Speech Emotion Recognition Based on Feature Fusion

Qi Shen [1, a], Guanggen Chen [2, b] and Lin Chang [2, c]

[1]School of Information Technology, Beijing University of Technology, Beijing 100124, China

[2]School of Software, Beijing University of Technology, Beijing 100124, China

[a]shenq@bjut.edu.cn, [b]stick360@163.com, [c]changlin@163.com

Abstract: Speech emotion recognition is mainly based on the differences of characteristics between different emotions. The traditional recognition method is based on the manual extracted features, such as MFCC and LPCC, etc., and also achieved well. But it is unclear what kind of feature are able to reflect the characteristics of human emotion from speech. With Convolution Neural Network (CNN) shows strong ability in the field of image classification, attracting more researchers to apply CNN to the learning of the spectrogram feature. However, the study of speech emotion either according to the characteristics of the traditional manual extraction or completely dependent on spectrogram of speech. There is still no combination of traditional features and spectrogram feature. In this paper, we propose a fusion neural network model combining the characteristics of traditional with spectrogram features. This multimodal CNN is trained with two stages. First, two CNN models pre-trained are fine-tuning respectively on the corresponding labeled audio datasets. Second, the outputs of the two CNN models are connected to a fusion network of fully-connected layers. The fusion network is trained to obtain a joint feature representation for emotion recognition. From the recognition results of emotional speech database, the proposed algorithm has higher speech emotion recognition rate and robustness.

Keywords: speech emotion recognition, convolution neural network, feature fusion.

1. Introduction

Speech is one of the important ways for human's communication. The speech signal contains not only the expressed speech meaning, but also the speaker's emotion information which always be ignored by the traditionally speech processing. Human-computer interaction as a research hotspot in the field of artificial intelligence has always been highly concerned by researchers. Although speech recognition has achieved the desired effect in human-computer interaction, we are still far from being able to naturally interact with machines, partly because machines difficulty understand our emotion states. Therefore speech emotion recognition [1] is an important subject for human-computer interaction.

The problems that need to be solved in human-computer interaction are the same as the important factors in the communication between between people and people. For speech emotion recognition, the most critical research lies in the emotional characterization method and the environmental applicability of the algorithm. The commonly used features of speech emotion recognition can be roughly summarized as prosodic features, based on spectral related features and sound quality characteristics, such as energy-related features, pitch frequency features, formant frequency, Zero Crossing Rate (ZCR), Mel-Frequency Cepstrum Coefficients (MFCC), Linear Prediction Cepstrum Coefficients (LPCC) and Linear Prediction Coefficients (LPC),etc. Based on these characteristics, speech emotion recognition research has made great progress. Many traditional pattern recognition methods such as Support Vector Machine (SVM), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Artificial Neural Network (ANN) is widely used in speech emotion recognition.

In recent years, with the deep learning makes great success in the field of voice, images. Many researchers apply the deep learning to the emotion recognition of speech. The correlation between the frequency and time domain of the signal plays an important role in speech emotion recognition. But the study of correlation between speech signals often concentrated in only frequency domain or time domain. Spectrograms as a voice of the time-frequency energy distribution of visual expression, the horizontal axis represents time, the vertical axis represents the frequency, the time-frequency united on a 2D image. Convolutional Neural Network [2] has a strong feature learning ability and classification ability, and used in the field of image recognition and classification widely. Therefore, researchers mainly apply CNN to the study of the spectrogram to make a recognition of emotion recognition, and make a variety of optimization from this. However, for the characteristics of the spectrogram, it is entirely dependent on the CNN learning ability and transparent to us.

Motivated by some researchers employ deep learning to a joint model for emotion recognition, including face, speech, and text modalities. However, they are all on a combination of different signals, there are still not a joint research on the different aspects of the same signal. Therefore we propose a fusion network model for emotion recognition of speech. Fig.1 shows the structure of the model. There are two steps in our multimodal training: (1)Generating a spectrogram and extracting the traditional characteristics from the voice signal at the same time, then pre-training the two individual CNN models respectively. (2)Integrated the result of trained model in a fusion network, with fine-tuning to perform speech emotion recognition tasks.

