (Qian *et al.*, 2017). For *openSMILE feature sets*, we use the expert-designed statistical functionals (can be referred to (Eyben, 2015)) applied to LLDs of ComParE, and eGeMAPS, respectively. The final dimensions of each feature set are listed in Table 5. Before feeding into the classifier, all the features were standardised from the information give in the train set, and applied into dev, or test set.

Table 5. Dimension of each feature set. Features with a varied dimension are marked by an asterisk.

	16 ms	32 ms	64 ms
CF	4	4	4
F0	4	4	4
PR_{800}	4	4	4
Formants	24	24	24
SFFs	44	44	44
SERs	32	32	32
MFCCs	52	52	52
ComParE	6373	6373	6373
eGeMAPS	88	88	88
WTE*	96	112	128
WPTE*	252	508	1020
WEF*	348	620	1148

3.3. Classifiers

To make a comprehensive study, we compare the performance of Support Vector Machine (SVM) (Cortes, Vapnik, 1995), k-Nearest Neighbours (k-NN) (Bishop, 2006), Linear Discriminant Analy-

Table 6. Main parameters setting grids for each classifier.

Classifiers	Main Parameters
SVM	kernels: 'linear', 'polynomial', 'radial basis function', 'sigmoid'; C-value: 10^{-5}, 10^{-4}, ..., 10^{4}, 10^{5}
k-NN	k-value: 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100; distance metrics: 'euclidean', 'cityblock', 'chebychev','correlation', 'cosine', 'hamming', 'jaccard', 'minkowski', 'seuclidean', 'spearman'
LDA	discriminant type: 'linear','diaglinear','pseudolinear'; gamma: 0:0.05:1.00
RF	number of trees: 2^{1}, 2^{2}, ..., 2^{9}, 2^{10}; fraction for the treebagger: 0.1:0.1:1.00
ELM	activation functions: 'signmoidal', 'sine', 'hardlim', 'tribas', 'radbas'; number of hidden neurons: 2^{1}, 2^{2}, $\cdots$, 2^{14}
KELM	kernels: 'radial basis function', 'linear', 'polynomial', 'wavelet'; regularization coefficients: 10^{-5}, 10^{-4}, ..., 10^{4}, 10^{5}
MLP	two hidden layers; neurons: 2^{1}, 2^{2}, ..., 2^{9}, 2^{10}
DNN	structured by two-layer stacked auto-encoders, neurons: [64 64], L_2: 10^{-3}, ..., 10^{3}; Sparsity Proportion: 0.1:0.1:0.9; Sparsity Regularization: 2

sis (LDA) (Bishop 2006), Random Forests (RF) (Breiman, 2001), Extreme Learning Machine (ELM) (Huang, 2006), Kernel-based Extreme Learning Machine (KELM) (Huang *et al.*, 2014), Multilayer Perceptrons (MLP) (Basheer, Hajmeer, 2000), and stacked auto-encoder based Deep Neural Network (DNN) (Vincent *et al.*, 2010). The experiments on comparing the effects by frame size and overlap are done with the classifier of SVM, which is a popular classifier for baseline testing. The main parameters of classifiers are optimised by the dev set and applied to test set. The grids for tuning parameters of each classifier are shown in Table 6.

4. Experimental results

In this section, we will show the experimental setup and the results. The discussion based on the experiments will be given in next section.

4.1. Experimental setup

The main experiments are done in a software environment of Matlab by Math Works. The feature sets ComParE, and eGeMAPS are extracted by openSMILE toolkit (Eyben *et al.*, 2013). The other feature sets are extracted by Matlab scripts used in (Qian *et al.*, 2017). The SVM classifier is implemented with the popular toolkit LIBSVM (Chang, Lin, 2011). The other classifiers are implemented by Matlab scripts.

