Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 10, No. 1, 2024

Tibetan Speech Emotion Recognition based on Capsule Network and Spatiotemporal Features

Rongzhao Huang [1], Ang Chen [1], Menglin Bai [1], Bawangdui Bian [1, 2]

[1] College of Information Science and Technology, Tibet University, Lhasa Tibet 850000, China
[2] National Experimental Teaching Demonstration Center for Information Technology, Lhasa Tibet 850000, China

Abstract: Speech emotion recognition is an important branch of natural language processing that aims to automatically recognize and classify emotional information in speech through computer technology. In the specific language environment of Tibetan, due to relatively limited research and some existing studies appearing cumbersome and complex in feature extraction steps, a new network model has been proposed. The model is based on a capsule network and achieves lightweight design. It only uses Mel Frequency Cepstral Coefficients (MFCC) as its input features, extracts the spatiotemporal information of MFCC through multiple convolutional layers, and sends it into the capsule network for deep analysis. The recognition rate of 81.52% was achieved on the self-built Tibetan language emotion corpus TBSEC001. Meanwhile, the method achieved an unweighted accuracy (UA) of 85.63% and 95.54% respectively on the EMO-DB and RAVDESS public corpora, demonstrating the method's effectiveness.

Keywords: Speech Emotion Recognition; Tibetan; Mel-frequency Cepstral Coefficients; Deep Learning; Capsule Net.

1. Introduction

As an important carrier in human communication, emotion has remarkable subjectivity and contextual relevance. Different individuals may adopt different ways to express and interpret emotions, which reflects the complexity and diversity of emotional communication [1] With the continuous progress of computer technology and the increasingly prominent demand of Human-Computer Interaction (HCI), people frequently communicate and interact with computers in daily life and work. If the computer can automatically recognize the users emotional state, it can respond to the user's needs in a more intelligent way, thereby improving the interactive experience [2]. The process of automatically recognizing human emotions by computers is called speech emotion recognition (SER).

The speech emotion recognition task includes many key steps, the core parts of which are: building a speech emotion corpus, extracting speech emotion features, model training and evaluation. The task first builds a corpus by collecting and labeling emotion-rich speech segments, and then extracts features that can reflect emotional states from these segments, sends them into the model as input for training and testing. In this process, the extraction of speech emotional features and the creation of efficient models constitute the main research contents of the SER task [3], it is great significance to improve the accuracy and robustness of emotion recognition.

Generally speaking, the affective features of speech are divided into three categories: prosodic features [4], spectrum features [5-7] and sound quality features [8]. In the early SER task, the above three types of features were mainly extracted as the input of the model, which are also called manually extracted Low-level Descriptors (LLDs). In fact, the extraction of these features involves a lot of knowledge of speech signal processing. For researchers, designing input emotional features is one of the difficulties of SER task. In recent years, with the improvement of computer computing power, deep learning has also risen. The proposal of deep features based on deep learning has reduced the difficulty for researchers to study speech emotion recognition to a certain extent.

In the field of deep learning, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) were initially applied to speech emotion recognition tasks and achieved good results. Subsequently, Long Short-Term Memory (LSTM) surpasses traditional speech recognition classifiers in performance by virtue of its stronger context association ability [9]. In terms of feature extraction, researchers began to try feature fusion experiments, by horizontally stitching LLDs of speech signals and sending them to the network. References [10] [11] [12] show that these fused features have achieved better performance in public corpus. Due to the problem that LSTM networks cannot parallelize computing, Google proposed the Transformer model in 2017. Its advent marks a great success of attention mechanism in the field of natural language processing (NLP), and this mechanism also has significant effects on SER tasks [13-15]. Literature [16] achieved better classification performance by combining attention mechanisms and LSTM networks and modifying their forgetting gates. Furthermore, reference [17] improved the Transformer model, and verified the feasibility and effectiveness of the method on multiple public data sets by adjusting the attention window in its multi-head attention mechanism.

