Archives of Acoustics Vol. 49, No. 2, pp. 169–178 (2024), doi: 10.24425/aoa.2024.148775

Research Paper

Snoring Sound Recognition Using Multi-Channel Spectrograms

Ziqiang YE(1), Jianxin PENG(1)*, Xiaowen ZHANG(2), Lijuan SONG(2)

(1) *School of Physics and Optoelectronics, South China University of Technology*
Guangzhou, China

(2) *State Key Laboratory of Respiratory Disease, Department of Otolaryngology-Head and Neck Surgery*
Laboratory of ENT-HNS Disease, First Affiliated Hospital, Guangzhou Medical University
Guangzhou, China

*Corresponding Author e-mail: phjxpeng@scut.edu.cn

(*received October 24, 2020; accepted April 2, 2020; published online January 29, 2024*)

Obstructive sleep apnea-hypopnea syndrome (OSAHS) is a common and high-risk sleep-related breathing disorder. Snoring detection is a simple and non-invasive method. In many studies, the feature maps are obtained by applying a short-time Fourier transform (STFT) and feeding the model with single-channel input tensors. However, this approach may limit the potential of convolutional networks to learn diverse representations of snore signals. This paper proposes a snoring sound detection algorithm using a multi-channel spectrogram and convolutional neural network (CNN). The sleep recordings from 30 subjects at the hospital were collected, and four different feature maps were extracted from them as model input, including spectrogram, Mel-spectrogram, continuous wavelet transform (CWT), and multi-channel spectrogram composed of the three single-channel maps. Three methods of data set partitioning are used to evaluate the performance of feature maps. The proposed feature maps were compared through the training set and test set of independent subjects by using a CNN model. The results show that the accuracy of the multi-channel spectrogram reaches 94.18%, surpassing that of the Mel-spectrogram that exhibits the best performance among the single-channel spectrograms. This study optimizes the system in the feature extraction stage to adapt to the superior feature learning capability of the deep learning model, providing a more effective feature map for snoring detection.

Keywords: obstructive sleep apnea-hypopnea syndrome; snoring; convolutional neural network; multi-channel spectrogram.

1. Introduction

Obstructive sleep apnea-hypopnea syndrome (OSAHS) is a sleep respiratory disease characterized by the repeated collapse and blockage of the upper airway during sleep, resulting in apnea or hypopnea (Strollo, Rogers, 1996). Obstructive breathing leads to instinctive body responses, such as brain arousal, sympathetic activation, and decreased blood oxygen saturation. Seriously interrupted and non-restorative sleep will occur, causing most patients with OSAHS to suffer from morning headaches and daytime somnolence. Long-term poor sleep can even lead to a series of complications, such as abnormal metabolism, neurocognitive dysfunction, and cardiovascular disease (Young *et al.*, 2002). Surveys show that the overall prevalence of OSAHS in the general adult population ranges from 6 to 17%, with the prevalence increasing significantly with age (Senaratna *et al.*, 2017).

Polysomnography (PSG) is the gold standard for diagnosing OSAHS patients (Ahmadi *et al.*, 2009; Mendonça *et al.*, 2019). Subjects are required to wear contact-type monitoring instruments throughout the night. The PSG signal obtained from these instruments is used by professional doctors to determine whether the subjects suffer from OSAHS. Although reliable results can be obtained, patients may have to bear the burden of expensive fees and endure discomfort from physically attached sensors (Mendonça *et al.*, 2019).

Therefore, there is an urgent need to seek a low-cost, easy-to-operate, and non-contact method to assist in the diagnosis of OSAHS. Snoring is the most distinctive clinical feature of OSAHS, occurring in 70–90% of patients with OSAHS (Karunajeewa *et al.*, 2008; Maimon, Hanly, 2010). The acoustic characteristics of snoring reflect changes in the structure of the upper airway. Moreover, snoring analysis offers the advantages of being non-contact, simple, and reliable, making it feasible to identify patients by analyzing the acoustic characteristics of snoring (Won *et al.*, 2012; Fiz *et al.*, 1996; Pevernagie *et al.*, 2010; Beck *et al.*, 1995; Ip *et al.*, 2002; Perez-Padilla *et al.*, 1993; Sola-Soler *et al.*, 2003; Ng *et al.*, 2008).

In order to improve the initial screening of OSAHS, an increasing number of scientists are dedicated to developing new technologies that can achieve a more accurate clinical diagnosis of OSAHS in a simpler manner (Yadollahi, Moussavi, 2010; Ankişhan, Ari, 2011; Ankişhan, Yilmaz, 2013). So far, there have been numerous studies on the identification technology of OSAHS. Duckitt *et al.* (2006) extracted 39-dimensional Mel-frequency cepstral coefficients (MFCC) from sleep sound recordings of six subjects and classified the signals into snoring, breathing, duvet noise, and other noises based on hidden Markov model (HMM). The recognition rate for snoring can reach the range of 82–89%. Cavusoglu *et al.* (2007) selected recording signals from 18 simple snorers and 12 OSAHS patients to cut the voiced segments by a double threshold method. Then, the authors calculated the sub-band energy distribution of the sound segments and used principal component analysis (PCA) for feature reduction. Finally, robust linear regression was used to classify these sound segments into snoring and non-snoring sounds with an accuracy of 90.2%.

