Bidirectional LSTM with Multiple Input Multiple Fusion Strategy for Speech Emotion Recognition

Yuanbo Fang, Hongliang Fu, Huawei Tao*, Xia Wang and Li Zhao

Abstract—**Speech emotion recognition is one of the important challenges in the field of artificial intelligence. In order to further improve the accuracy of speech emotion recognition, a double bidirectional long short-term memory network with multiple input multiple fusion strategy (DBL-MM) for speech emotion recognition is proposed. Firstly, log Mel spectrum feature and frame-level statistical feature are extracted from speech signals, and then the two types of frame-level features are input into two bidirectional LSTM networks for learning simultaneously. Secondly, attention pooling and average pooling are used to fuse the outputs of the two BiLSTM networks to obtain two high-level fusion features which are concatenated and batch normalized later. Finally, a softmax classifier is used to classify emotions. The proposed DBL-MM model simultaneously processes two different types of features and a multiple fusion strategy is used to better learn the subtle changes in emotion. The experimental results on two public datasets show the superiority of this method.**

Index Terms—**Speech Emotion Recognition, Bidirectional LSTM, Multiple input multiple fusion**

I. Introduction

As the most direct and efficient way of information transmission, human voice has obvious differences in the characteristics of timbre, quality, rhythm and volume when expressing different emotions. Therefore, it is feasible to use machine to mine the speech information and simulate the human emotion perception process to realize the speech emotion recognition. The research of speech emotion recognition has appeared for decades [1]. With the development of artificial intelligence, human beings have stepped into the era of human-computer interaction through speech [2]. It is of great significance for the computer to recognize the emotional state of the speaker by getting the information reflecting the emotional state from the speech signal.

Feature extraction is the first and most important step in speech signal processing. The quality of features directly determines the recognition effect. At present, the acoustic features used in speech emotion recognition can be roughly summarized as prosody features, spectrum based related features and voice quality features [3]. However, these artificial features are low-level which still cannot express the emotion in the discourse well. In recent years, deep learning has been widely used in speech emotion recognition. The neural network model trained by appropriate algorithm can extract more valuable features from the original data set and integrate feature learning into model construction. The models such as convolutional neural network and automatic encoder have achieved a certain degree of effect on speech emotion recognition. Shiqing Zhang et al. [4] proposed a discriminating temporal pyramid matching (DTPM) algorithm, which combines the deep features learned by CNN for speech emotion recognition. Deng et al. [5] proposed a sparse automatic coder for feature transfer learning of speech emotion recognition. In the proposed method, common emotion specific mapping rules are learned from a small set of marker data in the target domain. Then, the new reconstructed data is obtained by applying the rule to emotion specific data in different domains.

Although the above algorithms have been successfully applied to emotion recognition, most of the traditional machine learning algorithms and deep learning networks can only accept fixed dimension features as input. These features are often extracted in frames, but they participate in emotion recognition in the form of global feature statistics. The units of global statistics are generally auditory independent sentences or words, and the commonly used statistical indicators are extremum, extremum range, variance, etc. In this way, the variable length speech waveform is ignored. In addition, these features lose the time information of speech waveform in the process of extraction. In recent years, RNN in deep learning method has attracted attention in the field of speech emotion recognition, especially long short term memory (LSTM) [6], which proposed to solve the problem that the traditional RNN model has limited ability to process

This work was supported in part by National Natural Science Foundation of China(No.61673108 and 61601170), National Key R&D Program of China (No.SQ2019YFC200103), Henan Provincial Science and Technology Research Project(No.192102210101), Natural Science Project of Henan Education Department(No.19A510009), Start-up Fund for High-level Talents of Henan University of Technology(No.31401148).

Yuanbo Fang is a Master candidate of the Key Laboratory of Grain Information Processing and Control (Henan University of Technology), Ministry of Education, Zhengzhou, 450001 China. (e-mail: fyb1126@163.com).

Hongliang Fu is a Professor of the Key Laboratory of Grain Information Processing and Control (Henan University of Technology), Ministry of Education, Zhengzhou, 450001 China. (e-mail: jackfu_zz@163.com).

Huawei Tao is a Lecture of the Key Laboratory of Grain Information Processing and Control (Henan University of Technology), Ministry of Education, Zhengzhou, 450001 China and he is also a Postdoctor of the School of information science and engineering, Southeast University, Nanjing, 210096 China. (corresponding author to provide e-mail: thw@haut.edu.cn).