Although deep learning has made remarkable progress in the field of speech emotion recognition, the research of Tibetan speech emotion recognition is still relatively lagging behind. As a unique minority language, Tibetan faces multiple challenges in speech emotion recognition, including data scarcity, labeling difficulties and differences in emotion expression habits [18]. Although some researchers have been involved in the field of Tibetan speech emotion recognition, they mainly fuse multiple LLDs. However, these methods have some problems, such as too high emotional feature dimension, too cumbersome emotional feature extraction, complex network design and high computing power consumption [19]. To effectively address these challenges, an

innovative lightweight network model is proposed in this paper. This model combines the spatiotemporal feature extraction of Mel-Frequency Cepstral Coefficients (MFCC) with the powerful vectorization capabilities of capsule networks, aiming to achieve accurate recognition of Tibetan speech emotions. In order to comprehensively evaluate and verify the performance of the model, this paper conducts experiments in EMO-DB, RAVDESS corpus, and the self-built Tibetan speech emotion corpus TBSEC001, and finally achieves satisfactory classification results. The main research contents of this paper are as follows:

(1) A Tibetan speech emotion corpus TBSEC001 is constructed and improved, which provides data resources for related research.

(2) A lightweight network model is proposed, which fuses the spatiotemporal features of MFCC coefficients and the vectorization capability of capsule networks. The self-built corpus TBSEC001 was used for Tibetan speech emotion recognition, and the accuracy was 81.52%, and the unweighted average accuracy (UA) was 85.63% and 95.54% on EMO-DB and RAVDESS corpus, respectively.

2. Model Design

The network model structure proposed in this paper is shown in Figure 1. It is mainly composed of four parts: MFCC (input feature), spatiotemporal feature extraction backbone network (STFeatCNN-4L), Capsule network layer and Margin-Loss (loss function).

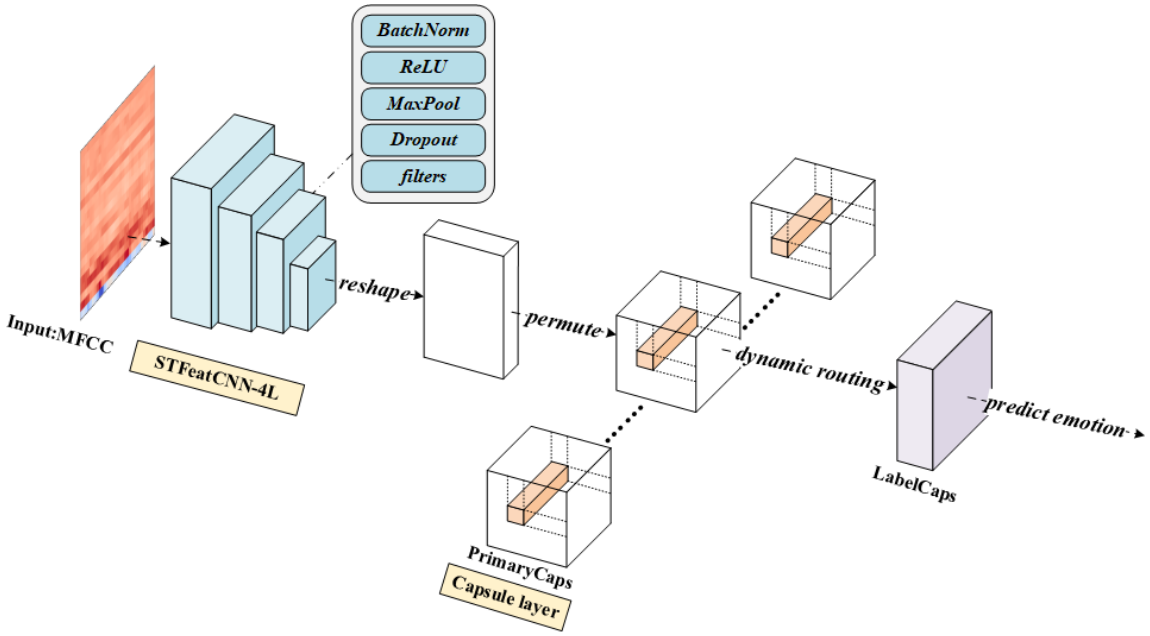


Fig 1. Overall structure of the model

First, the model receives and processes the MFCC as features, and these features are sent to the spatiotemporal feature extraction backbone network STFeatCNN-4L to extract spatiotemporal information; Then, the data is transmitted to the Capsule layer. With its unique dynamic routing update algorithm, the capsule network layer efficiently processes the input features and outputs the prediction results for emotion; Finally, Margin-Loss is used as the loss function to guide the model to adjust parameters to optimize the performance of the model, so as to complete the task of efficient and accurate speech emotion recognition.

