Audio Signals Encoding for Cough Classification Using Convolutional Neural Networks: A Comparative Study

Hui-Hui Wang, Jia-Ming Liu, Mingyu You* and Guo-Zheng Li
Department of Control Science and Engineering, Tongji University,
Shanghai, 201804, China
Email: myyou@tongji.edu.cn*

***Abstract*—Cough detection has considerable clinical value, which can provide an objective basis for assessment and diagnosis of respiratory diseases. Motivated by the great achievements of convolutional neural networks (CNNs) in recent years, we adopted 5 different ways to encode audio signals as images and treated them as the input of CNNs, so that image processing technology could be applied to analyze audio signals. In order to explore the optimal audio signals encoding method, we performed comparative experiments on medical dataset containing 70000 audio segments from 26 patients. Experimental results show that RASTA-PLP spectrum is the best method to encode audio signals as images with respect to cough classification task, which gives an average accuracy of 0.9965 in 200 iterations on test batches and a F1-score of 0.9768 on samples re-sampled from the test set. Therefore, the image processing based method is shown to be a promising choice for the process of audio signals.**

***Keywords*—cough detection; audio features; convolutional neural networks; image processing; cough classification**

I. Introduction

Cough is the body's way of removing foreign material or mucus from the lungs and upper airway passages or of reacting to an irritated airway. It is a common and important symptom in the evaluation and diagnosis of many respiratory diseases. Although cough acts as a protective mechanism, its increased frequency and intensity have adverse impact on quality of life and may cause organ injuries [1].

In clinical diagnosis and treatment of respiratory diseases, assessment of cough severity is the very first step [2]. Precise assessment results can contribute doctors to master the patients' condition and make a reliable diagnosis. However, in current clinical environment, measurements of cough severity are highly uncertain according to the doctors' working expertise, the patients' tolerance and human's subjective judgments. Thus, objective cough detection systems are necessary tools in clinical practice.

A typical objective cough detection method is to detect speech events and classify them to cough events and non-cough events. And then, cough events are counted to evaluate the cough severity for cough frequency. A few methods based on audio signal analysis have been proposed in recent years and shown good performance [1][3]. Sergio Matos et al. [1] presented a system – The Leicester Cough Monitor (LCM) for the automatic analysis of 24 hours continuous recordings of cough. In this work, Mel frequency cepstral coefficients (MFCCs), as well as the first and second-order derivatives of these coefficients, were chosen as features. This system got a median sensitivity value of 85.7%, median positive predictive value of 94.7%. Samantha J Barry et al. [3] proposed a method: The Hull Automatic Cough Counter (HACC). It used digital signal processing (DSP) to calculate spectral coefficients of sound events, which were then classified into cough and non-cough events by using a probabilistic neural network (PNN). HACC achieved a sensitivity of 80% and a specificity of 96% on an hour long recording.

Recently, motivated by the success of deep learning, several researchers, who try to apply deep learning methods to detect cough events, reported their achievements. Jia-Ming Liu et al. [2] proposed a deep learning based cough detection method. By introducing deep neural networks (DNN), a more accurate cough detection system was built, in which a HMM based segmentation method and a deep neural network based classifier were used. Compared with GMM-HMM baseline, 13.38% and 22.0% relative error reduction were achieved in speaker dependent test set and speaker independent test set respectively.

Inspired by the tremendous progress made by Convolutional Neural Networks (CNNs) in the field of computer vision, Ossama Abdel-Hamid et al. [4] reported their achievements on applying CNNs to speech recognition. In their work, hybrid CNN-HMM approach was proposed which employed mel-frequency spectral coefficients (MFSCs) as features. Compared with DNNs, CNNs reduced the error rate by 6%-10% on the TIMIT phone recognition and the voice search large vocabulary speech recognition tasks.

Following previous work, in this paper, we explore the optimal method to encode audio signals as images so that it can be fed into CNNs and obtain better performance in subsequent work. In this way, we aim to find the components between cough signals and non-cough signals which are both representative and distinctive in cough classification task. The rest of the paper is organized as follows. Five methods to encode audio signals as images are firstly described in Section II. Then, a review of CNNs and its application in speech recognition are given in Section III. Experiment details and results discussion are shown in Section IV. Finally, the

978-1-4673-6799-8/15/$31.00 ©2015 IEEE

conclusion of this study and the future work are given in Section V.

II. Encoding Audio Signals as Images

The input data to be fed into CNNs need to be organized as a number of feature maps [4]. Thus, the first step of using CNNs for cough classification is to encode audio signals as images.

