Speech Emotion Recognition Based on Convolutional Neural Network for Emergency System of Railway Station

Kaibei Peng¹,*, Liuyi Wu¹, Xiaoshu Wang¹ and Tianyun Shi²
¹Institute of Computing Technology, China Academy of Railway Sciences Corporation Limited, Beijing 100190, China
²China Academy of Railway Sciences Corporation Limited, Beijing 100190, China
*Corresponding author email: pengkaibei@126.com

Abstract. This paper proposed a speech emotion recognition model based on Convolutional Neural Network (CNN). The model first extracts the Mel Cepstral Coefficient (MFCC) feature of each speech, and then sends the extracted feature matrix to the convolution the neural network is trained, and finally the category of each voice is output by the network. In addition, a confidence setting is added to the output layer of the model, and it is believed that the probability of each voice belonging to a certain category is greater than 90%. Experimental results show that the model has a higher accuracy rate compared with Recurrent Neural Network (RNN) and Multilayer Perceptron (MLP). This method provides a certain reference for the application of deep learning technology in speech emotion recognition and early warning of dangerous situations in railway stations and other places.

1. Introduction
In recent years, Speech emotion recognition has been applied in many fields, such as smart cars, online calls and medical emergency treatment[1]. Speech emotion recognition mainly extracts the related features of different emotions from people's emotional speech, and then identifies and classifies different emotions according to the extracted features[2]. With the development of artificial intelligence, deep learning is widely used in emotion recognition. Kim[3] used CNN for text emotion classification for the first time, and the experiment showed that CNN was superior to other traditional machine learning algorithms. Zhao Xiaolei[4] et al. input the features extracted from artificial statistics and deep learning into support vector machine for recognition, and this method has a high recognition rate. The above methods only consider the performance of speech recognition in some public data sets, while ignoring the shortcomings of strong randomness and poor regularity of data sets in practical applications. This paper applies emotion recognition technology to railway stations, and provides some help for early warning of dangerous situations by identifying the positive and negative emotions of passengers.

2. Speech Emotion Recognition
2.1. Speech Pre-processing
In this experiment, six kinds of emotion data, anger, fear, happiness, neutrality, sadness and surprise, were divided into two categories. Anger, fear and sadness are negative voice emotional data, while neutrality, happiness and surprise are positive emotions. Therefore, a total of 4797 voices were collected, which were recorded by several people of different genders. In the pre-processing process, firstly, the empty sounds in each speech are carefully removed, so that each speech only contains...
useful features, but not useless features; then, the volume of each speech is unified, so that the volume of each speech is within the same range; finally, the quality of each speech is improved, so that each speech in the data set only contains high-quality speech features.

2.2. Speech Feature Extraction
This paper uses mel-frequency cepstral coefficients (MFCC) to extract speech features. After mel-frequency cepstral coefficients feature extraction, 259-dimensional features are extracted from each speech in the data set. The default values in the extracted feature matrix are supplemented to keep the integrity of the feature matrix. Then, the order of the feature matrix is disordered to reduce the sequential correlation between data and enhance the training effect of the model. Finally, the extracted and processed feature matrix is divided into training data sets according to the ratio of 8: 1: 1, which, mainly used for model effect testing and testing data sets, and used for model result testing.

2.3. Training
The task dimension classifies tasks, and the distribution of speech features has a clear hierarchical relationship. Therefore, this paper uses convolutional neural network to train the model[5]. The network structure developed in this paper is shown in Figure 1.

![Figure 1. Convolutional neural network structure.](image)

The convolution operation represents the hierarchical extraction of speech features, while the maximum pooling operation removes redundant information in the previous layer features, and the operation is simplified. An activation layer is set after each convolution layer, and the best activation function determined by experiments is Relu. Two fully connected neurons are set in the output layer to divide the speech signals into two categories. In addition, considering the complexity of the network structure, random Dropout is set after each hidden layer to prevent the network from over-fitting in the training process[6]. After adding Dropout, the neurons in each hidden layer will have a certain probability of not updating the weights in the training process, and the probability of each neuron not updating the weights is equal.

3. Simulation Results

3.1. Accuracy of Training Set and Verification Set
In this experiment, convolutional neural network (CNN) is used to train the speech data in the training set. Fig. 2 shows the accuracy of CNN model RNN model and MLP model training set and verification set. It can be seen from Figure 2 that with the increase of iteration times, the accuracy of training set and verification set tends to increase, especially the training set. Finally, after 200 iterations, the accuracy of the training set is over 97%, and that of the verification set is over 80%. CNN model developed in this paper has high accuracy than RNN and MLP in both training set and verification set.
3.2. Test Set Accuracy
In order to further verify the generalization of the developed model, a total of 214 voice signals recorded by different people were used to test the generalization of the model. It can be seen from Table 1 that there are 144 voices representing positive emotions in the classification of voices representing positive emotions, and 133 voices representing positive emotions in the classification results of CNN model proposed in this experiment, 80 voices in RNN model and 72 voices in MLP model. Among the classification results of CNN model proposed in this experiment, there are 66 voices representing positive emotions, 39 voices of RNN model and 36 voices of MLP model. It is proved that the CNN model proposed in this experiment has high accuracy in the classification of positive emotions, and it is higher than RNN and MLP.

Table 1. Classification results of three models in emotional speech.

| Algorithm | Positive | Negative | All |
|-----------|----------|----------|-----|
|           | Quantity | accuracy | Quantity | accuracy | Quantity | accuracy |
| CNN       | 133      | 92.36%   | 66       | 94.29%   | 199      | 92.99%   |
| RNN       | 102      | 70.83%   | 51       | 72.86%   | 153      | 71.50%   |
| MLP       | 99       | 68.75%   | 50       | 71.43%   | 149      | 69.63%   |

4. Conclusion
In order to reduce the harm caused by danger in crowded places such as railway stations, this paper applies speech emotion classification and recognition technology to the emergency treatment of railway stations, and proposes a CNN emotion classification model, which first extracts MFCC features of speech, then sends the extracted features to convolutional neural network, and finally outputs the category of each speech. Simulation results show that the accuracy of this model is higher than RNN and MLP models. The method proposed in this paper provides a certain reference for the danger warning of railway stations.

Acknowledgements
This work is supported by Major project of China Academy of Railway Sciences Group Co., Ltd. (2018YJ102).

References
[1] Langari S, Marvi H, Zahedi M. Efficient Speech Emotion Recognition Using Modified Feature Extraction[J]. Informatics in Medicine Unlocked, 2020. 20: 2352-9148.
[2] Wang G Y, Zhang G Y. Intelligent recognition of Chinese speech emotion information based on SVM multi classification algorithm[J]. Electronic Component and Information Technology, 2020, 4(07): 62-63+99.
[3] Kim Y. Convolutional Neural Networks for Sentence Classification[C]. Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1746–1751.
[4] Zhao X L, Xu X F. Speech emotion recognition based on shallow learning and deep learning model [J]. Computer Applications and Software, 2020, 37(12): 108-112+176.
[5] Seo M, Kim M. Fusing Visual Attention CNN and Bag of Visual Words for Cross-Corpus Speech Emotion Recognition[J]. Journal of Intelligent Systems, 2020. 20(19): 5559-5559.

[6] Zhang G X, Ma A Q. Prediction method of arrester resistive current based on random deactivation LSTM network[J]. Hydropower energy science, 2020, 38(10): 159-162+98.