Table 1. Parameter configuration of STFeatCNN-4L layer

Layer Name	Components	Kernel size	Stride	Input channel	Output channel
Conv1	Convolutional layer	3×3	1×1	1	96
	BN (Batch Normalization) + ReLU	-	-	-	
	Max-Pooling	2×2	1×1	-	
	Dropout	-	-	-	
Conv2	Convolutional layer	3×3	1×1	96	96
	BN + ReLU	-	-	-	
	Max-Pooling	2×2	1×1	-	
	Dropout	-	-	-	
Conv3	Convolutional layer	3×3	1×1	96	96
	BN + ReLU	-	-	-	
	Max-Pooling	2×2	1×1	-	
	Dropout	-	-	-	
Conv4	Convolutional layer	3×3	1×1	96	96 × Emotion Types
	ReLU				

(1) Backbone network STFeatCNN-4L

In order to extract the spatiotemporal features of MFCC, the backbone network STFeatCNN-4L is designed in this paper. In STFeatCNN-4L, the scaling invariance of CNN is adopted to capture and retain the key features among the input features [20], while multiple convolutional filters are used.

These filters are able to learn and extract different spatial features in MFCC, and by combining spatiotemporal features, the network can understand the spatial structure of the speech signal more comprehensively [21, 22], thus improving the recognition performance. In addition, batch normalization is introduced in each layer of the network to reduce the internal covariate shift and stabilize the training process; Using ReLU activation function to increase nonlinear expression ability; Using max-pooling to reduce dimensionality and preserve important features; Dropout is adopted to prevent overfitting and improve generalization performance. Regarding the specific parameter configuration of STFeatCNN-4L, Table 1 provides an explanation.

(2) Capsule network layer

The capsule network layer consists of the main capsule layer (PrimaryCaps) and the label capsule layer (LabelCaps). Compared with the convolutional neural network, its remarkable feature is that can convert the calculated neurons from a single value into a vector [23]. This vector not only reflects the existence of neurons, but also contains their directionality. In order to make full use of this feature and obtain spatiotemporal features, this paper will convolve and reshape the features after STFeatCNN-4L, and then use dimensional transformation to turn them into multiple capsule vectors, finally forming PrimaryCaps. The capsules in PrimaryCaps are output to LabelsCaps after updating the algorithm through dynamic routing. Table 2 shows the changes of features before and after shaping and dimension transformation, where L(length) and W(width) represent the dimensions of features after STFeatCNN-4 L.

Table 2. Changes in characteristic dimensions of capsule network

Operation	Input feature shape	Output feature shape
Original features	-	[Batch, 96 × Emotion Types, L, W]
Reshape	[Batch, 96 × Emotion Types, L, W]	[Batch, 96, Emotion Types, L × W]
Dimension permutation	[Batch, 96, Emotion Types, L × W]	[Batch, Emotion Types, L × W, 96]

The core component of capsule network is dynamic routing update algorithm. Assuming that the number of iterations is T, each capsule is represented by u_i, and initialize routing parameter $b_{ij} = 0$. In each iteration t $(t = 1,2,\ldots,T)$, the dot-multiplication operation with the weight matrix W_{ij} and u_i needs to be performed, and the value obtained by dot-multiplication is multiplied by the coefficient c_{ij} and accumulated to obtain s_j. At the same time, in order to represent whether the vector exists or not, it is *squash* normalized to obtain v_j. The specific formula is as follows:

$$\hat{u}_{i|j} = W_{ij} \cdot u_i \tag{1}$$

$$c_{ij} = softmax(b_{ij}) = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \tag{2}$$

$$s_j = \sum_i c_{ij} \hat{u}_{j|i} \tag{3}$$

$$v_j = squash(s_j) = \frac{\| s_j \|^2}{1+\| s_j \|^2} \cdot \frac{s_j}{\| s_j \|} \tag{4}$$

Then update the routing coefficients b_{ij}:

$$b_{ij} = b_{ij} + \hat{u}_{j|i} \cdot v_j \tag{5}$$

Repeat the above steps until T iterations are completed, the updated v_j is sent to the labelsCaps layer. In this layer, it is necessary to modulate each emotion v_j and use a *softmax* function to predict the possibility of each emotion, and finally realize the task of speech emotion recognition.

According to the core mechanism of the dynamic routing update algorithm, the capsule network will adjust the length of the "sum vector" and each prediction vector according to the similarity between them during the route update process, that is, the vector with high similarity will be elongated, while the vector with low similarity will become shorter. This mechanism ensures that vectors that are more similar to "sum vectors" can be sent to LabelCaps more, thus effectively achieving classification effects.

