ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Multi-ethnic Spring Festival Ballads in Gansu Province MFCC parameters extraction LSTM neural network training sentiment classification model

Yijia Ding[1,*]

[1] School of Music and Dance, Lanzhou University of Arts and Science, Lanzhou, 730000, Gansu, China

SUMMARY: *The development of digital humanities and intelligent audio analysis technology provides a new computational path for the emotional research of multi-ethnic folk songs. Taking the multi-ethnic Spring Festival ballads of Gansu Province as the object, this paper constructs a sentiment classification model by focusing on corpus collection, audio preprocessing, MFCC parameter extraction and LSTM neural network training. A total of 526 valid samples were sorted out, with a cumulative duration of 1149.2 minutes. The samples were labeled as four types of emotions: celebration, blessing, thinking and expressing, and narrative peace, and 39 dimensional MFCC temporal features were extracted as model input. Experimental results show that the model training loss decreases from 1.31 to 0.11, the training accuracy reaches 91.9%, and the validation accuracy reaches 85.4%. In the test set of 104 samples, the model correctly identified 90 samples, and the overall accuracy was 86.5%, the Precision, Recall and F1-score were 86.5%, 86.6% and 86.6%, respectively, which were better than SVM, CNN, RNN and GRU. The results show that the combination of MFCC and LSTM can effectively represent the emotional acoustic features in the Spring Festival songs, which provides technical support for the digital protection, emotional label construction and intelligent retrieval application of multi-ethnic Spring Festival songs in Gansu province.*

KEYWORDS: *Multi-ethnic Spring Festival songs; MFCC parameter extraction; LSTM neural network; Sentiment classification model*

1 Introduction

Spring Festival ballads are a kind of sound texts with the most life temperature in the folk oral tradition, which assume unique functions in seasonal rituals, emotional expression and group memory inheritance [1]. Gansu is located in a multi-ethnic cultural intersection area. Different regions and different nationalities have formed different styles of ballads during the Spring Festival, with distinct differences in melody direction, rhythm organization, sound mode and emotional color. Such sound materials not only contain festival wishes, family ethics, labor memory and regional aesthetics, but also retain traces of ethnic exchanges and cultural integration [2]. In the past, researches on Spring Festival ballads mostly focused on folklore, musicology and intangible cultural heritage protection, and paid attention to text sorting, melody tracing and cultural interpretation. However, quantitative analysis of their acoustic structures is still insufficient, especially in the automatic identification of emotional categories,

*dyj10013@163.com
https://doi.org/10.65102/is2026309
Published: 30 April, 2026

there is no technical path that takes into account both the characteristics of folk music and the ability of computational models [3, 4]. At the same time, there are still some differences in the acquisition standards, audio quality, label system and analysis dimensions of the existing relevant materials, which makes it difficult to form a unified digital research framework between Spring Festival ballads of different regions and different ethnic groups, and also limits the in-depth development of subsequent intelligent recognition, emotional retrieval and cross-regional comparative research. Based on this situation, we believe that it is necessary to carry out a more systematic quantitative research on multi-ethnic Spring Festival ballads in Gansu Province from the perspective of combining audio computing and emotion recognition.

With the development of digital humanities, speech signal processing and deep learning technologies, traditional folk song research is moving from static description to computable analysis [5]. A large number of details contained in audio signals, such as spectrum distribution, formant variation, energy fluctuation and timing characteristics, provide a new observation entrance for revealing the emotional expression mechanism of songs. MFCC parameter can well represent the short-time spectral envelope characteristics of sound, and has been widely used in speech recognition, speaker recognition and affective computing tasks. LSTM neural network has strong modeling ability for sequence data, can capture the continuous change of audio features in the time dimension, and has strong adaptability to deal with the common drawling, stress, stop connection and gradual change of emotion in ballad singing [6, 7]. Based on the above understanding, we try to combine MFCC parameter extraction with LSTM network training, hoping to improve the accuracy of sentiment classification of Spring Festival ballads and provide new method support for digital preservation and intelligent retrieval of folk music. For multi-ethnic symbiotic regions such as Gansu, we also hope to use audio computing technology to identify the emotional differences in songs, so as to provide a more operational analysis basis for regional cultural characteristics refinement, hierarchical arrangement of folk music resources and optimization of digital transmission methods.

However, Spring Festival ballads are not ordinary spoken speech, with strong melody, free rhythm, obvious dialect differences, and complex singing scenes and recording conditions. It is often difficult to obtain stable effects by directly applying general speech emotion recognition methods [8]. Based on this, we take the multi-ethnic Spring Festival ballads of Gansu Province as the research object, focus on corpus collection and arrangement, audio preprocessing, MFCC feature extraction and LSTM sentiment classification model training, and try to build a computational framework suitable for audio analysis of festival ballads. We hope to promote the deep integration of folk music research and artificial intelligence methods on the basis of preserving the cultural attributes of ballads, and provide referable technical solutions for the digital protection, emotional label construction and subsequent intelligent application of multi-ethnic Spring Festival ballads.

2 Theoretical basis and related research

2.1 Emotional Expression Characteristics and Audio computational analysis basis of Multi-ethnic Spring Festival Ballads in Gansu Province

The emotional expression of multi-ethnic Spring Festival ballads in Gansu Province does not simply rely on the literal meaning of the lyrics. The deeper appeal often comes from the joint effect between the melody direction, the rhythm tension, the speed change, the pitch fluctuation and the singing tone. As the Spring Festival is an important node in the transition

of the New Year, the emotions carried by the ballads are usually obviously complex: there is not only the bright atmosphere of welcoming the New Year, but also the eager expectation of family reunion, harvest vision and auspicious life. Some works also contain the solemn sense brought by local memory, clan identity and national etiquette and customs [9, 10]. Because of this, the same festival singing, different ethnic groups, different regions, different ways of singing the emotional color is not consistent, some of the more enthusiastic, some more implicit stretch, some emphasis on rhythm promotion, and some rely on dragging and revolving to form emotional extension.

From the perspective of audio computing, these differences will eventually be translated into acoustic features that can be analyzed. The intensity change, fundamental frequency fluctuation, spectrum center of gravity migration, formant distribution and duration structure of Spring Festival songs formed in the singing process actually constitute an important basis for emotion recognition [11]. Cheerful songs often show faster rhythm, higher energy and more obvious high-frequency activity. Lyrical or aspirational ballads tend to have longer notes, more gradual energy changes, and more stable melodic lines. The computational processing of these details can transform the emotional perception originally relying on empirical judgment into trainable, comparable and reproducible data expression, which is also the key premise for traditional music research to enter the intelligent analysis stage [12, 13].

Among the existing audio analysis methods, MFCC can preserve the spectral envelope information of the sound more intensively, and has a good description ability for the frequency band changes that are more sensitive to human perception, so it is suitable for characterizing the timbre differences and emotional cues in Chinese New Year songs. At the same time, ballad emotion is not a simple concatenation of static segments, but gradually unfolds in the flow of time, with continuity of syllables, cohesion of sentences and emotional advancement [14-16]. The LSTM model can track such continuous changes at the time level, which is more suitable for dealing with the phenomenon of trailing notes, grace notes and beat elasticity common in folk singing. It can be seen that the emotional expression research of multi-ethnic Spring Festival ballads in Gansu province connects the regional culture and folk music characteristics at one end, and relies on audio feature extraction and sequence modeling methods at the other end. The combination of the two provides necessary theoretical support for the establishment of subsequent emotional classification models.

