Education Theory: Teaching and Learning (ISSN Online: 2771-9030)
June 2022 Vol.1, No.3, pp.99-107
DOI: https://doi.org/10.55571/ettl.2022.06020

4D Publishing Group

Speech Emotion Recognition of Minnanese Based on Deep Learning

Xin-hua Guo* **and Miao Zhang**
Software College, Quanzhou University of Information Engineering, Quanzhou 362000, China

*Correspondence:
350087154@qq.com

Abstract. Minnanese is a representative branch of Fujian dialect.Due to objective reasons such as the late start of the research on emotion recognition of Minnanese speech and the immaturity of technology, there are few research achievements in this area, especially in the area of corpus resources of Minnanese language. Based on the establishment of Minnanese language emotion corpus, soft-Max is adopted as the algorithm of emotion recognition, and emotion recognition is carried out based on CNN model and BLSTM model. In view of the global and temporal characteristics of Minnanese language corpus, an algorithm model based on CNN-BLSTM fusion feature is proposed.Experimental results show that CNN-MOBLSTM fusion feature model is superior to CNN model and BLSTM model in recognition of Minnanese emotion, and it is an effective algorithm model.

Keywords. Deep Learning, Minnanese, Emotion recognition, CNN-MOBLSTME

1. Introduction

Speech emotion recognition has been around for more than three decades and has been a hot topic of artificial intelligence research in recent years [1]. Speech emotion recognition refers to a comprehensive technique that uses computers to process speech in frames, extract features of speech, simulate human perception, understand human emotions, and then derive a classification as a certain type of speech emotion [2]. Nowadays, dialect emotion recognition, as an important branch of speech emotion recognition, is also gradually being emphasized in the field of speech emotion recognition research. In recent years in China, some scholars have achieved some results on the research of dialect emotion recognition. For example, Zhang Ce et al. established a speech recognition system for Chongqing dialect based on HMM as an acoustic model [3]; Yang Bo established a speech recognition system for Guilu dialect based on RNN [4]; Zhu Guanglu established a speech synthesis system for Minnanese based on association rules [5]; Zhang Hongwei implemented a speech recognition system for Mongolian based on deep neural network [6]; Ji Changpeng implemented a speech recognition system based on improved BP- Adaboost and HMM hybrid model for sentiment recognition system of Yangtze dialect [7]. There are still few studies on deep learning for emotion recognition in Minnanese language.

Minnanese originated from ancient Chinese and is one of the Min languages and the first major dialect in Fujian. The research on speech emotion recognition of Minnanese is not only important for the study of Minnan culture and the protection of intangible cultural heritage, but also has a wide

Education Theory: Teaching and Learning (ISSN Online: 2771-9030)
June 2022 Vol.1, No.3, pp.99-107
DOI: https://doi.org/10.55571/ettl.2022.06020

application prospect in practical industries such as tourism, car navigation, medical and health care, and service industries.

At present, there are few studies on emotion recognition of Minnanese based on deep learning, mainly because of the lack of Minnanese emotion database, which makes it impossible to carry out the research on Minnanese emotion recognition. In this paper, a Minnanese emotion database is established by referring to the Chinese speech database (CASIA) and adopting the international standard speech database its production standard, with Xiamen dialect as the representative. Soft-max is used as the algorithm for emotion recognition, and three methods based on CNN model, BLSTM model, and fusion feature model of CNN-BLSTM are used for emotion recognition in the Minnanese emotion database.

2. Establishment of the emotional corpus of Minnanese

2.1. Design of text corpus

When building the Minnanese speech corpus, it is necessary to combine the characteristics of Minnanese to select representative text corpus contents, and also to make the established text material emotion classification as even and rich as possible. The text corpus mainly uses existing stable and mature emotion dictionaries, such as microblogs, online forums, news information, etc. The vocabulary involved is as lifelike as possible, which is better to highlight the phonetic characteristics of Minnanese. The content of the text material collected above is then screened by machine and proofread manually, the word frequency in the content cannot be too high or too low, and the sound-hypophone combination of the text covers the common pronunciation phenomena in life as much as possible, finally forming a text corpus.

2.2. Recording of speech corpus

In order to make the emotion type of speech clearer and the distribution of emotion more even, the pre-prepared text corpus is used for reading aloud to record the Minnanese speech corpus, and the steps are as follows.