Xia Wang is a Lecture of the School of Information Science and Technology, Nantong University, Nantong, 226000 China. (e-mail: wangxia9802@163.com).

Li Zhao is a Professor of the Laboratory of Underwater Acoustic Signal Processing, Southeast University, Nanjing, 210096 China. (e-mail: zhaoli@seu.edu.cn).

long-term sequential sequences through gating mechanism, and overcome the problem of gradients vanishing, so that the neural network can train for long-term sequential modeling. Although the LSTM network can use the features of speech sequence signal or speech frame to learn the time sequence information of emotion change, it does not consider the problem of unbalanced distribution of emotion in the speech. In the process of training, it will also learn the non-emotional information to reduce the performance of the model.

In order to solve these challenges, a method combining BiLSTM with multiple input multiple fusion strategy is proposed for speech emotion recognition, this study uses two BiLSTM to learn two types of frame-level features simultaneously, and two types of high-level fusion features are obtained by using multi-fusion method at the output of two BiLSTM. The multi-fusion method fully learns the emotional information contained in each frame from the two aspects of the weighted sum and average of the attention based on each frame. Finally, the two high-level features are concatenated and batch normalized, and then classified by softmax classifier. Experiments show that the algorithm achieves good recognition results based on EMO-DB corpus and SAVEE corpus.

The main contributions of this paper can be summarized as follows:

(1) The proposed DBL-MM model combines multiple input multiple fusion strategy. The input of the model uses two types of frame-level acoustic features to learn the emotional details under different features fully.

(2) At the output end of the DBL-MM model, attention pooling and average pooling are adopted to fully learn the emotional information contained in each frame of speech.

II. Related Work

Speech emotion recognition consists of two basic steps: feature extraction and classification. In recent years, many scholars have devoted themselves to these two aspects and made many breakthroughs. In speech emotion recognition, the commonly used speech emotional classifier models include Multilayer Perceptron (MLP) [7], Support Vector Machine (SVM) [8], K-Nearest Neighbor (KNN) [9] classification algorithm, etc.

In the aspect of feature extraction, people found and extracted a group of features related to emotional state closely from speech. The author extracted a group of 6373 feature sets using the short-time window sliding method. Eyben et al. [10] proposed a more concise and effective data set called the Geneva Minimal Acoustic Parameter Set (GeMAPS), which is composed of 62 features, and the GeMAPS dataset was extended to include 88 features. Spectral correlation features reflect the changes of vocal cords and channels in the process of voice production. It has a obvious effect in the task of speaker recognition and also plays a significant role in speech recognition. Some studies have shown that spectral correlation features can also play a role in emotion recognition. For example, the distribution of spectral energy in different areas of the spectrum of a speech has a significant relationship with the emotional information contained in the speech [7]. The commonly used spectral correlation features include Linear spectral features and Cepstrum features. Linear spectral features include Predictor Coefficient (LPC), Log-Frequency Power Coefficient (LFPC), Cepstrum features include: Linear Predictor Cepstral Coefficient (LPCC), Mel-Frequency Cepstral Coefficient (MFCC), etc.

With the development of deep learning in recent years, many scholars begin to use deep neural network to extract speech features directly. Deep learning feature extraction mainly uses CNN or LSTM network to extract high-level features of speech. Generally, the traditional acoustic features of speech are extracted as low-level features first, then the low-level features are input into deep neural network successively to obtain high-level features which are used to train emotion classifier in the end [11,12]. Because LSTM is more suitable for the input of sequence data, the research of LSTM is the focus of speech emotion recognition.

Long and short-term memory is a kind of time recurrent neural network, which has good performance in translation language, image recognition, handwriting recognition, click-through rate, stock, synthetic music and other fields. LSTM has also been applied in the field of speech emotion recognition. Wöllmer [13] applied LSTM to continuous emotion recognition firstly, and 4843 features of each utterance are extracted as input of LSTM. In his further work, static features are used as input of BiLSTM to predict the emotional expression of speakers. Schmidhuber [14] proposed a peephole connection using the state of historical unit as the input information to improve the ability of learning historical information. Shi et al. [15] proposed the Convolutional LSTM network to extract mixed time-frequency information in new features. However, these improved LSTM variants enhance memory information at the cost of computational complexity.

