Speech Emotion Recognition of Teachers in Classroom Teaching

Liang Jie[1], Zhao Xiaoyan*[1,3], Zhang Zhaohui[1,2]

1. School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China

2. Beijing Engineering Research Center of Industrial Spectrum Imaging, University of Science and Technology Beijing, Beijing 100083, China

3. Shunde Graduate School of University of Science and Technology Beijing, Fo Shan 528399, China
E-mail: zhaoxiaoyan@ustb.edu.cn

Abstract: With the development of information technology, speech emotion recognition technology was applied to the classroom evaluation, which is helpful to improve teaching quality by analyzing and quantifying evaluation indexes in real time. The paper studied teachers' speech signals and a set of emotion detection audio processing system was designed. The teachers' speech was used to judge their emotions. The recurrent neural network (RNN) algorithm was used to construct a speech emotion recognition classification model. Emotions were reclassified based on preprocessing of original data such as pre-weighting, frame-adding window and endpoint detection, so as to establish the speech emotion corpus of teacher evaluation system. By improving the traditional feature extraction process of Mel Frequency Cepstral Coefficents(MFCC), the second-order differential process was added to eliminate the convolution noise of MFCC. Especially, the 1-dimensional energy feature is added to the 39-dimensional MFCC coefficient for experiment, and the results showed that the average recognition rate of the 40-dimensional feature parameter improved 2.53% than the 39-dimensional parameter. Through experiments on the unit network structure of the classification model, the Long Short-Term Memory (LSTM) optimization model was obtained, and the average recognition rate of the five kinds of speech emotion classification reached 85.32%. Experiments showed that the improved MFCC feature value and neural network can improve the recognition rate of speech emotion more effectively than the traditional speech emotion recognition method, which can be used for speech emotion recognition in classroom teaching.

Key Words: Classroom evaluation, Speech emotion recognition, MFCC, RNN, Emotion classification

1 INSTRUCTIONS

In the 1980s, people carried out research related to speech emotions, which included many features related to emotion[1]. With the rapid development of artificial intelligence and the deepening of related research techniques, human-computer interaction has become as a hot research focus, and the ability of computers to recognize and express emotions as humans has become the goal of researchers. The traditional research on speech emotion recognition mainly includes the following aspects: the theoretical basis of speech emotion recognition and database selection, the preprocessing of speech signals, the feature extraction of speech emotion signals, and the construction of recognition model.

In terms of speech feature extraction, there are traditional methods such as prosodic feature, sound quality feature and spectral feature[2]. The literature[3] first proposed that there were many emotional factors in the prosodic features, and this conclusion was further studied in the literature[4]. The literature[5] divides prosodic features into pitch, energy, time series and pronunciation. Literature[6] proposed spectral features, in which emotional features in speech affect the distribution of energy in frequency spectrum. For example, in the speech spectrum of happy, there is higher energy in the high frequency part, while sadness has lower energy in the same frequency segment. These features are only studied in the time domain or frequency domain and cannot fully express the speech signal. Spectrogram has the advantage of combining time domain and frequency information, including a lot of information related to speech signal. Literature[7] proposed the study of speech emotion recognition based on the texture features of spectrogram, and proposed a new method for the study of speech emotion recognition by analyzing the speech emotion features with spectrogram. However, the spectrogram can not completely include the speech information and is computationally intensive. After that, the MFCC[8] is proposed, which is based on the auditory nervous system of human ear. MFCC mathematical relation formula is used to express the auditory nervous system of human ear, so as to obtain the spectral characteristics of voice emotional signals.

The designed classification recognition model is the key to speech emotion recognition. Traditional classification methods generally use Hidden Markov Model (HMM), Support Vector Machine (SVM), K Nearest Neighbor (KNN), Artificial Neural Network (ANN). HMM is a classical classification model that is widely used, but it is very computationally intensive. Corinna Cortes et al. explained the use principle of SVM in 1995. SVM mapped input vectors from low-dimensional space to high-dimensional space through kernel function, which makes the initial input realize linear classification. However, the training speed was very slow with the increase of training samples. Then the ANN model was proposed, which refers to the complex neural network formed by the

This work was supported by the Fundamental Research Funds for the Beijing Talent Training Joint Construction Project (No. GJ1804).