(1) Pre-preparation: i.e., preparation of the text corpus, recording equipment is a professional recorder, and the venue is a quiet laboratory, or classroom, etc.

(2) Selection of the recorders: five students of Xiamen domicile who are enrolled in our university and whose voices are bright and clean are selected.

(3) Recording: let the pronouncers use their lunch break to select the text material in (1), read it aloud and record it, and the recording time should not exceed 30 minutes per person each time.

After the recording, the materials of each pronouncer were named by pronouncer + material name (e.g. pronouncer 1 + text 1) and collected by category.

2.2.2. Speech pre-processing

In the preliminary obtained speech data, the tool Adobe Audition can be used to noise reduction processing, and then after kerning as well as finishing before use. In the order of text corpus annotation, the audio processing software tool CoolEdit Pro is used to cut the original speech data into several dialect speech corpus in sentence order.

The phonetic annotation was done using the "Pinyin scheme of Minnanese" in the common dictionary of Minnanese written by Baoqing Lin [9], using Praat software, with manual annotation as the main method, which needs to contain letters of information related to the pronouncer, phonetic data, phonetic-emotional features, and features of the syllable layer.

In the paper, a discrete description model is used to classify and store the annotated speech data with six categories of emotions with corresponding emotion names as file names, about 4711 speech data, and the insufficient ones are manually complemented by 89, totaling 4800 speech data. There are 800 sentiment data in each category, and the training set and test set are assigned by 3:1, i.e., there are 600 sentiment data in the training set and 200 in the test set. The composition of each type of sentiment corpus is shown in Table 2-1 below.

Education Theory: Teaching and Learning (ISSN Online: 2771-9030)
June 2022 Vol.1, No.3, pp.99-107
DOI: https://doi.org/10.55571/ettl.2022.06020

Table 2-1. Types of emotions Corpus' composition

Emotion classification	Anger	Happy	Fear	Neutral	Sadness	Surprised
Number of statements	800	800	800	800	800	800

3. Minnanese Emotion Recognition Related Technologies

3.1. Speech emotion features and algorithms

3.1.1. Energy and fundamental tone frequency

In speech signal endpoint testing, energy is often tested using two major types of patterns, namely short-time energy and short-time average over zero. Short-time energy not only shows the clear and turbid amplitude change of speech, signal strength, but also reflects the pause and accent. Short-time average over-zero rate, which refers to the number of times a speech signal waveform crosses the zero value per frame, is used to roughly reflect the spectral characteristics of speech signals, and is used to distinguish between clear and turbid tones and voiced and unvoiced characteristics in speech data.

The Fundamental Frequency of the Phonation (F0) [11] is the inverse of the fixed interbasic period for each start and shutdown of the vocal folds in human vocal tract rhymes. The basal frequency is generally used to discourse the information on the emotional activation of the part of speech. In the paper the most commonly used calculation method ACF method (own correlation method) is used.

3.1.2. Mel side spectral coefficients

The specific relationship between f_{mel} frequency and the actual frequency g is as follows.

$$f_{mel}(g) = 2595 \times log\left(1 + \frac{g}{700}\right) \quad (1)$$

The overall process of MFCC (Mel side spectral coefficient) feature extraction is as follows.

(1) The pre-processing of the signal contains pre-emphasis, frame splitting and windowing. The frame length can be set to 25ms if the signal is considered stable at 10ms~30ms.

(2) The FFT transform is performed to find the spectrum, i.e., the data signal of each frame is FFT transformed.

(3) Mel spectrum can be obtained by Mel filter set, i.e., the spectrum obtained in the previous step is obtained by operation.

(4) The Mel spectrum obtained in the previous step is subjected to logarithmic operation (Logarlithm);

(5) The value obtained from the logarithm operation is calculated again using the discrete cosine;

(6) The coefficients of the 2nd to 13th are retained, and the remaining 12 coefficients are used as the parameters of MFCC.

3.1.3. Speech emotion recognition algorithms

Many speech emotion recognition algorithms have been used as speech emotion recognition classifiers, among which the more classical ones are HMM (hidden Markov model, for temporal signals), GMM (Gaussian mixture model, which fits the data better) [12], ANN (artificial neural network, which approximates nonlinear relationships) [13], SVM (support vector machine, for small sample training), and Soft-max. Soft-max is a logistic (also a regression classifier) derived and supervised regression classifier, often used for multi-classification problems [14]. Therefore, Soft-max will be used as a regression classifier for deep learning in the paper.