Considering the natural unbalanced characteristic of SnS data (Hessel, de Vries, 2002; Fiz, Jane, 2012), we use the unweighted average recall (UAR) as the metric to evaluate the classification performance. The UAR (Schuller *et al.*, 2009) is defined as:

$$\mathrm{UAR} = \frac{\sum_{i=1}^{\kappa} \mathrm{Recall}_i}{\kappa}, \tag{1}$$

where κ is the total class of the data (here in our study $\kappa = 4$).

In this work, we calculate the mean and the std (standard deviation) values of the UARs of three experiments: train vs dev (the classifier is trained by train set and evaluated by dev set), train vs test (the classifier is trained by train set and evaluated by test set), and train+dev vs test (the classifier is trained by combination of train and dev sets, and evaluated by test set). This is to minimise the effects of some incident experimental results by the limitations of the number of segments in both train and dev sets (less than 1000).

4.2. Results and discussion

The results of varied frame size and overlap by each kind of feature set are shown in Tables 7–9, in which the classifier is SVM. When feeding into different classifiers, the frame size and overlap are chosen as the one with the best averaged performance. Tables 10–12 illustrate the results on each kind of feature set trained by varied classifiers.

This study is a further investigation of (Qian *et al.*, 2017), in which effects cause by frame size and overlap were not considered. We can find that, the pre-defined

Table 7. The Unweighted Average Recalls [%] achieved by conventional feature sets within varied frame sizes and overlaps (classifier: SVM). The parameters of classifiers are optimized by dev set and applied to test set.

	Frame Size	16 ms	16 ms	16 ms	32 ms	32 ms	32 ms	64 ms	64 ms	64 ms
	Overlap	25%	50%	75%	25%	50%	75%	25%	50%	75%
CF	train vs dev	39.8	41.8	39.6	39.1	39.1	39.5	35.7	36.1	36.6
	train vs test	41.2	35.4	41.6	40.4	39.8	37.3	36.1	33.4	37.2
	train+dev vs test	43.5	35.9	40.7	40.5	39.1	39.3	36.5	33.2	32.2
	mean	**41.5**	37.7	40.6	40.0	39.3	38.7	36.1	34.2	35.3
	std	±1.87	±3.56	±1.00	±0.78	±0.40	±1.22	±0.40	±1.62	±2.73
F0	train vs dev	44.7	43.2	43.1	46.3	40.6	41.3	41.6	36.2	34.2
	train vs test	25.9	25.8	27.6	25.6	25.0	29.5	28.3	30.0	32.9
	train+dev vs test	24.2	24.4	30.1	25.2	26.0	31.1	30.1	31.4	28.7
	mean	31.6	31.1	33.6	32.4	30.5	**34.0**	33.3	32.5	31.9
	std	±11.38	±10.47	±8.32	±12.07	±8.73	±6.40	±7.22	±3.25	±2.87
PR_{800}	train vs dev	40.6	39.2	37.5	36.5	35.6	36.0	36.3	33.7	33.6
	train vs test	30.4	34.5	32.6	34.4	33.2	35.5	34.0	32.0	32.4
	train+dev vs test	29.8	34.8	30.6	32.5	33.3	35.2	30.6	32.1	32.0
	mean	33.6	**36.2**	33.6	34.5	34.0	35.6	33.6	32.6	32.7
	std	±6.07	±2.63	±3.55	±2.00	±1.36	±0.40	±2.87	±0.95	±0.83
Formants	train vs dev	44.9	46.5	45.1	42.6	41.3	42.6	42.0	42.2	42.4
	train vs test	56.6	53.2	53.9	49.1	54.8	52.3	50.1	48.3	49.8
	train+dev vs test	50.1	49.7	53.3	45.4	47.6	41.7	48.4	44.5	49.9
	mean	50.5	49.8	**50.8**	45.7	47.9	45.5	46.8	45.0	47.4
	std	±5.86	±3.35	±4.92	±3.26	±6.75	±5.88	±4.27	±3.08	±4.30
SFFs	train vs dev	43.3	51.7	46.0	49.9	52.0	49.5	51.9	50.9	52.6
	train vs test	55.7	46.4	55.9	55.9	56.1	48.1	51.4	37.5	51.7
	train+dev vs test	56.3	30.5	51.1	57.2	56.4	42.7	53.9	33.8	49.8
	mean	51.8	42.9	51.0	54.3	**54.8**	46.8	52.4	40.7	51.4
	std	±7.34	±11.03	±4.95	±3.89	±2.46	±3.59	±1.32	±9.00	±1.43
SERs	train vs dev	44.3	44.7	46.1	46.1	46.9	47.2	46.3	46.3	45.7
	train vs test	58.0	57.3	58.1	66.0	56.8	61.2	63.8	63.9	59.8
	train+dev vs test	60.2	60.7	61.8	66.1	62.4	60.4	61.1	61.4	59.2
	mean	54.2	54.2	55.3	**59.4**	55.4	56.3	57.1	57.2	54.9
	std	±8.62	±8.43	±8.21	±11.52	±7.85	±7.86	±9.42	±9.52	±7.97
MFCCs	train vs dev	58.3	58.2	58.8	59.8	59.0	59.8	59.2	58.3	59.6
	train vs test	43.0	39.8	35.3	34.0	39.9	41.5	40.6	41.2	36.6
	train+dev vs test	36.0	37.0	33.4	38.3	42.3	40.7	41.2	42.0	43.4
	mean	45.8	45.0	42.5	44.0	47.1	**47.3**	47.0	47.2	46.5
	std	±11.40	±11.52	±14.15	±13.82	±10.40	±10.80	±10.57	±9.65	±11.82