978-1-7281-5855-6/20/$31.00 ©2020 IEEE

interconnection of a large number of neurons. It has the advantages of high classification accuracy and fast parallel distribution processing ability. However, the neural network requires a large number of parameters and the training process is cumbersome. Then, models such as convolutional neural network (CNN) and recurrent neural network (RNN) are proposed, and the recognition effect is better.

In order to highlight the uniqueness of teachers' speech signal emotion, this paper reclassified the speech emotion samples and established the database of teachers speech emotion. The traditional speech signal features can not express the emotional information well and the slow training speed and low accuracy of traditional classifier. This paper proposes a new method of speech emotion recognition based on MFCC and RNN-LSTM. This method improves the traditional method of extracting MFCC and adds one-dimensional energy features for feature extraction, which not only enhances the emotional information of speech signals, but also improves the accuracy of emotion classification by using energy features. After data preprocessing, it is placed in the RNN model for classification and identification, and LSTM is used for model optimization. The experiments show that the above model has a greater improvement in the effect of speech emotion recognition than the traditional method.

2 ESTABLISE EMOTION DATABASE

The high-quality emotion database is the precondition for the research of speech emotion recognition. At present, in the research of speech emotion recognition, there is no consistent criterion for the speech emotion database, and there are many classification forms. At present, there are prominent representative Chinese speech emotion databases, such as CASIA and ACCorpus. This paper refers to the above emotional database, and according to the characteristics of the teacher's lectures, reclassifies the emotions and establishes a teacher's speech emotion database with teaching as the background. The database is recorded by six male and six female teachers and proposed five different emotional features: angry, sad, happy, surprise, and neutral. Each Emotion have 100 sentences, a total of 6,000 sentences. The recorded speech is saved at a sampling frequency of 16000Hz, mono channel and. wav.

3 SPEECH EMOTION RECOGNITION BASED ON RNN

The flow chart of the speech emotion recognition design scheme is shown in Figure 1. Firstly, the speech emotion signal is preprocessed and the input speech emotion signal is pre-weighted. The purpose is to emphasize the high frequency part of the speech, remove the influence of lip radiation and enhance the high frequency resolution of the speech signal; According to the short-term stability of the speech signal, it can be divided into short segments to process, and the framing is implemented by using a movable finite-length window to weight; There is a gap in the waveform diagram of the speech signal, and the endpoint detection is used to eliminate the discontinuity. Generally, the detection is based on the zero-crossing rate and the short-time energy detection method to improve the quality efficiency of the feature extraction and reduce the interference of the irrelated data. Then the MFCC and energy characteristics of the speech signal are extracted. Finally, it is put into the RNN for training. The origin of the RNN is to better describe the relationship between the output signal of the current time in the time series and the signal between the moments.

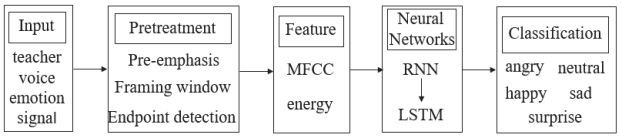

Fig 1 Flow chart of speech emotion recognition design

3.1 Feature Extraction

The MFCC uses mathematical calculations to represent the auditory nervous system of the human ear and uses a set of triangular filters to simulate the auditory nerve of a human ear, thus obtaining the spectral features of the speech emotion signal. However, the Mel frequency is non-linear and the distribution of triangular filters is concentrated in the low-frequency part, but sparse in the high-frequency and mid-frequency part. The low frequency part of the MFCC calculation accuracy and frequency resolution are high; However, the accuracy of MFCC calculation in the middle and high frequencies is poor, and some signals are lost. Therefore, we improve the traditional MFCC extraction process. The flow chart of the specific improvement is shown in Figure 2.

