Speech Emotion Recognition Using Deep Convolutional Neural Network and Simple Recurrent Unit

Pengxu Jiang, *Member, IAENG*, Hongliang Fu, Huawei Tao, *Member, IAENG*

***Abstract*—Speech emotion recognition is a frontier topic in human-machine interaction. To improve the accuracy of intelligent speech emotion recognition system, a speech emotion recognition method based on Deep Convolutional Neural Network and Simple Recurrent Unit is proposed. Firstly, log Mel-spectrograms are extracted from acoustic features set with static, delta, and delta-delta. The three channels of log Mel-spectrograms of each utterance are divided into several segments on the time axis as the DCNN input. Then the AlexNet pre-trained on the ImageNet dataset is employed to learn these features on each segment for fine-tuning. A Simple Recurrent Unit model aggregates these learned segment-level features. Finally, a SoftMax classifier is used to identify the types of speech emotion. The experimental results on the EMO-DB and CASIA database show that our model can effectively recognize the emotions contained in speech and performed better than the classifiers based individually on another kind of features.**

***Index Terms*—Convolutional Neural Network, Speech Emotion Recognition, Simple Recurrent Unit**

I. Introduction

Language, as one of the carriers of emotion, contains a wealth of emotional information. In the past decades, the related research of speech emotion recognition has made significant progress, and speech emotion recognition has broad prospects in many different research fields [1]. With the maturity of computer speech recognition technology and the continuous emergence of related research, speech emotion recognition begins to be more applied to the education, entertainment, and communications industries. Strengthening the recognition of speech emotions has become the focus of next-generation artificial intelligence development [2]. Given this, the research on speech emotion recognition has significant theoretical value and practical significance.

Feature extraction is the first and most crucial step in speech signal processing. So far, various features have been used by speech recognition researchers. The acoustic features widely used in emotion recognition can be roughly divided into three categories: prosody features, spectral features, and voice quality features. Traditional emotional features are mostly the fusion of these acoustic features — for example, the well-known International Speech Emotional Challenge Feature Set [3], [4]. And spectral features have more parameters to capture the instantaneous change of mood; therefore, these features are the most widely used in speech recognition [5], [6]. Although these features contribute a lot to speech emotion recognition, these hand-designed low-level features still do not distinguish subjective emotions very well.

In recent years, with the great success of deep learning in images, it has received extensive attention from domestic and foreign experts. In 2006, Hiton et al. [7] proposed using hierarchical abstraction instead of manual feature selection, automatic feature learning has been realized, and the difference of artificial feature selection has been eliminated. Deep learning provides a new way for us to acquire high-level features from shallow features, there are some recent studies on feature learning of speech signals based on convolutional neural networks (CNN) [8], [9]. However, the above work uses 1-D convolution to learn global features; 1-D convolution may learn fewer emotional details. And speech signals may have different durations, but most CNN models only accept the data with a fixed size, change the speech signal of varying durations to the same length may lose the temporal information of speech waveform.

To solve the above problems, a speech emotion recognition method based on deep convolutional neural network and simple recurrent unit (SRU) [10] is proposed. The framework of our model is shown in Fig 1. Firstly, 3-D log Mel-spectrograms (static, delta, and delta-delta) are extracted from acoustic features set as the DCNN model input to get more emotional details. Because speech signals may have different durations, to reduce the loss of emotional details in convolution, we divide global features into segment-level features of the same size. Then a DCNN model is employed to learn high-level features from segment-level features, in the process of learning, we use 2-D convolution. Compared with 1-D convolution, 2-D convolution contains more parameters to capture more detailed time-frequency correlations. To generate a global utterance-level feature representation based on the segment-level features, we employ a SRU model to integrate these high-level features. Because these segment-level features are time-dependent, as a variant of RNN, SRU model has a good integration effect on time-dependent sequences. Finally, a SoftMax classifier is used to classify emotions. Two public speech emotion databases will be used in this experiment.

This research project was founded in part by Natural Science Project of Henan Education Department (No. 19A510009), start-up Fund for High-level Talents of Henan University of Technology (No 31401148).

Pengxu Jiang is with the College of Information Science and Engineering, Henan University of Technology, China, e-mail: px20115c@163.com.

Hongliang Fu is with the College of Information Science and Engineering, Henan University of Technology, China, e-mail: jackfu_zz@163.com.

Huawei Tao School of Information Science and Engineering, Henan University of Technology, China, e-mail: thw@haut.edu.cn.