Spectrum, the most popular visual displays of audio signals, is the first choice to be considered. Another popular method to represent audio signals in picture form is known as RASTA-PLP, the abbreviation of Relative Spectral Transform - Perceptual Linear Prediction. PLP, originally proposed by Hynek Hermansky [5], was a way of warping spectra to minimize the differences between speakers while preserving the important speech information. RASTA is a separate technique that applies a band-pass filter to the energy in each frequency subband in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration [6]. In addition to the foregoing techniques, we plan to explore the performance of cepstral techniques in cough classification, whose feasibility has been demonstrated already by Noll, A. Michael [7]. So, the five different methods to be used to extract the image features of audio signals are original spectrum, RASTA-PLP power spectrum, RASTA-PLP cepstrum, 12th order PLP power spectrum without RASTA and 12th order PLP cepstrum without RASTA, which will be reviewed briefly in the following part of this section.

A. PLP

In the PLP technique, three concepts from psychophysics of hearing are used to derive an approximation of the auditory spectrum [5], which are

- the critical-band spectral resolution,
- the equal-loudness curve,
- the intensity-loudness power law.

Then,the resulting auditorylike spectrum is approximated by an autoregressive all-pole model.

In PLP technique, spectral analysis of speech segment is firstly carried out according to Equ.1 and Equ.2. A speech segment is weighted by the Hamming window defined as Equ.1, where N is the length of the window. Subsequently, the windowed speech segment is transformed into frequency domain by the Fast Fourier Transform (FFT). The short-term power spectrum $P(w)$, given in Equ.2, is the sum of the real and imaginary components of the short-term speech spectrum $S\,(w)$.

$$W(n) = 0.54 + 0.46\cos[2\pi n/(N-1)] \tag{1}$$

$$P(\omega) = \mathrm{Re}[S(\omega)]^2 + \mathrm{Im}[S(\omega)]^2 \tag{2}$$

1) Critical-band spectral resolution: The spectrum $P(w)$ is warped along its frequency axis into the Bark frequency Ω by

$$\Omega(\omega) = 6\ln\{\omega/1200\pi + [(\omega/1200\pi)^2 + 1]^{0.5}\} \tag{3}$$

The warped power spectrum $\Omega(\omega)$ is convolved with the power spectrum of the simulated critical-band curve $\Psi(\Omega)$. The result of convolution between $\Psi(\Omega)$ and $P(w)$ is shown in Equ.4.

$$\Theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega) \tag{4}$$

2) Equal-loudness pre-emphasis: Equ.5 represents the pre-emphasis of $\Theta[\Omega(\omega)]$, where $E(\omega)$, shown in Equ.6, is a function simulating the sensitivity of human hearing.

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)] \tag{5}$$

$$\begin{gathered} E(\omega) = [(\omega^2 + 56.8\times 10^6)\omega^4]/[(\omega^2 + 6.3\times 10^6)^2 \\ (\omega^2 + 0.38\times 10^9)(\omega^6 + 9.58\times 10^{26})] \end{gathered} \tag{6}$$

3) Intensity-loudness power law:

$$\Phi(\Omega) = \Xi(\Omega)^{0.33} \tag{7}$$

After the cubic-root amplitude compression operation given in Equ.7, the inverse DFT is applied to $\Phi(\Omega)$ to yield the auto-correlation function, which can be used to solve the autoregressive coefficients of all-pole model. Meanwhile the autoregressive coefficients could be further transformed into cepstral coefficients of the all-pole model [5].

B. RASTA-PLP

Motivated by the phenomenon that human listeners seem to pay little attention to a slow change of the communication environment or steady background noise, H. Hermansky et al. [6] replaced a conventional critical-band short-term spectrum in PLP speech analysis with a spectral estimation and named it "RASTA-PLP". The steps of RASTA-PLP are as described in [6].

Based on PLP and RASTA-PLP technology, We extracted original spectrum, 12th order PLP spectral features, 12th order PLP cepstral features, RASTA-PLP spectral features and RASTA-PLP cepstral features to compare their performance on cough classification. The features were extracted based on the RASTAMAT Matlab Toolkit [8].

III. Convolutional Neural Networks

CNNs were originally proposed for solving handwritten digit recognition by LeCun et al. After yielding unusually brilliant results in the ILSVRC2012 competition, CNNs were applied extensively in computer vision field and extended to other fields such as automatic speech recognition, natural language processing and bioinformatics where they have been shown to produce state-of-the-art results on various tasks.

Based on simple feed-forward Neural Networks structure, CNNs introduces three extra concepts between input layer and the first hidden layer: local filters, pooling, and weight sharing [9]. The other hidden layers in Neural Networks were treated as fully connected layers.