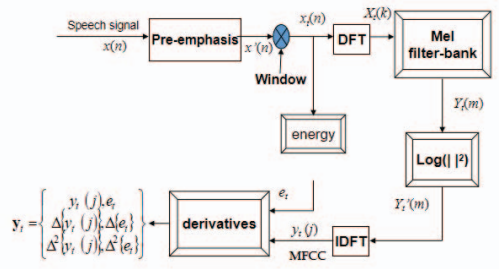

Fig 2 Improved MFCC extraction flowchart

1) Pre-aggravation means that the original speech emotion signal x(n) is changed to $x'(n)$ through a high-pass filter H(z), and the high-pass filter is expressed as:

$$H(z) = 1 - a \times z^{-1}, 0 < a \le 1 \tag{1}$$

In the above equation, the advantage of z is that the signal does not change at low frequencies; at high frequencies, the signal increases as the frequency increases.

2) Windowing is to use the short-term stationarity of the speech signal. In a short time, the characteristics of the signal are assumed to be unchanged. The speech is segmented into segments by a windowing method, these segments are called frames. Commonly used window functions are Hanning window and Hamming window. This paper uses Hanning window function to express:

$$w(n)=\begin{cases}0.54-0.46\cos(\frac{2\pi n}{L-1}), n=0,1,...,L-1\\ 0, others\end{cases} \quad (2)$$

3) The Discrete Fourier Transform (DCT) is a form in which the Fourier transform is discrete in both the time and frequency domains, transforming the time domain samples of the signal into the frequency domain samples of its DTFT. A Mel-scale based filter bank is applied, each filter output being the sum of its filtered spectral components.

4) The Mel filter bank process is similar to the human ear recognition process. We usually take 24 filter banks. Because the human auditory system is not aware of the phase, it take the absolute value and remove the negative sign. The reason for taking the log is that human feeling is relative, log is a special mathematical relationship.

5) After taking the IDFT, it return to the timeline. Finally, the differential is taken twice to obtain the change of the feature vector with time. Each time the derivative gets a 12-dimensional feature, a 36-dimensional feature is always obtained. Taking the differential operation, the convolution noise of the MFCC can be eliminated. m represents the number of filters, and the first and second order differentials are expressed as:

$$\Delta y_t(j)=\frac{\sum_{m=-p}^{p} m\cdot y_{t-m}(j)}{\sum_{m=-p}^{p} m^2} \quad (3)$$

$$\Delta^2 y_t(j)=\frac{\sum_{m=-p}^{p} m\cdot \Delta y_{t-m}(j)}{\sum_{m=-p}^{p} m^2} \quad (4)$$

6) The energy e_t of the signal is differentiated twice to obtain a 3-dimensional energy feature, which can more accurately represent the emotional features of the speech. The 39-dimensional MFCC features are finally obtained

3.2 RNN Structural Design

The nodes in the hidden layer in the RNN structure are connected with each other. The input signal of the hidden layer at each time point includes not only the input signal at the current time point but also the output signal at the previous time point. Figure 3 below shows a typical RNN structure.

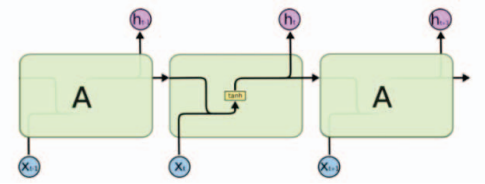

Fig 3 RNN structure diagram

As we can be seen from the above figure, the input of the recurrent neural network at all points in time is given an output in relation to the current state. The input to the main model A is both the input signal x_t and a state that gives the current time point. The RNN have an input x_t at each time point, and then provide an output h_t according to the current state, and then pass it to the next time point.