Table 8. Unweighted Average Recalls [%] achieved by openSMILE feature sets within varied frame sizes and overlaps (classifier: SVM). The parameters of classifiers are optimized by dev set and applied to test set.

	Frame Size	16 ms	16 ms	16 ms	32 ms	32 ms	32 ms	64 ms	64 ms	64 ms
	Overlap	25%	50%	75%	25%	50%	75%	25%	50%	75%
ComParE	train vs dev	25.0	48.4	51.2	25.0	25.0	49.5	25.0	25.0	25.0
	train vs test	25.0	59.5	55.3	25.0	25.0	55.3	25.0	25.0	25.0
	train+dev vs test	25.0	53.0	49.6	25.0	25.0	52.1	25.0	25.0	25.0
	mean	25.0	**53.6**	52.0	25.0	25.0	52.3	25.0	25.0	25.0
	std	±0.00	±5.58	±2.94	±0.00	±0.00	±2.91	±0.00	±0.00	±0.00
eGeMAPS	train vs dev	49.1	50.3	48.8	25.0	25.0	49.3	25.0	25.0	25.0
	train vs test	51.9	49.9	49.0	25.0	25.0	49.7	25.0	25.0	25.0
	train+dev vs test	48.7	49.3	50.0	25.0	25.0	51.5	25.0	25.0	25.0
	mean	49.9	49.8	49.3	25.0	25.0	**50.2**	25.0	25.0	25.0
	std	±1.74	±0.50	±0.64	±0.00	±0.00	±1.17	±0.00	±0.00	±0.00

Table 9. Unweighted Average Recalls [%] achieved by wavelet feature sets within varied frame sizes and overlaps (classifier: SVM). The parameters of classifiers are optimized by dev set and applied to test set.

