Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 11, No. 1, 2025

Bimodal Emotion Recognition based on Sichuan Dialect

Jia Wei *, Xiangguo Sun

Mechanical Engineering College, Sichuan University of Science and Engineering, Yibin, Sichuan, China
*** Corresponding author:** Jia Wei (Email: 18040480929@163.com)

Abstract: In view of the main technical challenges faced by dialect emotion recognition (data scarcity and low recognition rate), this paper discusses the Sichuan dialect emotion recognition technology, and proposes a multimodal emotion recognition model by constructing a high-quality dataset containing multiple emotion categories. The model adopts the dual-modal fusion strategy of Mel spectral features (MFCCs) and text features, uses dynamic convolutional network (ODConv) and convolutional block attention module (CBAM) to extract features, and combines the Text-CNN model for text sentiment analysis. Experimental results show that the proposed model has higher accuracy and robustness than the traditional speech recognition model CNN in Sichuan dialect emotion recognition task.

Keywords: Dynamic Attention; Channel Spatial Attention; Multimodal Fusion; CBAM.

1. Introduction

With the rapid development of artificial intelligence technology, the ways in which humans interact with intelligent systems have become more diverse. In this context, the application value of emotion recognition technology is becoming more and more prominent, especially for the recognition of dialect emotion. The integration of speech recognition and natural language processing technology, especially the breakthrough in the field of emotion recognition, provides strong support for personalized services in multiple scenarios. With the rapid rise of artificial neural networks, especially deep learning technology, the performance of emotion recognition models has been greatly improved, laying a solid foundation for emotion recognition in more complex language environments.

At present, a number of well-known speech emotion datasets in the world have promoted the in-depth development of research. However, much research has focused on Mandarin or the international lingua franca, while the study of dialects has often been neglected. For some specific scenarios, such as Sichuan dialect [1], the importance of emotion recognition lies in the fact that it cannot only promote more efficient cross-language and cross-cultural communication in Sichuan, but also provide more accurate support for local social needs.

The main content of the research in this dissertation:

In view of the problems existing in the current emotion recognition technology, this paper conducts an in-depth study. The main research contents include the following aspects:

(1) Phonetic dataset construction: In order to overcome the problem of insufficient database, this paper collects phonetic samples about Sichuan dialects, covering a variety of emotion categories. At the same time, speech recognition technology is used to transcribe speech into text to ensure that the dataset covers a wide range of sentiment categories and maintain the consistency of sentiment annotation. This step provides strong support for subsequent emotional feature extraction and model training.

(2) Emotion recognition model design: Most of the existing models are trained based on Putonghua or international common language datasets, which are difficult to be directly applied to dialect emotion recognition tasks. Therefore, it is necessary to optimize and verify the model for Sichuan dialect. In this paper, the current efficient network model is selected for feature extraction, which is helpful to improve the accuracy and robustness of emotion recognition.

(3) Aiming at the problem of low recognition rate of individual categories: this paper makes up for it by text emotion recognition, and adopts a dual-modal emotion recognition method based on the fusion of speech and text. By constructing a rich speech dataset, extracting comprehensive emotional features, and fusing multimodal information, the recognition of emotions is realized.

2. Dataset

The speech emotion corpus is the basis for speech emotion recognition, and the size and quality of the database directly determine the performance of the speech emotion recognition system trained based on the corpus. At present, the construction of the speech emotion database is based on the specific requirements of the task. The production of the dataset in this paper comes from two aspects: one is the use of recording software to record the human voice, which is performed by 3 men and 3 women according to the prepared text manuscript, which belongs to the spontaneous emotion dataset (that is, the emotion is not specified during the recording,); The second is to obtain audio and video through the network and edit and process them in Adobe Audition software. In this work, the speech data was extracted from Sichuan dialect film and television dramas. The main job of this part is to watch and check each set of videos, select videos that contain clear emotions and have no background music and noise, and then extract them from the video as audio files using Adobe Audition, and divide them into smaller audio according to the nature of the speech in a single audio, each audio duration is controlled at 3-5s, the sampling rate is 16000Hz, mono, and the bit depth is 16.

