Automatic Heart and Lung Sounds Classification using Convolutional Neural Networks

Qiyu Chen *[1], Weibin Zhang *[1], Xiang Tian †[1], Xiaoxue Zhang [2], Shaoqiong Chen [1]and Wenkang Lei [1]

South China University of Technology, GuangZhou, China[1]
Guangdong No.2 Provincial People's Hospital, Guangzhou, China[2]

chenqiyuscut@qq.com, eeweibin@scut.edu.cn, xtian@scut.edu.cn,
46683357@qq.com, 137741110@qq.com, lei.wenkang@mail.scut.edu.cn

***Abstract*—We study the effectiveness of using convolutional neural networks (CNNs) to automatically detect abnormal heart and lung sounds and classify them into different classes in this paper. Heart and respiratory diseases have been affecting humankind for a long time. An effective and automatic diagnostic method is highly attractive since it can help discover potential threat at the early stage, even at home without a professional doctor. We collected a data set containing normal and abnormal heart and lung sounds. These sounds were then annotated by professional doctors. CNNs based systems were implemented to automatically classify the heart sounds into one of the seven categories: normal, bruit de galop, mitral inadequacy, mitral stenosis, interventricular septal defect (IVSD), aortic incompetence, aorta stenosis, and the lung sounds into one of the three categories: normal, moist rales, wheezing rale.**

Keyword: heart sound classification, lung sound classification, Convolutional Neural Networks

I. Introduction

Sounds of hearts and lungs are one of the important human physiological signals. Heart sound is an acoustic signal generated by the heart beating and lung sound is an acoustic signal generated when we breath. Since Laennec firstly used a stethoscope to auscultate in 1816, sounds of hearts and lungs have been an important method for the diagnosis of heart diseases and respiratory diseases clinically. The abnormal sounds usually indicate some kinds of diseases of related organ. How to diagnose diseases from sounds of heart and lung has been one of the focuses of medical professional trainings. Currently, the diagnosis of heart and lung diseases based on sounds still needs to involve a well trained doctor, which is undoubtedly costly and inconvenient. A method that can accurately and automatically classify heart and lung sounds into different categories will be very meaningful. It can help discover potential threat at the very early stage conveniently, even at home without a doctor.

With the development of deep learning in recent years, Convolutional Neural Networks have been demonstrated as an effective end-to-end classifier in many fields, such as image[1], [2], [3], speech[4], [5], video[6] and music[7] . The key enabling factors behind these results are techniques for scaling up the networks to tens of millions of parameters and massive labeled datasets that can support the learning process [8]. CNNs are a biologically-inspired class of deep learning models that can automatically extract relevant information from the input features and output the classification results. Encouraged by the tremendous success of using CNNs in areas such as speech recognition[5] and music genre classification [7], we explore in this paper the effectiveness of using CNNs to automatically classify heart and lung sounds into different categories. We aim to build an automatic and high accurate diagnostic method to alleviate the burden of doctors.

There are many researches about heart and lung sounds analysis [9], [10], [11], [12], [13]. However, as far as we know, works on applying CNNs to the classification of human health related sounds are very little. This is probably because of the lack of enough training data. A. Kandaswamy [14] used a back propagation algorithm with wavelet coefficients to classify the lung sounds and the average accuracy was 92%.

As a typical classification problem, the first step to classify the heart and lung sounds would be to choose the meaningful features. These hand-crafted features will require a lot of prior expert knowledge. Luckily, CNNs offer us the ability to use very low level features and optimize them for the problem at hand during training, making the problem much easier for engineers who do not have a lot of medical knowledge. Thus we will use CNNs in our experiments in this paper. We use the Short Time Fourier Transform (STFT) magnitude spectrum as the input features of CNNs. Due to the limited amount of data collected, we found that a CNN architecture with two convolutional layers yielded the best results.