	Frame Size	16 ms	16 ms	16 ms	32 ms	32 ms	32 ms	64 ms	64 ms	64 ms
	Overlap	25%	50%	75%	25%	50%	75%	25%	50%	75%
WTE	train vs dev	49.2	51.0	50.4	52.4	48.7	48.7	47.8	47.7	49.5
	train vs test	63.0	52.2	50.1	55.0	43.8	64.2	41.2	63.7	66.4
	train+dev vs test	63.3	51.0	53.3	52.7	42.3	56.0	51.2	59.2	57.7
	mean	**58.5**	51.4	51.3	53.4	44.9	56.3	46.7	56.9	57.9
	std	±8.06	±0.69	±1.77	±1.42	±3.35	±7.75	±5.08	±8.25	±8.45
WPTE	train vs dev	49.7	50.4	49.5	52.1	51.2	52.8	50.4	50.5	50.8
	train vs test	67.3	70.9	67.4	59.9	70.6	58.2	55.4	60.9	65.4
	train+dev vs test	70.9	70.2	69.2	63.5	70.3	62.7	64.8	59.6	66.0
	mean	62.6	63.8	62.0	58.5	**64.0**	57.9	56.9	57.0	60.7
	std	±11.34	±11.64	±10.89	±5.83	±11.12	±4.96	±7.31	±5.67	±8.61
WEF	train vs dev	47.1	46.7	48.4	50.6	49.2	51.6	53.1	51.7	51.1
	train vs test	62.4	64.1	61.9	74.4	73.7	72.9	62.1	73.6	59.4
	train+dev vs test	67.6	65.4	65.3	71.2	72.7	70.2	69.3	72.3	63.4
	mean	59.0	58.7	58.5	65.4	65.2	64.9	61.5	**65.9**	58.0
	std	±10.66	±10.44	±8.94	±12.92	±13.87	±11.60	±8.12	±12.29	±6.27

Table 10. Unweighted Average Recalls [%] achieved by conventional feature sets within varied classifiers. The parameters of classifiers are optimized by dev Set and applied to test set.

	Classifiers	SVM	*k*-NN	LDA	RF	ELM	KELM	MLP	DNN
CF	train vs dev	39.8	40.6	34.3	37.9	41.9	39.4	43.5	34.6
	train vs test	41.2	42.2	41.7	34.5	38.3	40.6	39.1	34.9
	train+dev vs test	43.5	41.4	41.7	37.5	40.7	43.0	35.9	35.4
	mean	**41.5**	41.4	39.2	36.6	40.3	41.0	39.5	35.0
	std	±1.87	±0.80	±4.27	±1.86	±1.83	±1.83	±3.82	±0.40
F0	train vs dev	41.3	41.9	38.1	44.4	42.1	41.6	46.2	43.1
	train vs test	29.5	28.4	34.2	28.1	28.9	29.6	29.0	27.8
	train+dev vs test	31.1	29.0	28.2	27.6	29.1	29.9	30.0	28.2
	mean	34.0	33.1	33.5	33.4	33.4	33.7	**35.1**	33.0
	std	±6.40	±7.63	±4.99	±9.56	±7.56	±6.84	±9.65	±8.72
PR_{800}	train vs dev	39.2	33.6	30.9	33.8	36.3	36.0	36.1	34.9
	train vs test	34.5	35.3	33.4	32.6	35.7	35.7	34.2	32.1
	train+dev vs test	34.8	34.0	30.0	28.4	31.1	33.6	36.4	35.8
	mean	**36.2**	34.3	31.4	31.6	34.4	35.1	35.6	34.3
	std	±2.63	±0.89	±1.76	±2.84	±2.84	±1.31	±1.19	±1.93
Formants	train vs dev	45.1	45.3	43.1	43.3	45.3	45.1	46.5	45.0
	train vs test	53.9	63.4	59.4	66.6	60.7	59.4	62.5	59.0
	train+dev vs test	53.3	55.7	54.2	58.8	55.6	53.4	52.0	49.7
	mean	50.8	54.8	52.2	**56.2**	53.9	52.6	53.7	51.2
	std	±4.92	±9.08	±8.33	±11.86	±7.84	±7.18	±8.13	±7.12
SFFs	train vs dev	52.0	49.0	51.7	48.1	48.0	53.2	52.3	50.7
	train vs test	56.1	48.6	58.8	55.4	50.9	58.6	56.1	44.3
	train+dev vs test	56.4	37.1	53.9	53.0	46.7	53.9	50.1	35.5
	mean	54.8	44.9	54.8	52.2	48.5	**55.2**	52.8	43.5
	std	±2.46	±6.76	±3.63	±3.72	±2.15	±2.94	±3.04	±7.63
SERs	train vs dev	46.1	50.0	34.0	46.9	45.8	43.3	48.9	50.9
	train vs test	66.0	57.5	56.6	62.2	59.3	68.3	57.0	72.8
	train+dev vs test	66.1	59.6	57.1	60.6	62.8	70.0	56.9	70.5
	mean	59.4	55.7	49.2	56.6	56.0	60.5	54.3	**64.7**
	std	±11.52	±5.05	±13.19	±8.41	±8.98	±14.95	±4.65	±12.04
MFCCs	train vs dev	59.8	55.2	60.2	60.5	55.0	61.9	65.4	62.8
	train vs test	41.5	60.1	56.8	53.4	50.4	57.1	50.1	50.4
	train+dev vs test	40.7	50.9	54.6	48.4	48.9	52.3	49.0	57.1
	mean	47.3	55.4	**57.2**	54.1	51.4	57.1	54.8	56.8
	std	±10.80	±4.60	±2.82	±6.08	±3.18	±4.80	±9.17	±6.21

