Music Emotion Recognition and Analysis Based on Neural Network

Zhao Hanbing[1], Jin Xin[2]*, Guo Jinfeng[3]
College of Music, Beihua University, Jilin City,Jilin Province, People's Republic of China,132113[1]
Changchun University for the Aged, Changchun City, Jilin Province 130021, China[2]
Beijing Cuiwei Primary School Miyun Branch, China, Beijing, Miyun, 101500[3]

***Abstract*—The close connection between music and human emotions has always been an important topic of research in psychology and musicology. Scientists have proven that music can affect a person's emotional state, thereby possessing the potential for therapy and stress relief. With the development of information technology, automatic music emotion recognition has become an important research direction. The MultiSpec-DNN model proposed in this article is a multi-spectral deep neural network that integrates multiple features and modalities of music, including but not limited to melody, rhythm, harmony, and lyrical content, thus achieving efficient and accurate recognition of music emotions. The core of the MultiSpec-DNN model lies in its ability to process and analyze various types of data inputs. By combining audio signal processing and natural language processing technologies, the MultiSpec-DNN model can extract and analyze the comprehensive emotional characteristics in music files, thereby achieving more accurate emotion classification. In the experimental section, the MultiSpec-DNN model was tested on two standard emotional speech databases: EmoDB and IEMOCAP. The experimental results show that the MultiSpec-DNN model has a significant improvement in accuracy compared to traditional single-modal recognition methods, which proves the effectiveness of integrated features in emotion recognition.**

Keywords*—Music emotion recognition; multimodal fusion; audio signal processing; neural network; sentiment analysis; user experience

I. INTRODUCTION

Music is a powerful form of art that is closely linked to human emotions, capable of eliciting a range of emotional states from joy to sadness, and even neutral feelings. Scientific research since the 1950s has confirmed the ability of music to regulate emotions. When listening to music, people instinctively associate it with emotional labels, and this emotional effect is due to music containing key elements such as melody, rhythm, and timbre, which stimulate human emotions. Psychologists have extensively explored the impact of music on emotions and confirmed the connection between music and five basic emotions. Research reveals that different listeners have consistent emotional responses to the same piece of music, and most people have remarkably similar choices for the emotional type of music, thus the analysis of music emotions can be used to infer the psychological state of the listener. Accordingly, people also tend to seek out musical works that resonate with their own emotions when experiencing different emotional states.

In the Web 2.0 era, online listening to digital music has become extremely convenient, and most popular music works contain not only audio but also textual information such as lyrics. Studies show that lyrics can effectively influence emotional changes, sometimes even more effectively than audio. Therefore, sentiment analysis technology has shown its importance in many fields such as social networks, e-commerce feedback, and film reviews. Researchers in the field of music use various music features, including audio and text, to perform emotion classification and carry out automated music emotion recognition. A major challenge hindering music recognition at present is the lack of easily accessible basic ground truth data. To perform emotion recognition, it usually requires a large number of participants to listen to music and record their feelings, but this method is costly and inefficient. With the continuous advancement of sentiment analysis technology, we can now more accurately identify user emotions and provide more personalized services based on this. Having the ability to grasp user emotions is not only crucial for personal services but also has practical value on a broader scale.

The development of multimodal fusion technology has also brought new opportunities for music information retrieval [1], [2]. Studies have shown that combining audio and lyrical information can improve the accuracy of emotion classification. For example, methods such as combining Language Model Difference (LMD) and Bag of Words (BOW) model, and the transformation of psychological categories have enhanced classification efficiency [3][4]. The development of deep learning has further promoted research on neural network-based information fusion and emotion classification.

Project Number: JJKH20250841SK, this paper proposes a multimodal information fusion method for music emotion recognition, providing a new direction for research in automated music emotion recognition, aiming to cope with the ever-growing digital music library and new songs, to minimize manual annotation work, and to lay the foundation for practical application scenarios. The key to the method proposed in this study is that by combining the analysis results of audio features and lyrical content, a more comprehensive understanding of the emotional expression of music can be achieved. Multimodal fusion helps to improve the accuracy and robustness of emotion classification. Ultimately, this method provides the

*Corresponding Author.

possibility of developing efficient music emotion analysis tools that can be embedded into various applications, thereby enhancing user experience, such as providing more personalized music recommendations by identifying the types of music emotions favored by users, or selecting appropriate music based on the emotional state of patients in psychotherapy. With the continuous development of music digitalization and intelligent technology, the potential of automated music emotion recognition will be explored and applied more broadly.