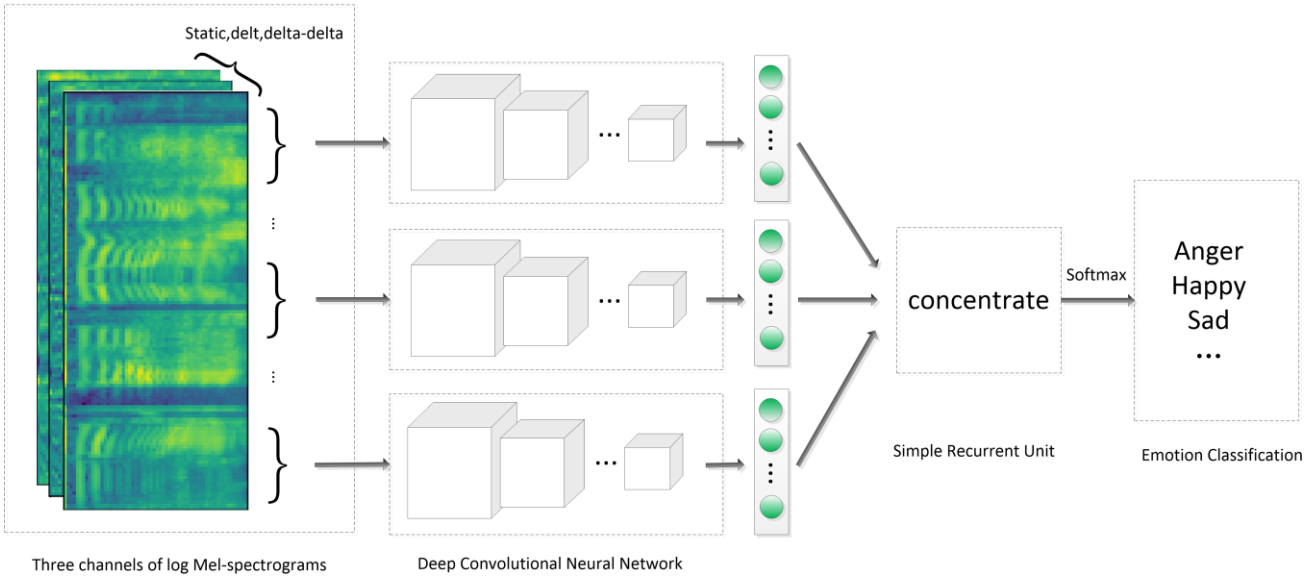

Fig. 1. The framework flow chart of speech emotion recognition model based on CNN and SRU.

In summary, the main contributions of this work to speech emotion recognition are three-fold:

- We extract the 3-D log Mel-spectrogram from each utterance and divide them into the same size to reduce the loss of emotional details in convolution.
- CNN with 2-D convolution is used to learn more detailed time-frequency correlation and integrate the learned high-level features with SRU model.
- Experimental results show that our method of speech signal segmentation with different durations has better recognition effect than other methods.

II. Proposed Methodology

A. Feature Extraction

Spectrogram feature is a popular feature in speech recognition nowadays. As a visual expression of time-frequency distribution of speech energy, the spectrum contains more parameters and stronger correlation. Considering the frequency axis and time axis, we can extract more emotional information. Therefore, we obtain log Mel-spectrograms from acoustic features set with static, delta, and delta-delta.

In recent years, with the great success of deep learning in images, as a model of supervised learning in deep learning, convolutional neural networks have made breakthroughs in many applications [11]. AlexNet [12] was designed by 2012 ImageNet competition winner Hinton and his student Alex. AlexNet is a feed-forward neural network model. Optimize the network structure by receptive field, sharing weight and pooling. We use 2-D convolution in AlexNet to learn high-level features from segment-level features. 2-D convolution can get more time-frequency correlation from log Mel-spectrograms and enhance the learning effect of the model.

B. Relevant Emotional Processing Model

For Emotion Classifier, the commonly used speech emotional classifier models include Multilayer Perceptron (MLP) [13], Support Vector Machine (SVM) [14], K-Nearest Neighbor (KNN) [15] classification algorithm, etc. Although these classifiers have contributed a lot to speech emotion recognition, the above classifiers still cannot distinguish different types of emotions better.

