Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 8, No. 3, 2024

Emotion Recognition in Lhasa Tibetan Speech based on Bi-LSTM Graph Convolutional Networks

Ang Chen, Rongzhao Huang, Tong Xi, Liang Wu, Wangdui Bianba *

School of Information Science and Technology, Tibet University, Lhasa, Tibet 850000, China
*** Corresponding author:** Wangdui Bianba

Abstract: Speech Emotion Recognition (SER) is a crucial component in the field of Human-Computer Interaction (HCI), with significant research and practical application implications. However, due to the complexity of the Tibetan language and the scarcity of datasets caused by the difficulty in collecting various dialects, there are not many research achievements in Tibetan speech recognition. Based on the foundation of constructing a TBLS1 dataset containing 6,000 Tibetan-language speech samples, an approach was devised for Tibetan speech emotion recognition. This approach leverages MFCC features and incorporates a Bi-directional Long Short-Term Memory (Bi-LSTM) network within a graph convolutional neural network. Finally, by comparing the performance of different models on this dataset, we demonstrated the feasibility of our model for Tibetan speech emotion recognition.

Keywords: Tibetan Speech Emotion Recognition; Mel-frequency Cepstral Coefficients (MFCC); Bidirectional Long Short-Term Memory (Bi-LSTM); Graph Convolution Network (GCN).

1. Introduction

In today's increasingly intelligent and automated technological context, speech emotion recognition plays an important role in human-computer interaction and is an important way for artificial intelligence to understand human emotions [1].

Compared to other languages, the research on Tibetan speech emotion recognition started later, with fewer participating scholars. Moreover, due to the large dialectal differences in different regions of Tibet and inconvenient transportation, it is difficult to collect datasets, resulting in fewer overall research results [2].

Among the Tibetan-speaking regions, Lhasa, as a populous area in Tibet, has a larger base of speakers of the Lhasa dialect than other Tibetan dialects. It is widely spoken, and it is relatively easy to collect data, making it easier to collect high-quality speech data. Therefore, this paper mainly focuses on the emotion recognition of the Lhasa dialect.

Datasets

Due to the lack of large-scale, high-quality public corpora in Tibetan, we built a Lhasa dialect emotion recognition dataset for this experiment. This dataset was recorded by 12 Tibetan students, including 6 males and 6 females. we selected some commonly used phrases and words in their daily lives for recording. Recording was done using Adobe Audition 22.0. Each student recorded voices of 5 different emotions (anger, fear, happiness, neutral, sadness). Each emotion includes 100 voice recordings, totaling 6000 voice recordings, which make up the dataset TBLS1 used in this experiment. The specific recording parameters are: sampling rate: 16kHz, quantization level: 16bit, single channel. The participants in the recording are not professional actors, so there may be a slight lack of authenticity in the voice data. However, this dataset is balanced in terms of gender and the number of data points for each emotion, so it is unlikely to be affected by the imbalance of voice data volume from a certain participant or emotion. Therefore, it can be considered that the model's performance on this dataset is highly credible.

2. Tibetan Speech Emotion Feature Extraction

2.1. MFCC

Mel-Frequency Cepstral Coefficients (MFCC) is a very widely used speech feature in the field of speech emotion recognition[3]. The Mel frequency is a frequency scale established based on the human auditory system's different acceptance levels for sound signals at different frequency bands, and it has a nonlinear relationship with frequency. The correspondence between the Mel frequency scale and the normal frequency is as follows:

$$Mel(f) = 2595 * \log_{10}(1 + f/100) \quad (1)$$

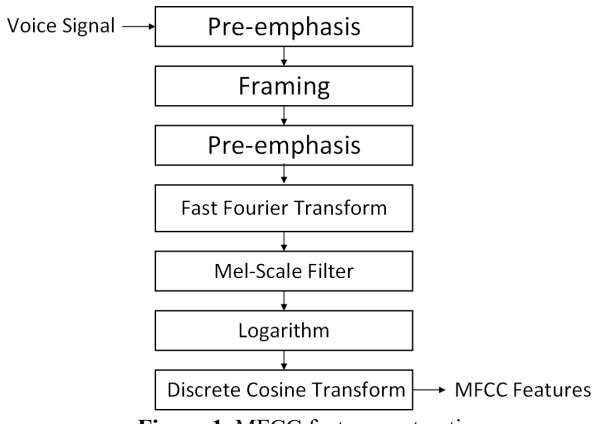

Figure 1. MFCC features extraction

MFCC features are the cepstral coefficients obtained by taking the logarithm of the Mel spectrum and then performing an inverse transformation. The process of extracting MFCC usually includes pre-emphasis, framing, windowing, etc. Pre-emphasis is to enhance the high-frequency part of the speech signal; while framing and windowing are to perform local analysis of the signal in the time domain. Subsequently, the data is subjected to FFT, passed through a Mel filter, and the