Table 1. The number of different emotional categories

category	Angry	Fear	Happy	Neutral	Surprise	Sad
number	1002	336	1001	1022	394	755

3. Related Work

Emotion recognition has been studied for decades, with scholars extracting features from audio data and then applying those features to a range of classifiers. These classifiers include: hidden Markov models, convolutional recursive networks, SVMs, hierarchical binary decision trees, Gaussian mixtures, neural networks, and more. Much of the above work relies on context to provide additional information to infer the emotional content extracted from the data.

In the process of developing emotion recognition, scholars have proposed a variety of innovative models. For example, attention-based networks, which are designed to align text and audio information and perform feature extraction. However, with the development of technology, Dynamic Convolutional Networks [2] has gradually become a research hotspot due to its efficient feature extraction ability and good processing ability of time series data. Some scholars have applied dynamic convolutional networks to speech emotion recognition, and they have achieved a more accurate capture of emotional information in audio data by adjusting the network structure and parameters. At the same time, Tang et al. [2]proposed a multimodality-based model as a benchmark, which uses a Multi-Modal to effectively fuse speech, text, and external knowledge. However, in the field of feature extraction, the Convolutional Block Attention Module [4] (CBAM) has attracted much attention due to its powerful attention mechanism. Another scholar introduced CBAM and dynamic convolution[4]models into the field of speech emotion recognition, and they used the attention mechanism of CBAM to enhance the recognition ability of key emotional information by combining the features of audio and text data. In addition, in terms of text recognition, the Text-CNN model is widely used because of its efficient text feature extraction ability. Combining the Text-CNN model with audio emotion recognition enables a more comprehensive analysis of sentiment information by extracting key information from the text and combining it with audio features.

4. Experimental Methods

In this section, we will introduce the dynamic convolutional model, CBAM model, and Text-cnn sentiment recognition model based on the previous ones, and the model we propose is based on the above model. The model uses two different data modalities as input sources: MFCCs and word vectors. Initially, each modality is handled individually. The overall model structure is shown in Figure 1. The architecture is a multimodal classification model, which combines the spectral features of speech signals (MFCC) and the text generated by speech recognition, extracts feature through two parallel paths and fuses them to complete the classification task. The audio path focuses on capturing the low-level acoustic features of speech, while the text path captures high-level semantic information. The two paths of the model are respectively for speech and text modalities, and the dynamic convolutional network, CBAM model and attention mechanism are used to extract their respective features, and finally, the fused features are generated by the fully connected layer and the classification module. Speech Modal Input: The input is a speech signal, and the acoustic signature is extracted by the Mel frequency cepstrum coefficient (MFCC). Text modal input: Speech signals are transcribed into text form by speech.

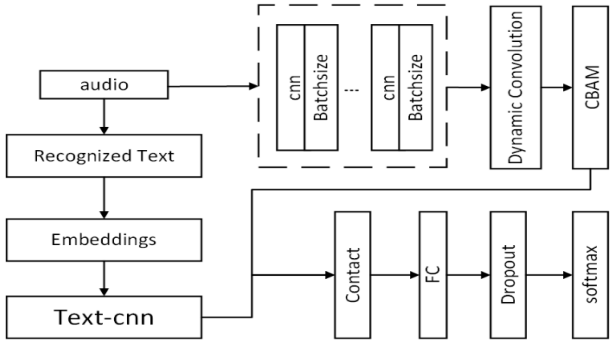

Figure 1. Flow chart of multimodal emotion recognition

4.1. Audio Modal Processing

4.1.1. MFCC Coefficient Extraction

The Mel frequency cepstrum coefficient (MFCC) is extracted from the input speech signal to obtain a two-dimensional feature matrix: is the number of time steps, $\mathbf{F}_{\text{mfcc}} = f_{\text{mfcc}}(A), \mathbf{F}_{\text{mfcc}} \in \mathrm{R}^{T\times d}$, T is the number of time steps, and d is the MFCC feature dimension. MFCC preserves the spectral properties of speech and is a low-dimensional feature that describes the physical properties of speech.