II. Literature Review

Music emotion recognition techniques utilize computers to extract and analyze musical features, forming mappings between music features and the emotional space, thereby achieving recognition of the process of emotional expression in music. Specifically, music emotion recognition techniques typically use audio signals as input, and then employ various algorithms and techniques to extract and analyze musical features, such as frequency, time domain, spectrum, and more. These features can be represented in the form of vectors or matrices, and compared with each point in the emotional space to determine their similarity. By calculating these similarities, an emotional score can be obtained to describe the emotions conveyed by the music. Below is related work on music emotion recognition.

A. Techniques Based on Acoustic Features

Techniques based on acoustic features analyze music using the acoustic characteristics of emotional speech. By simulating continuous audio signals that become discretized through sampling for computer processing, these sampling points are extracted for rhythm, spectrum, timbre, duration, speech rate, fundamental frequency, intensity, Mel-frequency cepstral coefficients (MFCC), Linear Predictive Coding (LPC), Chromagram, and other physical features related to music, using these features to represent the emotions in music.

Due to the complexity of emotional features, it is difficult to accurately describe a person's emotional state. Currently, there is no unified understanding in the academic world about the representation of emotions, nor is there a qualitative and quantitative measurement and evaluation standard. Therefore, how to extract effective feature parameters and use appropriate models to express the correlation between these feature parameters and emotions is a key issue that needs to be addressed [5]. Sordo et al. extracted multiple acoustic features from music, such as frequency domain features, time domain features, and higher-level genres and styles, mapping them to semantic features, and using the K-Nearest Neighbors algorithm (KNN) to complete the music emotion classification problem [6]. Yang et al. compared models for emotional classification of English and Chinese songs to explore the cultural characteristics of different countries [7]. Markov et al. researched the effects of different features (MFCC, LPC, timbre features, Chroma, etc.) and their combinations on emotion recognition using Gaussian Processes (GP) and Support Vector Machines (SVM). To solve the "semantic" gap between low-level audio signal features and high-level musical concepts, [8] Weninger et al. proposed an emotion recognition method based on Recurrent Neural Networks (RNN), first extracting low-level features from frame spectra, then calculating general features such as kurtosis, percentiles, and regression coefficients on their contours for multivariate regression to compute levels of pleasure and arousal. [9] Chin et al. built emotion recognition models for different genres, based on sparse representation of music to calculate genre indicators. Renato [10] Panda et al. advanced the latest music emotion recognition techniques by proposing novel emotion-related audio features, such as musical texture features, expressiveness features, etc. The ability of neural networks to extract excellent feature parameters is increasingly drawing attention, with more research directly feeding unstructured data into Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and other deep learning models. The input data passes through layers of networks to abstract the extracted low-level features for the final classifier layer to predict classification results. Research on emotional features is not just for improving the effectiveness of music emotion recognition; there is already application of music's acoustic fingerprint features in semantic-based cross-media music retrieval, modeling the potential semantic associations between text and music to explore their correlation.

B. Techniques Based on Temporal Variations

Emotions are behaviors that change over time; their evolution goes through a certain duration, thus the dependency of emotional information before and after is to be considered. Traditional dynamic models, such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs), have shown better recognition performance than static models due to their inherent properties for modeling temporal contextual information. However, these models consider only a short span of temporal information, which limits their effectiveness. Yang et al. [11] extracted emotional features based on a continuous psychological model of emotions in three dimensions: valence/pleasantness, arousal/intensity, and dominance/control. They used linear regression models to map the emotional state of music to a continuous emotional space and employed two fuzzy classifiers to measure emotional intensity for recognizing emotions in music. Schmidt et al. [12] established a connection between human emotional space and the acoustic signals of music, developing regression models to study emotional changes as they occur over time in music. Since different individuals may annotate the same piece of music with different emotions, Wang et al. [13] proposed that musical emotions should be represented as a probability distribution. They introduced the Audio Emotion Gaussian (AEG) model for the annotation of VA (Valence-Arousal) musical emotions, learning a VA Gaussian distribution for the latent feature class of each sound, and representing musical emotions through a weighted mixture of these VA Gaussian distributions. However, the assumption of a probability distribution for VA values does not necessarily hold in practice, so Wang et al. [14] proposed an HDM model to predict continuous features of music, dividing the VA space into a G×G grid of a two-dimensional histogram to predict musical emotions. To identify dynamic musical emotions, Li et al. [15] proposed a music dynamic emotion prediction method based on Deep Bidirectional Long Short-Term Memory (DBLSTM) networks, training multiple DBLSTMs on time series of various scales, and integrating multiscale DBLSTM results using an Extreme Learning

