Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 2593036, 9 pages
https://doi.org/10.1155/2019/2593036

Research Article

Audio-Textual Emotion Recognition Based on Improved Neural Networks

Linqin Cai, Yaxin Hu, Jiangong Dong, and Sitong Zhou

Key Laboratory of Industrial Internet of Things & Networked Control, Ministry of Education, Chongqing University of Posts and Telecommunications, Chongqing, China

Correspondence should be addressed to Linqin Cai; iamlqcai@163.com

Received 4 October 2019; Revised 17 November 2019; Accepted 4 December 2019; Published 31 December 2019

Academic Editor: Paolo Crippa

With the rapid development in social media, single-modal emotion recognition is hard to satisfy the demands of the current emotional recognition system. Aiming to optimize the performance of the emotional recognition system, a multimodal emotion recognition model from speech and text was proposed in this paper. Considering the complementarity between different modes, CNN (convolutional neural network) and LSTM (long short-term memory) were combined in a form of binary channels to learn acoustic emotion features; meanwhile, an effective Bi-LSTM (bidirectional long short-term memory) network was resorted to capture the textual features. Furthermore, we applied a deep neural network to learn and classify the fusion features. The final emotional state was determined by the output of both speech and text emotion analysis. Finally, the multimodal fusion experiments were carried out to validate the proposed model on the IEMOCAP database. In comparison with the single modal, the overall recognition accuracy of text increased 6.70%, and that of speech emotion recognition soared 13.85%. Experimental results show that the recognition accuracy of our multimodal is higher than that of the single modal and outperforms other published multimodal models on the test datasets.

1. Introduction

In the information era, social media such as Weibo and WeChat are becoming the main way of our daily communication. Many social media platforms produce a lot of textual and audio data which contain rich emotion information, and they play an increasingly important role in emotional recognition. Emotion recognition, as one of the basis of human-computer emotional interaction, affects the stable development of artificial intelligence technology. Nowadays, emotion mining based on the social media has become one of the most important tasks in optimizing human-computer interaction. The textual sentiment analysis published by social media such as microblog has also attracted widespread attention, and many related research studies have been conducted [1–4]. Nevertheless, the emotion information contained in text is limited, and there are several fetters on the identification of technical terms in specific fields.

One of the challenges in emotional recognition is that emotion information contained in a single mode is limited. With the increase in audio information on the social media, it is difficult to meet the needs of the current emotional recognition system to get the right emotional state only from single modal. Indeed, in textual sentiment analysis, the system can only judge the emotion expressed in communications by dealing with the words, phrases, sentences, and dependencies among them. It is usually not enough to understand the full emotional content. Text often accompanied by voice in everyday conversations and social media is closely related. As the most direct means of human communication, voice itself can transmit abundant information. Many researchers also have conducted in-depth researches on speech emotion recognition [5–8], and good progresses have been made. Considering the internal connection between text and speech, modal fusion can be utilized to optimize the performance of the social media emotional recognition system. The final emotional state would be determined by the output of both speech and text emotion analysis.

Another challenge is that traditional feature extraction methods, such as HMM (hidden Markov model) and GMM (Gaussian mixture model) [9, 10], only model a limited amount of context information. However, human emotions usually change slowly and depend heavily on context information. Contrasted with traditional feature extraction methods, deep neural networks (DNN) can effectively reveal the hidden inner structure between data and extract high-level abstract features which are useful for emotion classification in emotion recognition. It is believed that the deep learning method could be able to extract the intrinsic features from large-scale training data [11]. In recent years, LSTM (long short-term memory) [11, 12] and CNN (convolutional neural network) [13] have all been used in single mode emotion recognition to learn the high-level features from raw data, and their performances are obviously superior to traditional feature extraction methods. It is reasonable to believe that deep neural networks can still perform well in multimodal emotion recognition, and in fact, the relevant research studies are under way.

In this paper, a multimodal emotion recognition model from both acoustics and textual features on the IEMOCAP database was presented. The proposed model takes advantage of different modalities to get more comprehensive and precise information for emotion classification. To get more effective features for the multimodal emotional classification, a DNN network was put forward to learn the fused features from text and speech, which used the hand-crafted and high-level features. Moreover, according to the qualities of text and audio information, we set up the suitable model to detect features. It is worth mentioning that a CNN-Bi-LSTM-Attention (CBLA) model was applied to capture acoustic features. The CBLA contains dual channels CNN and Bi-LSTM networks and considers both the global and contextual temporal information in the data. The L_2 regularization was also used to optimize the model to avoid overfitting in model training.

2. Related Work

Deep learning is a subproblem of machine learning. Its main purpose is to automatically learn effective feature representation from data. In deep learning, the original data features are transformed into a feature representation through multistep feature transformation and further input into the prediction function to get the final result. In recent years, deep learning has developed rapidly and has achieved great success in many subfields of artificial intelligence, such as human action recognition [14] and emotion recognition. As the mainly used model in deep learning, neural networks such as CNN and LSTM have been widely used in emotion recognition and perform effectively. Traditional feature extraction methods, such as HMM (hidden Markov model) and GMM (Gaussian mixture model) [9, 10] could only model limited amount of context information, and did not take full advantage of the fact that human emotions usually change slowly and depend heavily on context information. Contrasted with traditional feature extraction methods, deep neural networks can effectively reveal the hidden inner structure in data and extract high-level abstract features that are useful for emotion classification in emotion recognition.

For instance, Li et al. [11] applied LSTM to achieve multiclassification for text emotional attributes and found that LSTM is better in analysing emotion of long sentences than the conventional RNN (recurrent neural network). In other studies, CNN [13] and LSTM [12] have all been used in speech emotion recognition to earn the high-level features from raw audio clips and performed well. LSTM is an improved RNN with powerful computing capacity and storage capabilities that can effectively solve the problem of gradient explosion or disappearance of simple RNN. Voice is a kind of nonlinear time-series transform signal, and text information is closely related to temporal context, both of them are time-related. Therefore, the LSTM network is suitable for acoustic and text feature extraction and learning that models in context and helps to learn the relevance of features. However, in LSTM, there is no intermediate nonlinear hidden layer that causes an increase in variation in the hidden state factors [15].