Dafna *et al.* (2013) adopted a feature selection algorithm to filter the 34 most discriminative features from 127 time-domain and frequency-domain features, and then used AdaBoost to construct a snoring recognition model, obtaining an average detection rate of 98.2%, a sensitivity of 98%, and specificity of 98.3% with a cross-validation method. In a study by Cheng *et al.* (2022), a multi-input model based on long short-term memory (LSTM) was proposed, which can accept various audio features to synthesize information for snoring identification. Furthermore, MFCC, Mel filter banks (Fbanks), linear prediction coefficient (LPC), and short-term energy were extracted as the input of the model, finally achieving 95.3% accuracy. With the development of the field of artificial intelligence, deep learning models are gradually applied to the classification task of snoring and non-snoring.

Khan (2019) developed a deep learning model for snoring detection and transferred it to an embedded system that can be connected to a smartphone app using home Wi-Fi. In Khan's study, 1000 sound samples were used to calculate the MFCC images, then the images were fed into a convolutional neural network (CNN) model, resulting in a snoring recognition rate of 96%. The spectrogram, Mel-spectrogram, and constant-Q transformation (CQT) spectrogram collected from the recordings of 15 subjects were used to classify snoring and non-snoring by Jiang *et al.* (2020). The results indicated that the accuracy of Mel-spectrogram in each group reached 95.07%. The advantage of the deep learning model is to learn increasingly complex data samples. Previous studies (Khan, 2019; Jiang *et al.*, 2020; Xie *et al.*, 2021) used single-channel spectrogram as input. However, it is important to note that different feature maps only contain limited frequency-domain information, which could potentially restrict the model's ability to learn diverse representations of audio recordings. Therefore, input features should provide more information about snoring.

In our work, a multi-channel feature map based on the fusion of Mel-spectrogram, spectrogram, and continuous wavelet transform (CWT) is proposed. Three spectrograms of each sound signal are employed as three channels of the red-green-blue (RGB) image to construct the feature map. A CNN model is utilized to perform the classification tasks. In addition, spectrogram, Mel-spectrogram, and CWT are used for comparative experiments. The comparison of the classification performance between the multi-channel spectrogram with that of the single-channel spectrogram is conducted to achieve higher resolution.

2. Methods

2.1. Data acquisition

This study was approved by the Ethics Committee of Guangzhou Medical University (Reference Number 2019-73), and informed consent was obtained from all participants.

Thirty subjects who underwent PSG at the First Affiliated Hospital of Guangzhou Medical University were selected to obtain snoring sounds throughout the night. The recording time for each subject's sleep snoring sounds was not less than 6 hours. The most important indicator for PSG detection to assess the severity of OSAHS is the apnea-hypopnea index (AHI), which is defined as the average number of sleep apnea or hypopnea per hour. It is divided into four categories: simple, mild, moderate, and severe, based on the following ranges: $\text{AHI} < 5$, $5 \leq \text{AHI} < 15$, $15 \leq \text{AHI} < 30$, and $\text{AHI} \geq 30$ (Maimon, Hanly, 2010). Table 1 lists statistical information on the subjects' gender, age, body mass index (BMI), AHI, and the severity of OSAHS for each participant. For recording snoring sounds, a digital audio recorder (Roland, Edirol R-44, Japan), with a frequency response range of 40–20 000 Hz and a microphone (RODE, NTG-3, Sydney, Australia) hanging

Table 1. Statistical information of subjects.

Parameter	Data
Male/female	27/3
Age (years)	44 ± 13 (range: 23–70)
BMI [kg/m^2]	26.7 ± 2.8 (range: 20.8–31.9)
AHI [events/h]	40.8 ± 28.3 (range: 3.2–91.1)
OSAHS [y/n]	28/2

vertically on the heads of patients, positioned about 45 cm above the mouth and nose were used. The original sleep sound signals were recorded by the microphone. PSG device (Alice-5, Pittsburgh, Pennsylvania, USA) was used to monitor PSG signals. The recorded sound was digitized at a sampling rate of 44 100 Hz and a resolution of 16 bits.

2.2. Feature extraction

2.2.1. Spectrogram

A snoring sound is a one-dimensional time-domain signal, making it challenging to observe the frequency conversion pattern. While the frequency distribution of the signal can be viewed by Fourier transform, time-domain information is lost. Many time-frequency analysis methods have emerged to address this problem. Short-time Fourier transform (STFT) is the most classical time-frequency analysis method in speech and audio processing applications and offers minimal calculation and low cost. First, the audio signal is framed into a short time window. In this work, the size of windows is 25 ms with 50% overlap. Next, the Hamming window is applied to each frame signal, and followed by the fast Fourier transform (FFT) to obtain its power spectrum (Rabiner *et al.*, 1975). Each frame is then spliced along the time dimension to form a two-dimensional signal map called the spectrogram.

2.2.2. Mel-spectrogram

While the frequency of the spectrogram is linearly distributed, the extracted features may not be useful for signals with an inhomogeneous frequency distribution. The Mel-scale filter banks are used to transform the spectrogram into the Mel-spectrogram (Peng *et al.*, 2019; Winursito *et al.*, 2018), where the Mel-scale describes the nonlinear characteristics of human ear frequency, and its relationship with frequency can be approximately expressed by the equation:

$$\mathrm{Mel}(f) = 2595 \times \log\left(1 + \frac{f}{700}\right). \tag{1}$$

In this study, features are calculated using frames of 25 ms frame size with 50% overlap. The Mel-spectrogram is computed using a group of 128 triangular filters in the Mel-scale based on the STFT, and the logarithm of the filtered signal is determined. Figure 1 shows the triangular filter banks used in this study.