In addition, the last moment output of LSTM is often used as the input to the next model in many LSTM applications which will lose the output of the intermediate nodes of the LSTM neurons [12, 16, 17], especially in the processing of long speech which is easy to cause information loss. In the task of emotion recognition, Keren [18] introduced the pooling operation of the convolutional neural network into the output of the LSTM. Tao [19] applied attention mechanisms to update the cell state of LSTM which focused on information between cells, considering more previous cell states. Xie [20] uses the self-attention algorithm to modify the LSTM forgetting gate, enabling memory cells to more utilize historical information efficiently while reducing the computational complexity of LSTM.

The above work mainly uses LSTM type neural network to learn single type features and extract emotion information from speech signals to identify emotions. However, the quality of features determines the recognition effect. Therefore, we propose a DBL-MM model, which uses two LSTM networks to process two types of features simultaneously to obtain better representational features. Finally, two feature fusion strategies are used to fuse the output high-level features of the two networks to improve the representativeness of the fused high-level features. In chapter 3, we will introduce the proposed DBL-MM model in detail.

III. Proposed Framework

The main part of the proposed model is composed of two BiLSTM modules and two types of pooling layers. This

chapter introduces the input of the model, each module of the model and two pooling strategies. Finally, the overall structure of the proposed model is introduced.

A. Generation of DBL-MM input

TABLE I
FRAME-LEVEL STATISTICAL FEATURES

Featuer ID	Feature name
1-13	MFCCS
14-26	Delta MFCCs
27-39	Second order delta MFCCs
40	ZCR
41	RMS Energy
42-45	Spectral Centroid, Spectral Band-width, Spectral Flatness, Spectral Roll-off

Frame-level statistical features and log Mel-spectrogram are used as two types of frame-level acoustic features to be input into the DBL-MM model. In order to obtain the same time and frequency resolution of acoustic features, the speech signal was processed by hanming windows size of 25 ms and 15 ms overlapping. As shown in Table 1, 45 acoustic low-level descriptors (LLDs) were extracted from each frame. In detail, LLDs have the MFCCs, and its first-order and second-order delta, a zero-crossing-rate (ZCR) from the time signal, a root mean square(RMS) of the frame energy, a center of gravity of the spectrum, the bandwidth of the spectrum, the flatness of the spectrum, and a roll-off frequency of the spectrum.

Compared with traditional manual design features, spectral features can consider both frequency and time axis to extract more emotional information. Log Mel-spectrum describes the detailed information in speech from another perspective, which not only contains rich time-frequency characteristics, but also can obtain the variation of subjective emotions in speech at different time frequencies. We used 64 Mel filters to obtain log Mel-spectrogram, each frame of which has the same degree of detail. The size of log Mel-spectrograms varies with the size of the frame-level feature.

B. Bi-directional Long Short-Term Memory

Speech-based emotion recognition algorithms rely on the dynamic process of language parameters heavily. Therefore, the feature should have context dependency, and the model should have the ability to learn the dependency. In this study, the context dependence between these features is established in the order of frames which can be learned through the BiLSTM.

LSTM is an improvement of recurrent neural network (RNN) which introduces three types of control gates: input gate, output gate and forgetting gate. It writes, reads and resets the hidden unit. One drawback of traditional LSTM is that it can only use the previous content from the forward sequence. In speech emotion recognition, the future content from the reverse sequence plays an important role in the judgment of emotion polarity. In this way, complementary information from the past and the future can be integrated for reasoning. Bidirectional LSTM network [21] is an improvement of the standard forward LSTM model, which can operate a series of features in both forward and backward directions.

The original LSTM state:

$$i_t = \sigma(w_{xi}x_t + w_{hi}h_{t-1} + w_{ci}c_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(w_{xf}x_t + w_{hf}h_{t-1} + w_{cf}c_{t-1} + b_f) \quad (2)$$

$$c_t = f_t c_{t-1} + i_t \tanh(w_{xc}x_t + w_{hc}h_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(w_{xo}x_t + w_{ho}h_{t-1} + w_{co}c_{t-1} + b_o) \quad (4)$$

$$h_t = o_t \tanh(c_t) \quad (5)$$

Among them, σ is the commonly used activation function of sigmoid, i, f, o, c represent the input gate, forgetting gate, output gate and memory cell respectively which are the same dimension size as the hidden layer vector h.