According to the characteristics of capsule network, this paper deeply studies the parameter of weight matrix W_{ij} in dynamic routing algorithm. When the weight matrix is assigned to each capsule separately for training, the model parameters will increase dramatically, and problems such as overfitting, gradient explosion and gradient disappearance will also occur. Therefore, this paper adopts the method of sharing weights, which reduces the complexity of the model and allows multiple capsules to share the same set of weights, so that the model can more accurately grasp the internal relationship between features and improve its generalization ability.

(3) Loss function

The loss function used in this paper is the Margin-Loss function. Compared with traditional loss functions such as cross entropy and mean square error, Margin-Loss focuses more on the relative distance and interval between samples, it can maximize the interval of different types of samples to optimize the model [24]. Margin-Loss is calculated as follows:

$$L_c = T_c \cdot max(0, m^+ - \| v_c \|^2) + \lambda(1 - T_c) \cdot max(0, \| v_c \| - m^-)^2 \tag{6}$$

Among them, T_c is the vector obtained by *one-hot* encoding the emotion label, v_c represents the modulo of each emotion entity,$\lambda \in (0,1)$.

3. Experimental Analysis

(1) Corpus

To realize Tibetan speech emotion recognition, this paper constructs a Tibetan speech emotion corpus: TBSEC001. At the same time, in order to enhance the comprehensiveness and accuracy of the research, this paper also selects two widely recognized public corpora as supplements: EMO-DB and RAVDESS [25].

The TBSEC001 corpus consists of speeches by 12 native Tibetan speakers, including 6 women and 6 men. All the participants in the recording are from Tibet University, and their mother tongues are ‘Weizang’ dialect. The database contains five emotional categories: angry, fear, happy, neutral and sad. During the recording process, 12 recording personnel not only participated in the recording work, but also checked each others emotional expression of the audio to ensure that each audio can accurately convey the preset emotion. The audio that does not meet the requirements is culled to ensure the overall quality of the corpus. Through repeated inspection

and recording, the TBSEC001 corpus finally included 6,000 emotional and high-quality Tibetan speeches. These speech samples can provide solid data support for the research of this paper. Table 3 lists the specific information of TBSEC001 sentiment corpus in detail.

Table 3. Introduction to TBSEC001 Speech Emotion Corpus

Corpus	TBSEC001
Language	Tibetan
Dialect	Weizang dialect
Type	Performance type
Types of emotions	5 kinds
Number of recordings	6 males 6 females
Tone	4
Characteristics of consonant pronunciation	Voiced consonants are cleared and complex consonants are simplified [26]
Characteristics of vowel pronunciation	Nasalized & non-nasalized vowels [27]
Duration	0.8 ~ 4 seconds
Number of corpus	1200 items for each of 5 emotions, 6000 items in total

In addition to some basic information about TBSEC001, Table 3 also illustrates some unique features of Tibetan pronunciation: literature [26] shows that Tibetan has a tendency to clear voiced consonants and simplify complex consonants; In terms of vowels, the vowels of unit sounds, especially the vowels of nasalized vowels, have increased, forming two kinds of complex vowel vowels: nasalized vowels and non-nasalized vowels. In addition, Tibetan has four tones, and the cadence of tones can help it better express a person's emotions. Overall, it is feasible to use TBSEC001 for Tibetan language speech emotion recognition.

EMO-DB Corpus, also known as Berlin Emotional Speech Database, is a performance corpus recorded by the Technical University of Berlin, Germany, containing 5 men and 5 women, cover seven different emotional types: angry, boredom, disgust, fear, happy, neutral and sad. In this paper, five emotions: angry, fear, happy, neutral and sad, are selected to evaluate the performance of the model.

The full name of RAVDESS corpus is Ryerson Audio-Visual Database of Emotional Speech and Song. It is a widely used speech and emotional corpus recorded by a total of 24 actors including 12 men and 12 women. The corpus is divided into speech part and song part. This paper uses the corpus of songs, and there are six emotions: angry, calm, fear, happy, neutral and sad.

(2) Experimental preparation and evaluation methods

In this paper, NVIDIA TESLA P40 is used as the hardware platform, and tensorflow is selected as the deep learning framework. All corpus files are read at a sampling rate of 16K, and the MFCC of audio segments are extracted by using librosa library. In addition, the data set is divided into training set and verification set according to the ratio of 4: 1. The entire training process is set to 300 rounds, and the Adam optimizer is used to optimize the model. In order to effectively deal with the gradient instability in the training process, this paper adopts a strategy of dynamically adjusting the learning rate: every 100 training rounds, the learning rate is halved. This strategy aims to maintain the convergence speed of the model while ensuring the stability and generalization ability of training. According to the characteristics of each corpus, different initial learning rates are applied. Tables 4 and 5 detail the usage details of the corpus, as well as the dimensions and related parameters of the extracted MFCC.