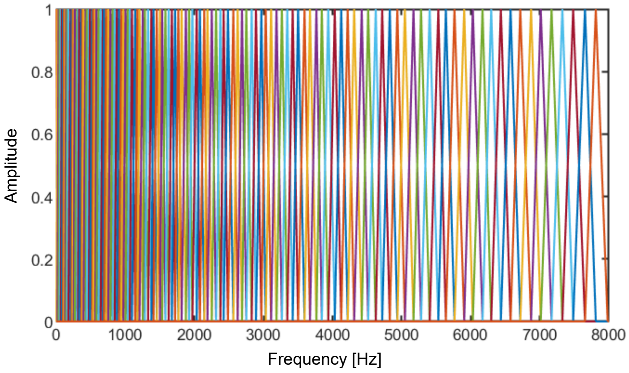

Fig. 1. 128 triangular filters in the Mel-scale applied to the STFT for obtaining the Mel-spectrogram.

2.2.3. Continuous wavelet transform

The time and frequency resolutions of STFT are determined by the size and time shift of the window. A small window size can lead to poorer frequency resolution. Compared to STFT, CWT has the characteristics of window adaptation, enabling high-frequency values to have high-frequency resolution and low time resolution (Qian *et al.*, 2019).

CWT uses wavelet basis functions to decompose signals, and is defined as:

$$\mathrm{CWT}(\tau s) = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} x(t)\psi\left(\frac{t-\tau}{s}\right) \mathrm{d}t, \tag{2}$$

where $x(t)$ is the audio signal, $\psi(x)$ is the mother wavelet (Morlet wavelet in this study), and τ and s, respectively, represent displacement and scale.

Usually, when analyzing time series, it is expected to obtain smooth and continuous wavelet amplitude, so a non-orthogonal wavelet function is more suitable. In addition, to include the information of both amplitude and phase of the time series, a complex-valued wavelet should be selected, because the complex-valued wavelet has an imaginary part and can express the phase very well. The Morlet wavelet is not only non-orthogonal, but also exponential complex-valued wavelet, so it is used in this experiment to obtain the information of both amplitude and phase.

2.2.4. Multi-channel spectrogram

Multi-channel spectrogram has been used in speech recognition with beneficial effects (Adavanne *et al.*, 2018; Xu *et al.*, 2018; Arias-Vergara *et al.*, 2021).

The spectrogram, Mel-spectrogram, and CWT, each with a size of 224 × 224 × 3, were extracted from each audio segment. Figure 2 shows the above three feature maps of a snore signal. Subsequently, they are normalized to fall between −1 and 1, serving as three channels of the RGB image to construct the multi-channel spectrogram with a size of 224 × 224 × 3. In this construction, the spectrogram is the first channel, the CWT is the second channel, and the Mel-spectrogram is the third channel. When the input data contains multiple channels, the number of input channels of the convolutional kernel in the model is the same as that of the input data. In this way, the convolutional kernel of different channels can perform cross-correlation operations with the input data of different channels, and the multi-channel input will enable CNN to supplement information from two other time-frequency representations.

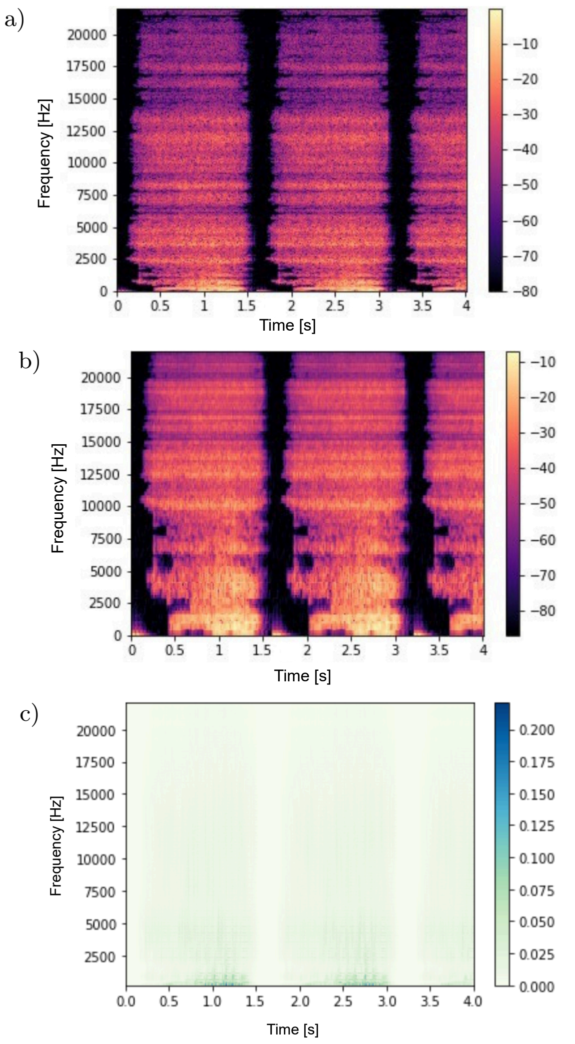

Fig. 2. Feature maps of a snore segment from an OSAHS patient: a) spectrogram; b) Mel-spectrogram; c) CWT.