Machine (ELM) method to determine the emotions in music. Currently, emotion recognition models based on deep learning have stronger non-linear modeling capabilities and have been widely applied in the field of emotion recognition. For instance, the Long Short-Term Memory (LSTM) model by Wang et al. [16] and the classic CNN-based models by Luz et al. [17] have achieved good results in the modeling process. However, these models assume the same contribution to emotion prediction for each frame, which is an unreasonable assumption; to address this issue, Chen et al. [18] introduced an attention mechanism that automatically learns the importance of different frames for emotion recognition through global contextual information to obtain matching weight coefficients, enabling more targeted emotion modeling.

C. Research Gaps

Over the past few decades, researchers have been exploring how to quantify and classify emotional states in music. Early studies mainly relied on the perception of sound timbre and manual annotations based on patterns to achieve emotion classification. However, as the emotional state in music differs from emotions in other contexts and media, this recognition remains a challenge. Specific issues include:

When discrete emotional space models are used, the recognition of musical emotions is treated as a classification task, which is more straightforward and simple compared to continuous emotional space. The goal is to tag unfamiliar music with emotional labels through classification models. Currently, there is a wide variety of extracted musical emotion features, but individual features have poor generalization capabilities and cannot adapt flexibly to different datasets. Secondly, deep learning networks are simple in construction, adept at extracting deep information, but musical emotions are more subjective, and overall feature analysis is also important. Therefore, how to better select musical emotion features and build deep learning networks, and how to extract both breadth and depth features are urgent problems to be solved.

Moreover, it is understood from the current research status that despite the abundance of various feature types, traditional manual acoustic features remain the richest set of features in terms of emotional content. Appropriate feature optimization and selection schemes are essential for achieving good emotional recognition performance when dealing with high-dimensional manual acoustic features.

On the other hand, spectrograms, as an important carrier of information in speech signals, represent an important avenue for improving music emotion recognition performance by analyzing and mining emotional features from them using image processing methods with the development of deep learning technology.

In summary, future research needs to develop new models and techniques to address these challenges in music emotion recognition, to truly deliver the most suitable music to listeners.

III. The Discrete Emotional Space of Music

This section first extracts GTF and MFCC as features for musical emotion, with MFCCs being weighted with the residual phase (RP) for compensation. Building upon the Word2Vec method, the Chord2Vec approach is proposed to extract chord information and train it into chord vectors as one of the input features, providing a clear representation of the musical content. These features are then fused together as input for the MultiSpec-DNN model to determine the contextual relationship of the music. The results from MultiSpec-DNN are fed into the enhanced nodes of the BLS (Broad Learning System), where they undergo mapping processing to form the output of the enhanced nodes.

A. Principle of Chord2Vec

The classification of musical emotions is different from other classification tasks. In the case of speech emotion classification, not only can commonly used signal features such as audio energy be chosen as emotional features, but textual information can also be processed through textual expression for feature calculation, making the feature selection multimodal. However, for most music without lyrics, due to the absence of universal textual or visual features, most people have to rely on listening to recognize and appreciate the music. Therefore, only auditory-related features can be selected, which results in suboptimal music emotion classification. Inspired by the principle of Word2Vec, this chapter proposes the Chord2Vec method, which converts chord information in music into musical chord vectors through the Skip-gram model, thus providing multimodal emotional features for the task of music emotion classification.

B. Extraction of Note Information

The expression of musical emotions can be achieved through the combination of different chords, rhythms, dynamics, and tempos. A chord refers to the vertical combination of three or more musical notes of different pitches. By setting reasonable rules, chords can form the "textual information" of music more than elements such as rhythm, dynamics, and tempo. Therefore, the order of notes within a chord and the intervals between each note are crucial for chord information. MIDI, as an audio format, can record information about notes, dynamics, positions, and durations. By using the read function in the musicpy library, it is possible to extract all note information for each piece of music. Due to the large amount of information, Table I only shows the note information for pure music of four different emotions from 1 minute 10 seconds to 1 minute 13 seconds (with note intervals preserved to two decimal places).