Table 4. Corpus emotion and quantity statistics table

Corpus	Angry	Calm	Fear	Happy	Neutral	Sad	Number of training	Number of testing	Total
TBSEC001	1200	-	1200	1200	1200	1200	4800	1200	6000
EMO-DB	129	-	69	72	99	135	404	100	504
RAVDESS	184	184	184	184	92	184	663	165	828

Table 5. MFCC dimensions and parameter configuration extracted from each corpus

Corpus	MFCC dimension	FFT window	Frame shift	Mel filter	Learning rate
TBSEC001	45	2048	512	45	0.005
EMO-DB	42	800	200	42	0.0001
RAVDESS	45	800	200	45	0.0001

The indexes used in the model evaluation are unweighted accuracy (UA) and weighted accuracy (WA). The formula is as follows:

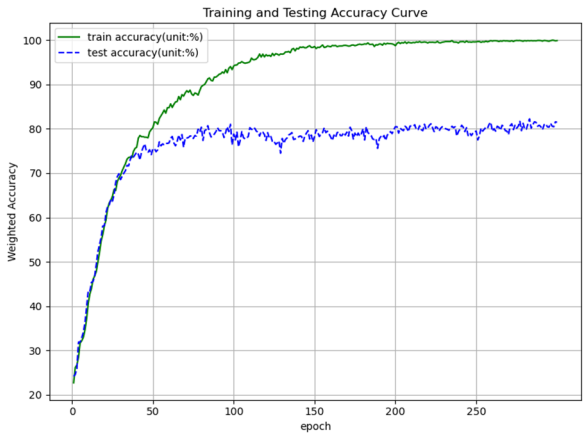
$$UA = \frac{T}{N} \tag{7}$$

$$WA = \frac{\sum \lambda_i T_i}{N} \tag{8}$$

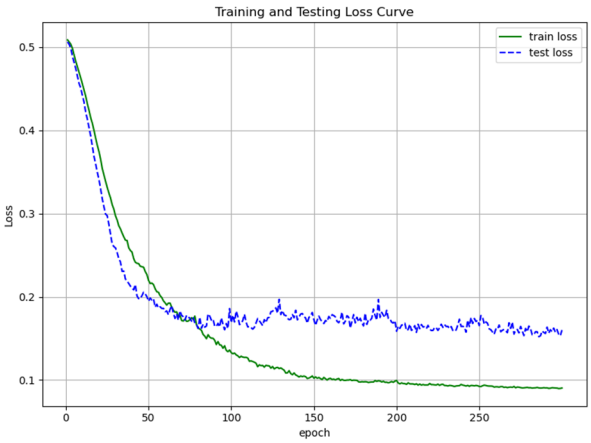
Among them, N represents the total number of samples, T represents the total number of samples with correct prediction, λ_i represents the proportion of Type i of emotion in the total samples, and T_i represents the number of samples with correct prediction for Type i of emotion.

(3) Experimental results and analysis

In the SER task experiment for TBSEC001 corpus, the training process of the model is shown in Figure 2. Within the first 100 rounds of training, the model effectively learns the emotional features in the data, and the test curve converges rapidly. Between 100-300 rounds, although there were some fluctuations in the test part, the overall performance was still good. The final model achieved an accuracy rate of 81.52%, and the loss value was stable around 0.16.



(a) TBSEC001 corpus training and testing accuracy curve



(b) TBSEC001 corpus training and validation loss curves

Fig 2. Accuracy and loss curve of TBSEC001 corpus

In order to explain the prediction of each emotion in more detail, the confusion matrix is drawn in this paper, as shown in Figure 3. Among them, the classification effect of angry emotions is the best, and a total of 83.92% of angry emotions are successfully identified. 11.61% of fear was identified as sad. Overall, TBSEC001 performs well on the model in this paper.