In using CNNs, the first step is to organize the input data as a number of feature maps. Then, the convolution and pooling layers were applied in sequence. A pair of convolution and pooling layers in succession is usually named as one CNNs

layer and a deep CNN consists of two or more of these pairs in succession [4].

CNNs have the property of translation invariance for many image processing tasks. Abdel-Hamid et al. [9] applied CNN to speech recognition. After using local filtering and max-pooling in frequency domain to normalize speaker variance, they achieved higher multi-speaker speech recognition performance. Cough classification is also a multi-speaker speech recognition task. Intuitively, the speaker invariance of CNNs will be in favor of performance improvement.

IV. Experiments

A. Dataset Preparation

All the data used in this paper were collected from 26 patients in Shanghai Tongji Hospital in China, who were diagnosed with various respiratory diseases, including community acquired pneumonia (CAP), bronchial asthma (BA) and chronic obstructive pulmonary disease (COPD). The data were recorded using a portable digital audio recorder (SONY ICD-LX30) and a microphone (ECMCS10) with 44.1kHz sampling frequency and 192 kbps bit rate. In order to obtain real and reliable clinic data, patients were encouraged to go through their daily routines as usual. All the data were annotated by several graduate students on Praat platform [10].

In the previous work [2], we applied GMM-HMM (Gaussian Mixture Model- Hidden Markov Model) model to detect the cough events and wrote the results to wav files. Then, all wav files were reviewed manually to pick out the false "cough" events from the predicted cough events and the true "cough" events from the predicted non-cough events. In this paper, the reviewed wav files were used as our dataset. It contains 70000 audio clips, 13804 of which are cough signals and the rest 56196 of which are non-cough signals. The original intention for us to choose the results of GMM-HMM as dataset is to explore whether CNNs are confused by the data as GMM-HMM model does and if there exists some crucial components which can be used to distinguish cough signals and non-cough signals decisively.

B. Experiments Setup

We employed 5 different methods to encode audio signals as images as described in Section II. Considering the computation cost, the frequency of original audio files were resampled as 16,000Hz. And, all of the encoded images in the data set were resized as 256x256.

For the learning rate of CNN, it was set to 0.01 in the beginning and was reduced by 0.1 after each 8000 iterations. The maximum number of iterations is set to 10000. The factor of weight decay is set to 0.0005.

The architecture of CNN, is based on AlexNet [11]. The net contains eight layers; the first five are convolutional and the remaining three are fully connected. The output of the last fully-connected layer is fed to a 2-way softmax layer which produces a distribution over the two class labels [11]. The CNN model was developed on the basis of CAFFE – a deep learning framework [12]. Furthermore, GPUs in a NVIDIA Tesla K20 were exploited in experiment in order to acceleration training, which contains 5GB of GDDR5 RAM and 2496 processing cores.

TABLE I
Average Accuracy in 200 Iterations on Test Set Batches.

Method	Average Accuracy
Original spectrum	0.938
RASTA-PLP spectrum	**0.9965**
RASTA-PLP cepstrum	0.8956
12th order PLP spectrum	0.9392
12th order PLP cepstrum	0.9302

C. Results and Discussions

Each of the five trained CNN models is tested for 200 iterations on test set batches separately to estimate the stability of the models. The accuracy distribution of 200 test iterations are shown in Fig.1 and the average accuracy of five methods are shown in Table.I. From results, we can easily draw the conclusion that RASTA-PLP spectrum is the optimal method to encode audio signals as images in cough classification task. Its accuracy is concentrated and 100% classification accuracy is obtained on most of the test batches. The model trained on RASTA-PLP Cepstrum features demonstrates poor stability for the reason that its accuracy distribution span is relatively large (ranging from 0.7 to 1) and the average accuracy is low compared with other four methods.

Furthermore, accuracy, precision, recall and F1-score are utilized to evaluate the performance of the five trained CNN models on the re-sampled samples (each sample in the test set was re-sampled to 10 samples: center, four corners and their mirrors). In order to make a compare with traditional cough classification methods, typical MFCCs features were extracted on the cough data set, and a popular Support Vector Machine (SVM) was chosen as classifier. Test results were recorded in Table.II. On one hand, these statistics show that performance of the combination of RASTA-PLP spectrum and CNN is outstanding with regard to all the four evaluation indexes especially in recall and F1-score. The accuracy obtained by RASTA-PLP spectrum and CNN is over 9 percentage point higher than the second place. And it achieved a complete victory in terms of recall and F1-score. On the other hand, the results reveal that there are three methods in five ways to encode audio signals as images which acquire better performance than the combination of MFCCs and SVM [13][14].