2.2 Research status of MFCC parameter extraction and LSTM sentiment classification methods

In recent years, with the continuous development of speech signal processing, machine learning and deep learning technology, the research on emotion recognition has gradually expanded from text analysis to audio analysis [17]. Researchers have begun to pay attention to the emotional cues contained in sound, hoping to use computable features to complete the automatic recognition of speech, songs, folk songs and other acoustic materials. In this process, acoustic feature extraction and temporal model construction have always been the focus of research, among which MFCC and LSTM have become a common combination in audio emotion classification research due to their high technology maturity and wide application range [18-20]. The related results mainly focus on speech emotion recognition, speaker state analysis, emotion judgment of music clips and multimodal emotion computing, which provides a more referable method basis for the research of emotion recognition of folk songs. Combining these studies, we can see that audio emotion classification has gradually formed a more complete technical route from feature extraction to time series modeling.

MFCC is a class of representative spectral features, and its calculation process usually

includes steps such as pre-emphasis, framing, windowing, fast Fourier transform, Mel filter bank processing and discrete cosine transform. After a series of transformations, the original audio signal can be compressed into the characteristic parameters that reflect the law of auditory perception. Previous studies generally believe that MFCC has strong description ability for sound texture, timbre differences and short-time spectral envelope changes, and has stable performance in speech recognition, emotion detection and music information retrieval [21]. For emotion recognition tasks, MFCC can better present the differences in energy distribution and frequency structure accompanied by emotional changes, so it has high value in constructing input features for classification [22-24]. Some studies will also combine auxiliary features such as first-order difference, second-order difference, short-time energy and zero-crossing rate to jointly improve the model's ability to distinguish subtle emotional differences [25]. From these results, we believe that MFCC has strong feasibility as the basic feature of Chinese New Year ballad emotion recognition, and can provide a more stable input representation for subsequent time series modeling.

In terms of model methods, traditional algorithms such as support vector machine, hidden Markov model and decision tree are mostly used in the early stage of sentiment classification. This kind of method has certain effect in small-scale tasks with clear feature dimensions, but it is often difficult to fully capture the dynamic correlation between the before and after frames when facing long audio sequences. The introduction of LSTM provides a new technical path for audio emotion classification. This model retains, updates and outputs historical information through the gated mechanism, and can learn the sequence change law in more detail, so it is suitable for processing audio data with progressive emotion over time [26]. Previous studies have shown that LSTM has good adaptability in continuous speech emotion recognition, singing voice sentiment analysis, and audio classification tasks in complex scenes. Especially in the case of sample length fluctuation and uneven distribution of local emotion features, LSTM has more obvious advantages in modeling context relations [27]. Combined with the characteristics of strong melody, flexible rhythm change, and continuous emotional advancement of Spring Festival ballads, we believe that LSTM has a good method matching in this task.

In general, the existing research has provided a mature theoretical and technical support for the combination of MFCC parameter extraction and LSTM sentiment classification method, but the special research on local, multi-ethnic and festive ballad audio is still relatively limited. Spring Festival ballads have distinct characteristics in melody structure, language habits, emotional expression intensity and singing environment, which put forward higher requirements for feature extraction accuracy, sample annotation quality and model generalization ability. The application of MFCC and LSTM to the emotional classification of multi-ethnic Spring Festival ballads in Gansu Province can not only extend the application boundaries of existing audio emotion recognition research, but also help to promote the digital arrangement and intelligent analysis of ethnic music resources. Based on the above research status, we further carry out the construction of sentiment classification model for multi-ethnic Spring Festival songs in Gansu Province, which has strong problem pertinency and research necessity.

3 Extraction of MFCC parameters of Chinese New Year Ballads LSTM neural network training sentiment classification model construction

3.1 Collection and arrangement of multi-ethnic Spring Festival ballads corpus and audio data preprocessing in Gansu Province

The audio corpus of Spring Festival songs used in this study mainly comes from the field collection data of multi-ethnic living areas in Gansu Province, local cultural digital archives, and the public arrangement of festival audio samples. In order to ensure that the corpus has better regional coverage and emotional identification, we take into account the ethnic origin, singing scene, recording clarity and emotional integrity when selecting samples, and eliminate the audio with repetitive content, heavy noise and short singing clips. The collected corpus is uniformly converted into.wav format, and the sampling rate is fixed to 16 kHz and the number of quantization bits is set to 16 bit to reduce the parameter offset caused by different source devices. The audio data acquisition and preprocessing process of multi-ethnic Spring Festival ballads in Gansu province is shown in Figure 1.

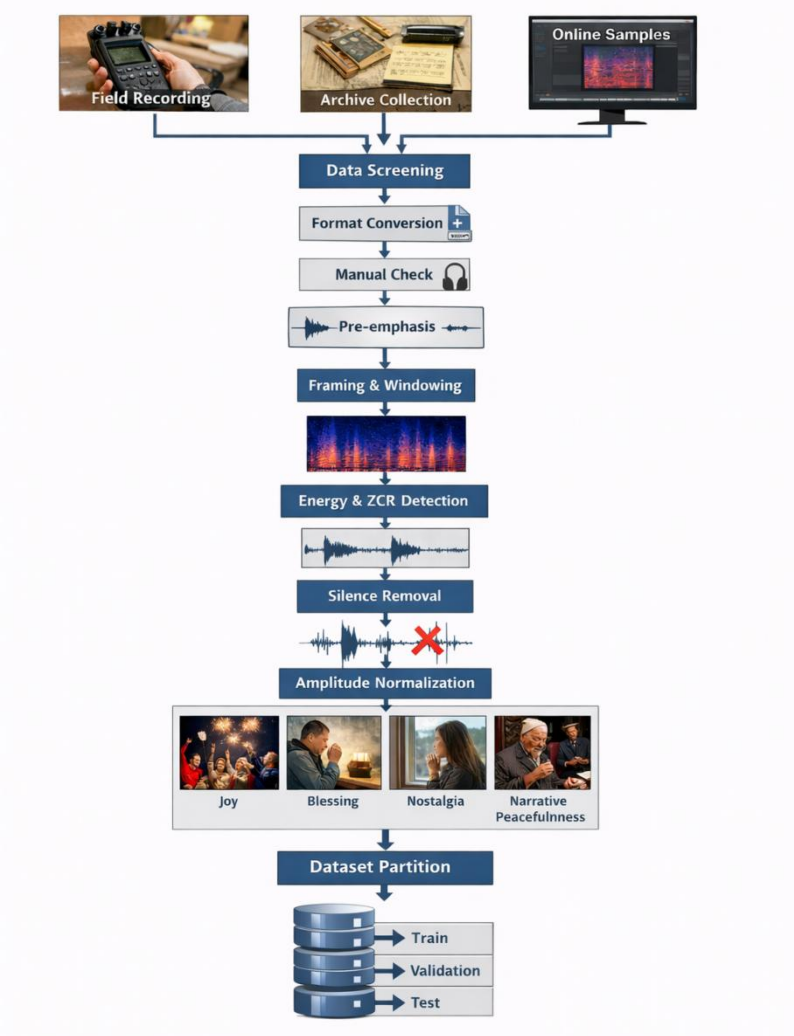

Figure 1: Flow chart of audio data acquisition and preprocessing of multi-ethnic Spring Festival ballads in Gansu Province

In the preprocessing stage, pre-emphasis is used to highlight high-frequency details and attenuate the vocal tract smoothing effect, which is expressed as follows:

$$y(n)=x(n)-\alpha x(n-1) \tag{1}$$

Here, x(*n*) is the original speech sampling point, y(*n*) is the signal after pre-emphasis, and α is taken to be 0.97. This processing helps to retain the finer fricative sound, air change sound and high frequency harmonic information in the Spring Festival songs, and provides a more stable input basis for subsequent MFCC feature extraction.

Considering the large differences in the recording environment of multi-ethnic ballads, some samples have background human voice, ceremonial site noise and space reverberation, the short-term energy judgment is introduced before segmentation. The short-time energy of frame m is defined as follows:

$$E_m=\sum_{n=0}^{N-1}[x_m(n)]^2 \tag{2}$$

where N is the length of a single frame and $x_m(n)$ is the NTH sampling point in the MTH frame. Frames with low short-term energy mostly correspond to silent segments or invalid background segments, which can be eliminated in the initial screening stage.

