The best input feature when using convolutional neural network for cough recognition

Yunan Cai *, Wenlong Xu
School of Information Engineering, China Jiliang University, Hangzhou, China

*Corresponding author e-mail: yunancai96@163.com

Abstract. In recent years, the use of convolutional neural networks has been successful in the task of cough recognition. This method mainly converts audio clips into the form of spectrograms, and then uses convolutional neural networks for classification, which prompts us to seek better input representation. For more effective training to achieve better performance, in this article, we use STFT, Mel spectrogram, Log-Mel spectrogram and MFCC four different features as the input of the convolutional neural network, in the case of the same parameters, compare their performance below, where the Mel spectrum is used as input to achieve 92.5% classification accuracy. Secondly, we compared the classification performance of different Mel spectra generated by different window sizes and frame shifts under the same convolutional neural network. For a 320ms cough segment, the window size is 64, and the frame shift is 32, which has the best performance.

1. Introduction
Cough is a common symptom of many respiratory diseases[1], factors such as the duration of coughing and the intensity of coughing are often used as important parameters for doctors to judge the condition and conduct treatment.

The World Health Organization recommends that individuals who have a cough that lasts for two weeks or more should be tested for tuberculosis[2]. However, in the actual medical process, patients cannot truly reflect the actual situation when describing the symptoms of cough[3]. Therefore, since the 1950s, researchers have been researching and developing cough monitoring instruments. Since the beginning of this century, with the rise of machine learning, scientists have begun to use traditional machine learning methods such as Hidden Markov Models to monitor coughs[4], but such methods require manual selection of sound features, so the results are very good and bad. Much depends on the quality of feature extraction. With the development of artificial intelligence in recent years, deep learning methods have been more and more applied to data processing. Convolutional neural networks have achieved very superior performance in image processing. At the same time, Convolutional Neural Networks[5] has also been proven to be used for cough detection. Figure 1 shows the audio file recognition process of traditional speech recognition methods and deep learning methods. This article aims to explore the best input data when using convolutional neural networks to recognize cough sounds by comparing different frequency spectrum inputs, so as to provide an algorithm basis for recognizing cough sounds from sound clips in the later stage.
2. Method

2.1. Data Collection
In order to build and evaluate the proposed cough sound recognition system, we created a data set containing 1000 cough sounds and 2000 hospital murmurs. These sound clips all come from the Freesound audio database. All audio files in the database are accompanied by uploading user titles and descriptions. We have 765 audio files with "cough" in the title or description and "hospital" in the title or description. "Cough" contains the cough sound data we need, and "hospital" contains the noise in the hospital, including footsteps, the sound of equipment in the hospital, the voice of patients, etc. We listened to the human ear and put the audio clips into the corresponding categories, because the duration of the collected cough clips was 230 milliseconds to a maximum of about 700 milliseconds. To ensure that the length of our training examples is the same, we also set the "hospital" The sound fragments in the 605 files in the file are divided into smaller fragments, and random duration is generated according to the Gaussian distribution of the duration of the cough example [6], which makes the average duration of all audio examples in the database be around 320 milliseconds.

2.2. Convolutional neural network architecture
Our convolutional neural network architecture is inspired by the LeNet-5 network structure [7]. This network structure has achieved very ideal results on the MNIST handwritten character data set, and compared with other network structures such as AlexNet, Googlenet, VGG-19 and other network structures [8,9,10], LeNet-5 is a relative Smaller network, our data set only contains about 3000 data, we have made certain improvements on the basis of LeNet-5 to adapt to the task of cough recognition. Our network consists of 6 layers: 2 convolutional layers, 3 fully connected layers and 1 softmax classification layer. The first convolutional layer takes a 64x64x3 spectrum segment as input and uses a 5x5 size filter, the stride is 1, use the tanh function to activate, and then use the maxpooling layer with a size of 2x2 and a stride of 2. The second convolution layer uses a filter with a size of 5x5, and the stride is set to 1, using tanh The function is activated, followed by a maxpooling layer with a filter size of 2x2 and a stride of 2; after the maxpooling layer, 3 fully connected layers are connected, and the softmax function is used to classify the input as a cough event or a noise event.