V. Conclusion and Future Work

In this paper, we apply five different ways to encode audio signals as images and explore the optimal way in cough detection task. The experiment results show that feature extracted by RASTA-PLP spectrum method can obtain better performance than other methods.

In addition, we establish a connection between audio signal processing and computer vision. We apply CNNs, a powerful

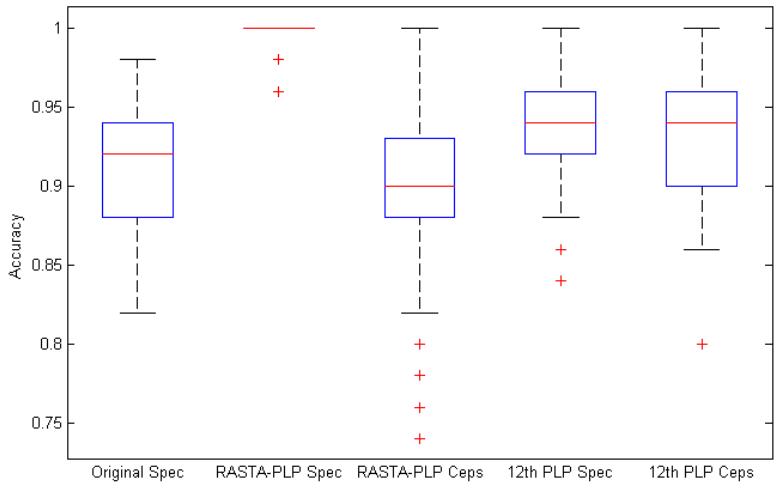

Fig. 1. Accuracy distribution of five different methods in 200 iterations on test set batches.

TABLE II
PERFORMANCE COMPARISON ON THE RE-SAMPLED DATA SET

Method	Accuracy	Precision	Recall	F1 score
Original spectrum+CNN	0.8947	0.9578	0.679	0.7947
RASTA-PLP spectrum+CNN	**0.9859**	**0.9709**	**0.9827**	**0.9768**
RASTA-PLP cepstrum+CNN	0.8928	0.9451	0.6823	0.7925
12th order PLP spectrum+CNN	0.8511	0.8286	0.635	0.719
12th order PLP cepstrum+CNN	0.6405	0.5315	0.4635	0.4954
MFCCs+SVM [13][14]	0.8691	0.7818	0.5737	0.6618

technology from computer vision, to the task of cough classification and get a surprising result.

However, there exists some content which is needed to be studied further in future works, such as other methods to encode audio signals as images, or the performance of these five methods on other data sets other than the cough data.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of China under grant no. 61273305 and 81274007, and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Sergio Matos, Surinder S Birring, Ian D Pavord, and David H Evans. An automated system for 24-h monitoring of cough frequency: the leicester cough monitor. *Biomedical Engineering, IEEE Transactions on*, 54(8):1472–1479, 2007.

[2] Jia-Ming Liu, Mingyu You, Zheng Wang, Guo-Zheng Li, Xianghuai Xu, and Zhongmin Qiu. Cough detection using deep neural networks. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 560–563. IEEE, 2014.

[3] Samantha J Barry, Adrie D Dane, Alyn H Morice, and Anthony D Walmsley. The automatic recognition and counting of cough. *Cough*, 2(1):8, 2006.

[4] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(10):1533–1545, 2014.

[5] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[6] Hynek Hermansky and Nelson Morgan. Rasta processing of speech. *Speech and Audio Processing, IEEE Transactions on*, 2(4):578–589, 1994.

[7] A Michael Noll. Short-time spectrum and cepstrum techniques for vocal-pitch detection. *The Journal of the Acoustical Society of America*, 36(2):296–302, 1964.

[8] Daniel P. W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab, 2005. online web resource.

[9] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4277–4280. IEEE, 2012.

[10] Paul Boersma and David Weenink. Praat, a system for doing phonetics by computer. 2001.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[12] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[13] Sandra Larson, Germán Comina, Robert H Gilman, Brian H Tracey, Marjory Bravard, and José W López. Validation of an automated cough detection algorithm for tracking recovery of pulmonary tuberculosis patients. *PLoS ONE*, 7(10), 2012.

[14] Brian H Tracey, Germán Comina, Sandra Larson, Marjory Bravard, Jose W Lopez, and Robert H Gilman. Cough detection algorithm for monitoring patient recovery from pulmonary tuberculosis. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 6017–6020. IEEE, 2011.