Table 11. Unweighted Average Recalls [%] achieved by openSMILE feature sets within varied classifiers. The parameters of classifiers are optimized by dev set and applied to test set.

	Classifiers	SVM	k-NN	LDA	RF	ELM	KELM	MLP	DNN
ComParE	train vs dev	48.4	49.0	49.3	55.1	51.3	51.0	50.7	51.6
	train vs test	59.5	48.6	50.6	64.6	59.5	50.6	55.3	53.0
	train+dev vs test	53.0	49.8	51.9	71.9	55.5	50.5	51.9	55.8
	mean	53.6	49.1	50.6	**63.9**	55.4	50.7	52.6	53.5
	std	±5.58	±0.61	±1.30	±8.42	±4.10	±0.26	±2.39	±2.14
eGeMAPS	train vs dev	49.3	48.3	51.7	51.7	46.2	49.0	51.1	49.1
	train vs test	49.7	42.2	47.7	54.7	49.9	46.1	47.1	36.1
	train+dev vs test	51.5	44.1	47.8	56.8	44.9	47.8	39.5	40.2
	mean	50.2	44.9	49.1	**54.4**	47.0	47.6	45.9	41.8
	std	±1.17	±3.12	±2.28	±2.56	±2.59	±1.46	±5.89	±6.65

Table 12. Unweighted Average Recalls [%] achieved by wavelet feature sets within varied classifiers. The parameters of classifiers are optimized by dev set and applied to test set.

	Classifiers	SVM	k-NN	LDA	RF	ELM	KELM	MLP	DNN
WTE	train vs dev	49.2	52.1	53.0	47.4	48.7	53.3	51.9	47.7
	train vs test	63.0	55.8	57.8	63.3	57.5	59.9	39.3	50.8
	train+dev vs test	63.3	55.4	55.4	67.2	54.8	59.4	25.0	48.2
	mean	58.5	54.4	55.4	**59.3**	53.7	57.5	38.7	48.9
	std	±8.06	±2.03	±2.40	±10.49	±4.51	±3.67	±13.46	±1.66
WPTE	train vs dev	51.2	50.6	48.6	49.6	54.3	54.0	57.7	55.9
	train vs test	70.6	64.5	63.1	59.8	61.5	54.7	54.8	52.3
	train+dev vs test	70.3	64.7	63.6	65.9	70.7	60.6	51.5	58.6
	mean	**64.0**	59.9	58.4	58.4	62.2	56.4	54.7	55.6
	std	±11.12	±8.08	±8.52	±8.24	±8.22	±3.63	±3.10	±3.16
WEF	train vs dev	51.7	52.4	47.3	49.8	52.7	48.5	56.6	55.5
	train vs test	73.6	68.3	61.4	59.5	57.9	65.7	57.4	59.5
	train+dev vs test	72.3	70.0	62.6	64.3	59.5	66.5	60.7	58.8
	mean	**65.9**	63.6	57.1	57.9	56.7	60.2	58.2	57.9
	std	±12.29	±9.71	±8.51	±7.39	±3.56	±10.17	±2.17	±2.14