2.3. Acoustic characteristics
We evaluated the cough detection performance of four audio features: STFT, MFB, Log-Mel, MFCC. These features are chosen because they are widely used in speech recognition systems and have achieved very good performance. Figure 2 is the MFCC feature extraction process, from which we will extract four audio features for comparison experiments.

![Figure 2. The specific extraction process of MFCC](image-url)
2.3.1. Short-time Fourier transform. The sound signal is usually a non-stationary signal. Using the entire time domain information of the signal to calculate the spectral value of each frequency component cannot analyze different frequency components well. The short-time Fourier transform (STFT) is to add to the original time domain signal. After the window, multiple time-domain signals of shorter and equal length are obtained, and then fast Fourier transform (FFT) or discrete Fourier transform (DFT) is performed on each signal, and finally the results are arranged in time to obtain a short-time spectrum. So STFT is a standard Fourier transform of window selection function. Its basic form is shown in equation (1), where $X_n$ is the coefficient after STFT transformation, $x(m)$ is the original signal, and $w(n - m)$ is the window function. Different STFT results are obtained after different window function sequences are applied.

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)w(n - m)e^{-j\omega m} \quad (1)$$

2.3.2. MFB. Since the frequency range of human hearing is usually 20-20000 Hz, auditory studies have shown that the cochlea has time-frequency analysis characteristics of sound [11]. When sound is transmitted from the outer ear to the middle ear, the movement of the stapes causes changes in the pressure of the fluid in the cochlea, which leads to the propagation of waves along the basement membrane. Since sounds of different frequencies produce different traveling waves, the difference is reflected in the peaks of traveling waves appearing in different positions on the basement membrane, which leads to different frequency responses corresponding to different positions of the basement membrane, that is, frequency selectivity. The Mel scale is an analogy at this point, the non-linearity based on the human hearing response characteristics is introduced, and the calculation formula is shown in equation (2).

$$\text{Mel}(f) = 2595 \times \log \left(1 + \frac{f}{700}\right) \quad (2)$$

When the frequency is small, Mel changes quickly with Hz; when the frequency is large, Mel rises slowly. These characteristics correspond to the human ear's sensitive perception of low-frequency tones and slower perception of high-frequency tones.

2.4. Evaluation

The indicators of our evaluation model are mainly sensitivity, specificity and accuracy. Sensitivity refers to the ratio of correctly identified cough sounds to the total number of coughs in the test group; specificity refers to the ratio of correctly identified non-cough sounds to the total number of non-cough sounds; accuracy is a comprehensive indicator of sensitivity and specificity, which refers to correct identification The ratio of samples (including coughing and non-coughing) to the total number of samples in the test set.

$$\text{SENS} = \frac{TP}{TP + FN}$$
$$\text{SPEC} = \frac{TN}{TN + FP}$$
$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}$$

$TP, FP, TN, FN$ represent the number of true and false positives and true and false negatives respectively.

3. Experiment and Results

3.1. Network training

When training the neural network, we divide the data set into two parts: 80% of which are used to build the model, 20% are used for testing, and then the data used to build the model is further divided into training set and verification at a ratio of 80:20 set[13]. First use the training set to train the neural network,
and then improve the trained model based on the results of the validation set to find the best parameters (such as the learning rate, the number of filters, etc.). After determining the network parameters, the model is rebuilt Training and final evaluation on the results of the test set. We use the stochastic gradient descent method to train the neural network, where the learning rate is 0.01, the batch size is 128, and the momentum size is 0.9. The experiment is run on the MATLAB2020a platform.