On the contrary, the CNN network can reduce variance in frequency of the input and captures local information but without considering the global features and context. In brief, the modelling capabilities of CNN and LSTM are both limited. Trigeorgis et al. [16] combined CNN with LSTM in order to automatically learn the best representation of the speech signal directly from the raw time representation. Based on the research, we proposed a similar model which combined CNN and LSTM for the speech emotion recognition.

Furthermore, multimodal emotion analysis is an emerging field recently. The research of multimodal emotion recognition based on effective fusion of single-modal information such as voice, expression, and text has made some good progress. Considering the internal correlation between different modes, we can fuse the features from different modes, such as text and speech, to get more effective emotional characteristics for emotion recognition. The deep neural network model can also be used for feature learning on the fused multimodal data. The original high-dimensional heterogeneous data can be transformed into abstract semantic expression in the same feature space through multiple nonlinear transformations and then used for multimodal feature extraction and multimodal feature selection.

For the research on multimodal emotional analysis, [17–20] all applied CNNs were as a trainable feature extractor to extract textual, visual, or audio features. Poria et al. [17] fed all the emotional features into the MKL (multiple kernel learning) classifier. Besides, they make use of the open source software openSMILE to extract low-level handcraft audio features and feed them into the classifier too. In [18], the author used a deep residual network of 50 layers to extract features from the visual information and then study automatic effect sensing in an end-to-end manner. In the model of [19], the final CNN layer computed the weighted sum of all the information extracted from the attention input. After extracting features from CNN networks, Gu et al. [20] made use of a three-layer deep neural network to fuse the multimodal features.

Yoon et al. [21] took advantage of RNN to extract features. They proposed a novel multimodal attention method to focus on the specific parts of a transcript that contains strong emotional information, conditioning on the audio information. Angeliki Metallinou et al. [22] adopted Bi-LSTM to classify both audio and visual emotion information contained in the IEMOCAP database. The experimental results show that Bi-LSTM framework prevail over traditional HMM framework. Wang et al. [23] applied a the recurrent attended variation embedding network (RAVEN) for the multimodal emotion recognition, and the LSTM is used to extract features from single mode. The multimodal emotion studies above utilized the deep neural network model, and the results outperformed the traditional methods.

All the studies mentioned above make full advantage of hand-crafted features in the textual or audio material. Nevertheless, some of them focus on single mode emotion recognition, and the improvement in the models are all limited; besides, in the multimodal model, they all utilize CNN or RNN as a trainable feature extractor, do not concern both temporal information and global information at the same time, and ignore the temporal features of voice and text information. In addition, it is noteworthy that the above research studies have not noticed the problem of model fitting.

3. Proposed Multimodal Model

This paper proposed the framework of multimodal emotional recognition based on speech and text, shown as in Figure 1. Firstly, we preprocess speech signal and text information to extract low-level handcrafted emotional features. Secondly, we feed the speech features into CBLA model to capture the local and global information, and the textual features are feed into Bi-LSTM neural networks to acquire high-level features. After that, the feature fusion method is resorted to merge the emotional features of audio and text. Finally, we use a deep neural network to learn and classify the fusion features. We improved traditional deep neural networks using the dual networks to model CBLA for audio feature extraction. For the multimodal model, we made use of a DNN network to train fused features and applied the regularization method to optimize the model. The experiment has proved the model's validity.

3.1. Data Preprocessing and Low-Level Feature Extraction. Firstly, we preprocess the raw audio signal from voice data in the database and then extract 34-dimensional low-level handcrafted acoustic features of time-domain, spectral-domain, and cepstral-domain characteristics, including zero crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, MFCCs (mel frequency cepstral coefficients), 12-dimensional chroma vector, and standard deviation of chroma vector [24]. We keep the maximum input of 100 frames and finally get (100, 34) vector for each utterance. In order to show the manual acoustic emotional features more intuitively, we visualize several features including chroma, zero crossing rate, MFCCs, energy, and spectral flux.

Word embedding is a widely used textual feature extraction method, which uses dense vectors to represent words and documents. GloVe [25] is one of the most commonly methods of word embedding. It was a new method of word matrix generation proposed by Stanford University in 2014. It synthesizes global and local statistical information of words to generate language models and vectorize expressions of words. For the textual data, the pretrained GloVe word embedding method is used to obtain the text emotional feature vectors. For each utterance in the IEMOCAP dataset, we use 300-dimension vectors pretrained GloVe embeddings with a maximum sequence length of 500 to obtain (500, 300) vectors.

3.2. Textual Feature Extraction. LSTM (long short-term memory) [26] relies on its three gates structure, which effectively solves the long-term dependence in the neural network, and avoids the gradient disappearance problem in the common recurrent neural network, and it is suitable for the modelling of speech temporal signals and text signals which are closely related to time. Figure 2 is the structure block of the LSTM network.

The first step in LSTM is to determine what information to lose from the cell through the forgetting gate r_1:

$$r_1 = \phi 1\left(W_1 \times x^*\right), \tag{1}$$

where $x^* \triangleq [x, s^{(\text{old})}]$ indicates that the current input sample is connected to the downstream time channel $s^{(\text{old})}$.

The next step is to decide how much new information is added to the cell state. The sigmoid layer of the input gate r_2 determines which information needs to be updated, and the tanh layer generates alternative content for updating:

$$r_2 = \phi 1\left(W_2 \times x^*\right) \cdot \phi 2\left(W_3 \times x^*\right). \tag{2}$$

The output gate r_3 uses a sigmoid layer to determine which cell states to output. Then, it is processed through tanh layer and then multiplied, and the final output determines the part of the output:

$$r_3 = s^{(\text{new})} = \phi 2\left(h^{(\text{new})}\right) \cdot \phi 1\left(W_4 \times x^*\right), \tag{3}$$

where $W_1 \sim W_4$ represents the weight matrix corresponding to each gate.