The rest of the paper is organized as follows. Section 2 presents our learning model in details. Sections 3 describes CASIA datasets and reports our experimental results. Conclusion and future work are discussed in Section4.

2. The Multi-CNN Architecture

Fig.1 shows the multi-CNN Architecture model, which is composed of two sets of CNN models, which are CNN networks with spectrogram as the input and CNN network with traditional features as the input. Then CNN models is fully connected to the fusion network (FN), and the softmax classifier is applied at the end.

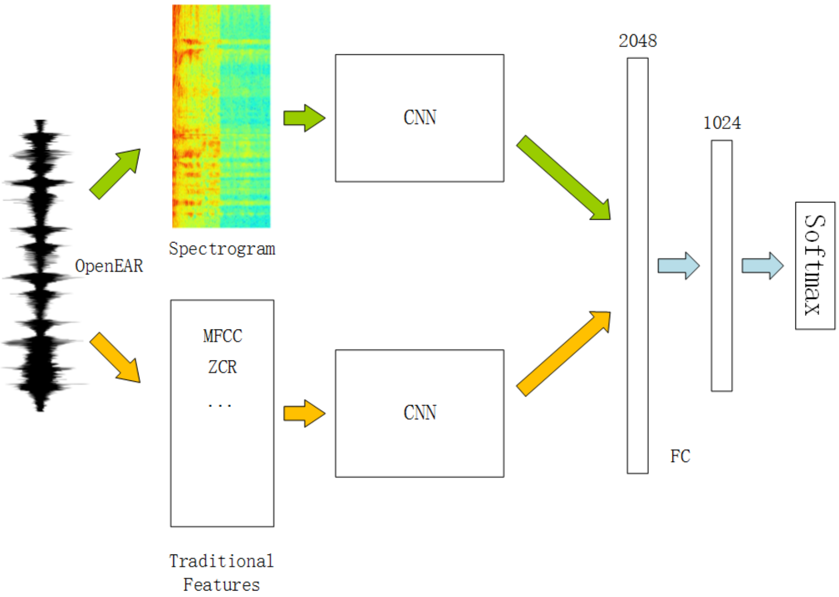

Fig. 1 The Multi-CNN Architecture

2.1 CNN model

Restricted by experimental equipment performance, we use the classical convolution neural network LeNet-5 [3] as the prototype. The LeNet-5 network model contains 3 convolution layers, 2 pooling layers and 1 full-connected layer. By modifying the model structure, the modified network model as our CNN model, the detailed description of the network structure parameters as shown in the Table 1.

For speech data, first need to be preprocessed to generate the appropriate input of the model. In detail, using 25ms Hamming window with a 10ms overlapping to deal with each speech data sampled at 16 kHz. For the spectrogram, the log Mel-spectrograms consist of 64 short-time Fourier transformation bins for each frame and the size of spectrogram segment level feature is set to 64 frames. Then we can obtain segments with size 64×64.

Table 1 CNN model structure parameters

Layer	Type	Channels	Width	Scheme 4	Kernel Size
0	Input	1	64	64	
1	Conv	6	60	60	5×5
2	Max Pool	6	30	30	2×2
3	Conv	16	26	26	5×5
4	Max Pool	16	13	13	2×2
5	Conv	32	9	9	5×5
6	FC	2048	1	1	
7	FC	1024	1	1	

For the traditional features, considering that the 384-dimensional basic feature set provided by INTERSPEECH 2009EC [4] has chosen the most widely used features and functions in prosodic features, sound quality characteristics and spectralcharacteristics, including 16 Low-Level Descriptors (LLDs) and 12 statistical functions as shown in the Table 2. And through the openEAR [5] open source toolbox to extract these 384 dimension features. Then we can obtain segments with size 384×64. In order to maintain the consistency of the two CNN models, then we resize the traditional features into a new size 64×64 with bilinear interpolation.