3.2. Experiment 1
In order to compare the performance of different input features for cough classification using convolutional neural networks, we used STFT, Mel spectrogram, Log-Mel spectrogram, and 40-dimensional MFCC as the input of the above neural network. In the process of generating the spectrogram, all sound fragments are resampled to 22.5kHz, and the generated picture size is 64x64x3, and then the image is input into the preset classifier, and the result is shown in Table 1 below.

| Feature | Accuracy | Sensitivity | Specificity |
|---------|----------|-------------|-------------|
| STFT    | 90.17%   | 90.00%      | 90.25%      |
| Mel     | 92.67%   | 93.00%      | 92.50%      |
| Log-Mel | 92.17%   | 92.50%      | 92.00%      |
| MFCC    | 89.17%   | 87.50%      | 90.00%      |

3.3. Experiment 2
It is obtained from experiment one that the Mel spectrum achieves the best performance among the four selected features. In experiment one, the window size used for feature extraction is 256 and the frame shift is 128. Because different frame shifts and window sizes will change the time resolution and frequency resolution of the results, the second experiment will use windows with sizes 256, 128, 64 and windows with sizes 128, 64, 32, 16. Frame shift to determine the best input for classification. The experimental results are shown in the table 2.

| Window size | Frame shift | Accuracy | Sensitivity | Specificity |
|-------------|-------------|----------|-------------|-------------|
| 256         | 128         | 92.67%   | 93.00%      | 92.50%      |
| 256         | 64          | 91.50%   | 89.00%      | 92.75%      |
| 128         | 64          | 91.33%   | 91.00%      | 91.50%      |
| 128         | 32          | 90.33%   | 89.50%      | 90.75%      |
| 64          | 32          | 91.00%   | 86.50%      | 93.25%      |
| 64          | 16          | 91.17%   | 87.50%      | 93.00%      |

4. Conclusion
In this study, we mainly compared the impact of different input features on the performance of cough recognition. As shown in Table 1 in the previous section, among the four input features we selected, the Mel spectrum achieved the best accuracy. And it has the best sensitivity, which may be largely due to the enhancement of certain features of cough during Mel conversion, which increases the accuracy of recognition. In experiment two, different window sizes and frame shifts also produced different results. For the data set used in this experiment, using a 256-size window and 128 frame shift has the best performance, but from the experimental results, In a certain interval, different window sizes and frame shifts do not have a particularly large impact on the recognition performance. Therefore, in the subsequent continuous speech recognition cough sound research, as long as a reasonable window size and frame shift size are determined, better results can be achieved. Performance. In follow-up research, we will continue to expand the number of data sets and improve the cough recognition system.
References

[1] S. S. Birring, S. Matos, R. B. Patel, B. Prudon, D. H. Evans, and I. D. Pavord, Cough frequency, cough sensitivity and health status in patients with chronic cough, Respiratory Med., vol. 100, pp. 1105–1109, Jun. 2006.

[2] World Health Organization (WHO), Early detection of tuberculosis: an overview of approaches, guidelines and tools, 2011.

[3] H. A. Bickerman and S. E. Itkin, The effect of a new bronchodilator aerosol on the air flow dynamics of the maximal voluntary cough of patients with bronchial asthma and pulmonary emphysema, Journal of chronic diseases, vol. 8, no. 5, pp. 629–636, 1958.

[4] Birring, T. Fleming, S. Matos, A. Raj, D. Evans, and I. Pavord, The leicester cough monitor: preliminary validation of an automated cough detection system in chronic cough, European Respiratory Journal, vol. 31, no. 5, pp. 1013–1018, 2008.

[5] J. Amoh and K. Odame, Deep neural networks for identifying cough sounds, IEEE transactions on biomedical circuits and systems, vol. 10, no. 5, pp. 1003–1011, 2016.

[6] Diego Augusto, Jose Stuchi, Ricardo Violato, and LusCuozzo, Exploring Convolutional Neural Networks for Voice Activity Detection, pp. 37–47, 07 2017.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,”Adv. Neural Inf. Process. Syst., pp. 1–9, 2012.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[10] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[11] J. M. Gutierrez-Arriola, R. Fraile, A. Camacho, et al. Synthetic sound event detection based on MFCC [J]. Integration, 2016, 1000(1):6

[12] F. Barata, K. Kipfer, M. Weber, P. Tinschert, E. Fleisch, T. Kowatsch. Towards device-agnostic mobile cough detection with convolutional neural networks, in: 2019 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, 2019, pp. 1–11.

[13] Igor D. S. Miranda; Andreas H. Diacon; Thomas R. Niesler. A Comparative Study of Features for Acoustic Cough Detection Using Deep Architectures*. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019.