logarithm is taken and then the required cepstral coefficients are obtained through the Discrete Cosine Transform (DCT)[4].

This experiment extracted the first 12 MFCC coefficients, because the first 12 cepstral coefficients correspond to the low-frequency band, which contains more emotion-related information compared to the high-frequency band [5]. In addition, this experiment combines the MFCC features with its first-order difference, zero-crossing rate, fundamental frequency, frame energy (short-time energy of speech), and other features as the final input features. The final feature set contains 35 features for each frame of speech.

2.2. Other Features

The zero-crossing rate is the number of times the sample values in each frame of the sound signal pass through zero. Since the zero-crossing rates of unvoiced and voiced sounds differ greatly, the zero-crossing rate is generally introduced as a distinguishing feature in the discrimination of these two. The zero-crossing rate of unvoiced sounds is generally significantly greater than that of voiced sounds. It should be noted that the zero-crossing rate of noise is close to that of unvoiced sounds, and the discrimination between unvoiced sounds and noise needs to be done through other features. The calculation process is as follows, ω(n) is the window function, we chose Hamming Window:

$$Z_n = \sum_{m=-\infty}^{+\infty} |\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]| * \omega(n-m) \quad (2)$$

The fundamental frequency contains a large number of basic features of the sound signal. The fundamental frequency is the lowest frequency component in the sound signal, corresponding to the period of vocal cord vibration. It is closely related to the speaker's tone, pitch, and emotional state. Generally speaking, a higher fundamental frequency usually corresponds to a higher pitch, while a lower fundamental frequency corresponds to a lower pitch. Changes in tone can convey emotional information, such as happiness, excitement, or depression. Short-time energy of speech also has certain value in speech emotion recognition. Generally, the energy of the speech signal is calculated frame by frame. This feature can better distinguish between unvoiced and voiced sounds, with the short-time energy of voiced sounds generally much higher than that of unvoiced sounds. In terms of emotion recognition, the energy of intense emotions such as anger and happiness are usually higher than that of neutral and sad emotions. The calculation of the short-time energy E_n corresponding to the window function starting from the nth point is as follows:

ω(n) is the window function, $h(n)=\omega(n)^2$;

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)\omega(n-m)]^2 = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m) = x^2(n) * h(n) \quad (3)$$

The features are concatenated to obtain the final feature set.

3. Speech Emotion Recognition Net

3.1. Graph Convolution Neural Networks

Compared to traditional neural networks, Graph Neural Networks (GNNs) have a unique structure and inter-data relationship, and they have shown impressive performance in the field of speech emotion recognition. They first need to convert the input data into a graph structure, where the smallest specific data units in the graph serve as nodes. Different nodes are connected by edges, which reflect the relationships between different nodes. The construction of the graph can be divided into cyclic graphs and chain graphs (closed-loop and open-loop). The difference is that in cyclic graphs, there is an edge connecting the first and last nodes, which is often used to handle data with cyclic relationships, and information can propagate in the loop. On the contrary, in chain graphs, the data cannot cycle from the end to the beginning, and the information propagation is unidirectional [6].

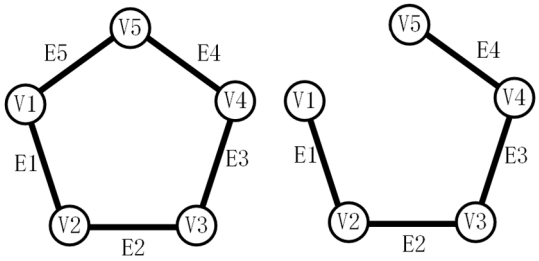

Figure 2. cycle graph and line graph

In this experiment, a simple frame-to-node conversion is used, and each frame node is connected by a unidirectional edge according to the time series. Since the speech data is not a cyclic data structure from beginning to end, this experiment constructs the speech data as a chain graph. Each speech is an independent chain graph, with frames as nodes, connected by edges in the order of the time series, and each node is associated with a node feature vector[7], The node feature vector contains the 35-dimensional feature information on that node frame.