In order to avoid mistakenly deleting weak songs only by the energy threshold, this study further combines the zero-crossing rate to complete the endpoint detection. The zero-crossing rate of frame m is written as follows:

$$Z_m=\frac{1}{2}\sum_{n=1}^{N-1}|\mathrm{sgn}[x_m(n)]-\mathrm{sgn}[x_m(n-1)]| \tag{3}$$

where sgn(·) is the sign function. When short-term energy and zero-crossing rate participate in boundary judgment, it can distinguish drawling, softly singing and real silence more carefully, and improve the effectiveness of corpus segmentation.

After the mute segment cleaning, amplitude normalization is also required to compress the amplitude fluctuations caused by different recording devices and singing distances. The normalized signal is denoted as follows:

$$x^*(n)=\frac{x(n)-\mu}{\sigma+\varepsilon} \tag{4}$$

Here, μ and σ represent the sample mean and standard deviation, respectively, and ε is a small constant to prevent the denominator from being zero. After this step, the samples of all ethnic groups are more consistent in the dynamic range, which is convenient for subsequent model training.

In the arrangement of emotional labels, combined with the semantic content, singing style and overall listening sense of Spring Festival ballads, the samples are divided into four categories: "jubilation and jubilation, wish and blessing, thinking and expressing feelings, and narrative peace". In order to reduce the subjective deviation caused by single labeling, the method of "initial labeling - review - negotiation and revision" was used to complete the label confirmation. The corpus composition statistics of multi-ethnic Spring Festival ballads are shown in Table 1.

Table 1: The corpus of multi-ethnic Spring Festival ballads constitutes a statistical table

Ethnic Group	Number of Samples / songs	Total Duration / min	Main Distribution Areas	Dominant Emotional Types	Sampling Rate / Hz
Han	132	286.4	Lanzhou, Tianshui, Dingxi	Festive Joy, Blessing and Prayer	16000
Hui	96	201.7	Linxia, Zhangye	Blessing and Prayer, Narrative Calmness	16000
Tibetan	88	214.3	Gannan	Homesickness and Emotional Expression, Blessing and Prayer	16000
Dongxiang	74	156.8	Dongxiang, Linxia	Narrative Calmness, Homesickness and Emotional Expression	16000
Yugur	58	129.5	Sunan, Zhangye	Festive Joy, Narrative Calmness	16000
Bonan	41	83.6	Jishishan	Blessing and Prayer, Homesickness and Emotional Expression	16000
Tu	37	76.9	Areas around the Hexi Corridor	Festive Joy, Narrative Calmness	16000
Total	526	1149.2	—	All four emotional categories are covered	16000

It can be seen from Table 1 that the sorted corpus has a certain degree of hierarchy in ethnic origin and emotional distribution. The total number of samples reaches 526 and the cumulative duration is 1149.2 minutes, which can provide a solid data foundation for subsequent MFCC feature extraction and LSTM emotional classification training. After unified sampling, silence elimination, normalization and label correction, the computability of the original ballad audio is significantly enhanced, and the data quality is more suitable for the model construction stage.

3.2 MFCC parameter extraction and feature representation Method for emotion recognition of Spring Festival Songs

After the completion of corpus screening, silent segment elimination and amplitude normalization, the Spring Festival ballads audio enters the acoustic feature extraction stage. The multi-ethnic Spring Festival ballads of Gansu Province have distinct differences in singing styles. Some samples have bright rhythms and large ups and downs, while some samples have long dragging passages and more melodic twists, which are often difficult to stably present emotion-related information. In order to enhance the computability and distinguibility of features, we choose MFCC as the core acoustic representation, and transform continuous audio into a temporal feature matrix suitable for subsequent neural network training through short-time analysis. Figure 2 shows the MFCC parameter extraction process of Spring Festival ballads.

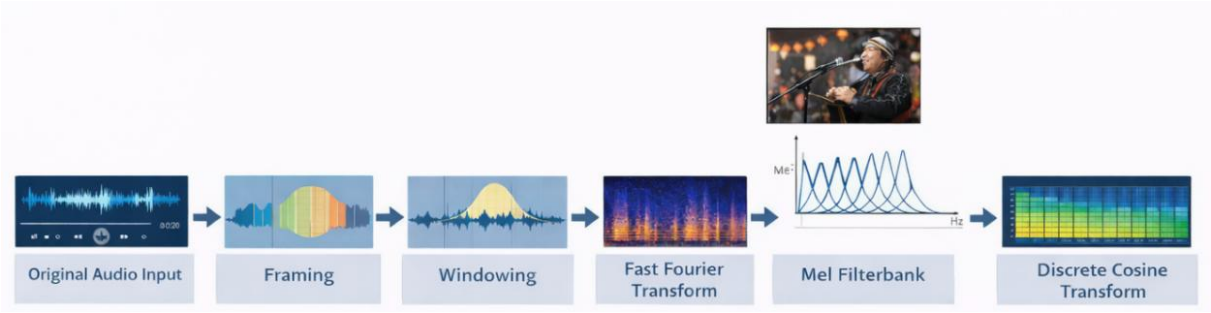

Figure 2: Flowchart of MFCC parameter extraction for Spring Festival ballads

The first step of MFCC extraction is framing. Since ballad audio can be approximately regarded as stationary signal in a relatively short time, the preprocessed audio is segmented into short-time frames according to a fixed window in this paper. Let the original discrete signal be x(n), then the MTH signal can be expressed as follows:

$$x_m(n)=x(n+mH), \quad 0\le n\le N-1 \tag{5}$$

where N is the frame length and H is the frame shift. Combined with the time domain fluctuation characteristics of the Spring Festival ballad audio, the single frame length is set to 25 ms and the frame shift is set to 10 ms in this paper, which can not only retain the local details in the singing, but also take into account the continuity between adjacent frames.

In order to reduce the spectral leakage caused by frame truncation, a window function should be applied to each frame after framing. Hamming window is used in this paper, which is expressed as follows:

$$w(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0\le n\le N-1 \tag{6}$$

After windowing, the MTH signal is written as follows:

$$s_m(n)=x_m(n)\cdot w(n) \tag{7}$$

This method can make the frame boundary transition smoother and help to maintain the stability of the timbre structure and local energy distribution of the ballad. In the frequency domain analysis stage, this paper performs fast Fourier transform on each frame signal to obtain the corresponding spectrum representation:

$$S_m(k)=\sum_{n=0}^{N-1} s_m(n)e^{-j2\pi kn/N}, \quad k=0,1,\ldots,N-1 \tag{8}$$

Considering the uneven perception of the human ear to frequency changes, the Mel scale will be introduced to complete the nonlinear mapping in the subsequent feature extraction, and the relationship is as follows:

$$M(f)=2595\log_{10}\left(1+\frac{f}{700}\right) \tag{9}$$

In this paper, 512 FFT points are set at 16 kHz sampling rate, and 26 Mel filters are used to aggregate the frequency domain energy in order to capture the frequency band distribution

related to emotional changes in Spring Festival ballads in a more detailed way.

The output energy of the RTH filter, weighted by the Mel filter bank, can be written as follows:

$$E_r=\sum_{k=0}^{N-1}|S_m(k)|^2H_r(k) \tag{10}$$

Here, $H_r(k)$ is the response value of the RTH filter at frequency point k. After taking the logarithm of the energy and applying the discrete cosine transform, the QTH MFCC coefficient can be obtained as follows:

$$C_q=\sum_{r=1}^{R}\log(E_r)\cos\left[\frac{\pi q}{R}\left(r-\frac{1}{2}\right)\right],\quad q=1,2,\ldots,Q \tag{11}$$

Considering the dynamic change requirements of emotion recognition tasks, we retain 13-dimensional static MFCC coefficients, and further extract the corresponding first-order difference and second-order difference features to form a 39-dimensional single frame feature vector. Then, the Z-score normalization method is used to uniformly process the whole feature sequence to reduce the scale shift caused by different recording conditions and singing intensity.