2.3. Model architecture

In order to obtain reasonable results, the classifier must be matched with a suitable input representation. Manual features such as MFCC were used with the traditional machine learning model, which effectively decorrelates features (Adavanne *et al.*, 2018). On the contrary, the advantage of CNN lies in their ability to learn spectral time characteristics of the spectrum through weight sharing and pooling technology. Previous studies have applied CNN to speech recognition with good effects (Abdel-Hamid *et al.*, 2012; 2014). For this experiment, a CNN model was designed, containing an input layer, three convolution layers with rectified linear unit (ReLu) activation functions. The size of the convolution kernel was multiplied layer by layer, leading to 256 neurons activated by ReLu, and the output layer was activated by a softmax function. The incorporated dropout layer will randomly discard some weights in the training process to suppress overfitting, and the dropout ratio is 0.5 (Hinton *et al.*, 2012). Figure 3 shows the process of feeding the multi-channel spectrogram into the CNN. The model parameters are presented in Table 2.

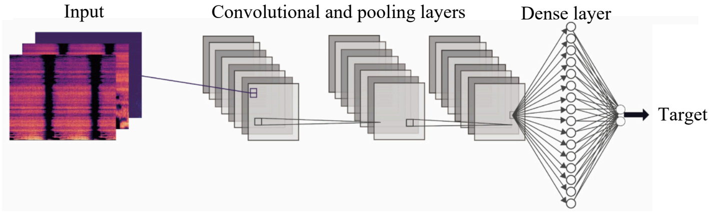

Fig. 3. Process of feeding the multi-spectrogram to a deep learning model (CNN).

Table 2. Structure of CNN.

Layer (type)	Input shape	Output shape	Params
Conv2D	(None, 224, 224, 32)	(None, 222, 222, 32)	896
MaxPooling2D	(None, 222, 222, 32)	(None, 111, 111, 32)	0
Conv2D	(None, 111, 111, 32)	(None, 109, 109, 64)	18 496
MaxPooling2D	(None, 109, 109, 64)	(None, 54, 54, 64)	0
Conv2D	(None, 54, 54, 64)	(None, 52, 52, 128)	73 856
MaxPooling2D	(None, 52, 52, 128)	(None, 26, 26, 128)	0
Flatten	(None, 26, 26, 128)	(None, 86 528)	0
Dense	(None, 86 528)	(None, 128)	11 075 712
Dense	(None, 128)	(None, 2)	258

For excellent training results, the Adam optimizer is used for training, with a learning rate of CNN set to 0.0001. In our experiments, categorical cross-entropy was chosen as the loss function, and each model was trained for 200 epochs on an NVIDIA GTX 1080Ti with a batch size of 128.

2.4. Validation method

In this study, the adaptive threshold method is used to segment the audio sounds from all recording subjects to obtain sound fragments that are subsequently labeled as either snoring or non-snoring under the guidance of ear-nose-throat (ENT) experts. Only sound segments less than 4 seconds long are retained, and two adjacent sound segments less than 0.02 seconds apart are merged. A total of 59 293 sound segments are obtained, consisting of 29 789 snore segments, and 29 504 non-snoring segments, which included sounds of footsteps, speech, breathing, coughing, door closing, and other environmental sounds. In order to evaluate the performance of different spectra, three experiments were designed: independent split training set and test set, leave-one-subject-out cross-validation (LOSOCV), and training set and test set containing all subjects. Table 3 shows the details of the data partition.

Experiment 1: the dataset of 30 subjects was divided into a validation set with 4 subjects, a test set with 4 subjects, and training set with the remaining 22 subjects, and the subjects in the training set, the test set, and the validation set were independent. For the purpose of eliminating the contingency of the experiment, five different partition methods were applied to the data set, and the model was trained on each divided dataset. Finally, the average and standard deviation were taken as the results.

Experiment 2: in a dataset containing 30 subjects, an independent test set and training set were constructed for each participant using the LOSOCV strategy. The data of one subject was selected as the test set, and the data of the remaining 29 subjects were used as the training set. This process is repeated 30 times and the average accuracy is calculated. This maximizes the use of data while ensuring that the subjects in the training set and the test set are from different independent subjects.

Experiment 3: the sound clips of all subjects are combined into a whole dataset, which is then divided into training, validation and test set, with a ratio of 6:1:3.

2.5. Model evaluation

The classification effect of each feature map can be evaluated by multiple indicators, including accuracy, precision, recall, $F1$-score, and the area under the curve (AUC) calculated from the receiver operating characteristic (ROC). Accuracy is the proportion of correct samples to the total number of samples. Precision relates to the ratio of the number of positive samples correctly classified by the classifier to the number of all positive samples classified by the classifier. Recall rate refers to the ratio of the number of positive samples correctly classified by the classifier to the number of all positive samples. $F1$-score is the harmonic mean of precision rate and recall rate. The AUC is meant by the area under the ROC curve, representing the probability that the predicted positive cases rank higher than the negative ones, ranging from 0.5 to 1. The calculation equation is:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{5}$$

$$F1_{\text{score}} = \frac{2\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{6}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

3. Results

To evaluate the classification performance, four different feature maps are imported into the model to compare which feature map is more discriminative for snoring. The CNN model is established by the validation set and evaluated on the test set. According to the data set division rules of experiment 1, the classification results are presented in Table 4. In terms of single-channel spectrograms, the classification performance of Mel-spectrogram was superior to those of spectrogram and CWT, with an accuracy of 91.58%, precision of 92.09%, sensitivity of 86.57%, $F1$-score of 88.85%, and AUC of 0.9614. The PPV of the spectrogram and Mel-spectrogram reached more than 90%,

Table 3. Data distribution of training, validation, and test sets in experiments.