3.2. Theoretical basis of deep learning

Deep learning, as a branch of machine learning, is a collection of framework algorithms based on multilayer neural networks combined with various machine learning algorithms to solve various

Education Theory: Teaching and Learning (ISSN Online: 2771-9030)
June 2022 Vol.1, No.3, pp.99-107
DOI: https://doi.org/10.55571/ettl.2022.06020

practical problems represented by images and texts. The core of deep learning is feature learning. Compared with traditional machine learning models, deep learning can extract multi-level feature structures, which can be more complex. Deep learning includes several important algorithms: CNN, RNN, Sparse Coding, DBN, RBM, AutoEncoder. Next, we introduce CNN (Convolutional Neural Network) and LSTM, which evolves from RNN.

3.2.1. Convolutional neural network

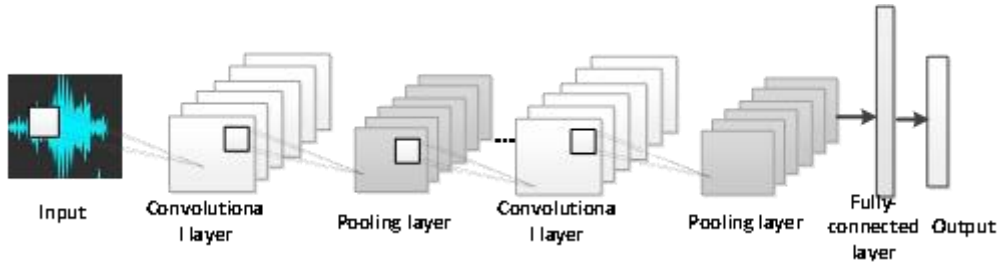

Fig. 3-1. Basic structure diagram for CNN

CNN (Convolutional Neural Network) is a layer network that changes the function and form of layers based on the traditional neural network.The structure of CNN contains data input layer, convolutional layer, pooling layer, fully connected layer and output layer [15], as shown in Figure 3-1. The convolutional layer contains at least one layer as the core part of CNN, and the computation of convolution is its important operation. Convolutional computation and pooling computation can largely reduce the complexity of the network structure, simplify the size of the parameter matrix, and thus achieve faster convergence.

3.2.2. Long and Short-Term Memory Networks and MOBLSTM

LSTM (Long Short Term Memory) network was proposed in 1997 [16] and evolved from RNN, which added Input Gate, Forget Gate and Output Gate to solve the problem of long-term dependence of short-term memory in RNN. The LSTM can only retain the previous information in the memory layer, but the future information cannot be retained, while the BLSTM, which is a two-way short-term memory network, can use the forward LSTM and the backward LSTM to extract the future information and the past information, which can better solve the above problem. Since each layer of BLSTM has output, if the output of each layer is transformed by fully connected layers to obtain the summation of the features of each layer, the output results will contain richer information of upper and lower layers. In this paper, MOBLSTM (Multiple Output Bidirectional Long S hort-Term Memory, MOBLSTM), a multilayer bidirectional short-term memory network, is proposed to improve BLSTM. If the BLSTM contains M layers, the real-time statistical features are input to the BLSTM, and the last output of each layer is selected to obtain 2* 2M feature vectors, and then the FC is transformed and summed to obtain a feature vector. the structure of MOBLSTM model is shown in Figure 3-5, and the figure shows 2 layers with 4 outputs.