This paper selects the TensorFlow machine learning framework and uses the basic RNN unit structure. The structural parameters of the experimental network are 40 input nodes, 6 output nodes, 20 hidden layer neuron nodes, the learning rate is 0.001, and the number of loop iterations is 1000. The first 60 samples of each emotion as the training date, and the last 40 samples as the test date. Figure 4 below is a flow chart of the training model of the RNN.

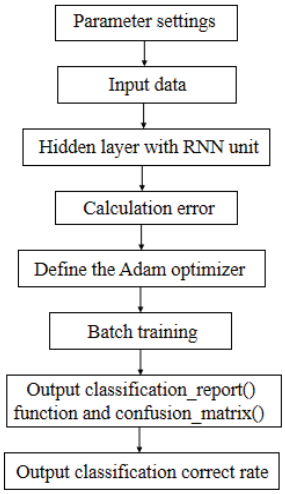

Fig 4 RNN training model flow chart

The main specificities of the network are: First, the Adam optimizer is created to reduce the error between the predicted value and the true value during the training. That is choose Adamoptimizer to minimize loss. Second, the classification_report() output classification evaluation function is used to analyze the accuracy, recall rate, and F1 value of different categories, so as to facilitate the accuracy and recall rate according to the category; The confusion_matrix() confusion matrix is a situation analysis table that summarizes the prediction results of the classification model in machine learning. The records in the data set are summarized in a matrix form according to the real category and the classification criteria predicted by the classification model.

Using the TensorFlow deep learning framework, the Python language extracts the features of the signal. The extracted features mainly use the following two data packets:

```
from python_speech_features import mfcc
from python_speech_features import delta
```

Among them, the following two functions are mainly used: def get_mfcc() function can be used to calculate the 13-dimensional MFCC coefficient and its second derivative; def get_energy() function is used to extract 1-dimensional energy features of speech emotion signals.

In order to test the influence of selecting different features on the result of speech emotion recognition, 39-dimensional MFCC features and 40-dimensional MFCC features are selected in this experiment.

Table 1. Identification results based on 39-dimensional MFCC+RNN model

Category	Percision(%)	Recall(%)	F1-score(%)
angry	89.19	94.29	91.67

sad	78.95	85.71	82.19
happy	61.22	85.71	71.43
neutral	72.09	88.57	79.49
surprise	76.74	94.29	84.62

Table 2 Identification results based on 40-dimensional MFCC+RNN model

Category	Percision(%)	Recall(%)	F1-score(%)
angry	94.12	91.43	92.75
sad	90.91	85.71	88.24
happy	63.83	85.71	73.17
neutral	55.17	91.43	68.82
surprise	86.84	94.29	90.41

It can be seen from Table 1 and Table 2 that the average accuracy of the five speech emotions in the 39-dimensional feature parameters is 75.64%, and the 40-dimensional is 78.17%. It can be seen that the recognition accuracy is improved 2.53% by adding one-dimensional energy feature. The short-time energy refers to the energy value of the speech signal in a short time, which is related to the intensity of sound vibrations, indicating that the energy feature enables speech signal to express emotion features more accurately.

3.3 RNN Model Optimization

Training the RNN over a long sequence of time will cause the early memory to be forgotten. In fact, as the data through the RNN, some information is lost at every moment. Soon after, there is no trace of the first input data in the RNN state. In response to the above problems, this paper introduces the optimization model of Long Short-Term Memory network (LSTM). Figure 5 below is a structure diagram of LSTM. The main difference between LSTM network and traditional RNN is that LSTM network has added three gate structures. Its unique gate structure enables LSTM unit to retain and obtain long-period speech signals.

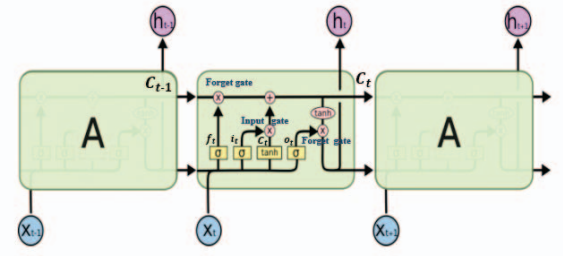

Fig 5 LSTM structure diagram

Gate is a way to selectively pass information. Each door is followed by an activation function and α represents various activation functions.