	Experiment 1			Experiment 2		Experiment 3		
	Train	Validation	Test	Train	Test	Train	Validation	Test
Subject	22	4	4	29	1		30	
Snore	23 767	3117	2905	LOSOCV		19 133	2872	7784
No-snore	21 971	4094	3439			16 443	3057	10 004

Table 4. Classification results of experiment 1.

Map type	Accuracy [%]	Precision [%]	Recall [%]	$F1$-score [%]	AUC
Mel-spectrogram	91.58	92.09	86.57	88.85	0.9614
Spectrogram	88.33	91.23	81.56	85.56	0.9448
CWT	85.24	81.78	85.10	83.00	0.9192
Multi-channel spectrogram	94.16	92.64	93.35	92.93	0.9730

indicating that the recognition of the snore fragments was reliable.

Figure 4 shows that the classification of the multi-channel spectrogram is significantly improved compared to that of the single-channel spectrogram, and it has an accuracy of 94.16%, which was 2.58% higher than that of Mel-spectrogram with the best effect in single-channel spectrograms. Other classification indexes were increased, respectively, by 0.55% (PPV), 6.78% (Recall), and 4.08% ($F1$-score). Although there was little difference in PPV between the two feature maps, the Recall of the multi-channel spectrogram classification was significantly higher than that of the Mel-spectrogram, which is beneficial for detecting the snoring segments of the patients throughout the entire night and further evaluating the severity of OSAHS patients.

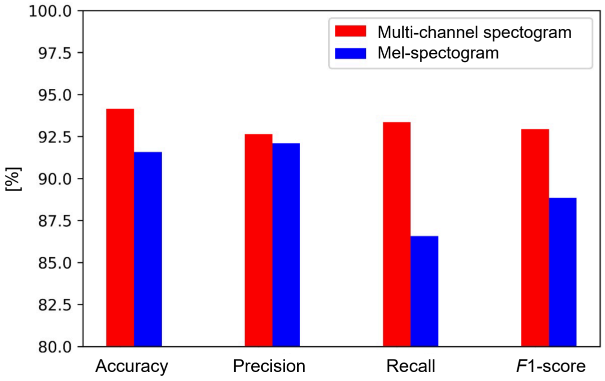

Fig. 4. Comparison histogram of Mel-spectrogram and multi-spectrogram in experiment 1.

Tables 5 and 6 show the classification results for experiments 2 and 3. The results show that the recognition effect of the multi-channel spectrogram is consistently better than that of the single-channel spectrogram when using different dataset partitioning methods.

4. Discussion

In this study, the performance of Mel-spectrogram, spectrogram, CWT, and multi-channel spectrogram in classifying snoring and the non-snoring sound was investigated. The results show that the Mel-spectrogram has the best recognition effect when the single-channel spectrograms were used as input, which is in agreement with the results of the study by Jiang (2020). The energy peak frequency of the snoring sounds mentioned in the study is 250 Hz, and most of the energy is distributed below 1000 Hz, while the energy of respiratory sounds and other noise is distributed above 1000 Hz (Pevernagie *et al.*, 2010; Jiang *et al.*, 2020). The frequency of the spectrogram is linear distribution, which leads to the insufficient frequency resolution in the low-frequency part, making it challenging to detect some weak snoring changes. The Mel-spectrogram converts the linear frequency into the Mel frequency, offering detailed representation of the low-frequency information and rough representation of the high-frequency information, which aligns with the energy distribution of the snoring spectrogram.

Apart from Spectrogram and Mel-spectrogram, which are computed based on STFT, the CWT commonly

Table 5. Classification results of experiment 2.

Map type	Accuracy [%]	Precision [%]	Recall [%]	$F1$-score [%]	AUC
Mel-spectrogram	90.51	90.83	93.08	91.94	0.9511
Spectrogram	89.36	93.18	88.85	90.97	0.9599
CWT	85.38	89.51	84.82	87.10	0.9191
Multi-channel spectrogram	93.10	92.28	96.66	94.42	0.9774

Table 6. Classification results of experiment 3.

Map type	Accuracy [%]	Precision [%]	Recall [%]	$F1$-score [%]	AUC
Mel-spectrogram	93.67	98.28	91.44	94.74	0.9817
Spectrogram	91.76	93.03	93.34	93.19	0.9717
CWT	88.99	91.74	89.84	90.78	0.9569
Multi-channel spectrogram	97.80	97.14	99.18	98.15	0.9979

used in speech recognition is also imported into the same CNN model. A study by Huzaifah (2017) proved that CWT performs significantly worse than spectrogram and Mel-spectrogram when employed in a CNN to classify various environmental sounds. The same result was obtained when the three feature maps were applied to snoring and non-snoring sound classification. It means that CWT cannot provide more snoring sounds details in the low frequency compared to the other two maps. However, it is premature to conclude that CWT is always inferior to the feature maps based on STFT, because the experiment may be influenced by parameter settings for map extraction and model structure.

It should be pointed out that the peak energy frequency of snoring sound among different people is not consistent, and even the snoring of the same person is different. Jiang *et al.* (2020) analyzed the energy distributions in snoring and non-snoring sub-bands of subjects and found that 60% of the snoring spectral energy was distributed between 100 and 300 Hz, and 40% of it was also distributed in each frequency band above 300 Hz. The information contained in a single-channel input may be restricted, which can limit the potential of the deep learning model to learn more complicated representations from snoring sound signals. The multi-channel map was used to overcome the limitation of a single-channel input in speech recognition. Various methods were used to construct multi-channel maps in such studies. Adavanne *et al.* (2018) proposed a method where multi-channel could be extracted from the same signal recorded by different microphones. Another approach by Fu *et al.* (2017) involved computing the real and imaginary parts of the STFT to form a 2D-channel spectrogram.