The bidirectional LSTM consists of two ordinary LSTM, a forward one which uses the past information and an inverse one that obtains the future information. In this way, the information at $t-1$ as well as at $t+1$ all can be used at time t. It would be more accurate than LSTM and can avoid the long-term dependence problem in features learning. Hence, it can be utilized for the textual emotion feature extraction. We feed the textual vectors into the bidirectional

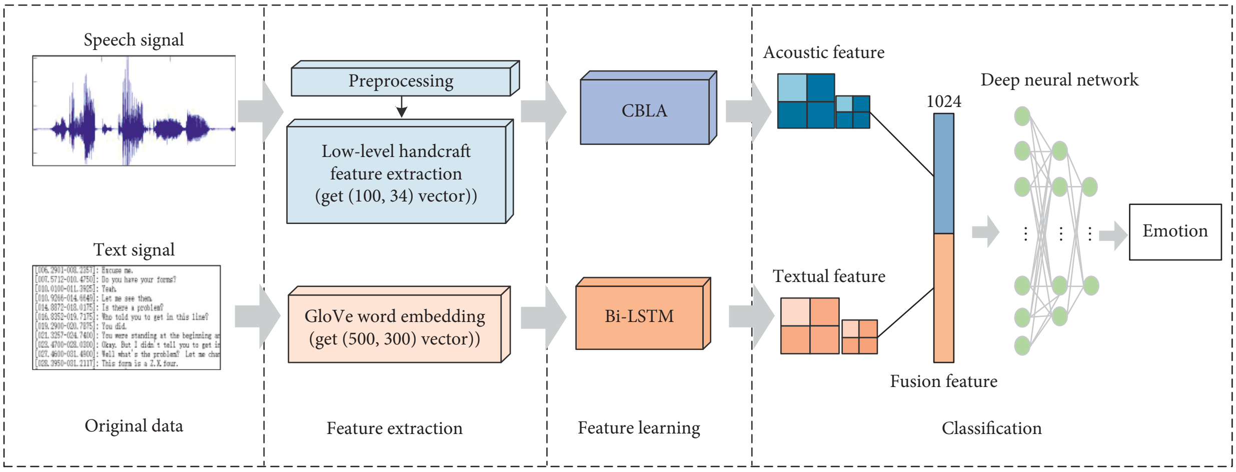


Figure 1: The structure of multimodal model.

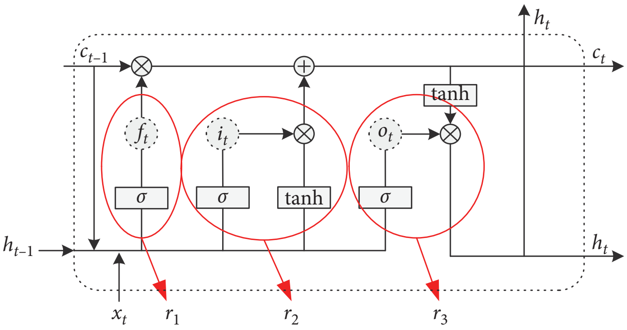


Figure 2: The structure block of the LSTM network.

long short-term memory network to extract high-level information. The structure of the Bi-LSTM network for textual features learning is shown in Figure 3.

3.3. Audio Feature Extraction. Convolutional neural network (CNN) is one of the common deep learning neural networks. The infrastructure of CNN includes the convolution layer, pooling layer, and dense layer. CNN can extract some advanced characteristics automatically. Convolution layer weight sharing reduces the complexity of the network model, alleviates overfitting, pooling operation reduces the number of neurons, and is more robust to the translation of input space. CNN is a process from local to global (local to global realization is in the dense layer), while the traditional neural network is the entire process.

The net input of the convolutional layer is computed as follows:

$$Z^p = W^p \otimes X + b^p = \sum_{d=1}^{D} W^{p,d} \otimes X^d + b^p, \tag{4}$$

where W represents the convolution kernel, X represents the feature mapping, and b represents the bias term.

The output characteristic mapping Y^p is obtained after the net input passes through the nonlinear activation function:

$$Y^p = f(Z^p), \tag{5}$$

where $f(\cdot)$ is the nonlinear activation function.

CNN network can reduce variance in frequency of the input [12] and captures local information but without considering the global features and context.

Voice is a kind of nonlinear time series signal; text information is closely related to temporal context, and they are all time-related. Therefore, it is the LSTM network which is suitable for acoustic and text feature extraction and learning that models in context and helps to learn the relevance of features. But in LSTM, there is no intermediate nonlinear hidden layer that causes the increase in variation in the hidden state factors [14]. In brief, the model capabilities of CNN and LSTM are both limited. CNN and LSTM networks were resorted in [27, 28] for speech emotion recognition, which became an improvement method commonly used in many recent research studies.

In the common CNN-LSTM network, the output of a set of CNN networks was thrown in LSTMs directly. From this way, we can get the high-level information which contains both local information and long-term contextual dependencies. However, CNNs focus on local information and discard a lot of data. To avoid valuable data losing, we constructed the model CBLA which uses binary channels of CNN and Bi-LSTM. In the CNN channel, we constructed four one-dimensional convolution layers with different filters' number and cropped for one-dimensional temporal input (audio features). Moreover, we employed the maximum pooling layer and global average pooling layer to carry out the maximum pooling operation and global average pooling operation for the data. In the Bi-LSTM channel, a set of Bi-LSTM cells with arguments of 256 were put up to extract long-term contextual dependencies information, and an attention mechanism was added to find more effective features. At last, the data from two channels were concatenated and the output was put into a dense layer. After the nonlinear change of the dense layer, the correlation between these features was extracted and finally mapped to the output space. The structure of the CBLA model is shown in Figure 4.