frame size and overlap for extraction of LLDs considerably effect the SnS classification (see Tables 7–9). Specifically, for features extracted by openSMILE, i.e., ComParE and eGeMAPS, the frame size of 64 ms can lead to a failed SVM classifier (chance level: 25.0%). Among the compared feature sets, five feature sets (CF, PR_{800}, Formants, ComParE, and WTE) achieved the best averaged performance within 16 ms frame size, and the other six feature sets (F0, SFFs, SERs, MFCCs, eGeMAPS, and WPTE) will be optimised by using 32 ms frame size. Only WEF reached its best performance by using 64 ms frame size within a 50% overlap, which was the configuration used in (Qian *et al.*, 2017). These results can help to build a prior knowledge to find the suitable frame size and overlap of the analysed audio chunk for extracting low-level descriptors.

In addition, we can learn that, the classification performance by each feature set varied among classifiers. As frequently-used classifiers, SVM, and RF can respectively reach the best averaged performance by four kinds of feature sets (CF, PR_{800}, WPTE, and WEF). As the state-of-the-art machine learning techniques, KELM, and DNN are suitable for SFFs, and SERs feature set respectively. In particular, the SERs feature set fed into DNN model can reach 72.8% UAR validated by test set, which is the highest performance among others. We calculate the significant levels (one sided Student's t-test (Spiegel *et al.*, 2009)) by comparing classifiers' averaged performance (see Table 13). We can see that, WPTE feature set can significantly outperform most other feature sets, except SERs and WEF. SERs feature set is ranked as the third place among all the feature sets. Formants, MFFCs, ComParE, and WTE share similar performance while in this study CF, SFFs, and eGeMAPS show limited classification capacity. F0 and PR_{800} are ranked as least efficient features in this work. Nevertheless, there are no significant differences between varied classifiers when fed with a same feature set in this study.

Table 13. Significance levels of the averaged UARs obtained from the statistical comparison (one sided Student's t-test) between different features by fed into varied classifiers.

Sign. Levels	CF	F0	PR_{800}	Formants	SFFs	SERs	MFCCs	ComParE	eGeMAPS	WTE	WPTE	WEF
CF				\	\	\	\	\	\	\	\	\
F0	\		\	\	\	\	\	\	\	\	\	\
PR_{800}	\	\		\	\	\	\	\	\	\	\	\
Formants					\	\	\	\		\	\	\
SFFs				\		\	\	\	\	\	\	\
SERs							\	\		\	\	\
MFCCs				\	\	\		\		\	\	\
ComParE				\	\	\	\			\	\	\
eGeMAPS				\	\	\	\	\		\	\	\
WTE				\	\	\	\	\			\	\
WPTE						\						\
WEF						\					\	

\ $p > 0.05$ $p < 0.05$ $p < 0.01$ $p < 0.001$

5. Conclusion

In this paper, we proposed a benchmark on machine listening for localisation of snore sound excitation. We found that, frame size, and degree of overlap for the analysed chunks being used for extraction of low-level descriptors, does effect the performance of snore sound classification. This helps us to establish prior knowledge and to set suitable frame size and overlap of snore sound when using sequential techniques like a recurrent neural network, or convolutional neural network. The wavelet packet transform energy, proved to be superior than most other feature sets. The subband energy ratios, when fed into a deep neural network can reach an unweighted average recall at 72.8% by independent subjects. Future work will be done on studying the sequence based learning techniques for VOTE SnS classification.