3.2. Graph Convolution

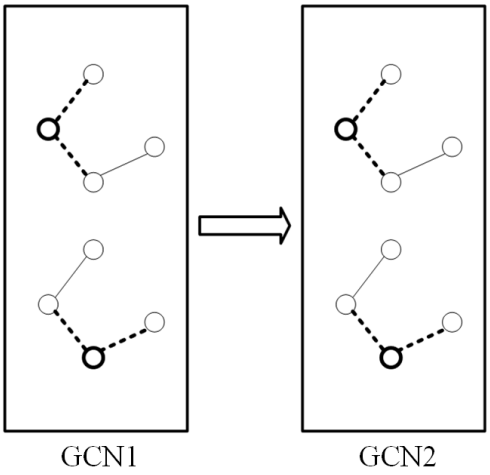

Figure 3. Node aggregation method in graph convolution layer

In the graph structure, the behavior of a node aggregating information from adjacent nodes with its own information is called graph convolution. Aggregation methods usually include taking the average, sum, or maximum. This paper uses the sum aggregation method. After performing a convolution, the node contains information from adjacent nodes. Based on this, repeating the graph convolution operation m times can achieve the effect of aggregating the information of nodes at a stride of m to one node [8], Therefore, the graph convolution structure has the advantage of being able to capture local information correlations well [9]. If the number of layers in the graph convolution is too high, it can lead to the node aggregating too much information from neighboring nodes and diluting its own information. Generally, the effect is best with 2 to 3 layers. In this experiment, a 2-layer graph convolution structure was set up.

The specific process is shown in the figure below, where in the same convolution layer, the aggregation operations of each node are performed simultaneously.

For the constructed graph structure data, suppose there are N nodes, and the dimension of the feature vector of each node is H. All node features can be composed into an N×H matrix X. The connection relationship between each node is an N×N relationship matrix, that is, the adjacency matrix, as follows [10], If the points are connected, the value is 1; if they are not connected, the value is 0. Since this experiment uses a chain graph, the first and last nodes are not connected.

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

Let D be the degree matrix of this graph, that is, a diagonal matrix whose diagonal elements are the number of nodes connected to that node. The Laplacian matrix $L = D - A$ is a commonly used matrix to describe the overall structure of the graph [11]. The more nodes a node is connected to, the lower the weight of the information passed from each connected node to that node during aggregation, Therefore, use $D^{-1}L$ to represent the weight of information each node should get from neighboring nodes. Because $D^{-1}L$ will make the feature value of the node with more neighboring nodes greater than that of the node with fewer neighboring nodes after aggregation, we use $D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$ as the node aggregation matrix after normalization. Therefore, for a GCN network with I layers, the propagation function between each layer is:

$$X^{(i+1)} = \sigma(D^{-\frac{1}{2}}LD^{-\frac{1}{2}}X^{(i)}W^{(i)}) \qquad (4)$$

$W^{(i)}$ is the weight matrix of the i-th layer; $\sigma(\cdot)$ is a nonlinear activation function, and the ReLU function is used in this experiment; $X^{(i)}$ represents the input information of the i-th layer.

3.3. Bi-LSTM

Long Short-Term Memory (LSTM) is a special type of recurrent neural network that selectively allows information to pass through gate units, enabling the network to maintain long-term memory of overall sequence features [12], It has good performance in capturing the overall features of the sequence[13].