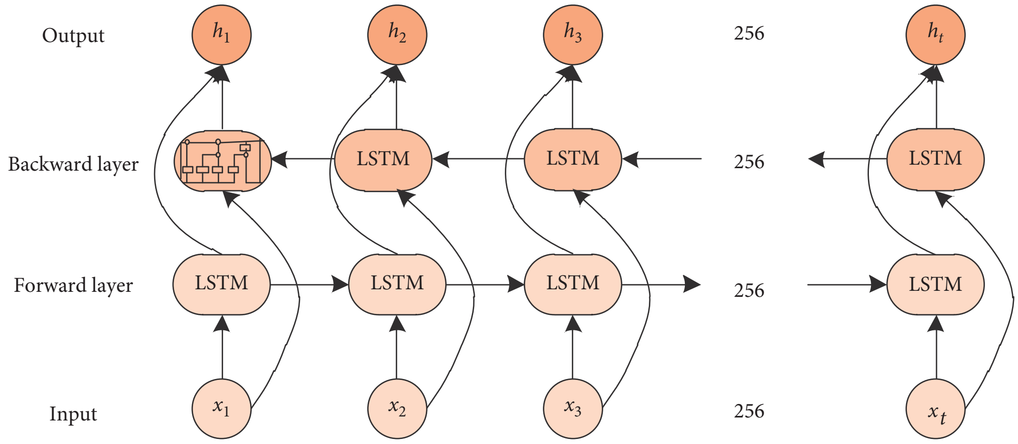


Figure 3: The structure of the Bi-LSTM network for textual feature learning.

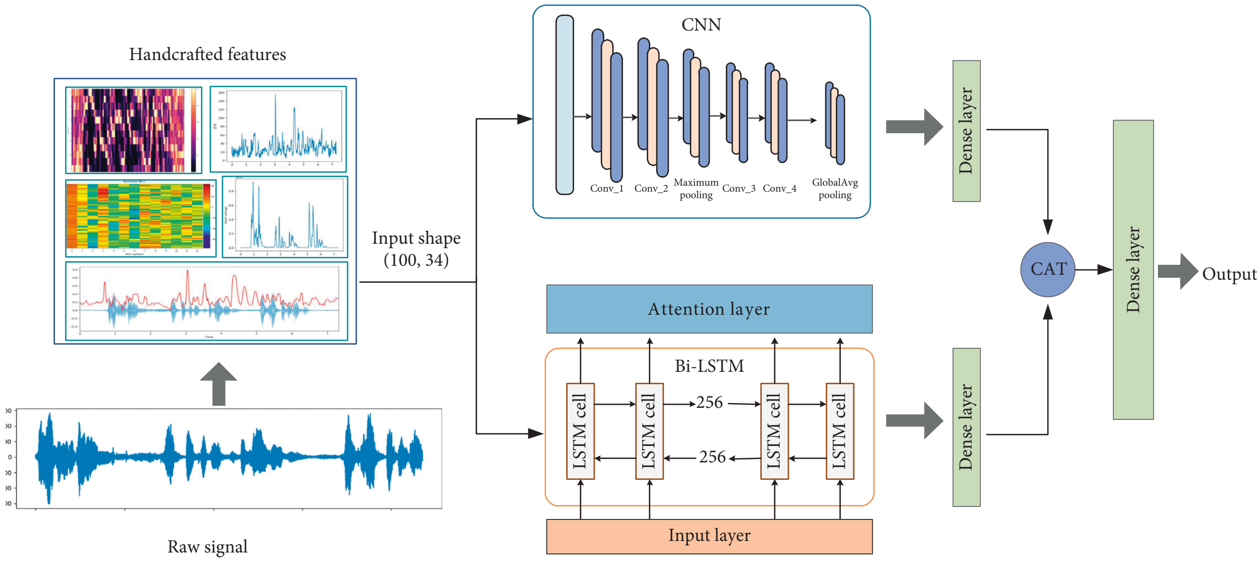


Figure 4: The structure of the CBLA model for acoustics feature learning.

3.4. Feature Fusion. After feeding acoustic features in to CBLA and textual into the Bi-LSTM network, we obtain the high-level textual features $H = \{h_1, h_2, \ldots, h_t\}$ and acoustics features $S = \{s_1, s_2, \ldots, s_t\}$ which consist of global and local information. In this paper, we adopt the feature-level fusion approach. The advantage is that emotional features extracted from different modes are directly related to the final decision, and the fusion results can retain the feature information needed in the final decision to the greatest extent.

The final descriptor of the multimodal emotion features vector $V = \{v_1, v_2, \ldots, v_t\}$ is created using an ordered concatenation of textual H and acoustics features S in the following equation:

$$V = [H, S]. \tag{6}$$

After that, we feed the fusion emotion feature vector V into a deep neural network (DNN) network containing three dense layers, in which the parameters are set as 1024, 512, and 4 and a softmax layer to capture the associations between the features from different modalities. The output of softmax represents the relative probability between different emotion classes:

$$p(x_a) = \text{softmax}(x_a) = \frac{\exp(x_a)}{\sum_{a'}^{A} \exp(x_{a'})}, \tag{7}$$

where a represents the emotion categories and $P(x_a)$ represents the probability of a-th category.

In addition, in order to avoid overfitting, we added regularization in the multimodal features training. The principle of regularization is to add an index to describe the complexity of the model in the loss function. The function used to characterize the complexity of the model is L_2 regularization, and the calculation formula is as follows:

$$R(w) = \|w\|_2^2 = \sum_i |w_i^2|, \tag{8}$$

where w represents the model weights.

4. Experiments

4.1. Experiment Setup

4.1.1. Database. IEMOCAP database [29] is an emotional database collected and recorded by Busso et al. at the SAIL Laboratory of USC. It contains about 12 hours of audiovisual data (video, audio and text, MOCAP etc.,) which consist of the improvised or the performance date. And it is one of the largest open multimodal databases which are available in emotional recognition. This paper references [30] and chooses the four most representative emotion states as the experimental emotional categories: anger, excite, neutral, and sad, which account for more than 70% of the dataset.