The Bidirectional LSTM state:

$$h_t = \left[\vec{h}_i \oplus \overleftarrow{h}_i\right] \quad (6)$$

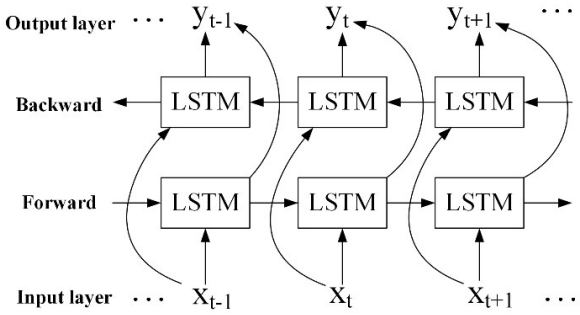

Fig. 1. Structure of BiLSTM

Two outputs of the BiLSTM network are connected to the same output node. Its structure is shown in Figure 1. Through this structure, the output layer can obtain both historical and future information. Therefore, compared with ordinary LSTM, BiLSTM does not need to wait until later time nodes to obtain future information. In addition, there is no shared state between LSTMS in these two different directions. In other words, the forward output state is only transmitted to the forward LSTM, while the reverse state is only transmitted to the reverse LSTM.

C. Mean-pooling

Frame-wise, Final-frame and Mean-pooling on time are three commonly used learning methods for emotional speech tags in LSTM networks. Experiments have proved that [22] Mean-pooling can learn the emotion contained in each frame more fully compared with the former two methods, Mean-pooling is to perform a sliding average over time for the output of LSTM, which is, to find the average value of all outputs:

$$O_{average} = \sum o(t) / T \quad (7)$$

D. Attention-pooling

Attention mechanism was first proposed in the field of visual image [23] which achieved good effects. The core idea is that the brain's attention to the whole picture is unbalanced,

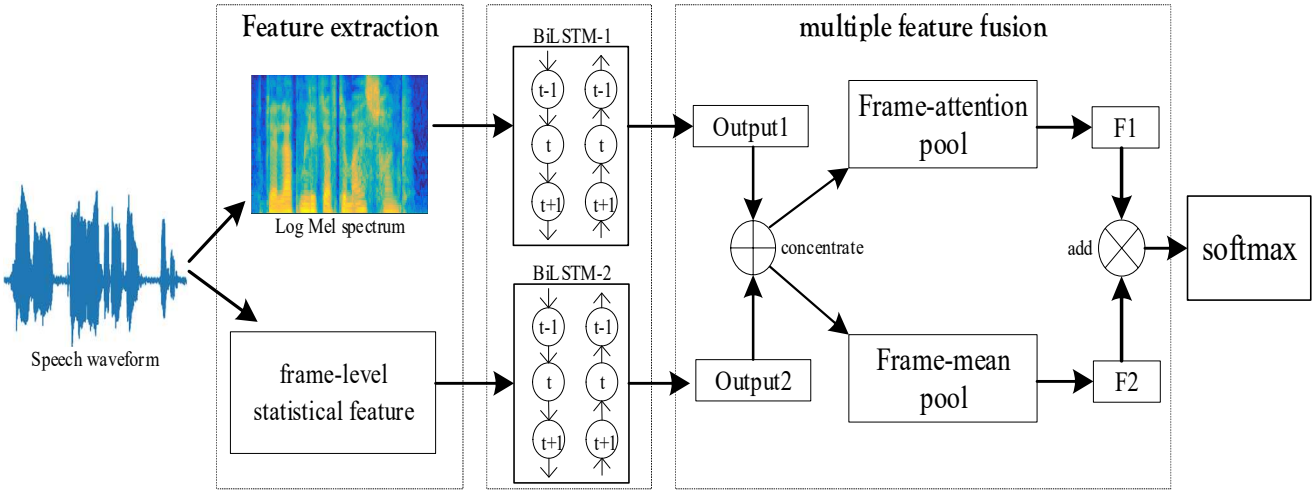

Fig. 2. Structure of DBL-MM.

and there are certain weight differences. Inspired by this, this paper applies the self-attention mechanism to the output calculation of the BiLSTM. Specifically, the time pooling technology for the BiLSTM model is implemented by calculating the weighted sum divided by the time [24]. The standard way for BiLSTM to use the attention mechanism is to choose a simple weighted sum similar to logical regression as the pooling layer. This weighted sum is the inner product between the frame direction output of BiLSTM(y_t) and the weight u which is the parameter vector in the attention model. In order to keep the weight sum unified, we apply the softmax function to the inner product.