We did not find any public data sets suitable for the problem we want to study at hand. Thus we collected our own database. Data were collected from online website or from volunteers who are potential patients. These data were then annotated by professional doctors. Though we have been doing our best to collect as more data as possible, only several hours of data have been collected when we wrote this manuscript. This process is very labor-intensive and very costly. Nevertheless, we were very excited to find that the data we collected are enough to train a not-so-deep CNN that can achieve very high recognition accuracy.

*Both authors contributed equally to this paper.

†The communication author.

The contributions of this paper can be summarized as follows:

- We have collected a data set, although very small, that can be used to study the effectiveness of automatic classification.
- Due to the lack of prior expert knowledge, we propose CNNs to automatically classify the heart and lung sounds, with simple STFT features as the CNNs input. We also experimentally demonstrate the feasibility of this approach.

The rest of this paper is organized as follows. In Section 2, we will describe the details of our methodology. Then the description of the data set, experimental setup, results follows in Section 3. In Section 4, we draw the conclusions.

II. Methodology

One difficulty of conventional signal processing is that sometimes it is difficult for us to know exactly what characteristics are decisive to data classification and how it makes data different. A bad characteristic may affect the results greatly. Machine learning, especially CNNs, deals with this problem better. We don't need to have too much prior knowledge when using a CNNs classifier since it will automatically adjust itself to better classify the input features. The input features of our networks are Short Time Fourier Transform (STFT) magnitude spectrum. Dtails will be described in Section 3.

As a supervised learning model, CNNs typically contain two parts: 1) several stacked convolutional and pooling layers to extract high level features. Through parameter sharing in the convolutional layers, it can help substantially reduce the number of parameters and thus alleviate the over-fitting problem. In addition, different feature maps can discover salient patterns locally for classification. Finally, the pooling operation helps again reduce the number of parameters by keeping only the important ones. 2) The convolutional and pooling layers are followed by fully connected layers that act as the classifier. All the parameters of a CNN (i.e. the convolutional, pooling and the fully connected layers) are tailored to the problem at hand during the training process. Thus CNNs usually performs much better than conventional classifiers with hand-crafted features.

However, the architecture and the parameters of the network will have a great impact on the performance of the model. For example, a model with more layers will be able to learn more abstract features, but at the same time it will also be more likely to become over-fitting. Since our data are not very large, we finally choose a CNN with seven layers in order to avoid over-fitting. In the following, we will discuss the architectures of the two networks used in heart and lung sounds classification, separately.

A. Heart sound classification

The architecture of the neural network used for heart sound classification is shown in Fig. 1. We use six layers, including two convolutional layers and two fully connected layers. The STFT spectrogram of the input audio signal, which contains 513 frequency bins for each frame, is used as the input features. As will be explained latter, each sound clip consists of about 45 frames. The first convolutional layer consists of 64 different filtering kernels with the same size. Each kernel surveys a fixed size of 513×30 of the input features, multiplying them with the associated weights in the kernel during the convolution operation. Then the kernel steps forward along the time dimension with a unit stride. Thus there are 64×1×16 features computed as the output of the first convolutional layer.

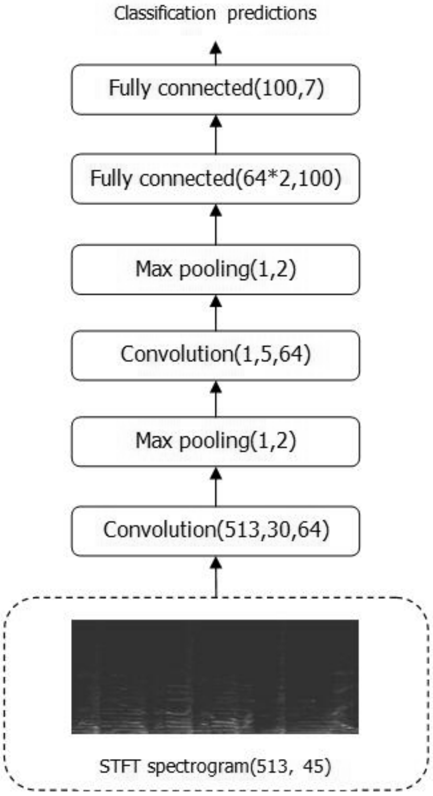

Fig. 1. The architecture of CNNs used for heart sound classification. The meaning of the parameters in the above figure is described in subsection of heart sounds classification

The CNNs structure can learn high level features from the spectrogram better since the kernel is shared inside a feature map. It allows useful features to be detected regardless of their position in the spectrum. In addition, different kernels can detect different feature patterns. The higher the convolutional layer is, the patterns detected will be more global.