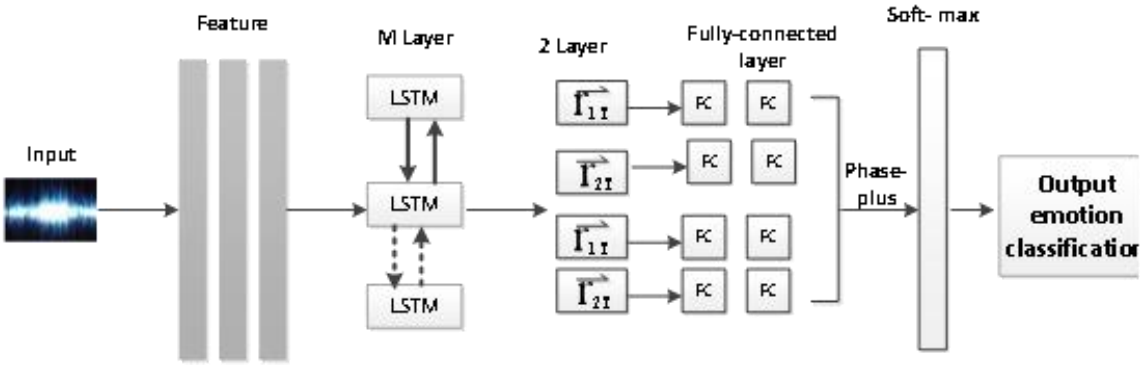

Fig. 3-5. structure diagram for MOBLSTM model

Education Theory: Teaching and Learning (ISSN Online: 2771-9030)
June 2022 Vol.1, No.3, pp.99-107
DOI: https://doi.org/10.55571/ettl.2022.06020

3.3. Evaluation Metrics

There are generally three representative model evaluation metrics in classification problems: accuracy, precision and recall. In the paper, the accuracy rate is used as the classification performance evaluation index with the following formula.

$$W_{AC} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad (2)$$

T_P 、T_N: the actual is consistent with the prediction when the prediction is positive or negative; F_P 、F_N: the actual is inconsistent with the prediction when the prediction is positive or negative.In other words, T_P and T_N and are predicted correctly, while F_P and F_N are predicted incorrectly. The accuracy rate is the percentage of the total number of correct predictions.

4. Feature Fusion based on CNN-MoBLSTM model for Speech emotion Recognition in Minnanese

In order to compare the influence of CNN, MOBLSTM and CNN-MOBLSTM on the speech emotion recognition of Minnanese, the three models were trained respectively and the experimental results were analyzed.

4.1. Speech emotion Recognition of Minnanese based on CNN model

In this paper, openSMILE software is used to extract the phonetic features of Minnanese. The log-MEL spectrogram was extracted as the input of CNN. The structure of CNN model is shown in Figure 4-1, and its specific process is as follows:

(1) The log-mel spectrogram is extracted from the original speech signal, and the specific process is described in Section 3.1.2. The parameters are set: the frame length and frame shift are set to 25ms and 10ms, respectively, and the number of filters in the mel filter set to 40, with a Hamming window and a frequency range of 0~8KHZ.

(2) The log-mel spectrogram obtained from (1) is used as the CNN input layer: the size of each spectrogram segment is 50*40.

(3) Then the data is fed into the convolution layer with the parameters: the filter is set to 32 and the size is 5*40. The output feature sequence can be obtained as 32 49-dimensional data.

(4) Send the result of (3) to the maximum pooling layer with the parameters: size 1*3. The output feature sequence is 32 47-dimensional data.

(5) Repeat (3)(4) three times, i.e., the convolution layer and the maximum pooling layer are operated three times.

(6) Feed the result of (5) into the average pooling layer and perform average pooling once.

(7) Send the result of (6) to the fully connected layer and perform the fully connected layer four times.

(8) Send the result of (7) through the softmax classifier for classification.

(9) Output the results.

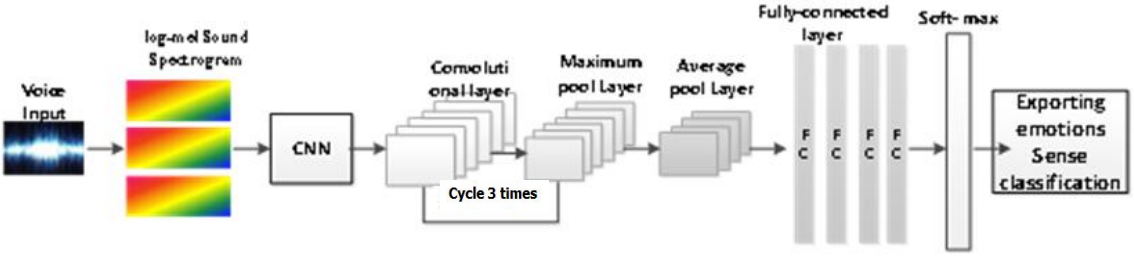

Fig. 4-1. Structure of CNN model

The CNN model training recognition results are shown in Table 4-1:

Tab. 4-1. CNN model recognition results

Education Theory: Teaching and Learning (ISSN Online: 2771-9030)
June 2022 Vol.1, No.3, pp.99-107
DOI: https://doi.org/10.55571/ettl.2022.06020

Results / Emotion	Anger	Fear	Happiness	Nature	Sadness	Surprise	Accuracy
Anger	193	0	2	0	3	2	96.50%
Fear	8	170	5	7	6	4	85.00%
Happiness	3	1	184	2	7	3	92.00%
Nature	5	2	4	183	4	2	91.50%
Sadness	1	8	10	0	172	9	86.00%
Accuracy	5	3	1	5	4	182	91.00%
Total Accuracy	90.3333%						

4.2. MOBLSTM Model-based Speech Emotion Recognition in Southern Fujian Dialect

4.2.1. Statistical features

In the paper, the InterSpeech Challenge 2010 features (IS10) are used to extract the speech features of Minnanese using openSMILE software. The input speech data were converted into frames, and the frame length and frame shift were set to 25ms and 10ms, respectively, using Hamming windows. It contains 1 dimension of over-zero rate, 1 dimension of energy, 1 dimension of fundamental frequency, 1 dimension of harmonic noise ratio, 3 dimensions of resonance peaks (mean, average rate of change and mean square deviation), 12 dimensions of MFCC coefficients, the above-mentioned 19-dimensional LLD is obtained; then the first-order difference calculation is performed to obtain 38-dimensional LLD; finally, based on these 38 dimensions, 12 statistical functions (maximum value, minimum value, mean value, standard deviation, etc.) are used to obtain a total of 456-dimensional speech emotion features. The structure of MOBLSTM model is shown in Figure 3-5 in Chapter 3, where the value of M is 2.

MOBLSTM network model with the following settings:

(1) Input layer: 456-dimensional statistical features from the InterSpeech Challenge 2010 features (IS10) are used as input.

(2) Recurrent layer: BLSTM is 2 layers, and the number of hidden units of BLSTM in each layer is set to 128. The output data of the final moment of each layer is selected and spliced to obtain 512-dimensional feature vectors.

(3) Using the final output of each layer of (2), the FC layer is transformed and then summed to obtain a feature vector.

(4) The results of (3) are classified by a softmax classifier.

(5) Output results.

The MOBLSTM model training recognition results are shown in Table 4-2:

Tab. 4-2. MOBLSTM model recognition results

Results / Emotion	**Anger**	**Fear**	**Happiness**	**Nature**	**Sadness**	**Surprise**	**Accuracy**
Anger	195	0	1	0	3	1	97.50%
Fear	4	186	5	3	0	2	93.00%
Happiness	3	1	183	2	8	3	91.50%
Nature	6	2	4	182	4	2	91.00%
Sadness	1	8	10	0	172	9	86.00%
Accuracy	4	1	0	1	3	191	95.50%
Total Accuracy	92.4167%						

Education Theory: Teaching and Learning (ISSN Online: 2771-9030)
June 2022 Vol.1, No.3, pp.99-107
DOI: https://doi.org/10.55571/ettl.2022.06020
4D Publishing Group