$$\alpha_t = \frac{\exp(u^H y_t)}{\sum \exp(u^H y_\tau)} \tag{8}$$

where α_t is the weight for the output at t-th time step y_t. Due to the memory ability of LSTM, the accumulated information is the most abundant in the output of the last moment. Theoretically, the last moment of LSTM networks should obtain a large weight. In this study, the last moment outputs of two BiLSTM modules is used as a reference to ensure that attention pooling can be used to obtain a large weight, the weight coefficients are applied to o_t on the time dimension and summed up in the time dimension as an output. The relevant calculation formula is:

$$o_{last\_time} = \left[o_{l1} \oplus o_{l2} \right] \tag{9}$$

$$\alpha_T = \text{softmax}(o_{last\_time} \times (o_t \times w_t)^H) \tag{10}$$

$$Z = \alpha_T \times o_t \tag{11}$$

Where o_{last_time} is the series connection of output o_{l1} and o_{l2} at the last moment of two BiLSTM modules. w_t is the weight for training. The result Z of attention pooling is obtained by multiplying attention weight α_T and o_t.

E. Structure of DBL-MM

At the output of the model, two feature fusion methods of average pooling and attention pooling are used to learn the emotional information contained in each frame of speech. The two feature fusion methods fuse the output of two BiLSTMs into two high-level fusion features, and then the two learned high-level features are concatenated and batch normalized. Finally, a softmax classifier is used to classify emotions. By using Batch Normlization, model learning can be promoted and makes the model less dependent on the initial weight value. In the process of training, different BiLSTM modules are trained simultaneously to ensure that the model can fully learn the emotional integrity of each speech. The simultaneous processing of features can ensure the consistency of the features, and ensure that the contribution of different features to the fusion feature is optimal during reverse training, rather than a simple feature stacking.

IV. Experiments

A. Speech emotion database

To showcase the performance of the suggested approach, we chose two different popular databases to avoid observations based on single corpus evaluation, including Berlin EMO-DB [25] German Emotional Voice Library and Surrey Audio-Visual Expressed Emotion (SAVEE) [26].

B. Evaluation methods

The Leave-One-Speaker-Out (LOSO) [27] cross-validation strategy was used in the experiment which is more realistic and challenging. In this strategy, a person's emotional speech sample data set is used as a test set selection experiment each time, and the remaining emotional speech samples are used as a training set, and each person's voice will be used as a test set. Finally, the average of several experiments was calculated as the result. In this paper, the evaluation criteria are Weighted Accuracy (WA) and Unweighted Accuracy (UA) as suggested in [27].

C. Experimental parameters

To get the emotional details contained in the speech better, we extract two different features of each speech as our DBL-MM inputs. Parallel operation of double BiLSTM in the model of DBL-MM, the parameters of the DBL-MM model are shown in Table 2.

To prevent over-fitting of the data during training, we added Dropout to the DBL-MM model and set it to 0.7. All experiments are carried out under the development environment with Tensorflow 1.4 version and GTX 1080Ti.

TABLE II
NETWORK PARAMETERS

Parameters	Value
Learning rate	0.0001
Batch size	32
Dropout	0.7
Inupt (BiLSTM-1)	64
Hidden (BiLSTM-1)	1024
Dense (BiLSTM-1)	2048
Inupt (BiLSTM-2)	45
Hidden (BiLSTM-2)	1024
Dense (BiLSTM-2)	2048
Mean-pooling	4096
Attention-pooling	4096
Output	4096

D. Model comparison

To illustrate the effect of our multiple input on the experimental results, experiments were carried out on single-feature model BiLSTM1 and BiLSTM2, that is, removing the one of parallel BiLSTM part of the model separately, and identifying emotions using single feature with the remaining parameters unchanged. Through 100 epoch iterative training, Table 3 provides detailed data comparison.

TABLE III
MODEL COMPARISON

Database	Model	WA	UA
EMO-DB	BiLSTM-1	80.80	78.37
	BiLSTM-2	73.14	70.05
	DBL-MM	**84.87**	**83.35**
SAVEE	BiLSTM-1	54.79	49.76
	BiLSTM-2	50.40	47.33
	DBL-MM	**60.63**	**57.62**

As shown in Table 3, the performance of DBL-MM was demonstrated on two databases; the WA is 84.87% and 60.63% respectively, and UA is 83.35% and 57.62% respectively, which are significantly improved compared with a single BiLSTM module.