Because of the low authenticity and exaggeration of the performance data, there is still a certain distance between the performance data and people's normal emotions in daily communication. Compared with performance data, improvised data are more authentic. Therefore, considering the authenticity of the data, we use improvised data for speech emotion recognition.

4.1.2. Evaluation Matrices. In this paper, confusion matrixes are used to analyse the states of four kinds of emotions. Each vertical row represents the real result, and each horizontal row represents the predicted result. Confusion matrix is a kind of visual tool from which we can find the recognition and confusion of each kind of emotion. In this paper, we combined weighted accuracy (WA) and unweighted accuracy (UA) with a confusion matrix to judge the validity of the model.

4.2. Experiment Settings

4.2.1. Text Emotion Recognition Models. We set up two models (T-LSTM and T-BL) to test the performance of the text emotion recognition model. The text-LSTM (T-LSTM) model used the LSTM layer. The text-Bi-LSTM (T-BL) model used the bidirectional LSTM layer. Both models employed the softmax layer for sentiment classification. The number of LSTM and Bi-LSTM cells is all set as 256.

4.2.2. Speech Emotion Recognition Models. Four models (S-CNN, S-DNN, S-CL, and S-CBLA) were put up for speech emotion recognition. The audio data came from impromptu conversation which contains lots of context-related information. Therefore, we utilized the dual channel networks of CNN and Bi-LSTM with attention mechanism in our Speech-CBLA (S-CBLA) model. In addition, as a comparison, we set up model Speech-DNN (S-DNN), Speech-CNN (S-CNN), and Speech-CNN-LSTM (S-CL). Model S-CL used common single-channel combination modes of CNN and Bi-LSTM. Model S-DNN was built with four dense layers with parameters of 512, 256, 128, and 32. Model S-CNN, whose structure was similar with the CNN channel in model CBLA, contained convolutional layers, max pooling layer, and global average pooling layer. The activation function was set as Leaky ReLU. Comparison of different models on accuracy is shown in Figure 5. From it, we can find that, with the improvement in the model, the performance becomes better on most emotional categories.

4.2.3. Multimodal Emotion Recognition Models. For the modal fusion, we adopted the feature-level approach mentioned in the previous section. The purpose of dense layers was to synthesize the features extracted from the front layers. Each node of the dense layer was connected to all the nodes of the upper layers. The arguments of units were set as 1024, 512, and 4.

The multimodal-CBLA-Bi-LSTM (M-CBLA-BL) model combined S-CBLA and T-BL for speech and textual feature processing. After fusing the high-level features from different modes, we used DNN to extract the correlation between the multimodal features by nonlinear change. Finally, we used the softmax layer to classify. Moreover, the multimodal-dense-Bi-LSTM (M-D-BL) model combined S-DNN and T-BL, and multimodal-CNN-LSTM-Bi-LSTM (M-CL-BL) model combined S-CL and T-BL. All the other models employed the same fusion approach as model M-D-BL. Model M-CL-BLR and M-CBLA-BLR added L_2 regularization on the basis of M-CL-BL and M-CBLA-BL. Comparison of different multimodal models on accuracy is shown in Figure 6.

5. Experimental Results

We trained the model using the dialogue data Sessions 1–4 in the IEMOCAP database, tested with Session 5, and finally obtained the results shown in Table 1. We set the number of epochs to 20, the batch size to 64, and ReLU as the activation function in the models. S-DNN, T-LSTM, and M-D-BL are the baseline models.

Comparing with baseline models, binary channels of CNN and Bi-LSTM performs best, verifying the effectiveness of the model S-CBLA we proposed in SER. Model S-CNN shows the superiority of the CNN network. Model S-CBLA shows that combining CNN with Bi-LSTM could improve the performances of the SER model. Model T-BL confirms Bi-LSTM is more accurate than LSTM in text emotion recognition. Combining two best models for speech and text emotion recognition, model M-CBLA-BL obtains the better accuracy and proves the validity of our fusion model. Model M-CBLA-BLR trained the fused multimodal features with a deep neural network and applied L_2 regularization to optimize the model. The comparison of model M-CBLA-BL and M-CBLA-BLR proves the utility of L_2 regularization. The specific experimental results are illustrated in Table 1.

After training the model S-CBLA, T-BL, and M-CBLA-BLR, we obtained the confusion matrixes in Figures 7 and 8. As shown in Figures 7(a) and 7(b), there is obvious emotional confusion in single mode. Excited is easily misjudged as angry in speech mode, while they are easy to distinguish in the text mode. From here, we could find that there is complementarity between text and speech. As shown in Figure 8, by merging acoustic and text emotional features, the accuracy of most emotion types improved, and the

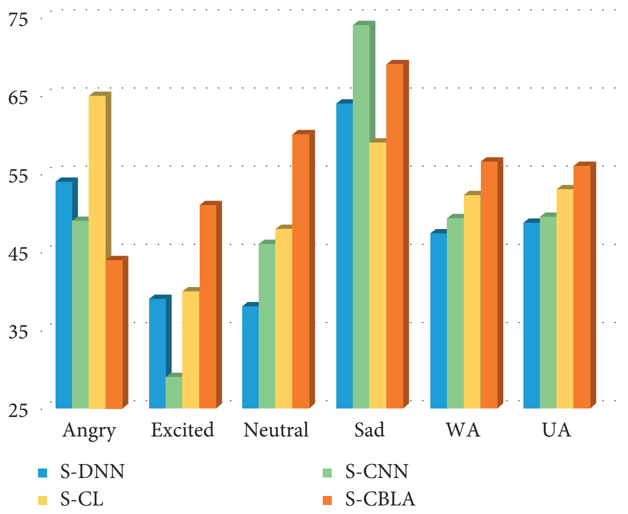


FIGURE 5: Comparison of different speech emotion recognition models on accuracy.