4.3. Emotion recognition of Minnanese speech based on CNN-MOBLSTM feature fusion model

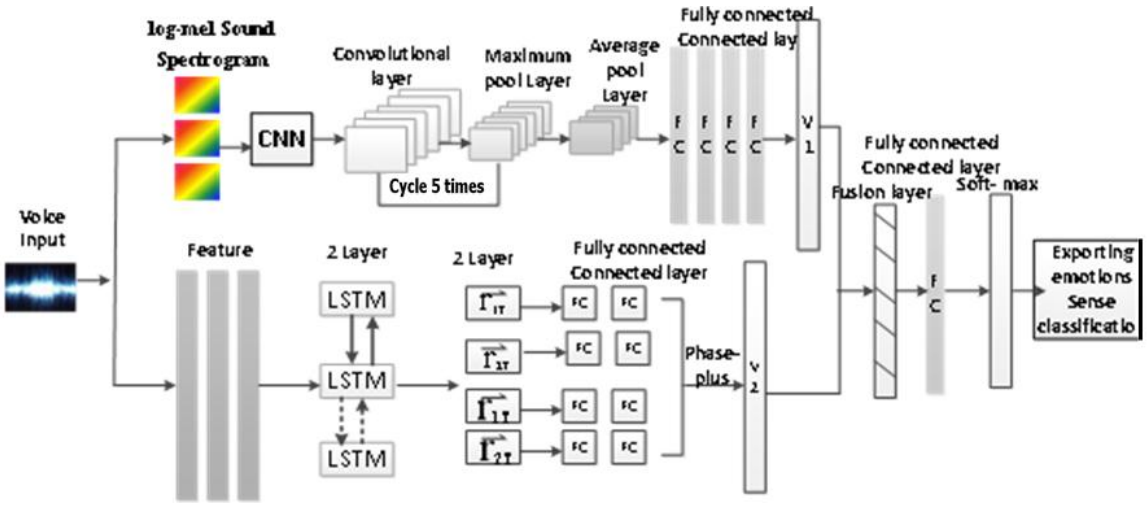

Fig.4-2. Structure of CNN model 's feature fusion

In the pre-training period, the temporal length and the number of temporal segments need to be determined as the input first. In the Minnanese speech corpus, each utterance generally does not exceed 3-5 seconds, and it is shown that the speech segments with not less than 250ms generally have more qualified emotional information, so 500ms is selected as the basic temporal length for cutting the original corpus in the paper. The number of fragments of an utterance segmented in the model is set to T. The feature fusion model structure of CNN-MOBLSTM is shown in Figure 4-2.

(1) In the CNN model, the log-mel acoustic spectrogram is extracted from the original speech signal, and the specific procedure is described in Section 3.1.2, and its parameters are set consistently.

(2) The [sr(1), sr (2),..., sr (T)] log-mel sound spectrum map (where sr(t) segments of the sound spectrum map, the number of segments is T) obtained from (1) is used as the CNN input layer: the size of each sound spectrum map segment is 50*40.

(3) Then three cyclic operations of convolution and maximum pooling are performed, where the filter is set to 32 with size 5*40 and the parameter size of the maximum pooling layer is 1*3, and the average pooling is performed once, and then the fully connected layer is performed four times.

(4) The feature vector sequence [sc(1), sc (2),..., sc (T)] is obtained from (3), and then the feature vector V1 is obtained by averaging all the CNN-based sc(t) segment features in a sentence of speech. its formula is as follows:

$$V1 = \sum_{t=1}^{T} \frac{sc(t)}{T}$$

(5) MOBLSTM input layer: 456-dimensional statistical features from the InterSpeech Challenge 2010 features (IS10) are used as input. The input sequence is noted as [cr(1), cr (2),..., cr (T)]

(6) A 2-layer BLSTM is used as the loop layer: the number of hidden cells of BLSTM in each layer is set to 128.

(7) The final output timing feature sequence is obtained using each layer of (6), and a feature vector V2 is obtained by summing up while transforming using the FC layer.

(8) Send V1 and V2 to the fusion layer for fusion, and then to the fully connected layer for connection.

(9) The result of (8) is classified by a softmax classifier.

(10) Output the results.

Education Theory: Teaching and Learning (ISSN Online: 2771-9030)
June 2022 Vol.1, No.3, pp.99-107
DOI: https://doi.org/10.55571/ettl.2022.06020

Tab. 4-3. CNN-MOBLSTM model recognition results

Results / Emotion	Anger	Fear	Happiness	Nature	Sadness	Surprise	Accuracy
Anger	198	0	1	0	0	1	99.00%
Fear	4	184	7	3	0	2	92.00%
Happiness	0	4	193	0	1	2	96.50%
Nature	6	2	4	183	3	2	91.50%
Sadness	1	6	10	0	176	7	88.00%
Accuracy	0	0	0	1	0	199	99.50%
Total Accuracy	94.4167%						

4.4. Experimental Analysis

The experimental results of the three models of CNN, MOBLSTM, and CNN-MOBLSTM feature fusion on the Minnanese database are shown in Table 4-4. It can be concluded that the CNN-MOBLSTM feature fusion model is higher than the CNN and MOBLSTM models in terms of accuracy. Among them, the accuracy rate is improved by 4.08% relative to the CNN model and 2% relative to the MOBLSTM model. The results show that the CNN-MOBLSTM feature fusion model does improve the accuracy of the six categories of emotion recognition in general. Since the CNN-MOBLSTM feature fusion model contains the emotion features linking contextual information V1 and spatial information V2, it can better improve the modeling ability of speech emotion recognition compared with the CNN model and MOBLSTM model alone.