4.1.2. Audio Modal Feature Extraction

(1) Convolutional Neural Network (CNN) feature extraction: MFCC features are processed through multi-layer convolutional networks to extract local temporal and spatial features: convolutional networks can capture the feature patterns of speech signals in different time windows. $\mathbf{F}_{\text{cnn}} = f_{\text{cnn}}(\mathbf{F}_{\text{mfcc}})$

(2) Dynamic convolution and CBAM (Attention Module): Dynamic convolution enhances the ability to model the change of speech signals over time. The Convolutional Block Attention Module (CBAM) focuses on the most important acoustic features through channel attention and spatial attention mechanisms. Final output processed audio characteristics: $\mathbf{F}_{\text{audio}} = f_{\text{cbam}}(\mathbf{F}_{\text{cnn}})$. ODConv[5]is created based on DyConv and CondConv[7]]. Unlike traditional convolution, which has fixed weights, ODConv dynamically generates kernel weights, allowing for more flexible and input-dependent feature extraction. This adaptive feature enables ODConv to capture more complex and diverse patterns in the data, especially when used in MFCC's data. Figure 2 illustrates the structure of ODConv, where $w_i$the convolutional kernel is represented and the compressed eigenvectors are mapped into a low-dimensional space at the FC layer. Compared with ordinary convolutional networks (CNNs), the ODConv has many advantages when identifying MFCC, and the dynamic convolutional kernel can be adaptively adjusted according to the local features of the input feature map. This adaptability allows the network to better capture features in different audio. The dynamic convolutional kernel enhances the robustness of the model to changes in the input data, while capturing both global featuresand local features.

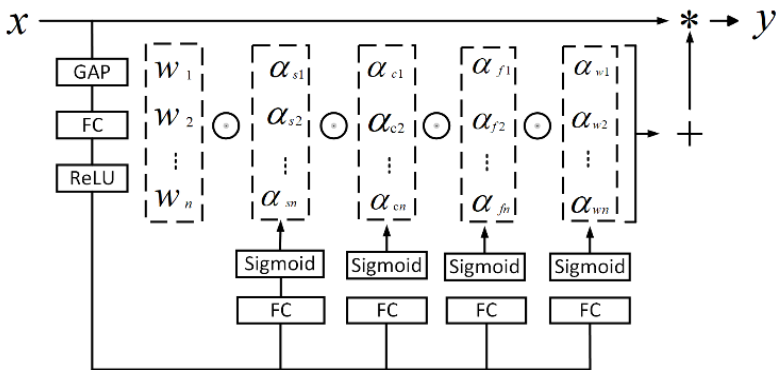

Figure 2. ODConv model diagram

where , represents a_{ω_i} the scalar attention factor of the convolution kernel; $a_{ci} \in R^{C_{in}}$, representing $a_{fi} \in R^{C_{out}}$, $a_{si} \in R^{k\times k}$ three new attention factors calculated along the spatial, output, and input dimensions, respectively. ⊗ represents element-by-element multiplication; * denotes a convolution operation.

(3) The CBAM model [9] mainly includes the channel attention mechanism as shown in Figure 3 and the spatial attention mechanism as shown in Figure 4. The main purpose of the channel attention mechanism is to calculate the attention weight in the channel dimension, so as to highlight the important channel information. CBAM can effectively enhance the model's ability to capture key information by calculating the attention weights in the channel and spatial dimensions respectively. The channel attention mechanism focuses on the importance of different channels, while the spatial attention mechanism focuses on the importance of different spatial positions. Through the combination of these two mechanisms, CBAM is able to significantly improve the performance of the model in a variety of tasks, especially those with complex features and backgrounds.

The formula for calculating the channel attention mechanism is as follows:

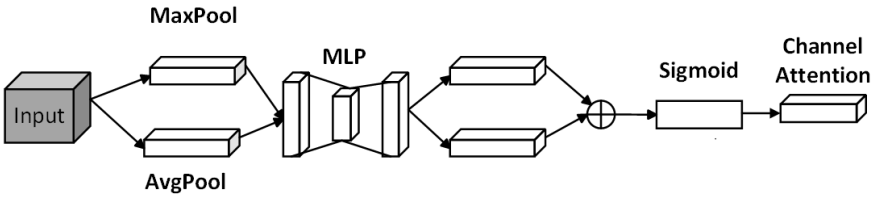

Figure 3. Channel attention mechanism

The input is a feature tensor $x \in R^{c\times H\times W}$, where C is the number of channels and H and W are spatial dimensions. The input feature maps are subjected to global average pooling and global maximum pooling, respectively, to extract spatial information and perform channel compression. Obtain two channel description vectors. The results of average pooling and maximum pooling are input into a two-layer MLP with shared weights, which is used to extract nonlinear relationships between channels. The outputs of the two MLPs are fused by element wise addition, and the fusion result is activated by a Sigmoid activation function to generate attention weights for each channel. The generated channel weights S are used to weight the input feature map, and the channel attention enhanced features are output.