Acknowledgment

This work was partially supported by the China Scholarship Council (CSC), the European Union's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu), No. 688835 (RIA DE-ENIGMA) and and the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

References

1. Abdel-Hamid O., Mohamed A.-R., Jiang H., Deng L., Penn G., Yu D. (2014), *Convolutional neural networks for speech recognition*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, **22**, 10, 1533–1545.

2. Agrawal S., Stone P., McGuinness K., Morris J., Camilleri A. (2002), *Sound frequency analysis and the site of snoring in natural and induced sleep*, Clinical Otolaryngology & Allied Sciences, **27**, 3, 162–166.

3. Aldrich M.S. (1999), *Sleep medicine*, Oxford University Press, New York, USA.

4. Basheer I., Hajmeer M. (2000), *Artificial neural networks: fundamentals, computing, design, and application*, Journal of Microbiological Methods, **43**, 1, 3–31.

5. Beeton R.J., Wells I., Ebden P., Whittet H., Clarke J. (2007), *Snore site discrimination using statistical moments of free field snoring sounds recorded during sleep nasendoscopy*, Physiological Measurement, **28**, 10, 1225–1236.

6. Bishop C.M. (2006), *Pattern recognition and machine learning*, Springer, New York, US.

7. Breiman L. (2001), *Random forests*, Machine Learning, **45**, 1, 5–32.

8. Chang C.-C., Lin C.-J. (2011), *LIBSVM: A library for support vector machines*, ACM Transactions on Intelligent Systems and Technology, **2**, 27:1–27:27, software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

9. Cortes C., Vapnik V. (1995), *Support-vector networks*, Machine Learning, **20**, 3, 273–297.

10. El Badawey M.R., McKee G., Marshall H., Heggie N., Wilson J.A. (2003), *Predictive value of sleep*

nasendoscopy in the management of habitual snorers, Annals of Otology, Rhinology & Laryngology, **112**, 1, 40–44.

11. Eyben F. (2015), *Real-time speech and music classification by large audio feature space extraction*, Springer International Publishing, Cham, Switzerland.

12. Eyben F. *et al.* (2016), *The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing*, IEEE Transactions on Affective Computing, **7**, 2, 190–202.

13. Eyben F., Weninger F., Gross F., Schuller B. (2013), *Recent developments in opensmile, the munich open-source multimedia feature extractor*, [in:] Proc. ACM MM, pp. 835–838, Barcelona, Catalunya, Spain.

14. Eyben F., Wöllmer M., Schuller B. (2010), *Opensmile: the munich versatile and fast open-source audio feature extractor*, [in:] Proc. ACM MM, pp. 1459–1462, Firenze, Italy.

15. Fiz J.A., Jane R. (2012), *Snoring analysis. A complex question*, Journal of Sleep Disorders: Treatment and Care, **1**, 1, 1–3.

16. Herzog M. *et al.* (2014), *Evaluation of acoustic characteristics of snoring sounds obtained during drug-induced sleep endoscopy*, Sleep and Breathing, pp. 1–9.

17. Hessel N., de Vries N. (2002), *Diagnostic work-up of socially unacceptable snoring. II. Sleep endoscopy*, European Archives of Oto-Rhino-Laryngology, **259**, 158–161.

18. Hill P., Lee B., Osborne J., Osman E. (1999), *Palatal snoring identified by acoustic crest factor analysis*, Physiological Measurement, **20**, 2, 167–174.

19. Huang G.-B. (2014), *An insight into extreme learning machines: random neurons, random features and kernels*, Cognitive Computation, **6**, 3, 376–390.

20. Huang G.-B., Zhu, Q.-Y., Siew C.-K. (2006), *Extreme learning machine: theory and applications*, Neurocomputing, **70**, 1, 489–501.

21. Kezirian E.J., Hohenhorst W., de Vries N. (2011), *Drug-induced sleep endoscopy: the vote classification*, European Archives of Oto-Rhino-Laryngology, **268**, 8, 1233–1236.

22. Marin J.M., Carrizo S.J., Vicente E., Agusti A.G. (2005), *Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study*, The Lancet, **365**, 9464, 1046–1053.