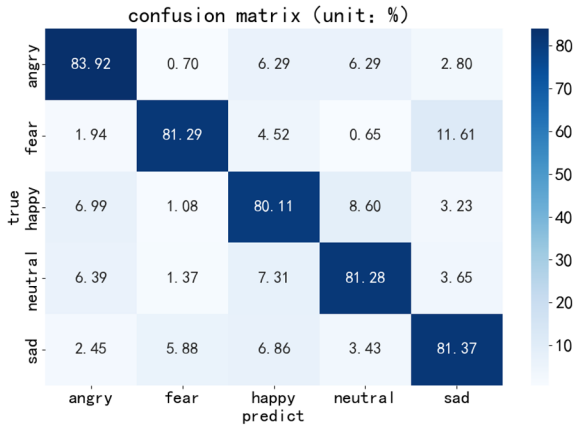
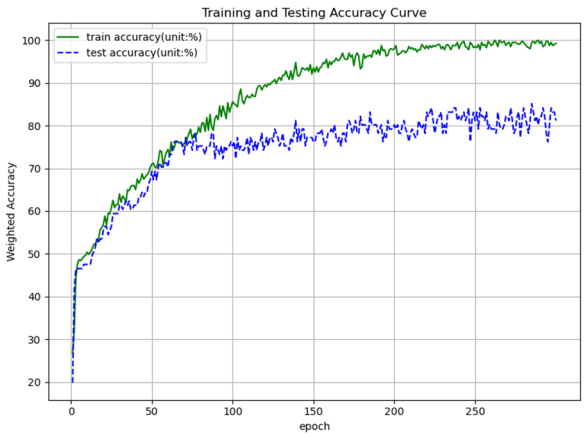
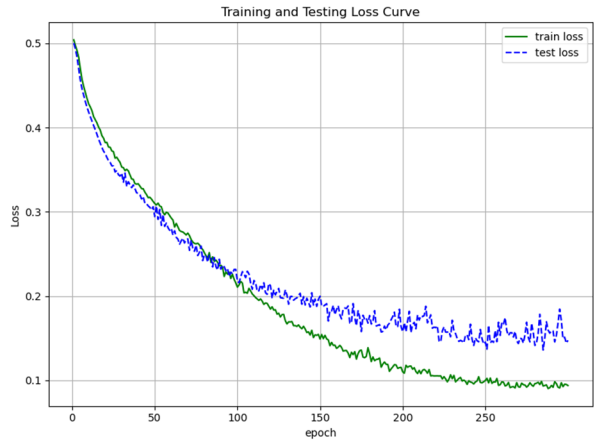


Fig 3. Confusion matrix of this model on TBSEC001 corpus

The training process of the EMO-DB corpus is shown in Figure 4. It can be seen from the figure that with the increase of training rounds, the UA gradually increases, and the maximum UA value of 85.63% is obtained when the training rounds reach about 270. Compared with the accuracy curve, the loss value of the loss curve fluctuates greatly, and the loss value fluctuates between 0.15 and 0.19.



(a) Training and validation accuracy curves of EMO-DB corpus



(b) Training and validation loss value curves of EMO-DB corpus

Fig 4. Accuracy and loss curve of EMO-DB corpus

The confusion matrix in Fig. 5 describes the classification of EMO-DB corpus by the model in this paper, in which the recognition effect of neutral and sad emotions is better, reaching the accuracy rates of 95.00% and 96.88% respectively; However, the recognition ability of fear and happy emotions is low, which may be due to their fewer corpus files, which leads to the better classification effect of other emotions.

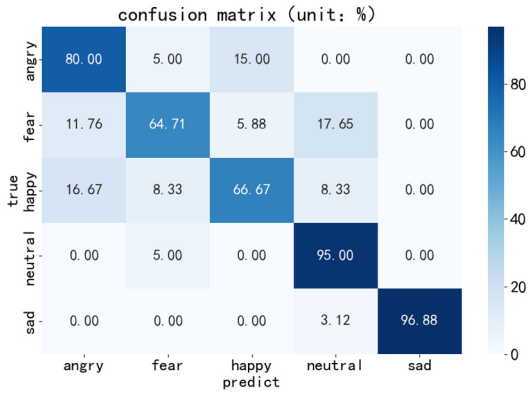
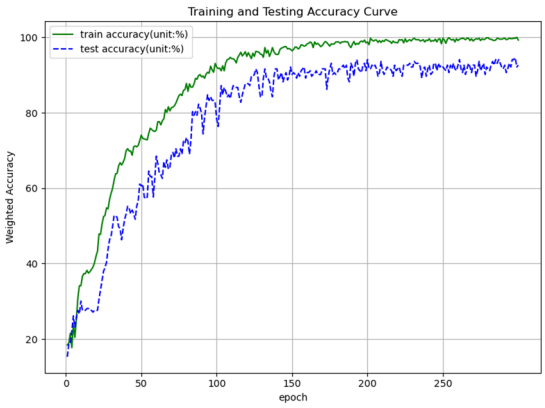
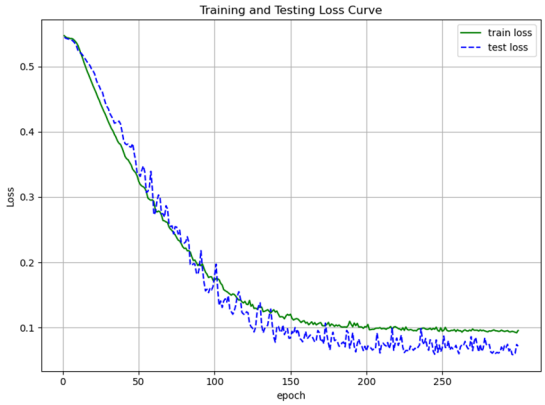