In order to study the improvement of the recognition effect of the two databases by different modules, the significance test of the experimental results is carried out. T-test results are shown in Table 4. When P-value is less than 0.05, the difference between the data is significant. As we can see from Table 4, the P-value of both databases are less than 0.05. Therefore, compared with a single BiLSTM module with a single type feature, the performance of DBL-MM model has improved significantly.

TABLE IV
T-TEST ON TEST RESULTS

Database	Models	P-Value
EMO-DB	(DBL-MM, BiLSTM-1)	< 0.0001
	(DBL-MM, BiLSTM-2)	< 0.0001
SAVEE	(DBL-MM, BiLSTM-1)	< 0.0001
	(DBL-MM, BiLSTM-2)	< 0.0001

Furthermore, the performance of the proposal is compared with some works as well. Table 5 shows a comparison between our proposed method and other methods.

TABLE V
RECOGNITION ACCURACY (%) OF DIFFERENT METHODS

Database	Refs	Features	WA	UA
EMO-DB	[28]	HuWSF	81.74	/
	[29]	Acoustic	81.90	79.10
	[30]	CNN-SRU	82.50	80.60
	Ours	DBL-MM	**84.87**	**83.35**
SAVEE	[31]	GA-BEL	44.18	/
	[28]	HuWSF	50.00	/
	[32]	RDBN	53.60	/
	Ours	DBL-MM	**60.63**	**57.62**

From Table 5, we can see that the best performance of our method in the SAVEE database and our method obviously outperforms [31], [28], [32]. The recognition rate is at least 7.03% higher than that of all comparative experiments. On the EMO-DB dataset, our method also clearly outperforms all the three compared works. The experimental results in two different databases demonstrate the feasibility of using features of different functions as model input simultaneously. And the recognition effect of high-level fusion features extracted by the proposed DBL-MM model outperforms other comparative experiments.

Fig. 3. Confusion matrix of DBL-MM on the EMO-DB dataset.

Fig. 4. Confusion matrix of DBL-MM on the SAVEE dataset.

To further investigate the recognition accuracy, we present the confusion matrix to analyze the performances of our DBL-MM model. Fig.3 shows that on the EMO-DB dataset,

anger and sadness are classified with accuracy are 95.32% and 94.96%, respectively. The classification accuracy of fear and neutral are both higher than 85%. The recognition rate of the other three emotions is less than 80%. Fig.4 shows that on the SAVEE dataset, neutral is identified with the highest accuracy of 81.67%, anger is identified with the accuracy of 70.00% and the accuracy of other five emotions was less than 70%.

From the confusion matrix, we can see that the difference between the highest recognition rate neutral and the lowest recognition rate disgust in SAVEE data-base is 48.34%. In EMO-DB database, there would not be such a big difference. The reason may be that LOSO strategy is adopted in this experiment, speed, voice line, and style of speech of different speakers also have different effects on the experiment. In addition, the sample number of SAVEE database is relatively small, and the neutral with the highest recognition rate accounts for the highest proportion of the sample number of SAVEE database.

V. Conclusion

In this work, we present our recognition framework in the speech emotion recognition. Different from the previous research on speech emotion recognition based on LSTM network, the proposed DBL-MM model is composed of two parts: log Mel-spectrograms features with BiLSTM and frame-level statistical features with BiLSTM. The parallel structure of the model can process both types of frame-level features simultaneously to extract high-level features with better representability. In addition, in order to better learn the emotional information contained in each frame from the output of the model, two fusion methods with attention-pooling and mean-pooling were adopted simultaneously. The experiment shows that the DBL-MM model can effectively mine the emotional information from these frame-level features, and experiments on two databases show that the proposed method has good performance.

References

[1] Van Bezooijen R, Otto S A, Heenan T A, Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics, Journal of Cross-Cultural Psychology, vol. 14, no. 4, pp. 387-406, 1983.

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

[3] Han WJ, Li HF, Ruan HB, Ma L, Review on speech emotion recognition, Ruan Jian Xue Bao/Journal of Software, vol. 25, no. 1, pp. 37-50, 2014.

[4] Zhang S, Zhang S, Huang T, et al, Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, IEEE Transactions on Multimedia, vol. 20, no. 6, pp. 1576-1590, 2017.

[5] Deng J, Zhang Z, Marchi E, et al, Sparse autoencoder-based feature transfer learning for speech emotion recognition, 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. pp. 511-516, 2013.