The recurrent neural network is the most the essential network model for dealing with natural language tasks and the preferred network for temporal data. Standard RNN models include Long Short-Term Memory (LSTM) [16] and Gated Recurrent Unit (GRU) [17]. As a variant of RNN, SRU model trains faster than other variant models and is also good at time-dependent processing sequences. So we employ an SRU model to generate global utterance-level features representation based on the segment-level features learned by CNN.

C. Our CNN-SRU model

The framework of our model is shown in Fig 1. The proposed model can not only make full use of the advantages of the CNN network in processing image features but also fully exploit the benefits of SRU processing time-related sequences.

First, we preprocess the speech emotion data. Mel-spectrograms are extracted from acoustic features set with static, delta and delta-delta. To reduce the loss of emotional details in convolution, each utterance X is divided into N ($1 \le N \le 5$) segments $\mathbf{x_N}$ on the time axis as the CNN input:

$$X = (x_1, x_2 \dots x_N) \in R^{d \times N} \quad (1)$$

with feature dimensionality d = 4096, Each emotional fragment corresponds to an emotional label. These segment-level features will be processed in the first step of our model.

Since Alexnet is pre-trained on ImageNet, the model has certain requirements for the size of the input. We need to use bilinear interpolation to resize the speech emotion segment into 227 × 227 × 3. The main structure of our CNN model is shown in Figure 2. There are eight learning layers in the model: five convolution layers and three fully connected layers. The convolution layer is equivalent to the filter. The convolution method of local connection is used between layers; convolution layer uses 2-D convolution to learn the time-frequency correlations of features better. Each neuron no longer connects with all the neurons in the upper layer, but

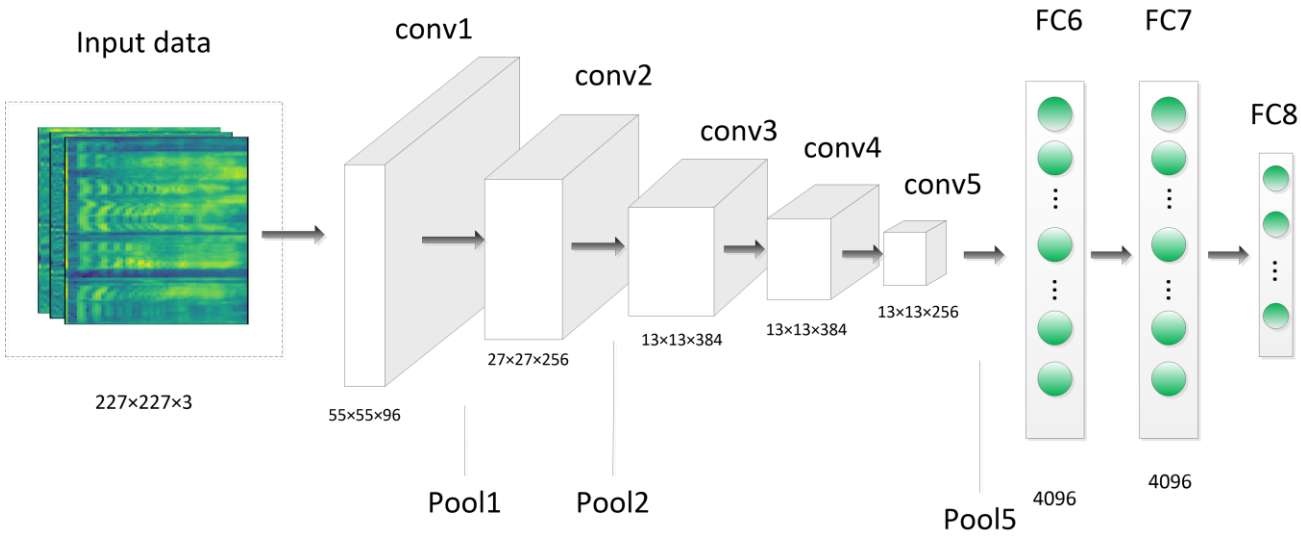

Fig. 2. A framework flow chart of DCNN model.

only with a few neurons, which significantly reduces the parameters. In the process of convolution, weight sharing is adopted. Each group of connections shares a weight, instead of having different weights for each connection, which reduces many parameters. The number of FC-8 in the last layer of AlexNet is 1000, but the number of emotions that our model outputs are 6 or 7, so we need to change the structure of FC-8 in the last layer. Finally, we use the 4096-D features in the FC7 layer as the high-level feature of each segment-level feature. For every utterance X:

$$f(X) = (f(x_1), f(x_2) \ldots f(x_N)) \quad (2)$$

$f(\cdot)$ represents an operation of extracting high-level features from shallow segment-level features using DCNN model, $f(x_t)$ is a 4096-dimensional high-level feature.