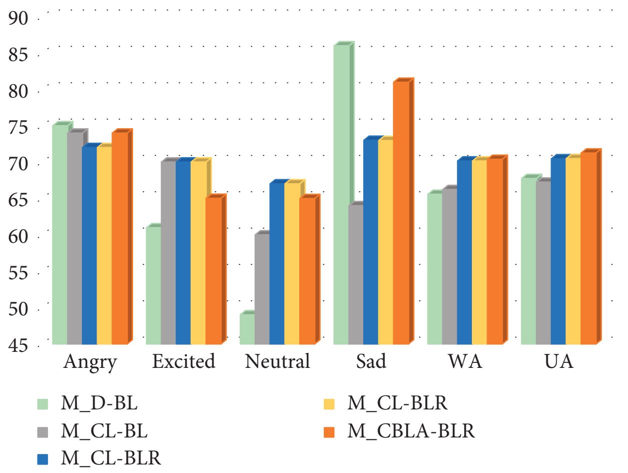
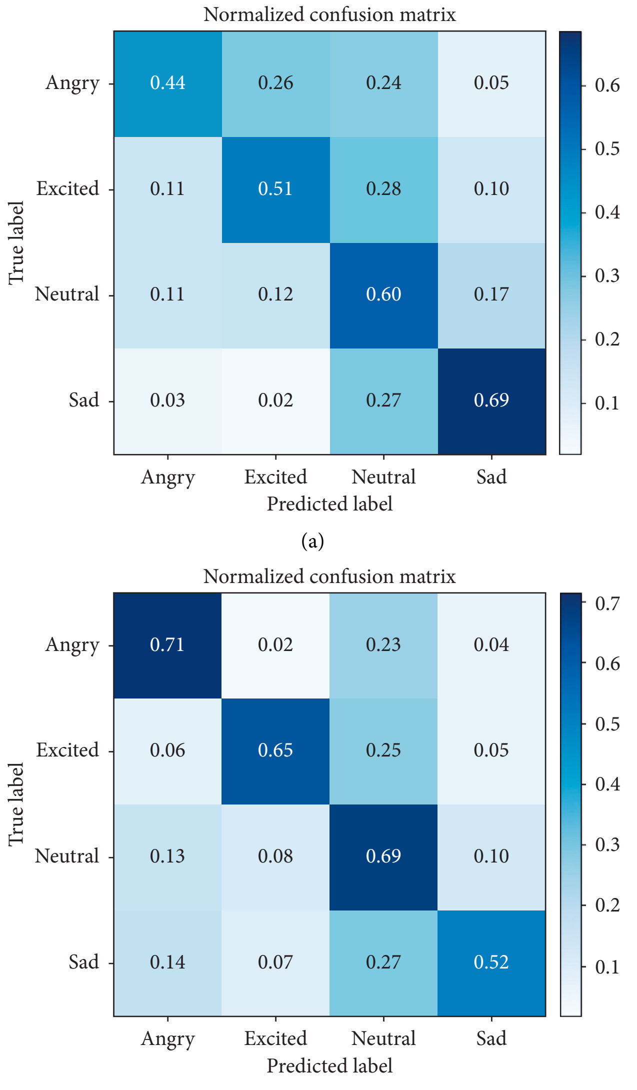


FIGURE 6: Comparison of different multimodal models on accuracy.

TABLE 1: The comparison of the experimental results.

Model	Accuracy (%)					
	Angry	Excite	Neutral	Sad	WA	UA
S_DNN	54	39	38	64	47.37	48.75
S_CNN	49	29	46	74	49.26	49.50
S_CL	65	40	48	59	52.23	53.00
S_CBLA	**44**	**51**	**60**	**69**	**56.55**	**56.00**
T_LSTM	68	59	65	61	63.56	63.25
T_BL	**71**	**65**	**69**	**52**	**63.70**	**64.25**
M_D_BL	75	61	49	86	65.59	67.76
M_CL_BL	74	70	60	64	66.26	67.25
M_CL_BLR	72	70	67	73	70.18	70.50
M_CBLA_BL	77	75	57	77	69.37	71.50
M_CBLA_BLR	**74**	**65**	**65**	**81**	**70.40**	**71.25**

confusion of emotion has been alleviated. It indicates the validity of modal fusion. The experimental results show that modal fusion could effectively reduce emotional confusion and improve emotional recognition rate.



(a)

(b)

FIGURE 7: The confusion matrixes of single mode models: (a) model S-CBLA; (b) model T-BL.

In order to test the performance of multimodal emotion recognition model proposed in this paper, we compared it with other different models on the IEMOCAP database. Soujanya Gu et al. [20] applied CNN-LSTM to process the speech date and CNNs for the textual features learning; finally, they integrated all features and trained them with a three-layer deep neural network. They adopted the feature fusion method which we also referenced. Shah et al. [31] improved traditional restricted Boltzmann machine (RBM). They set up RSM (replicated softmax model) by changing the binary variables in visual unit of RBM to softmax. The RSM model was used to analyse and get features from the data of speech and text on IMOCAP. For multimodal classification, they employed SVM classifier for a decision-level fusion approach.

Compared with [20, 31] based on the IMOCAP database, the model presented in this paper performs better, as shown in Table 2. For the modal fusing, the feature-level and decision-level fusion method are all useful. In view of the

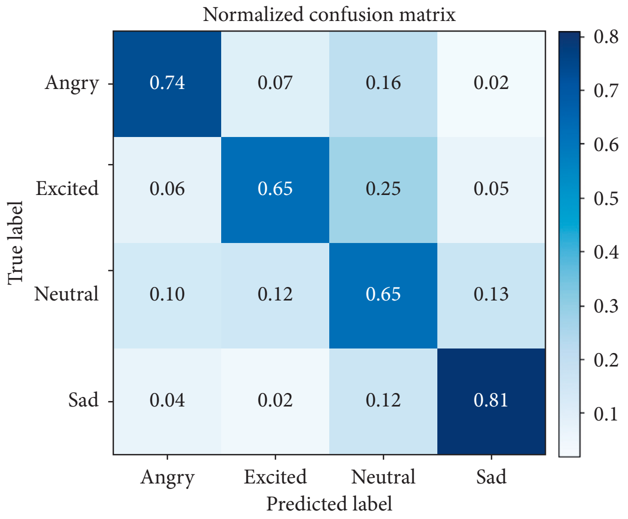


Figure 8: The confusion matrixes of model M-CBLA-BLR.