Tab. 4-4. Comparison of recognition results of different models

Models	Anger	Fear	Happiness	Nature	Sadness	Surprise	Accuracy
CNN	96.50%	85.00%	92.00%	91.50%	86.00%	91.00%	90.3333%
MOBLSTM	97.50%	93.00%	91.50%	91.00%	86.00%	95.50%	92.4167%
CNN-MOBLSTM	99.00%	92.00%	96.50%	91.50%	88.00%	99.50%	94.4167%

5. Summary and Prospect

In this paper, we realized the establishment of Minnanese emotion corpus, used Soft-max as the algorithm of emotion recognition, and used three models of CNN, MOBLSTM, and CNN-MOBLSTM fusion features to recognize emotion in Minnanese emotion corpus respectively, and the experiments showed that the CNN-MOBLSTM fusion feature model was better than CNN and MOBLSTM The CNN-MOBLSTM fusion feature model outperforms both CNN and MOBLSTM models. The study of speech emotion recognition for Minnanese in this paper is not only meaningful for the study of Minnan culture and the preservation of intangible cultural heritage, but also for the application in practice. Of course, there are shortcomings in the paper: (1) the corpus size is small, and further research can be conducted in the future through automatic expansion, dimensional expansion, etc., to continuously improve the depth and breadth of the Minnanese emotion corpus; (2) the main extracted features are all commonly used features, and later we can try to use emerging speech features to conduct in-depth research on the Minnanese emotion corpus.

References

[1] Han WJ, Li HF, Ruan HB, et al. (2014). A review of the progress of speech emotion recognition research. Journal of Software, 25(1):37-50.
[2] Sun Xiaohu,Li Hongjun. (2020). A review of speech emotion recognition. Computer Engineering and Applications, 56(11):1-9.

[3] Zhang Ce,Wei Pengcheng,Lu Xiaoyan,Shi Xi. (2018). Design and implementation of Chongqing dialect speech recognition system. Computer Measurement and Control, 26(01):256-259+263.
[4] Yang Bo. (2019). Research on speech recognition system based on RNN for Gui-Liu dialect. Modern Computer, (31):6-9+14.
[5] Zhu Guanglu. (2017). Design and implementation of a speech synthesis system for Minnan dialect. Nanjing University of Technology.
[6] Zhang Hongwei. (2017). Research on acoustic model of Mongolian speech recognition system based on deep neural network. Inner Mongolia University.
[7] Ji Changpeng. (2019). Dialect emotion recognition based on improved BP-Adaboost and HMM hybrid model. Journal of Chengdu University of Information Engineering, 34(5):495-500.
[8] Chen Liran. (2001). A preliminary study on the use of words in Minnan dialect. Jinan University.
[9] Lin Baoqing. (2007). A Dictionary of Commonly Used Mandarin Minnan Dialects. Xiamen: Xiamen University Press.
[10] Ortony A, Turner TJ. (1990). What's basic about basic emotions. Psychological Review, 97(3):315-331. doi: 10.1037/0033-295X.97.3.315.
[11] Qin, Xiji. (2007). Research on text-independent speaker recognition. Guangxi Normal University.
[12] H. Hu, M. Xu and W. Wu. (2007). GMM supervector based SVM with spectral features for speech emotion recognition. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP 07, Honolulu, HI, 2007, IV -413-IV-416.
[13] Han K, Yu D, Tashev I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. Interspeech.
[14] Chen S-C. (2013). Deep learning algorithm and application research based on convolutional neural network. Zhejiang University of Technology and Business.
[15] Li Yandong, Hao Zongbo, Lei Hang. (2016). A review of convolutional neural network research. Computer Applications, 36(9): 2508-2515, 2565.
[16] Hochreiter S, Schmidhuber J. (1997). Long Short-Term Memory. Neural computation, 9(8):1 735-1780.