23. Miyazaki S., Itasaka Y., Ishikawa K., Togawa K. (1998), *Acoustic analysis of snoring and the site of airway obstruction in sleep related respiratory disorders*, Acta Oto-Laryngologica, **118**, 537, 47–51.

24. Mokhlesi B., Ham S., Gozal D. (2016), *The effect of sex and age on the comorbidity burden of osa: an observational analysis from a large nationwide us health claims database*, The European Respiratory Journal, **47**, 4, 1162–1169.

25. Pancoast S., Akbacak M. (2012), *Bag-of-audio-words approach for multimedia event classification*, [in:] Proceedings of INTERSPEECH, pp. 2105–2108, Portland, Oregon.

26. Peppard P.E., Young T., Barnet J.H., Palta M., Hagen E.W., Hla K.M. (2013), *Increased prevalence of sleep-disordered breathing in adults*, American Journal of Epidemiology, **177**, 9, 1006–1014.

27. Peppard P.E., Young T., Palta M., Skatrud J. (2000), *Prospective study of the association between sleep-disordered breathing and hypertension*, New England Journal of Medicine, **342**, 19, 1378–1384.

28. Pevernagie D., Aarts R.M., De Meyer M. (2010), *The acoustics of snoring*, Sleep Medicine Reviews, **14**, 2, 131–144.

29. Qian K., Fang Y., Xu Z., Xu H. (2013), *Comparison of two acoustic features for classification of different snore signals*, Chinese Journal of Electron Devices, **36**, 4, 455–459.

30. Qian K. *et al.* (2017), *Classification of the excitation location of snore sounds in the upper airway by acoustic multi-feature analysis*, IEEE Transactions on Biomedical Engineering, **64**, 8, 1731–1741.

31. Qian K., Janott C., Zhang Z., Heiser C., Schuller B. (2016), *Wavelet features for classification of vote snore sounds*, [in:] Proc. IEEE ICASSP, pp. 221–225, Shanghai, China.

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

33. Qian K., Xu Z., Xu H., Wu Y., Zhao Z. (2015), *Automatic detection, segmentation and classification of snore related signals from overnight audio recording*, IET Signal Processing, **9**, 1, 21–29.

34. Roebuck A. *et al.* (2014), *A review of signals used in sleep analysis*, Physiological Measurement, **35**, 1, R1–R57.

35. Sak H., Senior A.W., Beaufays F. (2014), *Long short-term memory recurrent neural network architectures for large scale acoustic modeling*, [in:] Proceedings of INTERSPEECH, pp. 338–342, Singapore.

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

37. Schuller B., Steidl S., Batliner A. (2009), *The interspeech 2009 emotion challenge*, [in:] Proc. INTERSPEECH, pp. 312–315, Brighton, UK.

38. Schuller B. *et al.* (2013), *The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism*, [in:] Proc. INTERSPEECH, pp. 148–152, Lyon, France.

39. Spiegel M.R., Schiller J.J., Srinivasan R.A., LeVan M. (2009), *Probability and statistics*, McGraw-Hill, New York, NY, USA.

40. Strollo Jr P.J., Rogers R.M. (1996), *Obstructive sleep apnea*, New England Journal of Medicine, **334**, 2, 99–104.

41. Vincent P., Larochelle H., Lajoie I., Bengio Y., Manzagol P.-A. (2010), *Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion*, Journal of Machine Learning Research, **11**, 3371–3408.

42. Yaggi H.K., Concato J., Kernan W.N., Lichtman J.H., Brass L.M., Mohsenin V. (2005), *Obstructive sleep apnea as a risk factor for stroke and death*, New England Journal of Medicine, **353**, 19, 2034–2041.

43. Young T., Palta M., Dempsey J., Skatrud J., Weber S., Badr S. (1993), *The occurrence of sleep-disordered breathing among middle-aged adults*, New England Journal of Medicine, **328**, 17, 1230–1235.