Through the above steps, the original Spring Festival ballad audio is converted into a temporal feature matrix with clear structure and stable dimension. Such a representation can more effectively represent the timbral differences, rhythm changes and emotional ups and down in the singing of the ballad, and provide reliable input for the design of the LSTM neural network training sentiment classification model in the next section.

3.3 Design of LSTM neural network training sentiment classification model for Multi-ethnic Spring Festival ballads

After the MFCC feature extraction of Spring Festival ballads, the input data has been transformed from the original continuous waveform into a feature matrix with time order. The emotional expression of multi-ethnic Spring Festival ballads in Gansu Province often has the characteristics of continuous expansion. Although the acoustic state in the local frame is important, what really determines the category is more reflected in the coherent changes between multiple time segments. Based on this feature, we choose LSTM neural network to construct the sentiment classification model of Spring Festival songs, and model the context relationship in the long sequence through the gated memory mechanism to improve the recognition ability of the overall emotional trend. Figure 3 shows the structure of the LSTM neural network training sentiment classification model.

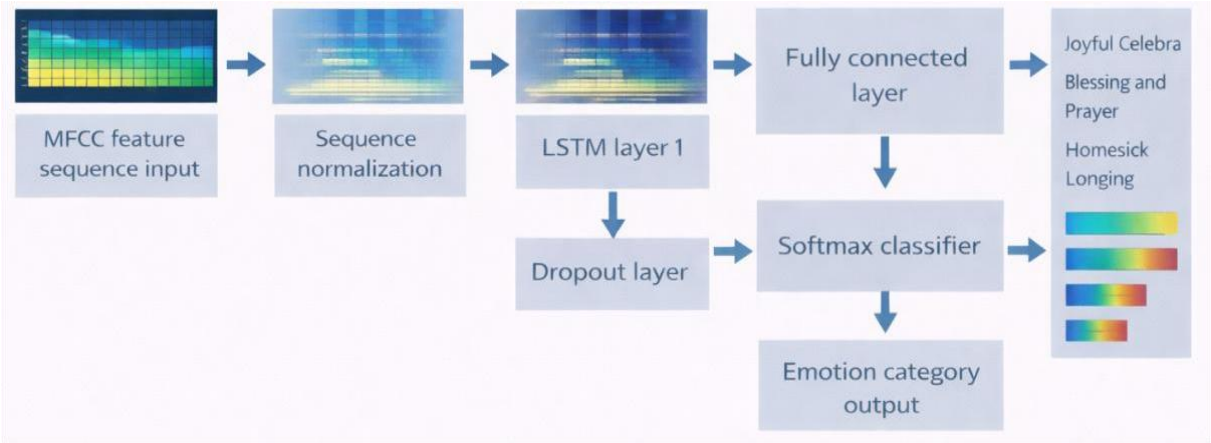

Figure 3: Structure diagram of the LSTM neural network training sentiment classification model

The input of the model is the MFCC time series feature matrix corresponding to each Spring Festival ballad. According to the Settings in the previous section, each frame is composed of 13-dimensional static MFCCS, first-order differences and second-order differences together, and the total feature dimension is 39. After inputting consecutive frames into the network in chronological order, LSTM first adjusts the retention ratio of historical states through the forget gate, and the calculation formula is written as follows:

$$f_t=\sigma(W_f x_t+U_f h_{t-1}+b_f) \tag{12}$$

where x_t represents the input vector at the current time, h_{t-1} represents the hidden state at the last time, W_f, U_f and b_f represent the corresponding weight matrix and bias term respectively. The function of the forgetting gate is to weaken the historical information that has low contribution to the current sentiment discrimination, so that the model can maintain a good memory selection ability in long sequence processing.

The input gate is used to control the degree to which new information is written to the memory cell at the current time step and is expressed as follows:

$$i_t=\sigma(W_i x_t+U_i h_{t-1}+b_i) \tag{13}$$

The corresponding candidate memory states are as follows:

$$\tilde{c}_t=\tanh(W_c x_t+U_c h_{t-1}+b_c) \tag{14}$$

Through this set of operations, the model is able to encode the emotion change trend contained in the current frame into new candidate information. Connected notes, slow singing and emotional accumulation often occur in the singing of Spring Festival ballads, and these information gradually appear in the frames before and after. Therefore, the dynamic writing of candidate states by LSTM is more suitable for this kind of audio feature modeling.

Under the joint action of the forget gate and the input gate, the state of the memory unit is updated as follows:

$$c_t=f_t\odot c_{t-1}+i_t\odot\tilde{c}_t \tag{15}$$

Here, $\odot$ denotes element-wise multiplication. On the one hand, the update process inherits the effective information related to the emotion category in the previous moment, and

on the other hand absorbs the new changes in the current feature sequence, so that the model can establish a stable emotional trajectory representation in a longer range of time. For the samples with compact rhythm and high energy, such as "jubilation and jubilation", the model is easier to form reinforcement memory from the high dynamic fluctuations of consecutive frames. For the samples of "thinking and expressing feelings" or "narrative peace", the slow fluctuation and continuous low-frequency distribution can also be effectively captured in the time accumulation.

The output gate determines the projection strength of the current cell state to the hidden state, which is expressed as follows:

$$o_t=\sigma(W_o x_t+U_o h_{t-1}+b_o) \tag{16}$$

The final hidden state can be expressed as follows:

$$h_t=o_t \odot \tanh(c_t) \tag{17}$$

After obtaining the high-level temporal representation of the whole audio, this paper sends the output of the second layer of LSTM into the fully connected layer, and completes the probability estimation of the four types of emotion through the Softmax classifier. The predicted probability of sentiment class k is written as follows:

$$p_k=\frac{\exp(z_k)}{\sum_{j=1}^{K}\exp(z_j)} \tag{18}$$

Among them, z_k represents the score of the output vector of the fully connected layer on the KTH class, K=4, corresponding to the four emotional labels of jubilation, wish and blessing, thinking and expressing, and narrative peace respectively. In the model training phase, the cross-entropy loss function is used to measure the difference between the predicted distribution and the true label:

$$L=-\sum_{k=1}^{K} y_k \log p_k \tag{19}$$

where y_k is the true label in one-hot encoded form. This loss form is suitable for multi-class classification tasks, and can steadily push the model parameters to update in the right direction.

In the network structure setting, this paper uses a two-layer stacked LSTM. The number of hidden units in the first layer was set to 128, which was used to extract richer shallow temporal patterns. The number of hidden units in the second layer is set to 64, which is used to compress and aggregate the sentiment representation in higher layers. A Dropout mechanism is added between the two layers, and the deactivation ratio is set to 0.30 to alleviate the excessive dependence on local sample features during training. The dimension of the fully connected layer is set to 64, which can not only maintain the discriminative ability of the sentiment representation, but also avoid the training instability caused by too large parameter scale. The Adam algorithm was used for model optimization, with the initial learning rate set to 0.001, batch size set to 32, and training rounds set to 100. The dataset is divided into training set, validation set, and test set with a 7:1:2 ratio to ensure a clear division of labor between model training and tuning for the final evaluation.

In conclusion, the LSTM sentiment classification model can better adapt to the temporal

structure characteristics of Spring Festival ballads audio. The input dimension is consistent with the previous MFCC feature extraction results, and the network depth and parameter scale are controlled in a relatively moderate range, which not only meets the requirements of emotion continuous modeling, but also takes into account the training stability and computational efficiency. Through this design, the emotion recognition process of Spring Festival songs changed from traditional subjective hearing to a computable, trainable and verifiable classification framework, which laid a model foundation for the subsequent analysis of experimental results.