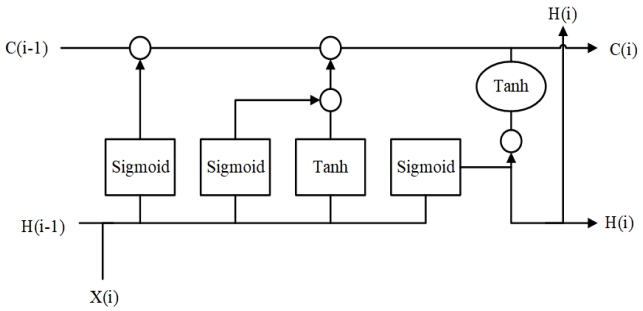

Figure 4. LSTM network

The Bi-directional Long Short-Term Memory (Bi-LSTM) consists of two independent LSTMs, which process the forward and backward information of the data sequence respectively. This network design allows the model to learn the context information of the past and future simultaneously [14]. Through this bi-directional processing mechanism, Bi-LSTM enhances the model's ability to capture information in scenarios where understanding the entire data stream is required.

GCN networks are good at aggregating information from neighboring nodes and capturing the connection of local features. If you add too many GCN layers in an attempt to aggregate information from distant nodes to obtain global information, it will instead cause the node information to be overly aggregated and lose its original features. Therefore, using a GCN with fewer layers to capture local information, while combining it with a Bi-LSTM network to obtain the connection of global features, can compensate for the deficiency of the GCN network's receptive field[15].

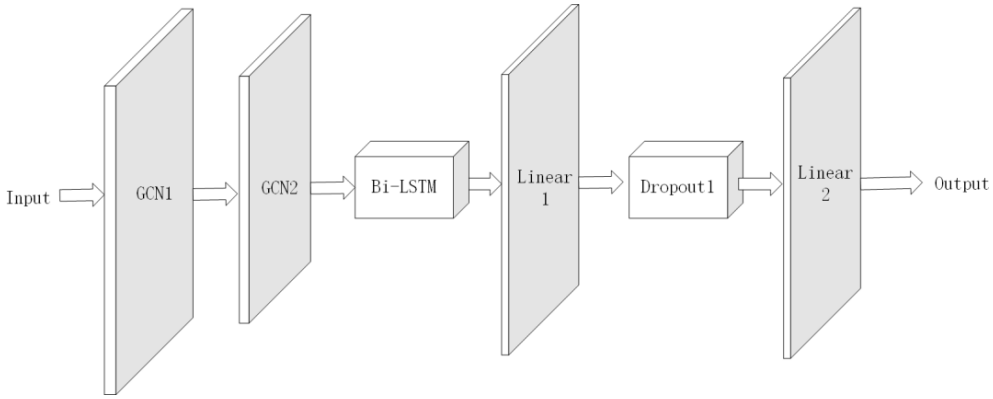

Figure 5. Overall network structure

3.4. Loss Function

In terms of loss function, previous speech emotion recognition tasks often use Cross Entropy Loss (CE), which has always performed well in multi-class tasks. In order to further capture the emotional features of the Tibetan language and more accurately direct the model training, a joint loss function of Cross Entropy and Center Loss is used. The center loss function sets a center vector for each emotion label. During the training process, the center loss will calculate the Euclidean distance between the current sample feature vector and the center vector of its label in each iteration, and update the center vector of the label through backpropagation. The inclusion of the center loss function theoretically helps the model better extract the unique features of each emotion. In this experiment, we add the Cross Entropy loss function and

the center loss function with weights to construct a joint loss function.

4. Experiment and Analysis

In this experiment, the Tibetan language speech emotion recognition dataset TBLS1 is evenly divided into 5 groups, with one group being used as the validation set in turn, and the rest being used as the training set. The Early Stop function is added, with a patience index set to 10 epochs. Each epoch includes 50 pieces of data. To ensure the amount of training, it is set that the final result can only be output after at least 300 epochs. The performance of different networks on this dataset is as follows:

Table 1. Comparison of performance of different models

Network	Loss Function	Accuracy (%)
LSTM	Cross Entropy Loss	45.5726
CRNN	Cross Entropy Loss	50.2564
CNN	Cross Entropy Loss	57.0854
CNN+LSTM	Cross Entropy Loss	59.2897
GCN	Cross Entropy Loss	59.3970
GCN+Bi-LSTM	Cross Entropy Loss	61.7092
GCN+Bi-LSTM	Cross Entropy Loss + Center Loss	**63.8897**

It can be seen that compared with the previous convolutional models, the Graph Convolutional Network (GCN) performs better in the task of Tibetan speech emotion recognition, and the accuracy of GCN is more than 2 percentage points higher than that of traditional convolutional networks. Moreover, the LSTM module can indeed compensate for the deficiency of GCN in extracting global features, thereby improving the performance of the model and increasing the accuracy by about 2%. In terms of loss function, the joint loss function of Cross Entropy and Center Loss captures the optimization direction and features of Tibetan speech emotion recognition better than the single Cross Entropy loss function used in the past. On the best-performing GCN+Bi-LSTM network, using the joint loss function can increase the accuracy by about 2% compared to using single Cross Entropy.