Table 2: The comparison of different methods.

Method	WA	UA
CNN-LSTM + CNN (Gu et al.) [20]	60.40	59.88
RSM + SVM (Shah et al.) [31]	—	61.96
Our M_D_BL	65.59	67.76
Our M_CL_BL	66.26	67.25
Our M_CL_BLR	70.18	70.50
Our M_CBLA_BL	69.37	71.50
Our M_CBLA_BLR	**70.40**	**71.25**

characteristics of speech and text itself, we proposed the more suitable models to extract high-level features for the multimodal emotion recognition. Compared with [20], Bi-LSTM is more effective than CNN for text emotion mining. CBLA takes full use of both global and local features of audio and avoid effective data losing. Considering that all the studies above did not consider the problem of overfitting of the model, we adopt the L_2 regularization to optimize the model. Our multimodal model is better than [20] and may be related with the use of the regularization method. The experimental results prove that the use of L_2 regularization is useful, and the proposed model is effective.

6. Conclusion and Future Work

With the development in social media, the research of multimodal emotion recognition has attracted the attention of many scholars. Since the emotion information contained in the single mode is limited, in this paper, a multimodal emotion recognition model based on speech and text on the IEMOCAP database was proposed. The model used the dual channel of CNN and LSTM to learn acoustic emotion features. It also applied Bi-LSTM to extract textual features. Moreover, a deep neural network was employed to learn the fusion features. The model used the deep learning networks, took advantage of both hand-crafted and high-level features, and considered the global and contextual temporal information in the data too. In addition, the L_2 regularization was also used to optimize the model. Experimental results showed that the recognition accuracy of our multimodal model is higher than that of the single modal and outperforms other published multimodal models on the test datasets. In the future, it is meaningful to explore more effective feature fusion methods to improve the multimodal model performance. Besides, we will try to optimize the speech modal model, hoping to get more effective audio emotion features. In the further study, we will also consider others modal in the multimodal emotion recognition.

Data Availability

The IEMOCAP database used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Key R&D Program of China (2017YFE0123000) and the Innovation Project of Graduate Research in Chongqing (no. CYS19273).

References

[1] M. Hasan, E. Rundensteiner, and E. Agu, “Automatic emotion detection in text streams by analyzing twitter data,” *International Journal of Data Science and Analytics*, vol. 7, no. 1, pp. 35–51, 2019.

[2] M. A. Razek and C. Frasson, “Text-based intelligent learning emotion system,” *Journal of Intelligent Learning Systems and Applications*, vol. 09, no. 1, pp. 17–20, 2017.

[3] C.-H. Chen, W.-P. Lee, and J.-Y. Huang, “Tracking and recognizing emotions in short text messages from online chatting services,” *Information Processing & Management*, vol. 54, no. 6, pp. 1325–1344, 2018.

[4] J. K. Rout, K.-K. R. Choo, A. K. Dash, S. Bakshi, S. K. Jena, and K. L. Williams, “A model for sentiment and emotion analysis of unstructured social media text,” *Electronic Commerce Research*, vol. 18, no. 1, pp. 181–199, 2018.

[5] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, “Emotion recognition from speech using global and local prosodic features,” *International Journal of Speech Technology*, vol. 16, no. 2, pp. 143–160, 2013.

[6] W. Zheng, M. Xin, X. Wang, and B. Wang, “A novel speech emotion recognition method via incomplete sparse least square regression,” *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 569–572, 2014.

[7] S. Gupta, A. Mehra, and Vinay, “Speech emotion recognition using SVM with thresholding fusion,” in *Proceedings of the 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 570–574, Noida, India, February 2015.

[8] J.-S. Park, J.-H. Kim, and Y.-H. Oh, “Feature vector classification based speech emotion recognition for service robots,” *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, pp. 1590–1596, 2009.

[9] K. Lu and Y. D. Jia, "Audio-visual emotion recognition with boosted coupled HMM," in *Proceedings of the Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pp. 1148–1151, Tsukuba, Japan, November 2012.

[10] S. S. Narayanan, S. Lee, and A. Metallinou, "Audio-visual emotion recognition using Gaussian mixture models for face and voice," in *Proceedings of the 2008 Tenth IEEE International Symposium on Multimedia*, pp. 250–257, Berkeley, CA, USA, December 2008.

[11] D. Li and J. Qian, "Text sentiment analysis based on long short-term memory," in *Proceedings of the 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI)*, pp. 471–475, Wuhan, China, October 2016.

[12] L. L. Chao, J. H. Tao, M. H. Yang, Y. Li, and Z. Wen, "Long shot term memory recurrent neural network based on encoding method for emotion recognition in video," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2752–2756, Shanghai, China, March 2016.

[13] J. Zhao, X. Mao, and L. Chen, "Learning deep features to recognise speech emotion using merged deep CNN," *IET Signal Processing*, vol. 12, no. 6, pp. 713–721, 2018.

[14] L. Q. Cai, X. L. Liu, F. L. Chen, and M. Xiang, "Robust human action recognition based on depth motion maps and improved convolutional neural network," *Journal of Electronic Imaging*, vol. 27, no. 5, Article ID 051218, 2018.

[15] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4580–4584, Brisbane, Australia, April 2015.

[16] G. Trigeorgis, F. Ringeval, R. Brückner et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200–5204, Shanghai, China, March 2016.

[17] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 439–448, Barcelona, Spain, December 2016.