4 Experimental results and analysis

4.1 Experimental environment Configuration and sentiment classification dataset Construction

In order to verify the applicability of the constructed model in the emotion recognition task of multi-ethnic Spring Festival ballads in Gansu province, we complete audio feature extraction, network training and result testing in a unified experimental environment. The experimental platform is implemented by Python language, the deep learning framework is TensorFlow, the audio processing relies on Librosa to complete reading, segmentation and MFCC feature extraction, and NumPy and Pandas are used for matrix operation and data sorting. The hardware environment is configured with Intel Core i7 processor, 16 GB memory and NVIDIA RTX 3060 graphics card, and the operating system is Windows 11. The experimental environment can meet the training requirements of medium-scale audio samples under multiple iterations, and ensure that the loss convergence process and classification output process have good stability.

The experimental data set comes from the multi-ethnic Spring Festival ballad corpus of Gansu Province, which is organized in the previous section. A total of 526 valid audio samples are included, with a cumulative duration of 1149.2 minutes. According to the ethnic origin statistics, 132 songs were written by Han nationality, 96 by Hui nationality, 88 by Tibetan nationality, 74 by Dongxiang nationality, 58 by Yugur nationality, 41 by Bao 'an nationality and 37 by Tu nationality. All audio was uniformly converted into a.wav file with 16 kHz sampling rate and 16 bit quantization accuracy. After denoising, mute segment elimination, amplitude normalization and feature standardization, it was converted into a 39-dimensional MFCC time series feature sequence suitable for model input. Combined with the semantics of lyrics, singing tone and overall listening sense, the dataset is finally labeled as four emotions: jubilation, blessing, thinking and expressing, and narrative peace. Among them, there were 148 celebrations, 136 wishes and blessings, 121 thoughts and feelings, and 121 narratives and peace. The four types of samples were relatively similar in overall scale, which could effectively reduce the interference of class imbalance on training results.

The results of dataset partitioning are shown in Table 2. In this paper, all samples are divided into training set, validation set and test set according to the ratio of 7:1:2. The training set is used for model parameter learning, the validation set is used to monitor overfitting and adjust hyperparameters, and the test set is used for final classification performance evaluation. After hierarchical partitioning, there were 369 songs in the training set, 53 songs in the validation set, and 104 songs in the test set. Specifically, the jubilation class was divided into 104 songs in the training set, 15 songs in the validation set, and 29 songs in the test set. There were 95, 14 and 27 blessings, respectively. There were 85, 12 and 24 songs in the category of thoughts, feelings and feelings respectively. There are 85, 12 and 24 songs in the narrative peaceful category, respectively. This division method not only retains the distribution

characteristics of the four types of emotional samples, but also enhances the comparability of subsequent model training and testing results.

Table 2: Statistical table of data set partition and emotion category distribution

Emotion Category	Total Samples / songs	Training Set / songs	Validation Set / songs	Test Set / songs
Festive Joy	148	104	15	29
Blessing and Prayer	136	95	14	27
Homesickness and Emotional Expression	121	85	12	24
Narrative Calmness	121	85	12	24
Total	526	369	53	104

In the setting of training parameters, the model adopted a two-layer LSTM structure, the number of hidden units was set to 128 and 64, the batch size was 32, the training rounds were 100, Adam was selected as the optimizer, the initial learning rate was set to 0.001, and the cross-entropy was used as the loss function. The evaluation indicators are selected as accuracy, precision, recall and F1 value, and the recognition performance of different emotion categories is refined and analyzed in the following sections combined with the confusion matrix. Through the above experimental environment configuration and specific data set construction method, the model training has a clear data boundary and evaluation basis, and also provides reliable support for subsequent performance analysis.

4.2 Analysis of MFCC feature extraction effect and LSTM model training process

After completing the data set division and model parameter setting, we further analyze the constructed Spring Festival ballad sentiment classification model from two levels of feature expression quality and training convergence state. The goal of MFCC feature extraction is to convert the timbral changes, short-term energy distribution and emotional ups and downs in the original audio into a temporal representation for the network to learn. LSTM training process analysis is used to observe whether the model can form stable emotion discrimination ability in successive iterations. Figure 4 shows the training loss change curve.

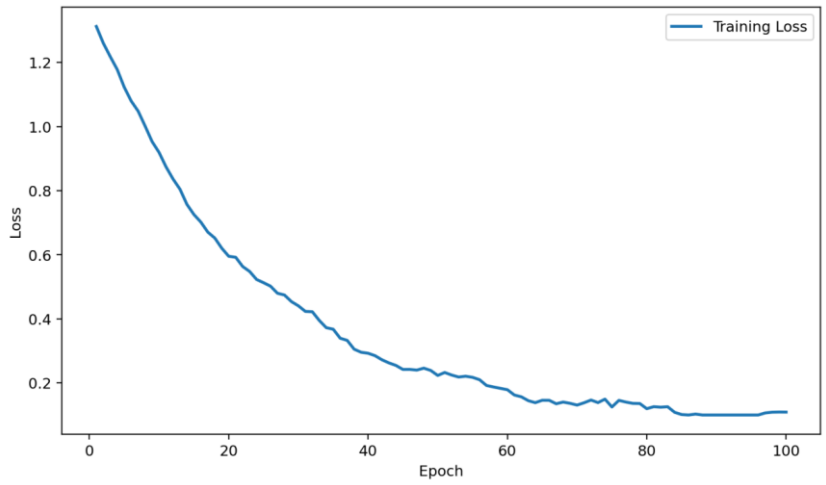

Figure 4: Plot of training loss variation

It can be seen from Figure 4 that the loss of the model decreases rapidly in the early iteration stage, which indicates that the MFCC feature sequence has a good basic representation ability for the emotional differences of Spring Festival ballads. With the increase of training rounds, the loss curve gradually turns from a rapid decline to a stable convergence state, indicating that the network has been able to effectively learn the boundary relationship between different emotional categories. The loss is about 1.31 in the first round of training, drops to about 0.44 in the 30th round, further drops to 0.18 in the 60th round, and stabilizes around 0.11 in the 100th round. Although there is a slight fluctuation in the second half of the curve, the overall amplitude is small and there is no obvious rebound, indicating that the model training process is relatively stable and the direction of parameter update is consistent with the task goal. In summary, after combining MFCC features with LSTM network, a distinguishable emotional mapping relationship can be quickly established, and a good convergence quality can be maintained in the middle and late stages.

Figure 5 shows the variation trend of accuracy between training set and validation set.

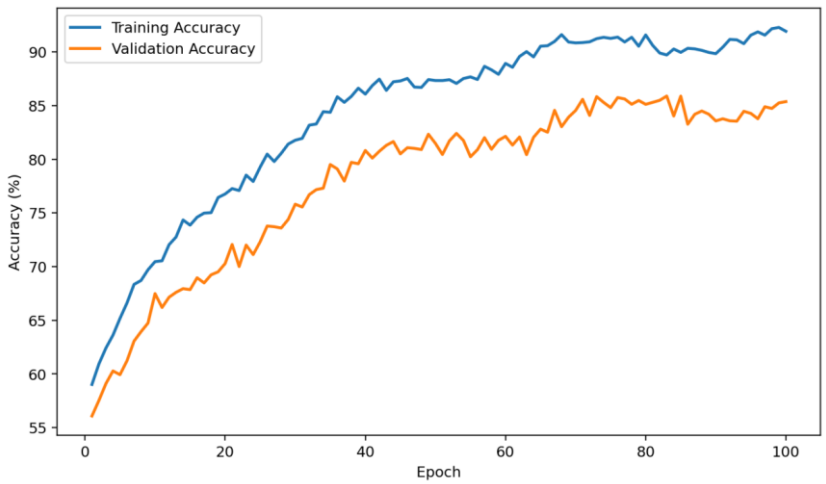

Figure 5: Plot of accuracy change between training set and validation set

As can be seen from Figure 5, both curves show a continuous upward trend, the training accuracy is gradually improved from 59.0% in the initial stage to 91.9%, and the validation accuracy is improved from 56.1% to 85.4%. In the first 30 rounds, the synchronous rise of the two curves is obvious, indicating that the model can quickly extract effective emotional clues in the Spring Festival ballad audio in the feature learning stage. In the middle and late stages, the accuracy of the training set still improved slightly, and the accuracy of the validation set increased slowly, but did not show a sharp decline, indicating that Dropout and parameter control measures inhibited overfitting to some extent. By the end of training, the accuracy difference between the training set and the validation set is about 6.5 percentage points, which is within a relatively acceptable range, reflecting the good generalization ability of the model.