Table 3 specific calculation process of LSTM

Step 1: The sigmoid layer of the "Forgetting Gate" determines what information is lost from the cell state.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Step 2: Decide what information to store in the cell state, in two steps:

(1) The Sigmoid layer of the "input gate" determines which values are updated and adds new memories from the current state;

(2) A tanh layer creation candidate vector C_t is added to the cell state.

$$i_t = \sigma(W_i \cdot [[h_{t-1}, x_t] + b_i)$$

Step 3: The temporary status of the current moment is:

$$\widetilde{C_t} = \tanh(W_C \cdot [[h_{t-1}, x_t] + b_C)$$

Step 4: The real state of the current moment is expressed as "forgotten gate" and "input gate" as:

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t}$$

Step 5: The Sigmoid layer of the "output gate" determines the output information:

$$o_t = \sigma(W_o \cdot [[h_{t-1}, x_t] + b_o)$$

Step 6: The final output is h_t:

$$h_t = o_t * \tanh(C_t)$$

In the above table, the weight matrix of each gate is represented W_*, the deviation of each gate is represented b_*, the Sigmoid activation function is represented σ, and the tanh() activation function is represented tanh.

In order to obtain a higher accurate recognition rate while taking a shorter time, it is necessary to optimize the structural unit of the neural network model. The comparison results of the network model optimization are shown in Table 5 below.

The unique gate structure of LSTM solves the problem of gradient disappearance or gradient explosion in the traditional RNN time dimension, achieves the reservation and transmission of speech signals in a larger range, and improves the accuracy of the experiment. According to the above experiments, when the neural network is LSTM, the number of hidden layers is 2, the number of hidden layers is 20, the window size is 10, the learning rate is 0.001, and the number of training times is 500, the recognition rate of the model reaches the maximum of 85.32%.

Table 5 Model optimization comparison of network element structure

Neural networks	Hidden layer	Neural node	Window size	Learning rate	Training number	Accuracy(%)
RNN	1	20	7	0.001	500	81.14
RNN	1	20	10	0.001	500	82.71
LSTM	1	10	10	0.001	500	82.24
LSTM	1	20	10	0.001	500	83.46
LSTM	1	30	10	0.001	500	83.28
LSTM	2	20	10	0.001	500	84.89

LSTM	2	20	10	0.001	600	85.32
LSTM	2	20	10	0.002	600	84.23
LSTM	2	20	10	0.005	600	84.19

4 Comparison of Experimental Models

The text is completed on the GPU version of the TensorFlow1.12.0 deep learning framework. In the experiment, RNN is selected for comparison with CNN and traditional classifier HMM and SVM. SVM classifier is implemented in MATLAB2012b with the help of Libsvm toolbox. In order to verify the authenticity of the network model, it is first compared with the existing speech emotion recognition algorithm on the CASIA public speech database. The experimental results are shown in Table 6. Then, the recognition effect of the algorithm is compared with the teacher's speech database. Table 7 shows the speech emotion recognition results of the teacher's speech database under different models.

Table 6 Identification results of CASIA database under different models (%)

Emotion category	HMM	SVM	RNN	LSTM
angry	77.96	78.26	90.98	93.78
sad	78.12	85.32	89.85	91.23
happy	73.15	75.19	80.23	82.98
neutral	66.98	68.34	69.87	71.23
fear	85.78	84.48	90.12	92.87
surprise	66.12	70.09	82.16	85.72
Average recognition rate	74.68	76.95	83.87	86.31

Table 7 Identification results of teacher speech database under different models (%)

Emotion category	HMM	SVM	RNN	LSTM
angry	87.96	88.26	90.98	93.78
sad	85.78	84.48	90.12	92.87
happy	77.15	75.19	80.23	82.98
neutral	66.98	68.34	69.87	71.23
surprise	76.12	80.09	82.16	85.72
Average recognition rate	78.79	79.27	82.67	85.32

As shown in the above table, the network model can effectively classify speech emotion recognition. The results of the classification model for each emotion recognition rate are shown in Figure 7.