After every convolutional layer, there is a max pooling layer. Each pooling neuron will survey a non-overlap 1×2 region from the input features and keep only the maximum. When the input is a 1×16 matrix, the output will be a 1×8 feature map. The pooling operation is an important factor to the success of our classification task. It helps keep only the salient features and reduce the number of parameters to be learnt.

The second max pooling layer will output 64×1×2 feature. The outputs are then reshaped to a 128 dimension vector,

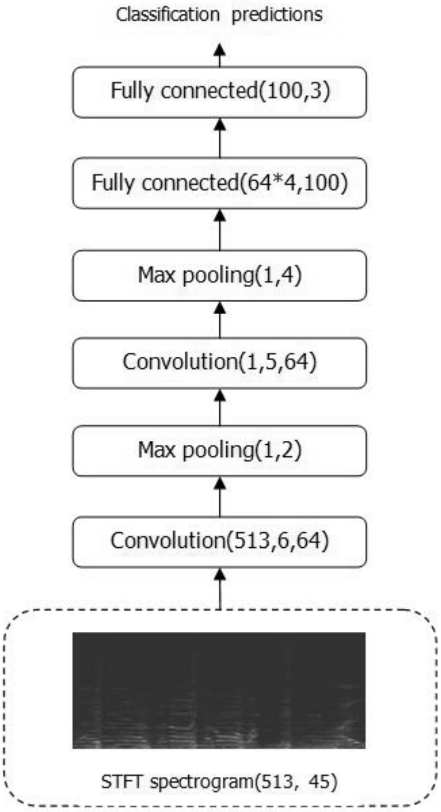

Fig. 2. The architecture of CNNs used for heart sounds classification. The meaning of the parameters in the above figure is described in subsection of lung sound classification

which can be viewed as the highest level features of the heart sound learned by the network. The vector will be fed into the fully connected feed forward networks. The fully connected layers are used as a classifier to automatically classify the sounds into different categories. The input of the first fully connected layer is a 128-dimension vector and it outputs another 100-dimension vector. The dimension of the output of the final layer equals the number of categories of the sounds.

The activation function used in our networks (except the last layer where the softmax function is used) is the rectified linear units (ReLUs) [15], [16]. Compared with other activation functions such as sigmoid and tanh, ReLUs can help accelerate the convergence of the models. During training, other techniques such as Dropout is also used to prevent the model from over-fitting.

B. Lung sound classification

The CNNs architecture of neural network used for lung sound classification is shown in Fig. 2, which is very similar to the Fig. 1. Again the networks have six layers. Some parameters have been changed to achiever better results. The kernel size of the first convolutional layer is changed to 513×6. The output of the first convolutional layer are 64 maps with dimension 1×40. The pooling size of the second layer is 1×4.

III. Experiments and Results

In this Section, we report the experiments used to evaluate the methodologies described in Section 2.

A. Dataset

The dataset used in the experiments is mainly collected from online websites and volunteers who are potential patients. All the collected data were annotated by professional doctors. The data set contains 393 recordings of heat sounds (about half an hour in total). All the recordings of heart sound are labelled into one of the following categories: normal, bruit de galop, mitral inadequacy, mitral stenosis, interventricular septal defect (IVSD), aortic incompetence and aorta stenosis. There are also 236 recordings of lung sounds (about 2 hours in total). All the recordings of lung sound are labelled into one of the following categories:normal, moist rales and wheezing rale. The duration of a recording ranges from several seconds to as long as a minute. All the recordings are sampled at 8000 Hz and 16 bits.