In order to further observe the expression effect of MFCC features on the emotional information of Spring Festival ballads, the feature matrix of typical samples is visualized in this paper, and the heat map of MFCC features is shown in Figure 6.

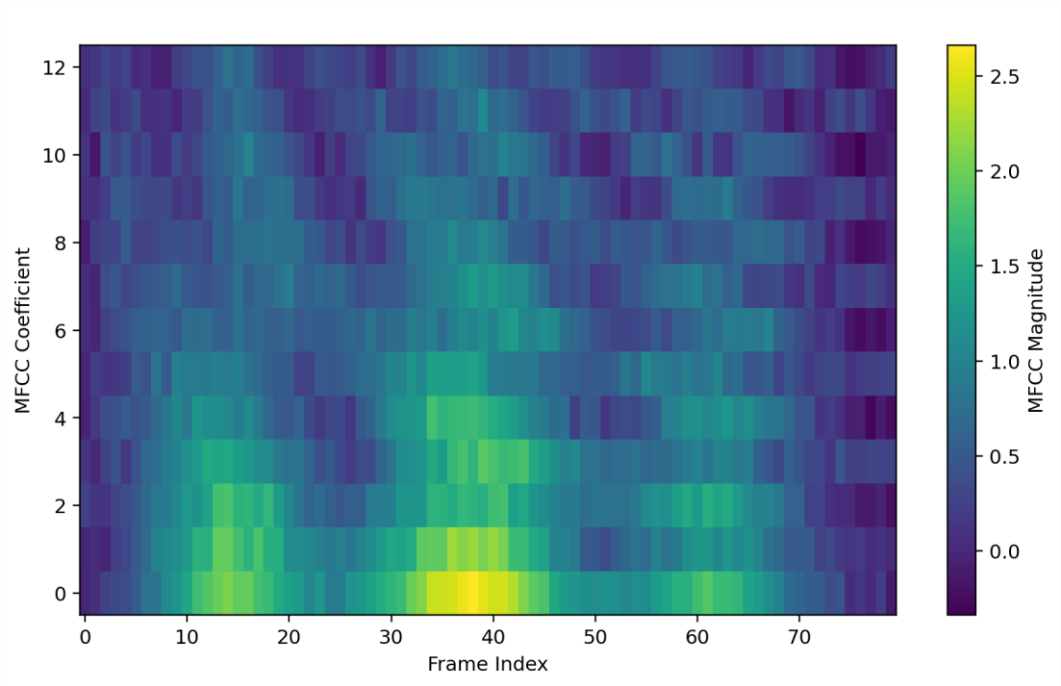

Figure 6: Heat map of MFCC features

It can be seen from Figure 6 that the low-order cepstral coefficient region shows a more continuous and obvious distribution of high response, indicating that it is more sensitive to the dominant timbre and emotional ups and downs in ballad singing. In the heatmap, the strong response was mainly concentrated near the first five coefficients, and the peak area appeared around the 39th frame, and the response intensity reached 1.32. The overall amplitude of the high-order coefficients is significantly weakened, and the mean value of the 11th to 13th order coefficients is about 0.35, indicating that this region reflects more local detail disturbance, and its contribution to the overall emotion discrimination is relatively weak. At the same time, the heat map shows the structural characteristics of continuous enhancement and decay alternately in multiple time slices, indicating that the emotional expression of Spring Festival songs has obvious time advancement characteristics, and also confirms the necessity of using LSTM for sequence modeling.

In general, the MFCC parameter extraction method adopted in this paper can clearly present the key emotional features in the Spring Festival songs audio, and the LSTM network achieves stable iterative convergence and high classification learning efficiency on this basis. At the later stage of training, the loss value has been compressed to near 0.11, the training accuracy has reached 91.9%, and the validation accuracy has reached 85.4%. At the same time, the low-order features in the heat map maintain a strong response, which indicates that the method has good applicability in both feature expression and model learning, and provides a reliable basis for the comparison and analysis of classification results in the next section.

4.3 Sentiment classification results of multi-ethnic Spring Festival ballads and comparative analysis of model performance

After completing the model training and convergence state analysis, we further evaluate the emotional recognition effect of multi-ethnic Spring Festival songs from three levels: the distribution of classification results, the recognition differences of different emotional categories, and the overall performance comparison of the model. This part focuses on the specific discrimination of four types of emotional samples in the test set, and combines the

experimental results of other common classification methods to analyze the advantages and limitations of the proposed model in the sentiment classification task of Spring Festival songs. The confusion matrix of sentiment classification of multi-ethnic Spring Festival ballads is shown in Figure 7.

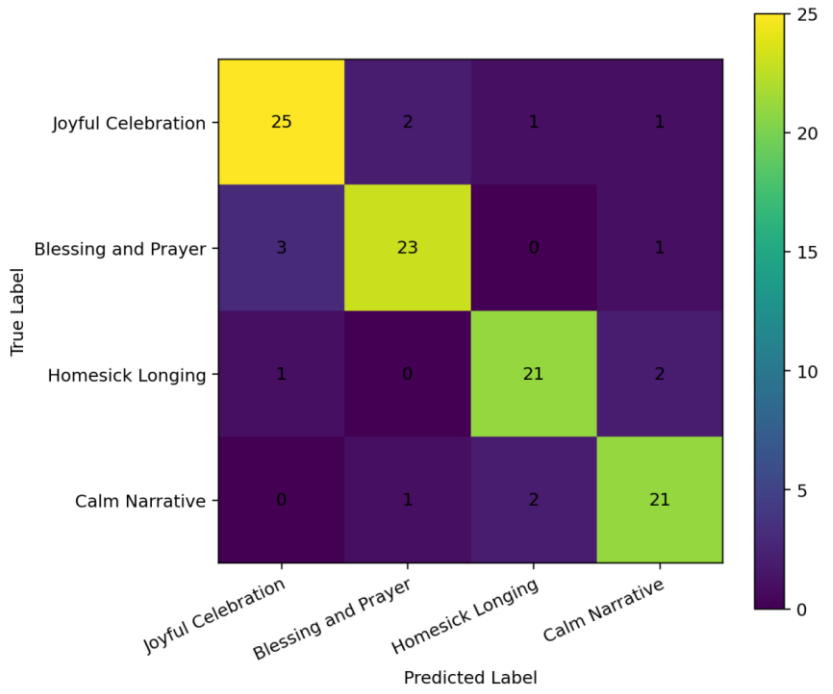

Figure 7: Confusion matrix plot

As can be seen from Figure 7, the recognition results of the proposed model on the four types of emotions are relatively balanced as a whole, and the values in the diagonal region are significantly higher than those in the off-diagonal region, indicating that most test samples can be correctly classified into the target category. In the test set of 104 samples, the model correctly identified 90 samples, and the overall classification accuracy reached 86.5%. Specifically, a total of 29 jubilant samples were identified, and 25 were correctly identified, with a recognition rate of 86.2%. Twenty-seven samples of blessing and blessing were correctly identified, and the recognition rate was 85.2%. Of the 24 samples, 21 were correctly identified, and the recognition rate was 87.5%. Of the 24 narrative and peaceful samples, 21 were correctly identified, and the recognition rate was also 87.5%. This result shows that LSTM neural network has strong adaptability in modeling the temporal emotional changes of Spring Festival songs, and can more stably capture the differences between different emotional categories in singing rhythm, timbral structure and melody advancement.