Then, we need to integrate these processed high-level features. Segment-level features of each utterance are trained with a label. We use SRU models to incorporate these

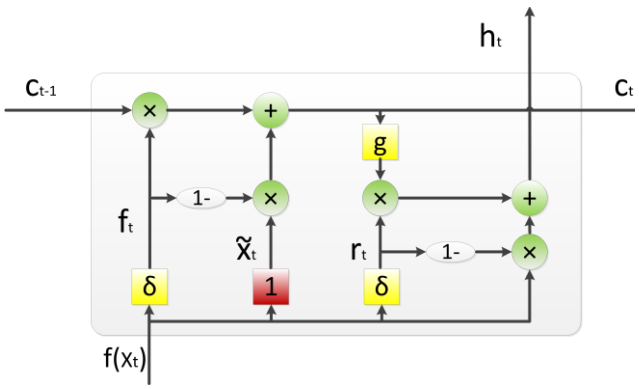

Fig. 3. The structure of Simple Recurrent Unit for Speech emotion recognition,

high-level features. The SRU model we used to integrate segment-level features is shown in Figure 3.

For each 4096-dimensional feature $f(x_t)$ ($1 \le t \le 5$), we first compute the output state c_t of the first 4096 -dimensional feature.

$$\tilde{x}_t = W f(x_t) \quad (3)$$

$$f_t = \delta(W_f f(x_t) + b_f) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + (1 - f_t) \odot \tilde{x}_t \quad (5)$$

W and b represent weight and bias respectively, and the output h_t is calculated once for each segment level feature.

$$r_t = \delta(W_r f(x_t) + b_r) \quad (6)$$

$$h_t = r_t \odot g(c_t) + (1 - r_t) \odot f(x_t) \quad (7)$$

δ and g represent the activation functions Sigmoid and tanh, respectively. Because these segment-level features of each

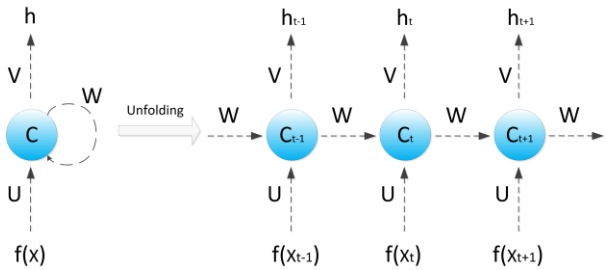

Fig. 4. A recurrent neural network in the forward expansion calculation in time, W, V, and U represents weight.

utterance are time-dependent, we integrate each segment -level feature as shown in Figure 4.

Finally, a SoftMax classifier is employed to classify emotions. For each utterance output feature h in our model, we use SoftMax classifier to normalize output feature.

$$f(h_i) = \frac{e^{h_i}}{\sum_j e^{h_j}} \quad (8)$$

We can calculate the proportion of each h_i, the sum of all h_i is 1. Compare the output label with the correct label and calculate the loss between them. We use the training set to update the parameters of the model. The test set is used to test the accuracy of our proposed speech emotion model.

III. Experiments

A. Simulation Database

We test the proposed method on Berlin EMO-DB [18] German Emotional Voice Library and CASIA [19] Chinese Emotion Prediction database.

The EMO-DB database is a German-language speech emotion database recorded by the University of Berlin. The database consists of 535 utterances that displayed by ten professional actors (five males and five females) with seven different emotions.

The CASIA database was recorded for the Institute of Automation of the Chinese Academy of Sciences. The database consists of 1200 utterances that displayed by four professional actors (two males and two females) with six different emotions.

B. Evaluation Methods

The experiment used the strategy of Leave One Speaker Out (LOSO) as suggested in [3], each time an emotional speech sample of one person is selected from the data set as a test set of experiments, the remaining emotional speech samples are used as a training set. Each person takes turns as a test set. Finally, the average of several trials is calculated as a result.

The evaluation criteria are Weighted Accuracy (WA) and Unweighted Accuracy (UA) as suggested in [20].