Arias-Vergara *et al.* (2021) computed CWT, Mel-spectrogram, and gammatone spectrogram from the audio signal and combined them into a 3D-channel spectrogram. Compared with single-channel maps, the performance of these multi-channel maps with a CNN model was improved. In our work, when a multi-channel spectrogram was used as the model input to identify snoring sounds, the result was consistent with the expectation, which was better than the Mel-spectrogram with the best classification effect of single-channel feature maps. This suggests that the multi-channel spectrogram contains more spectrum information than a single spectrum. The CNN model can capture more feature information from the fusion map than from a single-channel feature map through multi-layer convolution layers.

Many researchers have proposed a variety of experimental methods to classify snoring and non-snoring. Table 7 compares the research methods in related fields with the current experiment. Khan (2019) collected online snoring resources as datasets, extracted MFCC images, and input them into a CNN model training and obtained a 96% accuracy. However, the number of experimental samples was only 1000, and the source of snoring sound was singular. In our experiment, 59 293 sound samples were extracted from 30 subjects with better generalization ability, and three different verification methods were used to evaluate the performance of the feature map, resulting in the generalization of the results. Jiang *et al.* (2020) used two classifiers, CNNs-DNNs and CNNs-LSTMs-DNNs, to identify snores from sound fragments, including spectrogram, Mel-spectrogram, and CQT-spectrogram. The results demonstrate that the combination of Mel spectrogram and CNNs-LSTMs-DNNs was well suited for the task. However, the input images contained limited information from single-channel spectrogram. Moreover, the data of the training set and the test set are not independent and using this model to detect individual snore fragments throughout entire night may lead to deviation. Cheng *et al.* (2022) designed a multi-input

Table 7. Summary of previous studies on snoring detection.

Author	Subjects	Datasets	Features	Methods	Result [%]
Khan (2019)		1000	MFCC image	CNN	Accuracy: 96
Jiang *et al.* (2020)	15	12 457	Mel-spectrogram	CNN+LSTM+DNN	Accuracy: 95.07 PPV: 94.62 Sensitivity: 95.42
Cheng *et al.* (2022)	43	15 520	MFCC, Fbanks, Short-time average energy, LPC	A multi-input model based on LSTM	Accuracy: 95.3 PPV: 95.7 Sensitivity: 94.9
Dafna *et al.* (2013)	67	281 953	Time-related features, Spectral-related features	AdaBoost	Accuracy: 98.2 Sensitivity: 98.1
Cavusoglu *et al.* (2007)	30	9000	Average normalized energy in each subband	Robust linear regression	Accuracy: 90.2 PPV: 98.7
Sun *et al.* (2022)	24	36 938	Bark sub-band feature, MFCC, LPC, etc.	XGBoost	Accuracy: 87.22 PPV: 95.09 Sensitivity: 87.16
This work	30	59 293	Multi-spectrogram	CNN	Accuracy: 94.16 PPV: 92.64 Sensitivity: 93.35

model based on LSTM and extracted MFCC, Fbanks, short-term energy, and LPC as four branches of the input layer. After integration, ANN was used as the classifier, and finally, a 95.3% snoring recognition rate was obtained, an improvement compared with a single feature processing network. Nevertheless, the model's input layer has multiple parallel input branches, and the network structure is relatively complex.

In their experiment, the fusion feature maps were employed in feature extraction, and only one entry was needed for model input. In Dafna *et al.* (2013), 127 features from both the time domain and frequency domain were extracted. Using a feature selection method, 34 most effective features were selected objectively, and the AdaBoost classifier was used and yielded a 98.2% recognition rate. However, the extraction process involved various features, making the process of feature extraction complicated.

Cavusoglu *et al.* (2007) divided the frequency range of snoring sounds (0–7500 Hz) into 500 Hz subbands and calculated the average normalized energy in each sub-band to obtain spectral characteristics. The linear regression model was used and a 90.2% accuracy was obtained. However, the energy distribution of snoring was mainly concentrated in the low frequency and the band division of equal intervals may lead to insufficient low-frequency resolution. Sun *et al.* (2022) proposed a snoring detection algorithm based on acoustic features and XGBoost. Various training and test data splitting methods were used to evaluate model performance, and the results showed that when the training set and test set are from all subjects, the classification performance was better than that of the training set and test set from different independent subjects.

In terms of experimental accuracy, the method proposed in this work is significantly improved compared with 90.2% reported by Cavusoglu *et al.* (2007) and 92.78% obtained by Sun *et al.* (2022). However, it is important to acknowledge that different research samples are distinct, the subjective standards of labeled samples are different, and the methods of splitting data sets are also different. It is therefore difficult to compare the classification results to make a unified judgment. The multi-channel spectrogram proposed in this study has more than 92% in all evaluation indexes on the CNN model, indicating that this method can effectively detect snoring sound.

5. Conclusion

This study explored a classification method for distinguishing between snoring and non-snoring using a CNN model with a focus on a multi-channel spectrogram with a CNN model. Mel-spectrogram, spectrogram, and CWT were used as three channels for constructing multi-channel maps. The four feature maps of the snoring sound signals of 30 subjects were extracted for training and testing, and the results demonstrate that the classification performance indicators of the multi-channel spectrogram are improved compared with single-channel spectrograms. The main contribution of this work lies in proposing a multi-channel spectrogram based on the fusion of a single-channel spectrogram for snoring detection. The study also compared the classification performance of each feature map under the same network model.