$$g_{avg} = MLP(f_{avg}) \quad (1)$$

$$g_{max} = \mathrm{MLP}(\mathrm{f_{max}}) \quad (2)$$

$$MLP(f) = w_2\sigma(\mathrm{w_1 f}) \quad (3)$$

Among them, $\omega_1 \in R^{C\times\frac{C}{r}}$, $\omega_1 \in R^{\frac{C}{r}\times C}$ .For the weights of two fully connected layers, where r is the compression factor for the number of channels. σ is the activation function, and ReLU is usually chosen.

Global Mean Pooling (AvgPool): Averages feature maps in the spatial dimensions (H and W). Global MaxPool: Takes the maximum value in the spatial dimension of the feature map the pooled features are passed through a shared multilayer perceptron (MLP), which usually contains a hidden layer, and the dimension of the hidden layer can be compressed by the number of channels. where the sum is the weight of the MLP, and σ is the activation function (sigmoid).W_1W_2 By adding the average pooling features and the maximum pooling features of MLP, the fusion features are generated by the sigmoid activation function, and each channel of the input feature map is multiplied by the corresponding attention weight to achieve channel enhancement.

As shown in Figure 4, the main purpose of the spatial attention mechanism is to calculate the attention weights in the spatial dimension, so as to highlight important spatial regions. The main calculation formula is as follows.

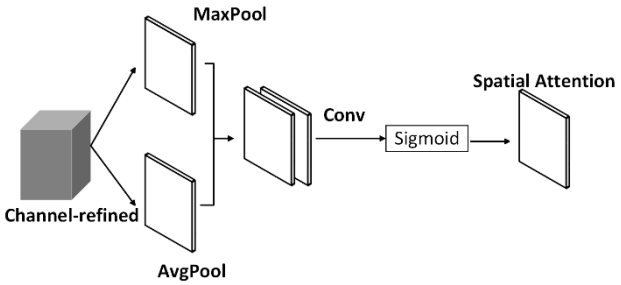

Figure 4. Spatial attention mechanisms

$$F_{avg}(i,j) = \frac{1}{C}\sum_{c=1}^{C} F^c(i,j) \quad (4)$$

$$F_{max}(i,j) = \max_{c=1}^{C} F^c(i,j) \quad (5)$$

$$F_{cat}(i,j) = F_{avg}(i,j) + F_{max}(i,j) \quad (6)$$

$$M_s = \sigma(\mathrm{convolution}(F_{cat})) \quad (7)$$

$$F_c(i,j,c) = A_c^c \cdot F(i,j,c) \quad (8)$$

$$F_{out}(i,j,c) = A_s(i,j) \cdot F_c(i,j,c) \quad (9)$$

Perform a channel-by-channel global average pooling operation on each channel of the input feature map F to obtain a spatially averaged feature map F_{avg}. Then, a channel by channel global max pooling operation is performed on each channel of the input feature map F to obtain a spatial dimension maximum feature map F_{max}. Add the two feature maps, F_{avg} and F_{max}, element by element to obtain a comprehensive feature map, F_{cat} Perform convolution operation on the comprehensive feature map F_{cat} to reduce the number of channels and increase nonlinearity. σ is the Sigmoid activation function.

4.2. Text Modal Processing

4.2.1. Text Data Preprocessing

Before you can feed text data into your model, we need to do some preprocessing steps to ensure that the data is suitable for model training and inference. Start by removing noise from the text (HTML tags, special characters, punctuation, numbers, etc.). Use the Jieba tool to tokenize text fragments and split each text fragment into a word list. Converts the segmented text into numeric form. Using a word vector model more suitable for the Chinese context, Word2vec, the word sequence is mapped to a high-dimensional vector to obtain a text vector.

$$\mathbf{E}_{\text{text}} \in \mathrm{R}^{N\times d} \tag{10}$$

where N is the length of the text sequence (i.e., the number of words in the text) and d is the embedding dimension.