C. Selection of Network

For every utterance, we use 60 Mel-filter banks to extract the whole log Mel-spectrogram with 25ms Hamming window and 10ms overlapping. Then we can compute the log Mel-spectrogram with delta and delta-delta. After we obtain three channels of log Mel-spectrogram, we divide it into N ($1 \leqslant N \leqslant 5$) speech emotion segments. Then, resize the speech emotion segment into $227 \times 227 \times 3$ by bilinear interpolation as the input of our model. Because in CASIA and EMO-DB data set, the shortest speech duration about 955 ms, if each speech segment is divided into six parts, the shortest duration of the divided segments about 159 ms. The divided segments with 160 ms used for speech recognition. So it's reasonable for us to split each utterance into five parts.

Fine-tuning the emotional speech feature on DCNN, the AlexNet pre-trained on the ImageNet dataset is employed. We adopt Gradient Descent Optimizer with a learning rate of 0.001, the dropout rate is set as 0.5, and the maximum number of epochs is set as 400 with a batch size of 30.

We used AdamOptimizer with a learning rate of 0.001 to train in SRU model; the weight and bias are set as 0.01 and 0.1, respectively. The maximum number of epochs is set as 100 with a batch size of 30. The hidden size is set as 4096. Then use the SoftMax classifier for classification.

D. Experimental Results and Analysis

This paper uses Python software based on TensorFlow platform with GPU mode for the experiment. To evaluate the performance of our model proposed in this paper, the following experiment is carried out on the proposed network model:

1) The non-segmented feature is used as the input of our model on the EMO-DB and CASIA database.
2) Divide each feature into two to five segments on the time axis as the input of the model on the EMO-DB and CASIA database.

The experimental results are compared with the existing literature to verify the feasibility of our model.

We first experiment on our model with non-segmented speech emotion data. The test results are shown in TABLE I.

TABLE I
Recognition Rate (%) of non-segmented speech emotion data

Database	WA	UA
EMO-DB	78.9	76.1
CASIA	41.0	41.0

It can be seen from the experimental results that our model used in this experiment can recognize the EMO-DB data set and CASIA data set, but the accuracy is not high. This is probably because the proposed model has no visible recognition effect when dealing with non-segmented speech emotion data.

TABLE II
Recognition Rate (%) Effect of Our Model
(N represents the number of segments)

Database	N	WA	UA
EMO-DB	2	78.2	76.3
	3	**82.5**	**80.6**
	4	77.5	77.5
	5	76.8	73.3
CASIA	2	41.0	41.0
	3	45.5	45.3
	4	45.3	45.2
	5	**51.3**	**51.3**

Then we experimented on our model with the segmented speech segments. The test results are shown in TABLE II; TABLE III is a comparison with existing recognition feature sets.

As shown in TABLE II, we observe that our model achieved the best performance on EMO-DB and CASIA database when each utterance was divided into three and five segments. The result shows that the best experimental effect of the model heavily depends on the size and type of input data; different data sets should adopt different data processing methods in speech emotion recognition. This is of considerable significance to the training of speech emotion recognition model with different emotional data sets.

In our model, a cascade connection between CNN and SRU models, In order to verify the superiority of this connection method, we compare the experiment with a single model, that is, with a single model of CNN or GRU, Using CNN or SRU to extract high-level features directly from low-level features, and then classify emotions. The experimental results are shown in TABLE III.

TABLE III
Recognition Rate (%) of Different model

Database	Model	WA	UA
EMO-DB	CNN	72.1	70.1
	SRU	64.2	61.5
	CNN+SRU	82.5	80.6
CASIA	CNN	33.1	33.1
	SRU	42.4	42.4
	CNN+SRU	51.3	51.3

It can be seen from the comparative experiment of two databases. Our cascade model also has distinct advantages over the two single models; recognition rate has been greatly improved. The recognition rate in the two databases has been improved by 10.4% and 8.9% respectively. We can also see from the table. Without segmenting the original speech, the recognition effect of our cascade model is still higher than that of a single model. This may be because our cascade model has more parameters to mine deep emotional information than other models; this is also of considerable significance to the selection of speech emotion recognition model.

A significance test was carried out on the experimental results to investigate the improvement of the recognition effect of different models on two databases. The results of T-test are shown in TABLE IV; it uses t-distribution theory to infer the probability (P-Value) of difference occurrence to judge whether there is a significant difference between two groups of data. When P-Value is less than 0.05, the difference between data is substantial. As we can see from TABLE IV, all P-Values are less than 0.05 on two databases. Therefore, compared with CNN and SRU model, the performance of our CNN+SRU has a significant improvement.