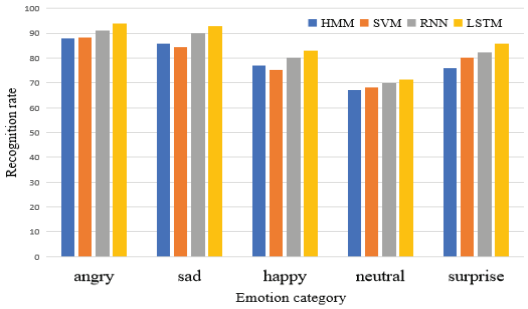

Fig 7 classification model test comparison effect

As shown in the above table, the average recognition rate of the five speech emotion classifications of HMM is the lowest, which is 78.79%; the lower SVM is 79.27%; the highest LSTM is 85.32%. Among them, when the emotion is angry, the recognition rate reaches the highest, which is 93.78%, which indicates that the emotional feature information of teachers in this state is more prominent. Since the HMM only relies on each state and its corresponding observation object, the model is less effective; The SVM trains all the samples so that when solving the quadratic programming problem, the training speed will slow down as the number of samples increases. The unique gate structure of LSTM solves the problem of gradient disappearance or gradient explosion in the traditional RNN time dimension, and the accuracy is improved.

5 CONCLISIONS

In this paper, the teaching quality evaluation system of teacher speech signal is studied in the classroom scene, and a speech emotion recognition model combining MFCC and RNN-LSTM is proposed. The speech signal of the instructor is collected for emotional reclassification, and the teacher speech emotion database is established. The process of MFCC feature extraction is improved, and the second-order differential process is added to eliminate the convolution noise of MFCC. One dimension energy frame is added into the 39 dimension MFCC parameters, and a total of 40 dimension characteristic input is obtained. Experiments show that the energy frame makes the feature parameters more effective in expressing speech emotional features. By further optimizing the structure of the recurrent neural network unit, the LSTM network model is proposed. The accuracy of the model classification prediction is 85.32%. The performance is better than the traditional speech emotion classification model, which provides the possibility to automatically evaluate the classroom teaching

quality in real time. How to express the teacher's emotions more accurately can also consider facial expressions and body movements. Future work will focus on the impact of other different characteristics on teacher emotions.

REFERENCES

[1] HUANG ZH W. Research on Speech Emotion Feature Learning Method Based on Depth Learning [Dissertation]. Zhenjiang: Jiangsu University, 2015.

[2] LI Hong, XU Xiaoli, WU Guoxin, *et al*. Research on the extraction of speech emotion feature based on MFCC[J]. *Journal of Electronic Measurement and Instrumentation*, 2017,31(3):448-453.

[3] Busso C, Lee S and Narayanan S. Analysis of emotionally salient aspects of fundamental frequency for emotion detection[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2009, 17(4): 582-596.

[4] Cowie R and Cornelius R R. Describing the emotional states that are expressed in speech[J]. *Speech communication*, 2003, 40(1): 5-32.

[5] Murray I R and Arnott J L. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion[J]. *The Journal of the Acoustical Society of America*, 1993, 93(2): 1097-1108.

[6] Lee C M and Narayanan S S. Toward detecting emotions in spoken dialogs[J]. *IEEE Transactions on Speech and Audio Processing*, 2005, 13(2): 293-303.

[7] TAO Huawei, CHA Cheng, LIANG Ruiyu, *et al*. Feature extraction algorithm for speech emotion recognition[J]. *Journal of Southeast University(Natural Science Edition),* 2015, 45(5): 817-821.

[8] ZHANG Deliang. Implementation of Deep Neural Network in Chinese Speech Recognition System [Dissertation]. Beijing: Beijing Jiaotong University, 2015.