According to the misjudgment distribution, the errors of the model are mainly concentrated in the adjacent emotional categories. There is an obvious cross between the jubilation class and the blessing class, and there is also a certain confusion between the thinking and lyricism class and the narrative peace class, which indicates that the positive emotion samples with strong festival and the gentle emotion samples with strong lyricism still have confusing characteristics in local pitch transition, energy fluctuation and melody advance mode. In addition to the above major misclassifications, there are also a small number of cross-class misclassifications in the test set, indicating that some samples have compound and transitional emotional expressions. In general, misjudgments mainly appear between classes with similar acoustic features, indicating that the model has formed a

relatively stable classification boundary, but there is still room for further improvement in the discrimination ability of edge samples.

The comparison of recognition rates for different emotion categories is shown in Figure 8.

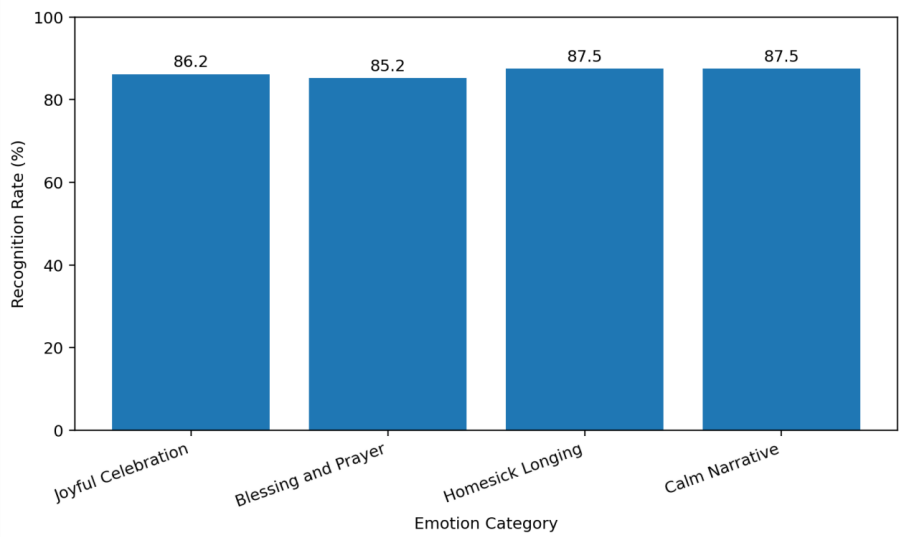

Figure 8: Bar chart of comparison of recognition rates for different emotion categories

As can be seen from Figure 8, the proposed model maintains a high recognition level for the four types of emotions in Spring Festival songs, among which the recognition rates of thoughts and feelings and narrative peace classes are both 87.5%, the recognition rates of joy and exclamation classes are 86.2%, and the recognition rates of blessing classes are 85.2%. The difference between the highest value and the lowest value is only 2.3 percentage points, indicating that the model has a relatively balanced recognition ability for different emotional categories, and there is no obvious bias due to a slightly higher number of samples in a certain category or stronger emotional characteristics. From the perspective of category performance, the overall recognition rate of the thinking and expressing emotion class and the narrative peace class is slightly higher, indicating that LSTM has a better modeling effect on the emotional expression with strong continuity and relatively smooth melody change. The low value of the blessing category reflects that the rhythm and acoustic characteristics of this category of ballads overlap with those of jubilation and jubilation, which increases the difficulty of classification.

In order to further verify the effectiveness of the proposed method, SVM, CNN, RNN and GRU are selected as comparison models, and the performance tests are carried out under the same data set, the same feature input and the same evaluation conditions. The performance comparison of different models is shown in Table 3.

Table 3: Performance comparison table of different models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM	78.8	79.6	78.8	78.9
CNN	81.6	82.1	81.6	81.5
RNN	83.4	83.8	83.4	83.2
GRU	84.7	85.0	84.7	84.6
Proposed Model	86.5	86.5	86.6	86.6

It can be seen from Table 3 that the proposed model achieves the best results in the four indicators of Accuracy, Precision, Recall and F1-score, reaching 86.5%, 86.5%, 86.6% and 86.6% respectively. Among them, the four indexes of SVM model are 78.8%, 79.6%, 78.8% and 78.9% respectively, which shows that the ability of traditional shallow classifiers to capture long-distance emotional association is still insufficient when dealing with complex time-series audio such as Spring Festival ballads. The Accuracy of the CNN model is improved to 81.6%, and the Precision is 82.1%, indicating that the convolutional structure can identify local feature patterns, but it still lacks sufficient expression for continuous emotional evolution across time segments. The Accuracy and F1-score of the RNN model reach 83.4% and 83.2%, respectively, and the GRU model is further improved to 84.7% and 84.6%, indicating that the gated recurrent network has better adaptability in this task. In contrast, the proposed model is 7.7 percentage points higher than SVM, 4.9 percentage points higher than CNN, 3.1 percentage points higher than RNN, and 1.8 percentage points higher than GRU, reflecting the good synergy between MFCC features and LSTM structure in Chinese New Year ballad emotion recognition.

In general, the LSTM sentiment classification model of multi-ethnic Spring Festival songs constructed in this paper performs well in terms of overall recognition effect, class balance and multi-index evaluation. In the test set, 90 out of 104 samples were correctly identified, and the overall accuracy reached 86.5%. The recognition rates of four categories were all above 85.0%, with the highest recognition rate of 87.5% and the lowest recognition rate of 85.2%, and the difference between categories was only 2.3 percentage points. Compared with other comparison models, the proposed method achieves the best results in Accuracy, Recall and F1-score, indicating that it can not only fully extract the emotional acoustic features in Spring Festival songs, but also stably learn the dynamic law of emotion evolution over time. Therefore, the model has good application potential in the sentiment classification task of multi-ethnic Spring Festival ballads, and also provides feasible technical support for the subsequent research on digital protection and intelligent retrieval of ethnic ballads.

5 Conclusion

(1) Focusing on the emotion recognition task of multi-ethnic Spring Festival ballads in Gansu Province, this paper constructs a complete technical process covering corpus collection and processing, audio preprocessing, MFCC parameter extraction and LSTM neural network training. A total of 526 valid audio samples were sorted out, with a cumulative duration of 1149.2 minutes, involving multi-ethnic ballad resources such as Han, Hui, Tibetan, Dongxiang, Yugur, Bao 'an and Tu, and the samples were uniformly converted into 16 kHz, 16 bit audio format. Combined with the semantics of lyrics, singing tone and overall listening sense, we divide the data into four types of emotions: joy and joy, blessing and blessing, thinking and expressing feelings, and narrative peace. On this basis, the data foundation suitable for the emotional analysis of Spring Festival ballads is formed, which provides a stable sample support for subsequent model training and experimental verification.

(2) At the method implementation level, we use MFCC to extract the short-time spectral envelope features of Spring Festival songs, and retain the static coefficients, the first-order difference and the second-order difference information. Finally, a 39-dimensional temporal feature vector is formed and input into the LSTM network. The two-layer stacked LSTM was set with 128 and 64 hidden units respectively, and the Dropout mechanism was added in the middle. Adam optimizer and cross-entropy loss function were combined to complete the model training. The experimental results show that the training loss continues to decrease from about 1.31 in the initial stage to 0.11 in the 100th round. The training accuracy is

improved to 91.9%, the validation accuracy is 85.4%, and the difference between the training set and the validation set is about 6.5 percentage points. The low-order coefficients in the MFCC heat map maintain a strong response, and the feature distribution near the first five coefficients is more concentrated, indicating that the feature combination can more clearly depict the timeliness changes and emotional ups and downs in the Spring Festival songs, and LSTM can also better learn the continuous law of emotional expansion over time.