TABLE IV
T-test on test results

Database	Models	P-Value
EMO-DB	(CNN+SRU,CNN)	< 0.0001
	(CNN+SRU,SRU)	< 0.0001
CASIA	(CNN+SRU,CNN)	< 0.0001
	(CNN+SRU,SRU)	< 0.0001

To verify the performance of our method, we selected several features for comparison, including PLP feature, MFCC feature, and HuWSF feature. Besides, The 2009 Emotional Challenge Feature Set (IS09) [3] and The INTERSPEECH 2010 Paralinguistic Challenge Set (IS10) [4] are selected as the comparative features. Among the above features, as shown in [21], the mean, standard deviation, maximum, minimum, kurtosis, skewness, range and median of the first and second-order difference of PLP, MFCC and HuWSF are calculated, which constitute the features of PLP, MFCC, and LPCC. IS09 contains 16 Low-Level Descriptor (LLD) and calculates 12 statistical values of 16 LLDs and their first-order variances, a total of 384-dimensional features. IS10 contains 1582-dimensional features which result from a base of 34 LLDs with 34 corresponding delta coefficients appended, and 21 functionals applied to each of these 68 LLD contours (1428 features). Also, 19 functionals are applied to the 4 pitch-based LLD and their four delta coecient contours (152 features). Finally, the number of pitch onsets (pseudo syllables) and the total duration of the input are appended (2 features). These features are classified using SVM classifier.

TABLE V
Recognition Rate (%) of Different features

Database	Feature	WA	UA
EMO-DB	PLP	73.3	/
	MFCC	57.6	/
	HuWSF	74.1	/
	IS09	75.5	74.0
	IS10	78.1	76.7
	Ours	**82.5**	**80.6**
CASIA	PLP	45.0	/
	MFCC	36.1	/
	HuWSF	42.5	/
	IS09	37.7	37.7
	IS10	47.6	47.6
	Ours	**51.3**	**51.3**

TABLE V shows a comparison between our proposed method and other methods. First, compared with PLP, MFCC, and HuWSF features, the WA values of our method are increased by 9.2%, 24.9%, and 8.4% respectively on the EMO-DB database. The WA values of our approach are increased by 6.3%, 15.2%, and 8.8% respectively on the CASIA database. Effects of the experiment are conspicuous. Emotion challenge feature are the most representative speech emotion recognition features at present. Compared with IS09 and IS10 features, the experimental effect of our method has also been significantly improved, the WA values of our method are increased by 7% and 4.4% respectively, and the UA values are increased by 6.6% and 3.9% respectively on the EMO-DB database, the accuracy of our approach are increased by 13.6% and 3.7% respectively on the CASIA database. The above experimental results further show that the deep features extracted by our model have excellent recognition performance.

To further investigate the recognition effect of our model, we present the confusion matrix, as shown in Fig. 5, and Fig. 6, we observe that on both EMO-DB and CASIA datasets, ‘anger’ obtains the highest recognition rate. On the EMO-DB database, ‘anger’ and ‘sadness’ are classified with accuracies more senior than 90%, the classification accuracy of ‘boredom’ and ‘happiness’ is less than 70%, ‘fear’ and ‘neutral’ are distinguished with accuracies higher than 80%, ‘disgust’ is identified with accuracies of 74%. On the CASIA database, ‘anger’ and ‘neutral’ have the highest classification accuracy, the classification accuracy of ‘happy’ is only 27%, and the other three emotions can be recognized with accuracies of 48%, 59%, 43%, respectively.

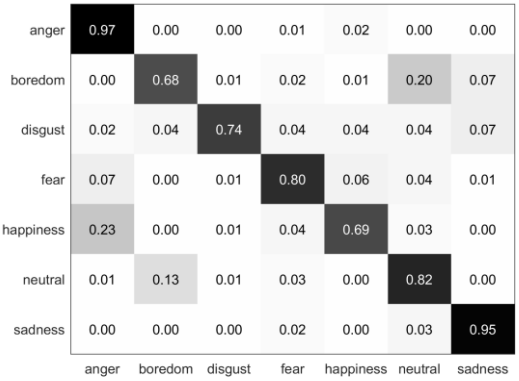

Fig. 5. Confusion matrix on the EMO-DB database (N=3).