We divided all the data (heart sound and lung sound were separated) into five batches to do a 5-fold cross validation. Every time we chose one of the five batches as the validation set. The remaining data were used as the training data. The validation set was further divided into a development data set (50%) and a testing data set (50%). The number of recordings of different categories in the train, development and testing sets is balanced. All the models were tuned on the development data set and tested on the testing data set. The results reported below were averaged over five runs.

B. Experimental Setup

As the normal practice in music genre classification[17], we cut every recording into shorter segments (3 seconds) with an overlap of 50%. We purposely chose three seconds so that the duration is at least two or three times of the period of heart beating or human breathing. These information is very important for the classification. Then, to get the STFT magnitude spectrum, we compute the spectrum on frames of length 1024 with an overlap of 50% and get the distribution of the energy along the frequencies. For every segment, we can finally get 45 frames spectrum, each frame is a 513 dimension feature vector.Each spectrum will be marked with a label.

When training the network, we use RMSPropop as our optimizer, the initial learning rate was set to 0.001. Every spectrum of the training set will be sent to the input layer, and then the network was trained by optimizing the cost function. We also used the dropout technique with 0.2 dropout rate to alleviate the over-fitting problem. During testing, different segments from the same recording will be counted and finally we select the label with maximum count as the predicted result.

C. Results

The model complexity is very important. We optimized the model complexity by varying the number of convolutional layers. The results for heart sound classification are shown in

Table 1. As can be seen, a CNN with only one convolutional layer might under fit the data. Due to the limited size of the training data, we found that the CNN with two convolutional layers achieves the best results. Table 2 shows the classification accuracy of lung sound. Since the data of the lung sound are much more than that of the heart sound, we achieved much better classification accuracy on this data set.

TABLE I
THE CLASSIFICATION RESULTS OF HEART SOUNDS WITH DIFFERENT NUMBER OF CONVOLUTIONAL LAYERS

Architectures	Accuracy
1 convolutional layer	94.80%
2 convolutional layer	95.49%
3 convolutional layer	92.47%

TABLE II
THE CLASSIFICATION RESULTS OF LUNG SOUNDS WITH DIFFERENT NUMBER OF CONVOLUTIONAL LAYERS

Architectures	Accuracy
1 convolutional layer	94.93%
2 convolutional layer	97.80%
3 convolutional layer	89.83%

Even though the training data are very limited, the best classification accuracies for the heart and lung sounds are all higher than 95%. We think that it is very promising to apply the proposed methods in real applications for automatic diagnosis.

IV. Conclusions

Acoustic signal generated by the heart beating and lung breathing contains important information for the diagnosis of heart disease and respiratory disease. Automatic diagnosis based on heart and lung sounds are very attractive since it can help discover disease at early stage without a doctor. In addition, automatic diagnosis based on sounds can be very cost effective since recording sound is very cheap.

In this paper, we study the effectiveness of using convolutional neural networks (CNNs) to automatically detect abnormal heart and lung sounds and classify them into different classes. We collected a data set containing normal and abnormal heart and lung sounds. These sounds were then annotated by professional doctors. CNNs based systems were implemented to automatically classify the heart sounds into one of the seven categories: normal, bruit de galop, mitral inadequacy, mitral stenosis, interventricular septal defect (IVSD), aortic incompetence, aorta stenosis, and the lung sounds into one of the three categories: normal, moist rales, wheezing rale. Our experimental results show that the classification accuracy can be as high as 97.80%, even though the training data are very limited. We believe that it is very promising to apply this method in real applications.

In the future, we want to collect more data. We are also interesting in exploring deeper or more advanced neural networks(e.g. [18]) for this problem.

V. Acknowledgement

This work is supported in part by the Science and Technology Planning Project of Guangdong Province, China (2014B010111003 and 2014B010111006), the Guangzhou Key Lab of Body Data Science (201605030011), the Fundamental Research Funds for the Central Universities (2015ZJ009) and the National Natural Science Foundation of China (61601187).