[6] Hochreiter S, Schmidhuber J, Long short-term memory, Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[7] Ramchoun H, Idrissi M A J, Ghanou Y, et al, New modeling of multilayer perceptron architecture optimization with regularization: an application to pattern classification, IAENG International Journal of Computer Science, vol. 44, no. 3, pp. 261-269, 2017

[8] Hashmi M F, Hambarde A R, Keskar A G, Robust Image Authentication Based on HMM and SVM Classifiers, Engineering Letters, vol. 22, no. 4, pp. 183-193, 2014.

[9] Chen P Z, Zhang X, Ru Y. Emotion Recognition System Based on Enhancement of KNN Algorithm, Science Technology & Engineering, vol. 17, no. 19, pp.197-200, 2017.

[10] Eyben F, Scherer K R, Schuller B W, et al, The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, IEEE transactions on affective computing, vol. 7, no. 2, pp.190-202, 2015.

[11] Zheng W Q, Yu J S, Zou Y X. An experimental study of speech emotion recognition based on deep convolutional neural networks, 2015 international conference on affective computing and intelligent interaction (ACII), pp. 827-831, 2015.

[12] Mao Q, Dong M, Huang Z, et al, Learning salient features for speech emotion recognition using convolutional neural networks, IEEE transactions on multimedia, vol. 16, no. 8, pp. 2203-2213, 2014.

[13] Wöllmer M, Eyben F, Reiter S, et al. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies, Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia, pp. 597-600, 2008.

[14] Gers F A, Schmidhuber J, Recurrent nets that time and count, Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, vol. 3, pp. 189-194, 2000.

[15] Xingjian S H I, Chen Z, Wang H, et al, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, Advances in neural information processing systems, pp. 802-810, 2015.

[16] Sainath T N, Li B, Modeling time-frequency patterns with LSTM vs. convolutional architectures for LVCSR tasks, Interspeech, pp. 813-817, 2016.

[17] Yoo J H, Large-scale video classification guided by batch normalized LSTM translator, arXiv:1707.04045, pp. 1-7, 2017.

[18] Keren G, Schuller B, Convolutional RNN: an enhanced model for extracting features from sequential data, 2016 International Joint Conference on Neural Networks (IJCNN), pp. 3412-3419, 2016.

[19] Tao F, Liu G, Advanced LSTM: A study about better time dependency modeling in emotion recognition, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2906-2910, 2018.

[20] Xie Y, Liang R, Liang Z, et al, Speech emotion classification using attention-based lstm, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 11, pp. 1675-1685, 2019.

[21] Su B H, Yeh S L, Ko M Y, et al. Self-Assessed Affect Recognition Using Fusion of Attentional BLSTM and Static Acoustic Features, Interspeech, pp. 536-540, 2018.

[22] Chen X, Han W, Ruan H, et al. Sequence-to-sequence modelling for categorical speech emotion recognition using recurrent neural network, 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), pp. 1-6, 2018.

[23] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on pattern analysis and machine intelligence, vol. 20, no. 11, pp. 1254-1259, 1998.

[24] Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention, 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227-2231, 2017.

[25] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech, Ninth European Conference on Speech Communication and Technology, Interspeech, pp. 1517-1520, 2005.

[26] Haq S, Jackson P J B, Edge J. Speaker-dependent audio-visual emotion recognition, AVSP, pp. 53-58, 2009.

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

[28] Sun Y, Wen G, Wang J. Weighted spectral features based on local Hu moments for speech emotion recognition, Biomedical signal processing and control, vol. 18, pp. 80-90, 2015.

[29] Stuhlsatz A, Meyer C, Eyben F, et al. Deep neural networks for acoustic emotion recognition: Raising the benchmarks, 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5688-5691, 2011.

[30] Jiang P, Fu H, Tao H. Speech Emotion Recognition Using Deep Convolutional Neural Network and Simple Recurrent Unit, Engineering Letters, vol. 27, no. 4, pp. 901-906, 2019.

[31] Liu Z T, Xie Q, Wu M, et al. Speech emotion recognition based on an improved brain emotion learning model, Neurocomputing, vol. 309, pp. 145-156, 2018.

[32] Wen G, Li H, Huang J, et al. Random deep belief networks for recognizing emotions from speech signals, Computational intelligence and neuroscience, pp. 1-9, 2017.