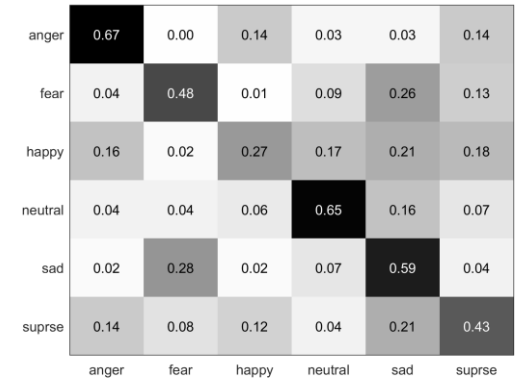

Fig. 6. Confusion matrix on the CASIA database (N=5).

IV. Conclusion

To further improve the recognition ability of speech emotion recognition system, this paper presents a speech emotion recognition network model based on convolutional neural network and simple recurrent unit. We have studied from the perspective of in-depth emotional features extraction in the speech by using the built model. We first extract three channels of log Mel-spectrograms from the speech emotion data set as input, to reduce the loss of emotional information in the learning of convolutional neural networks, segment-level features are used as input, and the effects of different segment sizes on the experiment were also discussed. Then we use a Simple Recurrent Unit to integrate segment-level features, which are time-dependent. Tests on the CASIA and EMO-DB databases show the superiority of our proposed method with some previous works.

REFERENCES

[1] Mencattini A, Martinelli E, Ringeval F, et al. Continuous Estimation of Emotions in Speech by Dynamic Cooperative Speaker Models[J]. IEEE Transactions on Affective Computing, 2016, PP (99):1-1.

[2] Song P, Zheng W, Ou S, et al. Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization[J]. Speech Communication, 2016, 83(C):34-41.

[3] Schuller, Steidl, Batliner. The Interspeech 2009 Emotion Challenge [J]. Interspeech, 2009:312--315.

[4] Christian Müller. The INTERSPEECH 2010 Paralinguistic Challenge [J]. Proc Interspeech, 2010:2794-2797.

[5] Mohamed A R , Sainath T N , Dahl G E , et al. Deep Belief Networks using discriminative features for phone recognition[J]. 2011.

[6] Deng L, Yu D, Platt J. Scalable stacking and learning for building deep architectures[C]// IEEE International Conference on Acoustics. IEEE, 2012.

[7] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. [J]. Science, 2006, 313(5786):504-507.

[8] Abdel-Hamid O , Mohamed A R , Jiang H , et al. Convolutional Neural Networks for Speech Recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(10):1533-1545.

[9] Mao Q, Dong M, Huang Z, et al. Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks [J]. IEEE Transactions on Multimedia, 2014, 16(8):2203-2213.

[10] Lei T, Zhang Y, Wang S I, et al. Simple Recurrent Units for Highly Parallelizable Recurrence [J]. 2017.

[11] Zhang S, Zhang S, Huang T, et al. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching[J]. IEEE Transactions on Multimedia, 2017, PP (99):1-1.

[12] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[13] H. Ramchoun, M A. Janati Idrissi, Y. Ghanou, and M. Ettaouil, "New Modeling of Multilayer Perceptron Architecture Optimization with Regularization: An Application to Pattern Classification," IAENG International Journal of Computer Science, vol. 44, no.3, pp261-269, 2017.

[14] Hambarde, Aaditya R. , M. F. Hashmi , and A. G. Keskar. "Robust Image Authentication based on HMM and SVM Classifiers". Engineering Letters, vol. 22, no.4, pp183-194, 2014.

[15] Chen P Z, Zhang X, Ru Y. Emotion Recognition System Based on Enhancement of KNN Algorithm [J]. Science Technology & Engineering, 2017.

[16] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 2012, 9 (8): 1735-1780.

[17] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer Science, 2014.

[18] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech[C]// INTERSPEECH 2005 - Eurospeech, European Conference on Speech Communication and Technology, Lisbon, Portugal, September. DBLP, 2005:1517-1520.

[19] Institute of Automation, Chinese Academy of Sciences. CASIA Chinese Emotional Corpus [DB/OL].

[20] Schuller B, Vlasenko B, Eyben F, et al. Acoustic emotion recognition: A benchmark comparison of performances[C]// Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on. IEEE, 2010:552-557.

[21] Sun Y, Wen G , Wang J . Weighted spectral features based on local Hu moments for speech emotion recognition [J]. Biomedical Signal Processing and Control, 2015, 18:80-90.