References

[1] Lo, Shih-Chung B and Chan, Heang-Ping and Lin, Jyh-Shyan and Li, Huai and Freedman, Matthew T and Mun, Seong K, "Artificial convolution neural network for medical image pattern recognition." in *Neural networks,* 1995:1201–1214.

[2] Ciresan, Dan C and Meier, Ueli and Masci, Jonathan and Maria Gambardella, Luca and Schmidhuber, Jürgen, "Flexible, high performance convolutional neural networks for image classification." in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence,* 2011:1237.

[3] Lawrence, Steve and Giles, C Lee and Tsoi, Ah Chung and Back, Andrew D, "Face recognition: A convolutional neural-network approach." in *Neural Networks, IEEE Transactions on,* 1997:98–113.

[4] Tóth, László, "Combining time-and frequency-domain convolution in convolutional neural network-based phone recognition." in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on,* 2014:190–194.

[5] Sainath, T and Parada, Carolina, "Convolutional neural networks for small-footprint keyword spotting." in *Proc. Interspeech,* 2015.

[6] Zha, Shengxin and Luisier, Florian and Andrews, Walter and Srivastava, Nitish and Salakhutdinov, Ruslan, "Exploiting image-trained cnn architectures for unconstrained video classification." in *arXiv preprint arXiv:1503.04144,* 2015.

[7] Nakashika, Toru and Garcia, Christophe and Takiguchi, Tetsuya, "Local-feature-map Integration Using Convolutional Neural Networks for Music Genre Classification." in *INTERSPEECH,* 2012:1752–1755.

[8] Karpathy, Andrej and Toderici, George and Shetty, Sanketh and Leung, Thomas and Sukthankar, Rahul and Fei-Fei, Li, "Large-scale video classification with convolutional neural networks." in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition,* 2014:1725–1732.

[9] Reed, Todd R and Reed, Nancy E and Fritzson, Peter, "Heart sound analysis for symptom detection and computer-aided diagnosis." in *Simulation Modelling Practice and Theory,* 2004:129–146.

[10] Rakovic, P and Sejdic, E and Stankovic, LJ and Jiang, J, "Time-frequency signal processing approaches with applications to heart sound analysis." in *Computers in Cardiology,* 2006:197–200.

[11] Baughman, Robert P and Loudon, Robert G, "Lung sound analysis for continuous evaluation of airflow obstruction in asthma." in *CHEST Journal,* 1985:364–368.

[12] Wang, Zhen and Jean, Smith and Bartter, Thaddeus, "Lung sound analysis in the diagnosis of obstructive airway disease." in *Respiration,* 2008:134–138.

[13] Kandaswamy, A and Rajkumar, S and Kumar, AS and Jayaraman, S, "Respiratory system diagnosis through lung sound processing." in *J. Systems Sci. Eng,* 1999:32–36.

[14] Kandaswamy, A and Kumar, C Sathish and Ramanathan, Rm Pl and Jayaraman, S and Malmurugan, N, "Neural classification of lung sounds using wavelet coefficients." in *Computers in Biology and Medicine,* 2004:523–537.

[15] Nair, Vinod and Hinton, Geoffrey E, "Rectified linear units improve restricted boltzmann machines." in *roceedings of the 27th International Conference on Machine Learning (ICML-10),* 2010:807–814.

[16] Glorot, Xavier and Bordes, Antoine and Bengio, Yoshua, "Deep sparse rectifier neural networks." in *International Conference on Artificial Intelligence and Statistics,* 2011:315–323.

[17] Wenkang Lei, Weibin Zhang, Xiangmin Xu and Xiaofeng Xing, "Improved musice genre classification with convolutional neural networks." in *Interspeech,* to appear in Interspeech 2016.

[18] Fernández, Santiago and Graves, Alex and Schmidhuber, Jürgen, "An application of recurrent neural networks to discriminative keyword spotting." in *Artificial Neural Networks–ICANN,* 2007:220–229.