4.2.2. Feature Extraction

In the text classification task, Text-CNN (text convolutional neural network) effectively extracts features from text data through a series of convolution and pooling operations. The core mechanism lies in the use of convolutional layers, which are able to slide filters (or convolutional kernels) of various sizes over the input word embedding matrix to capture local semantic features in the text.

Specifically, the convolutional layer of Text-CNN is configured with multiple filters of different sizes, using filters of 3, 4, and 5 word lengths, which are specifically used to identify and extract the combined features of consecutive n words. In this way, the model has the flexibility to capture phrases and structures of different lengths, which is essential for understanding the semantics of the text. In the feature extraction process, the convolution operation first slides these filters across the word embedding matrix to generate a series of feature maps, each corresponding to a specific type of local feature captured by one filter. After the nonlinear activation function (ReLU) is processed, these feature maps are passed to the pooling layer to reduce the dimensionality and select the features generated by the convolution.

For the text transcription information of a given speech segment, Jieba tool is used for preprocessing such as segmentation and cleaning of the text segment. Each text segment is divided into a word list $w = [w_1, w_2, \cdots, w_n]$, and the maximum word length of the text is set to N. The excess part of the text longer than N words will be discarded, and the text less than N words will be filled with zeros. Then, a 300 dimensional word vector is used for representation embedding, resulting in a text vector $t = [\bar{w_1}, \bar{w_2}, \cdots, \bar{w_n}]$. Then, by encoding the text data, the text feature f_t is obtained.

4.3. Multimodal Fusion

The text features and acoustic features were fused at the feature layer, firstly, the acoustic features f_a obtained by the acoustic composite features encoded by the model and the text features f_t obtained by word embedding and network encoding were fused. Perform series connection to obtain features.

$$f = \text{concat}(f_a . f_t) \tag{11}$$

where is the concatenation function.concat

Then, the function is applied to predict sentiment classification, and the distribution probability of the prediction target is calculated as follows.Softmax$\boldsymbol{E}$

$$E = \text{Softmax}(f^T M + b) \tag{12}$$

where sum is the learned parameter matrix and bias vectorMb

5. Experimental Results and Evaluation Indicators

In this study, we implemented a prototype of a multimodal emotion recognition algorithm and evaluated the performance of the dataset. To train this dataset, the model went through 50 steps of training. The optimizer used is Adam. We set the learning rate to 1e-5 and the weight decay to 1e-3.

We used all samples for 6 categories, including ang, happy, neutral, and sad, fear, surprise. This is a common setting for emotion recognition. We use accuracy to comprehensively measure the performance of the model. The Confusion Matrix is a table used to evaluate the performance of classification models. It displays the correspondence between the predicted results of the model and the actual labels, helping us to gain a detailed understanding of the model's performance in different categories. We split the dataset into 8/1/1 for training/val/test settings. We trained our model in the training section, which includes 80% of the data in the dataset. The final model was selected based on their performance in the 10% validation section. To demonstrate the effectiveness of the dataset and evaluate the performance of the model, we implemented three other models for comparison (i.e. Table 2). Table 2 shows the results of different models, compared with our model, which achieves improved multimodal fusion through enhanced attention mechanism. Below are our proposed models (Custom CNN) and Text CNN, as well as the accuracy after modal fusion. Figure 5 shows the loss of two models on the training set, and Figure 6 shows the training and testing performance of the model after modal fusion.

6. Conclusion

In this paper, we propose a sentiment recognition dataset for Sichuan dialects. Compared with the existing Chinese datasets, the proposed datasets are richer and more diverse in terms of content, and these samples are all from everyday conversations and are closer to real-life scenarios. In addition, this paper proposes a multimodal emotion recognition model, which uses dynamic convolution and attention structure for multimodal fusion. In the future, we will try to improve the quality of the model and improve the accuracy, and we will also recognize audio fragments without noise or distortion, and identify emotional data without language barriers, and contribute to improving sound quality.