(3) From the comparison of classification results and model performance, the proposed method shows good accuracy and balance in the emotion recognition task of multi-ethnic Spring Festival songs. In the test set of 104 samples, the model correctly identified 90 samples, and the overall accuracy reached 86.5%. The recognition rate of the four categories of emotion is 86.2% for jubilation, 85.2% for blessing, 87.5% for expressing thoughts, and 87.5% for narrative peace. The maximum difference between the categories is only 2.3 percentage points. Compared with SVM, CNN, RNN and GRU, the proposed model achieves the best results in Accuracy, Precision, Recall and F1-score. The Accuracy is increased by 7.7 percentage points compared with SVM, and 4.9 percentage points compared with CNN. It is 3.1 percentage points higher than RNN and 1.8 percentage points higher than GRU. The above results show that the collaborative modeling method of MFCC and LSTM can effectively adapt to the audio features of Spring Festival ballads with strong melody, free rhythm and rich emotional levels. Subsequent studies can continue to expand the coverage of ethnic samples and integrate spectrogram, attention mechanism or multimodal information to further improve the recognition accuracy and model generalization ability in complex emotional scenes.

Funding

This work was supported by The Research on the Artistic Forms of Spring Festival Culture of Various Nationalities in Gansu Province, a funded project for the introduction of doctoral research start-up funds by Lanzhou University of Arts and Sciences in 2024, with the project number 89090054.

References

[1] Wani T M, Gunawan T S, Qadri S A A, Kartiwi M, Ambikairajah E. A Comprehensive Review of Speech Emotion Recognition Systems[J]. IEEE Access, 2021, 9: 47795-47814. DOI:10.1109/ACCESS.2021.3068045.

[2] Wen S, Sun Y, Liu G. Speech Emotion Recognition Using Transformer Networks[J]. IEEE Access, 2021, 9: 129293-129304. DOI:10.1109/ACCESS.2021.3100932.

[3] Ancilin J, Milton A. Improved speech emotion recognition with Mel frequency magnitude coefficient[J]. Applied Acoustics, 2021, 179: 108046. DOI:10.1016/ j.apacoust.2021.108046.

[4] Shahin I, Hindawi N, Nassif A B, Alhudhaif A, Polat K. Novel dual-channel long short-term memory compressed capsule networks for emotion recognition[J]. Expert Systems with Applications, 2022, 188: 116080. DOI:10.1016/j.eswa.2021.116080.

[5] Jothimani S, Premalatha K. MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network[J].

Chaos, Solitons & Fractals, 2022, 162: 112512. DOI:10.1016/j.chaos.2022.112512.

[6] Bhangale K B, Kothandaraman M. Speech emotion recognition using the novel PEmoNet (Parallel Emotion Network)[J]. Applied Acoustics, 2023, 212: 109613. DOI:10.1016/j.apacoust.2023.109613.

[7] Pham N T, Nguyen S D, Nguyen V S T, Pham B N H, Dang D N M. Speech emotion recognition using overlapping sliding window and Shapley additive explainable deep neural network[J]. Journal of Information and Telecommunication, 2023, 7(3): 317-335. DOI:10.1080/24751839.2023.2187278.

[8] Mao K, Wang Y, Ren L, Zhang J, Qiu J, Dai G. Multi-branch feature learning based speech emotion recognition using SCAR-NET[J]. Connection Science, 2023, 35(1): 2189217. DOI:10.1080/09540091.2023.2189217.

[9] Guo Y, Zhou Y, Xiong X, Jiang X, Tian H, Zhang Q. A multi-feature fusion speech emotion recognition method based on frequency band division and improved residual network[J]. IEEE Access, 2023, 11: 86013-86024. DOI:10.1109/ACCESS.2023.3299822.

[10] Dal Rì F A, Ciardi F C, Conci N. Speech emotion recognition and deep learning: An extensive validation using convolutional neural networks[J]. IEEE Access, 2023, 11: 116638-116649. DOI:10.1109/ACCESS.2023.3326071.

[11] Li Y, Wang Y, Yang X, Im S K. Speech emotion recognition based on Graph-LSTM neural network[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2023, 2023(1): 40. DOI:10.1186/s13636-023-00303-9.

[12] Akinpelu S, Viriri S, Adegun A. Lightweight deep learning framework for speech emotion recognition[J]. IEEE Access, 2023, 11: 77086-77098. DOI:10.1109/ACCESS.2023.3297269.

[13] Li G, Hou J, Liu Y, Wei J. MPAF-CNN: Multiperspective aware and fine-grained fusion strategy for speech emotion recognition[J]. Applied Acoustics, 2023, 214: 109658. DOI:10.1016/j.apacoust.2023.109658.

[14] Dar G H M, Delhibabu R. Speech databases, speech features, and classifiers in speech emotion recognition: A review[J]. IEEE Access, 2024, 12: 151122-151152. DOI:10.1109/ACCESS.2024.3476960.

[15] Akinpelu S, Viriri S. Deep learning framework for speech emotion classification: A survey of the state-of-the-art[J]. IEEE Access, 2024, 12: 152152-152182. DOI:10.1109/ACCESS.2024.3474553.

[16] Jiang X, Zhang Y, Lin G, Yu L. Music emotion recognition based on deep learning: A review[J]. IEEE Access, 2024, 12: 157716-157745. DOI:10.1109/ACCESS.2024.3484470.

[17] Min D J, Kim D H. Speech emotion recognition via sparse learning-based fusion model[J]. IEEE Access, 2024, 12: 177219-177235. DOI:10.1109/ACCESS.2024.3506565.

[18] Zhang X, Xiao H. Enhancing speech emotion recognition with the Improved Weighted Average Support Vector method[J]. Biomedical Signal Processing and Control, 2024, 93: 106140. DOI:10.1016/j.bspc.2024.106140.

[19] Flower T M L, Jaya T. A novel concatenated 1D-CNN model for speech emotion recognition[J]. Biomedical Signal Processing and Control, 2024, 93: 106201. DOI:10.1016/j.bspc.2024.106201.

[20] Kang H, Xu Y, Jin G, Wang J, Miao B. FCAN: Speech emotion recognition network based on focused contrastive learning[J]. Biomedical Signal Processing and Control, 2024, 96: 106545. DOI:10.1016/j.bspc.2024.106545.

[21] Mishra S P, Warule P, Deb S. Speech emotion recognition using a combination of variational mode decomposition and Hilbert transform[J]. Applied Acoustics, 2024, 222: 110046. DOI:10.1016/j.apacoust.2024.110046.

[22] Yu L, Xu F, Qu Y, Zhou K. Speech emotion recognition based on multi-dimensional feature extraction and multi-scale feature fusion[J]. Applied Acoustics, 2024, 216: 109752. DOI:10.1016/j.apacoust.2023.109752.

[23] Zhao H, Huang N, Chen H. Knowledge enhancement for speech emotion recognition via multi-level acoustic feature[J]. Connection Science, 2024, 36(1): 2312103. DOI:10.1080/09540091.2024.2312103.

[24] Li T. Music emotion recognition using deep convolutional neural networks[J]. Journal of Computational Methods in Sciences and Engineering, 2024, 24(4-5): 3063-3078. DOI:10.3233/JCM-247551.

[25] Li J, Soradi-Zeid S, Yousefpour A, Pan D. Improved differential evolution algorithm based convolutional neural network for emotional analysis of music data[J]. Applied Soft Computing, 2024, 153: 111262. DOI:10.1016/j.asoc.2024.111262.

[26] Zhao H, Jin L. IoT-based approach to multimodal music emotion recognition[J]. Alexandria Engineering Journal, 2025, 113: 19-31. DOI:10.1016/j.aej.2024.10.059.

[27] Hao X, Li H, Wen Y. Real-time music emotion recognition based on multimodal fusion[J]. Alexandria Engineering Journal, 2025, 116: 586-600. DOI:10.1016/j.aej.2024. 12.060.