Fig 5. Confusion matrix of this model on EMO-DB corpus

Figure 6 depicts the training of the model in this paper on the RAVDESS corpus. It can be seen from the figure that the accuracy of the model increases with the increase of training rounds, and the maximum UA value is 95.54% around the 290th round; Loss values fluctuate between 0.05 and 0.1.



(a) Training and validation accuracy curves of RAVDESS corpus



(b) Training and validation loss value curves of RAVDESS corpus

Fig 6. Accuracy and loss curve of TBSEC001 corpus

In the confusion matrix of Figure 7, calm, happy and neutral emotions perform best, and their accuracy rates all reach 100.00%. Fear was the worst, and 6.98% of the corpus were predicted to be angry and sad.

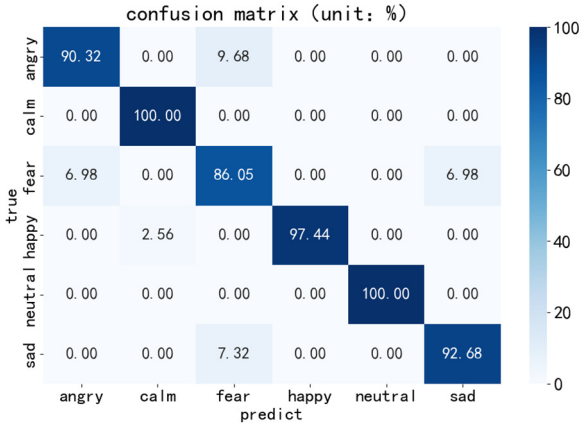


Fig 7. Confusion matrix of this model on RAVDESS corpus

(3) Comparative experiment

In this paper, the proposed model is compared with the speech emotion recognition models published in the past three years, and comparative experiments are carried out on TBSEC001, EMO-DB and RAVDESS corpus. See Table 6 for details.

It can be seen from Table 8 that although the model in this paper does not perform well enough on TBSEC001 and EMO-DB corpus, only the single feature of MFCC is used as input when dealing with TBSEC001 corpus. Compared with the complex feature set of multiple LLDs and high-level statistical functions (HSFs) in literature [19], its feature dimensions are greatly simplified. On the RAVDESS corpus, this model has achieved the best results.

Table 6. Comparison between the model in this paper and the existing model

Model	Trait	TBSEC001 (Accuracy)	EMO-DB (UA)	RAVDESS (UA)
Literature [19]	LLDs & HSFs	**88.40%**	84.10%	74.30%
Literature [28]	MFCC	-	**95.17%**	91.93%
Literature [29]	Mel Spectrogram	-	90.20%	85.80%
The model of this paper	MFCC	81.52%	85.63%	**95.54%**

4. Conclusion

This paper focuses on the field of Tibetan speech emotion recognition, and constructs a lightweight speech emotion recognition model for the problems of relatively few research and complex feature extraction methods. The model takes MFCC as input features. By extracting the spatiotemporal features of MFCC and combining the dynamic routing update algorithm of capsule network, Margin-Loss is used to guide the model to update parameters. Satisfactory recognition results are achieved in the self-built corpus TBSEC001 and public corpora EMO-DB and RAVDESS. In addition, compared with the existing excellent models, the experiment has a good performance, and also shows strong competitiveness and generalization ability. In future work, based on this model, we will combine the knowledge of deep learning and machine learning, new algorithms and models will be proposed to better handle speech emotion recognition tasks.