5. Conclusion

In the task of Lhasa Tibetan speech emotion recognition, the GCN+Bi-LSTM network performs better than traditional models, sacrificing a bit of convergence time to achieve higher accuracy. Moreover, the joint loss function combining CrossEntropy loss and Center loss also performs better in this task than when only using CrossEntropy as the loss function. Considering that many models did not achieve high accuracy in this experiment, there are currently several areas for improvement based on the analysis:

1. The Tibetan speech recognition dataset is insufficient, and more high-quality speech data needs to be collected.

2. Tibetan, as a language with unique pronunciation habits, has not yet found features that can efficiently represent emotions. The features extracted are mostly commonly used in speech emotion recognition.

3. There is still room for improvement in terms of network complexity.

"In future Tibetan speech emotion recognition tasks, graph neural networks may have more applications and development.

References

[1] Guzeyue, Bianbawangdui, Qi Jindong. Tibetan Speech Emotion Recognition Based on Multi-Feature Fusion[J]. Modern Electronic Technology, 2023, 46(21): 129-133. DOI: 10. 16652/j.issn.1004-373x.2023.21.024.

[2] Cai Youxin, Bianbawangdui. Tibetan Speech Emotion Recognition Based on Bidirectional GRU Model[J]. Information Technology and Informatization, 2023, (10): 209-213.

[3] Ding Nan. Research on Speech Emotion Recognition Based on Feature Learning[D]. Nanjing University of Posts and Telecommunications, 2023. DOI: 10.27251/d.cnki. gnjdc. 2023. 001294.

[4] Akçay M B, Oğuz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers[J]. Speech Communication, 2020, 116: 56-76.

[5] Huang Xiyang, Du Qingzhi, Long Hua, et al. Speech Emotion Recognition Algorithm Based on MFCC Feature Fusion[J]. Journal of Shaanxi University of Technology (Natural Science Edition), 2023, 39(04): 17-25.

[6] Shirian A, Guha T. Compact graph architecture for speech emotion recognition[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6284-6288.

[7] Liu J, Wang H. Graph Isomorphism Network for Speech Emotion Recognition[C]//Interspeech. 2021: 3405-3409.

[8] Shirian A, Guha T. Compact graph architecture for speech emotion recognition[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6284-6288.

[9] Liu J, Wang H, Sun M, et al. Graph based emotion recognition with attention pooling for variable-length utterances[J]. Neurocomputing, 2022, 496: 46-55.

[10] Li Zijing, Chen Ning. Speech Emotion Recognition Model Based on Multi-Modal Fusion of Graph Neural Network[J]. Computer Application Research, 2023, 40(08): 2286-2291+2310. DOI: 10.19734/j.issn.1001-3695.2023.01.0002.

[11] Zhang S, Tong H, Xu J, et al. Graph convolutional networks: Algorithms, applications and open challenges[C]// Computational Data and Social Networks: 7th International Conference, CSoNet 2018, Shanghai, China, December 18–20, 2018, Proceedings 7. Springer International Publishing, 2018: 79-91.

[12] Xu Huanan, Zhou Xiaoyan, Jiang Wan, et al. Speech Emotion Recognition Algorithm Based on Self-Attention Temporal and Spatial Features[J]. Acoustic Technology, 2021, 40(06): 807-814. DOI: 10.16300/j.cnki.1000-3630.2021.06.011.

[13] Peng Z, Dang J, Unoki M, et al. Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech[J]. Neural Networks, 2021, 140: 261-273.

[14] Li Y, Wang Y, Yang X, et al. Speech emotion recognition based on Graph-LSTM neural network[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2023, 2023(1): 40.

[15] Su B H, Chang C M, Lin Y S, et al. Improving Speech Emotion Recognition Using Graph Attentive Bi-Directional Gated Recurrent Unit Network[C]//INTERSPEECH. 2020: 506-510.