This work focused on improving the feature extraction stage, extracting the feature maps containing more time and frequency domain information, to adapt to the strong fitting ability of the deep learning model. Future work can be carried out in different directions. Firstly, a comparison of diverse types of multi-channel spectrograms combined with various classification networks could be explored to further improve the accuracy of current snoring detection algorithms. Another direction is to explore how snoring sound detection contributes to the task of detecting OSAHS. This experiment can be used as the first step in OSAHS detection because snoring events are closely related to apnea. In addition, the snoring sound identified by this model could be further used to quantitatively evaluate the severity of OSAHS.

However, the snoring data collected in this experiment is limited to a hospital environment. Different recording environments have different background noise, which cannot guarantee the performance of the model in other recording settings. Therefore, more recording data in diverse environments (bedroom, dormitory, hotel, etc.) is needed to obtain a more reliable snoring recognition model and make it more robust and generalized. In addition, it is necessary to pay attention to the computational efficiency and memory overhead of the model to ensure that model meets the requirements for mobile deployment.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant no. 11974121) and the National Youth Foundation of China (grant no. 81900927).

Author contributions

All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Informed consent

Informed consent was obtained from all individual participants who were included in the study.

Ethical approval

Consent was obtained from all study participants prior to their participation in the study. Ethical approval was obtained from the Ethics Committee of Guangzhou Medical University (Reference no. 2019-73).

References

1. Abdel-Hamid O., Mohamed A., 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, doi: 10.1109/TASLP.2014.2339736.
2. Abdel-Hamid O., Mohamed A., Jiang H., Penn G. (2012), Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition, [in:] *2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4277–4280, doi: 10.1109/ICASSP.2012.6288864.
3. Adavanne S., Politis A., Virtanen T. (2018), Multichannel sound event detection using 3D convolutional neural networks for learning inter-channel features, [in:] *2018 International Joint Conference on Neural Networks*, pp. 1–7, doi: 10.1109/IJCNN.2018.8489542.
4. Ahmadi N., Shapiro G.K., Chung S.A., Shapiro C.M. (2009), Clinical diagnosis of sleep apnea based on single night of polysomnography vs. two nights of polysomnography, *Sleep Breath*, **13**(3): 221–226, doi: 10.1007/s11325-008-0234-2.
5. Ankişhan H., Ari F. (2011), Snore-related sound classification based on time-domain features by using ANFIS model, [in:] *2011 International Symposium on Innovations in Intelligent Systems and Applications*, pp. 441–444, doi: 10.1109/INISTA.2011.5946113.
6. Ankişhan H., Yilmaz D. (2013), Comparison of SVM and ANFIS for Snore related sounds classification by using the largest Lyapunov exponent and entropy, *Computational and Mathematical Methods in Medicine*, **2013**: 238937, doi: 10.1155/2013/238937.
7. Arias-Vergara T., Klumpp P., Vasquez-Correa J.C., Nöth E., Orozco-Arroyave J.R., Schuster M. (2021), Multi-channel spectrograms for speech processing applications using deep learning methods, *Pattern Analysis and Applications*, **24**(2): 423–431, doi: 10.1007/s10044-020-00921-5.
8. Beck R., Odeh M., Oliven A., Gavriely N. (1995), The acoustic properties of snores, *European Respiratory Journal*, **8**(12): 2120–2128, doi: 10.1183/09031936.95.08122120.
9. Cavusoglu M., Kamasak M., Erogul O., Ciloglu T., Serinagaoglu Y., Akcam T. (2007), An efficient method for snore/nonsnore classification of sleep sounds, *Physiological Measurement*, **28**(8): 841–853, doi: 10.1088/0967-3334/28/8/007/.
10. Cheng S. *et al.* (2022), Automated sleep apnea detection in snoring signal using long short-term memory neural networks, *Biomedical Signal Processing and Control*, **71**(Part B): 103238, doi: 10.1016/j.bspc.2021.103238.
11. Dafna E., Tarasiuk A., Zigel Y. (2013), Automatic detection of whole night snoring events using non-contact microphone, *PLOS ONE*, **8**(12): e84139, doi: 10.1371/journal.pone.0084139.
12. Duckitt W.D., Tuomi S.K., Niesler T.R. (2006), Automatic detection, segmentation and assessment of snoring from ambient acoustic data, *Physiological Measurement*, **27**(10): 1047–1056, doi: 10.1088/0967-3334/27/10/010.
13. Fiz J.A. *et al.* (1996), Acoustic analysis of snoring sound in patients with simple snoring and obstructive sleep apnoea, *European Respiratory Journal*, **9**(11): 2365–2370, doi: 10.1183/09031936.96.09112365.
14. Fu S., Hu T., Tsao Y., Lu X. (2017), Complex spectrogram enhancement by convolutional neural network with multi-metrics learning, [in:] *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing*, pp. 1–6, doi: 10.1109/MLSP.2017.8168119.
15. Hinton G.E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R.R. (2012), Improving neural networks by preventing co-adaptation of feature detectors, *ArXiv*, doi: 10.48550/arXiv.1207.0580.
16. Huzaifah M. (2017), Comparison of time-frequency representations for environmental sound classification using convolutional neural networks, *ArXiv*, doi: 10.48550/arXiv.1706.07156.
17. Ip M.S., Lam B., Ng M.M., Lam W.K., Tsang K.W., Lam K.S. (2002), Obstructive sleep apnea is independently associated with insulin resistance, *American Journal of Respiratory and Critical Care Medicine*, **165**(5): 670–676, doi: 10.1164/ajrccm.165.5.2103001.
18. Jiang Y., Peng J., Zhang X. (2020), Automatic snoring sounds detection from sleep sounds based on deep learning, *Physical and Engineering Sciences in Medicine*, **43**(2): 679–689, doi: 10.1007/s13246-020-00876-1.
19. Karunajeewa A.S., Abeyratne U.R., Hukins C. (2008), Silence-breathing-snore classification from snore-related sounds, *Physiological Measurement*, **29**(2): 227–243, doi: 10.1088/0967-3334/29/2/006.
20. Khan T. (2019), A deep learning model for snoring detection and vibration notification using a smart wearable gadget, *Electronics*, **8**(9): 987, doi: 10.3390/electronics8090987.
21. Maimon N., Hanly P.J. (2010), Does snoring intensity correlate with the severity of obstructive sleep apnea?, *Journal of Clinical Sleep Medicine*, **6**(5): 475–478, doi: 10.5664/jcsm.27938.
22. Mendonça F., Mostafa S.S., Ravelo-García A.G., Morgado-Dias F., Penzel T. (2019), A review of obstructive sleep apnea detection approaches, *IEEE Journal of Biomedical and Health Informatics*, **23**(2): 825–837, doi: 10.1109/JBHI.2018.2823265.