References

[1] Mencattini A, Martinelli E, Ringeval F, et al. Continuous estimation of emotions in speech by dynamic cooperative speaker models [J]. IEEE transactions on affective computing, 2016, 8 (3): 314-327.

[2] Hashem A, Arif M, Alghamdi M. Speech emotion recognition approaches: A systematic review [J]. Speech Communication, 2023: 102974.

[3] Al-Dujaili M J, Ebrahimi-Moghadam A. Speech emotion recognition: a comprehensive survey [J]. Wireless Personal Communications, 2023, 129 (4): 2525-2561.

[4] Atmaja B T, Akagi M. The effect of silence feature in dimensional speech emotion recognition [J]. arXiv preprint arXiv: 2003.01277, 2020.

[5] 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.

[6] Fahad M S, Ranjan A, Yadav J, et al. A survey of speech emotion recognition in natural environment [J]. Digital signal processing, 2021, 110: 102951.

[7] Jain M, Narayan S, Balaji P, et al. Speech emotion recognition using support vector machine [J]. arXiv preprint arXiv: 2002. 07590, 2020.

[8] Aouani H, Ayed Y B. Speech emotion recognition with deep learning [J]. Procedia Computer Science, 2020, 176: 251-260.

[9] Wang J, Xue M, Culhane R, et al. Speech emotion recognition with dual-sequence LSTM architecture [C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 6474-6478.

[10] Liu G, He W, Jin B. Feature fusion of speech emotion recognition based on deep learning [C]//2018 International conference on network infrastructure and digital content (IC-NIDC). IEEE, 2018: 193-197.

[11] Guo L, Wang L, Dang J, et al. A feature fusion method based on extreme learning machine for speech emotion recognition [C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 2666 − 2670.

[12] Bandela S R, Kumar T K. Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC [C]//2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2017: 1-5.

[13] Lieskovská E, Jakubec M, Jarina R, et al. A review on speech emotion recognition using deep learning and attention mechanism [J]. Electronics, 2021, 10 (10): 1163.

[14] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning [J]. Neurocomputing, 2021, 452: 48-62.

[15] Vaswani A. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017.

[16] Ma J, Tang H, Zheng W L, et al. Emotion recognition using multimodal residual LSTM network [C]//Proceedings of the 27th ACM international conference on multimedia. 2019: 176-183.

[17] Xie Y, Liang R, Liang Z, et al. Speech emotion classification using attention-based LSTM [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27 (11): 1675-1685.

[18] Pengmao Tashi, Cai Zhijie, Cai Rang Zhuoma. Construction of Tibetan emotional speech database [J]. Journal of Peking University (Natural Science Edition), 2023, 59 (05): 773-781. DOI: 10.13209/J.0479-8023.2022. 121.

[19] Gu Zeyue, 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.

[20] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules [J]. Advances in neural information processing systems, 2017, 30.

[21] Lian Z, Liu B, Tao J. DECN: Dialogical emotion correction network for conversational emotion recognition [J]. Neurocomputing, 2021, 454: 483-495.

[22] Li S, Xing X, Fan W, et al. Spatiotemporal and frequential cascaded attention networks for speech emotion recognition [J]. Neurocomputing, 2021, 448: 238-248.

[23] Chen J, Liu Z. Mask dynamic routing to combined model of deep capsule network and u-net [J]. IEEE transactions on neural networks and learning systems, 2020, 31 (7): 2653-2664.

[24] Wen X C, Ye J X, Luo Y, et al. Ctl-mtnet: A novel capsnet and transfer learning-based mixed task net for the single-corpus and cross-corpus speech emotion recognition [J]. arXiv preprint arXiv: 2207.10644, 2022.

[25] Livingstone S R, Russo F A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English [J]. PloS one, 2018, 13 (5): e0196391.

[26] Qu Aitangs "Research on Tibetan Finals" [M] Qinghai Ethnic Publishing House. July 1991

[27] Yang Jie, Li Yonghong, Hu Axu, et al. Study on tone and laryngeal plug rhyme perception in Tibetan Lhasa dialect [J]. National Languages, 2023, (04): 101-110.

[28] Ye J, Wen X C, Wei Y, et al. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition [C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[29] Sadok S, Leglaive S, Séguier R. A vector quantized masked autoencoder for speech emotion recognition [C]//2023 IEEE International conference on acoustics, speech, and signal processing workshops (ICASSPW). IEEE, 2023: 1-5.