[18] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[19] L. C. Woo, K. Y. Song, J. Jeong, and W. Y. Choi, "Convolutional attention networks for multimodal emotion recognition from speech and text data," 2018, https://arxiv.org/abs/1805.06606.

[20] Y. Gu, S. Chen, and I. Marsic, "Deep multimodal learning for emotion recognition in spoken language," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, April 2018.

[21] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 112–118, Athens, Greece, December 2018.

[22] A. Metallinou, A. Katsamanis, M. Wöllmer, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification (Extended abstract)," *International Conference on Affective Computing and Intelligent Interaction*, pp. 463–469, 2015.

[23] Y. S. Wang, S. Ying, L. Zhun, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: dynamically adjusting word representations using nonverbal behaviors," 2019, https://arxiv.org/abs/1811.09362.

[24] T. Giannakopoulos, "PyAudioAnalysis: an open-source python library for audio signal analysis," *PLoS One*, vol. 10, no. 12, Article ID 0144610, 2015.

[25] J. Pennington, R. Socher, and C. Manning, "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] R. Ge, C. H. Wang, X. Xu et al., "Action recognition with hierarchical convolutional neural networks features and bi-directional long short-term memory model," *Control Theory & Applications*, vol. 34, no. 6, pp. 790–796, 2017.

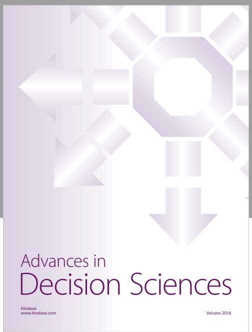
[28] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

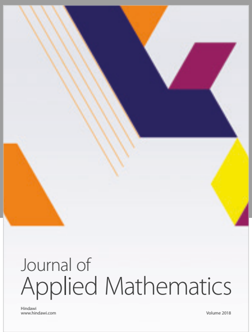
[29] C. Busso, M. Bulut, C.-C. Lee et al., "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[30] V. Chernykh and P. Prihodko, "Emotion recognition from speech with recurrent neural networks," 2017, https://arxiv.org/abs/1701.08071.

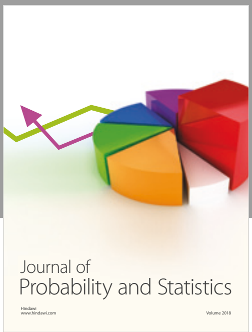
[31] M. Shah, C. Chakrabarti, and A. Spanias, "A multi-modal approach to emotion recognition using undirected topic models," in *Proceedings of the 2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 754–757, Melbourne, Australia, June 2014.

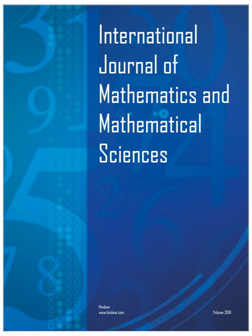
Advances in Operations Research


Advances in Decision Sciences

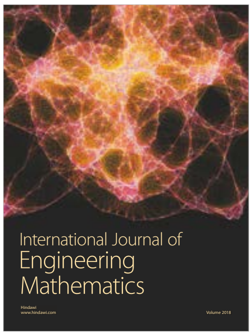

Journal of Applied Mathematics

The Scientific World Journal

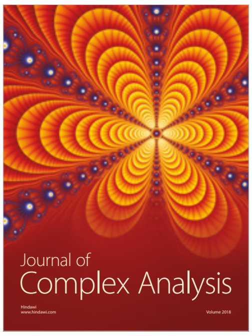

Journal of Probability and Statistics

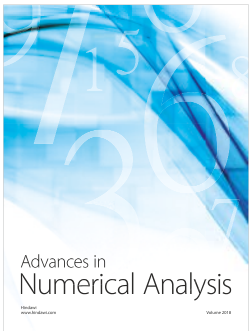

International Journal of Mathematics and Mathematical Sciences

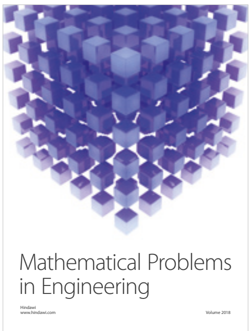
Journal of Optimization

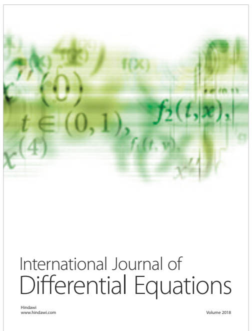

International Journal of Engineering Mathematics

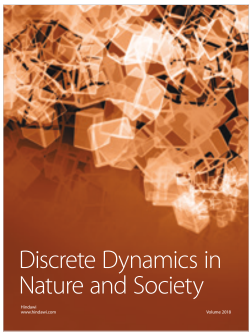
International Journal of Analysis

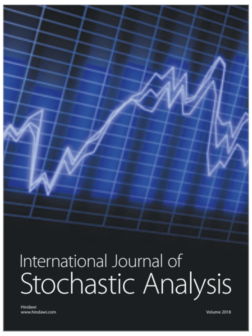

Journal of Complex Analysis


Advances in Numerical Analysis


Mathematical Problems in Engineering

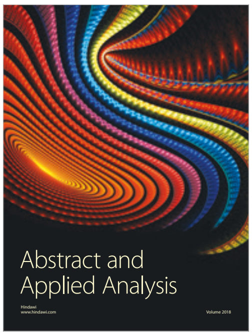

International Journal of Differential Equations

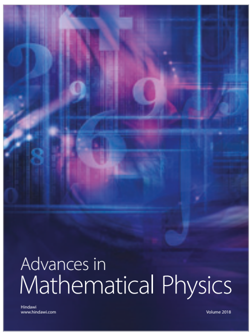

Discrete Dynamics in Nature and Society


International Journal of Stochastic Analysis

Journal of Mathematics

Journal of Function Spaces


Abstract and Applied Analysis


Advances in Mathematical Physics