23. Ng A.K., Koh T.S., Baey E., Lee T.H., Abeyratne U.R., Puvanendran K. (2008), Could formant frequencies of snore signals be an alternative means for the diagnosis of obstructive sleep apnea?, *Sleep Medicine*, **9**(8): 894–898, doi: 10.1016/j.sleep.2007.07.010.

24. Peng P., He Z., Wang L. (2019), Automatic classification of microseismic signals based on MFCC and GMM-HMM in underground mines, *Shock and Vibration*, **2019**: 5803184, doi: 10.1155/2019/5803184.

25. Perez-Padilla J.R., Slawinski E., Difrancesco L.M., Feige R.R., Remmers J.E., Whitelaw W.A. (1993), Characteristics of the snoring noise in patients with and without occlusive sleep apnea, *American Review of Respiratory Disease*, **147**(3): 635–644, doi: 10.1164/ajrccm/147.3.635.

26. Pevernagie D., Aarts R.M., De Meyer M. (2010), The acoustics of snoring, *Sleep Medicine Reviews*, **14**(2): 131–144, doi: 10.1016/j.smrv.2009.06.002.

27. Qian K. *et al.* (2019), A Bag of wavelet features for snore sound classification, *Annals of Biomedical Engineering*, **47**(4): 1000–1011, doi: 10.1007/s10439-019-02217-0.

28. Rabiner L.R., Gold B., Yuen C.K. (1975), Theory and application of digital signal processing, *IEEE Transactions on Systems, Man, and Cybernetics*, **8**(2): 146–146, doi: 10.1109/TSMC.1978.4309918.

29. Senaratna C.V. *et al.* (2017), Prevalence of obstructive sleep apnea in the general population: A systematic review, *Sleep Medicine Reviews*, **34**: 70–81, 10.1016/j.smrv.2016.07.002.

30. Sola-Soler J., Jane R., Fiz J.A., Morera J. (2003), Spectral envelope analysis in snoring signals from simple snorers and patients with obstructive sleep apnea, [in:] *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, **3**: 2527–2530, doi: 10.1109/IEMBS.2003.1280430.

31. Strollo P.J., Rogers R.M. (1996), Obstructive sleep apnea, *New England Journal of Medicine*, **334**(2): 99–104, doi: 10.1056/NEJM199601113340207.

32. Sun X., Peng J., Zhang X., Song L. (2022), Effective feature selection based on Fisher Ratio for snoring recognition using different validation methods, *Applied Acoustics*, **186**: 108483, doi: 10.1016/j.apacoust.2021.108429.

33. Winursito A., Hidayat R., Bejo A. (2018), Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition, [in:] *2018 International Conference on Information and Communications Technology*, pp. 379–383, doi: 10.1109/ICOIACT.2018.8350748.

34. Won T.B. *et al.* (2012), Acoustic characteristics of snoring according to obstruction site determined by sleep videofluoroscopy, *Acta Oto-Laryngologica*, **132**: 13–20, doi: 10.3109/00016489.2012.660733.

35. Xie J. *et al.* (2021), Audio-based snore detection using deep neural networks, *Computer Methods and Programs in Biomedicine*, **200**: 105917, doi: 10.1016/j.cmpb.2020.105917.

36. Xu K. *et al.* (2018), Mixup-based acoustic scene classification using multi-channel convolutional neural network, [in:] *Advances in Multimedia Information Processing – PCM 2018*, pp. 14–23, doi: 10.48550/arXiv.1805.07319.

37. Yadollahi A., Moussavi Z. (2010), Automatic breath and snore sounds classification from tracheal and ambient sounds recordings, *Medical Engineering & Physics*, **32**(9): 985–990, doi: 10.1016/j.medengphy.2010.06.013.

38. Young T., Peppard P.E., Gottlieb D.J. (2002), Epidemiology of obstructive sleep apnea: A population health perspective, *American Journal of Respiratory and Critical Care Medicine*, **165**(9): 1217–1239, doi: 10.1164/rccm.2109080.