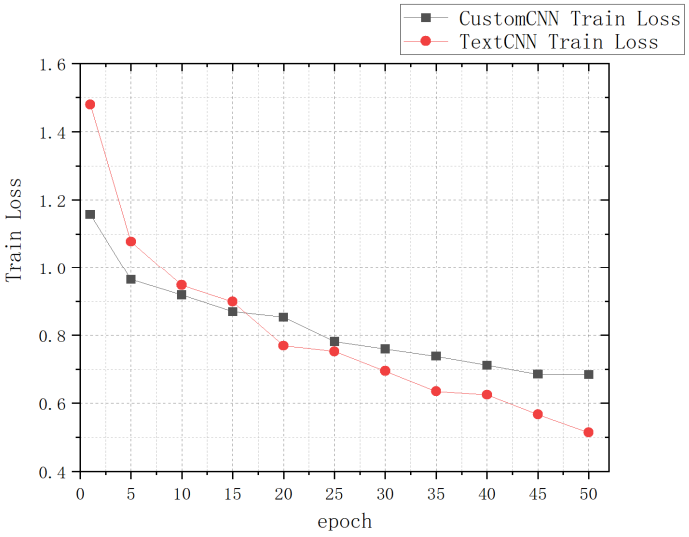

Figure 5. Training loss of two modal models

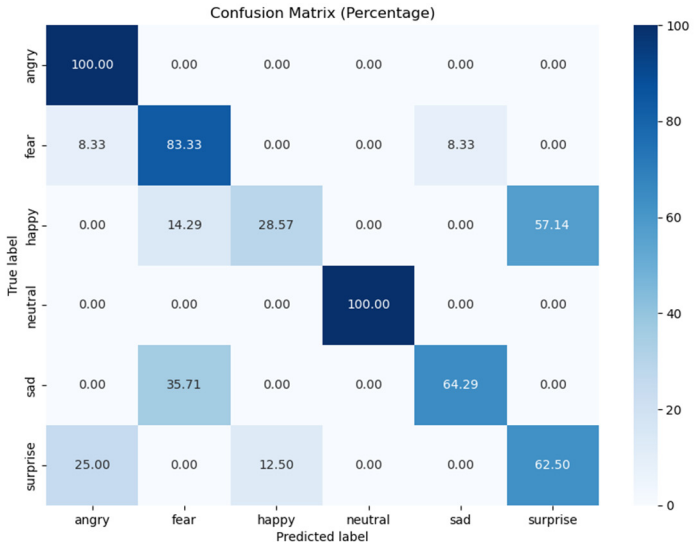

Figure 6. Performance on test set after modal fusion

References

[1] Xie Jinhong, Wei Xia. Sichuan dialect speech recognition based on ResCNN-BiGRU [J]. Modern Electronic Technology, 2024, 47 (01): 89-93. DOI: 10.16652/j.issn.1004-373x.2024. 01. 016.

[2] Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., & Liu, Z. (2020). Dynamic convolution: Attention over convolution kernels. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11030 11039).

[3] Jinghua Tang, Liyun Zhang, Yu Lu. VCEMO: Multi-Modal Emotion Recognition for Chinese Voiceprints. arXiv preprint arXiv:2408.13019 (2024).

[4] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).

[5] Quang-Anh N.D., Manh-Hung Ha, Thai Kim Dinh. EMOTIONAL VIETNAMESE SPEECH-BASED DEPRESSIONDIAGNOSIS USING DYNAMIC ATTENTION MECHANISM. arXiv preprint arXiv:2412. 08683 (2024).

[6] Li, Y., Xin, Y., Li, X., Zhang, Y., Liu, C., Cao, Z., ... & Wang, L. (2024). Omni-dimensional dynamic convolution feature coordinate attention network for pneumonia classification. Visual computing for industry, biomedicine, and art, 7(1), 17.

[7] Li, C., Zhou, A., & Yao, A. Omni-dimensional dynamic convolution. arXiv preprint arXiv:2209.07947 (2022).

[8] Yang, B., Bender, G., Le, Q. V., & Ngiam, J. (2019). Condconv: Conditionally parameterized convolutions for efficient inference. Advances in neural information processing systems, 32.

[9] Mao Xueli Research on Language Recognition Methods Based on Convolutional Networks and Attention Mechanisms [D]. Xinjiang University, 2021. DOI: 10.27429/d.cnki. gxjdu. 2021. 000429.

[10] Wang Mingtian Research on Speech Emotion Recognition Based on Text and Acoustic Features [D]. Shandong University, 2022. DOI: 10.27272/d.cnki.gshdu.2022.001740. Research on Language Recognition Methods Based on Convolutional Networks and Attention Mechanisms.