Table 2 INTERSPEECH 2009EC basic feature sets

LLDs (16×2)	Statistical Functions (12)
ZCR	arithmetic mean
RMS energy	standard deviation
HNR	maximum、minimum、range、relative position of max,min
MFCC (1-12)	Linear regression: offset、slope、error

2.2 FN model

Because both the spectrogram and the traditional features, it will lose the emotional related characteristics which haven't yet been cognitive. Therefore, we put forward a model to combine the two characteristics in depth for improving the speech emotion recognition rate. As shown in the Fig.1, each group of CNN models output 1024 dimension feature vectors, then we use fusion network to combine these two group vectors. And further enhance the generalization ability of the model with dropout. Finally through the softmax classifier to output the emotion result.

3. Evaluation

In our experiments the database we use is Chinese emotional database, which released by the Chinese Academy of Sciences Institute of Automation (CASIA) [6]. The corpus which contains 1200 sentences in total is recorded by four actors (two females and two males) with 6 kinds of emotional states: fear, happy, sad, angry, surprise and neutral. And each one read 50 different texts respectively with these 6 emotions.

Our multi-CNN model trains with a mini-batch size of 50 and using stochastic gradient descent (SGD) algorithm to update parameters. The learning rate is set to 0.001, the epochs are set to 300, and fine-tuning respectively. We first train the two group CNNs individually, including computer the loss and update the network parameters to minimize the loss with BP algorithm [7]. Then connect the output of the two groups of CNN to the fusion network. And optimizing the fusion network parameters with BP algorithm. Additionally, adopting the dropout [8] method to reduce over-fitting.

The experiment adopts 3 times cross validation method, which randomly disrupts all the data in the CASIA database and divides into 5 copies, then take turns to choose one for the testing data, and others for training data. Finally, take the average accuracy for the recognition rate. For comparison, we still experiments only with spectrogram or traditional features based on the above CNN model.

The final results are shown in Table 3. We use T to represent traditional features and use S to represent spectrogram.

Table 3 Comparison of speech emotion recognition results

Emotion	Fear	Happy	Sad	Angry	Surprise	Neutral	Average
Only T (%)	77	81	80	77	81	81	79.5
Only S (%)	76	78	82	76	80	78	78.3
T and S (%)	78	82	84	83	84	82	82.2

Table 3 shows that the combination of traditional features and spectrogram in speech emotion recognition performs better than traditional features or spectrogram separately. This clearly shows that our multi-CNN Fusion Network model can learn more efficient emotion features representation, which are more comprehensive.

4. Conclusions and Future Work

In this paper we present a new method for speech emotion recognition with multi-CNN Fusion Network. Experimental result on the CASIA database shows that our proposed model has a good feature representation ability with the combination of traditional features and spectrogram, which performs better than single feature sources in speech emotion recognition tasks. In the future work, we will combine lexical features to further improve the performance of our proposed model. And further improve the model to better adapt to the speech emotion recognition.

5. References

[1] M. El Ayadi, M.S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011.

[2] G. E. Hinton and R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science, vol. 313, pp. 504-507, 2006.

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, november 1998.

[4] B. Schuller, S. Steidl, and A. Batliner, The INTERSPEECH 2009 Emotion Challenge, in Proc. Interspeech, Brighton, UK, 2009, pp. 312–315.

[5] F. Eyben, M. Wollmer, and B. Schuller, openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit, in Proc. ACII, Amsterdam, Netherlands, 2009, pp. 576–581.

[6] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li. Speech emotion recognition using fourier parameters. IEEE TAC, 6(1):69–75, 2015.

[7] Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. Nature 323, 533–536 (1986).

[8] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929–1958, 2014.