TABLE I. Music Note Information

Music Name	Emotion	Note Combination	Note Intervals
Kiss The Rain	Joyful	D4,G4,E4,D4,C4,D 4,E4, F4,E4,D4,C4,D4	0.13,0.12,0.12,0.13,0.13, 0.24,0.14, 0.13,0.13,0.12,0.52,0.05
Canon	Sad	D5, D5, F5, F#5, E5, D5, B4, G4, G4, A4, B4	0.22,0.02,0.09,0.13,0.04, 0.33,0.03, 0.14,0.63,0.09,0.08
Victory	Excited	A4,E4,E4,G4, A4, B4, A4,A4, F4,E4	0.13,0.26,0.06,0.13,0.63, 0.13,0.13, 0.13,0.13,0.06
Dust	Tense	D4, A4, G4, G4, D4, F4 , G4, G4, G4,A4	0.12,0.79,0.12,0.11,0.03, 0.22,0.11,0.25, 0.12, 0.25

In the note combination, the suffix number after the same pitch level indicates the pitch height, increasing by one for every octave higher. The sharp sign "#" as a suffix denotes raising the basic pitch level by a semitone. Note intervals are expressed as the play interval between two consecutive notes, with the measure as the unit. A value of 0 indicates that the two notes are played together; a value of 1 means there is a one measure interval between the play of two notes; a value of 0.25 means there is a 1/4 measure interval between the play of two notes, and so on.

C. Chord Segmentation

Beats are the most basic elements in the composition of music, and measures, as units of beats, directly affect the overall melody of the music and the emotions the composer wishes to convey. Assume that after playing the primary melody note, the next note requires a time duration of 1 beat before being played; even if the composer intended to treat the current note as part of the primary melody, the audience may have difficulty perceiving a coherent melody. This is because, in essence, a melody is a series of notes with relatively similar pitch, contrasting with chords that have a greater pitch difference and are recognized as melody. The duration of a measure (in seconds) is related to the beats per minute (BPM), as shown in Eq. (1), where B represents BPM, and X represents the number of beats per measure.

$$Y = \frac{60}{B} \cdot X \tag{1}$$

Table II presents the results of chord segmentation for a 3-second note combination in Pachelbel's Canon, segmented according to different musical beats.

TABLE II. CHORD SEGMENTATION RESULTS (CANON)

Musical Beat	Chord Segmentation Results
4/4 Beat	D5 D5 F5 F#4 E5 D5/B4 G4 G4/A4 B4
3/4 Beat	D5 D5 F5 F#5 E5 D5 B4 G4 G4/A4 B4
2/4 Beat	D5 D5 F5 F#5 E5 D5 B4 G4 G4/A4 B4
6/8 Beat	D5/ D5 F5 F#5 E5 D5 /B4 G4 G4/A4 B4

Music can be composed of different musical beats, and for the sake of data uniformity in experiments, a 4/4 musical beat is adopted, which means each measure has 4 beats, and the duration of a measure is 240/B seconds. If the interval between successive notes is greater than or equal to 1 beat, or 0.25 measures, then the previous note is judged to be the last note of the preceding chord combination, and the following note is the first note of the subsequent chord combination.

D. Chord Vector

Let us assume that the chord information matrix G after chord segmentation consists of N pieces of music $\{\alpha_1, \alpha_2, \dots, \alpha_N\}$, with each piece having t chord combinations $\{\beta_1, \beta_2, \dots, \beta_t\}$. Thus, the music chord information set can be represented as in Eq. (2):

$$\mathbf{G} = \begin{matrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{matrix} \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{t1} \\ \beta_{12} & \beta_{22} & \cdots & \beta_{t2} \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{1N} & \beta_{2N} & \cdots & \beta_{tN} \end{bmatrix} \tag{2}$$

Here, β_{ij} represents the i-th chord combination of the j-th song, and N is the number of pieces of music.

Suppose after the Skip-gram model the number of chord features is V, and each piece of music contains M chord combinations, then the information matrix S of that piece of music can be represented as in Eq. (3):

$$\mathbf{S} = \begin{matrix} 1 \\ 2 \\ \vdots \\ \mathrm{V} \end{matrix} \begin{bmatrix} \mathrm{F}_{11} & \mathrm{F}_{21} & \cdots & \mathrm{F}_{M1} \\ \mathrm{F}_{12} & \mathrm{F}_{22} & \cdots & \mathrm{F}_{M2} \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{F}_{1V} & \mathrm{F}_{2V} & \cdots & \mathrm{F}_{MV} \end{bmatrix} \tag{3}$$

By weighting the V features of the M chord combinations, we obtain $[\mathrm{Z}_1, \mathrm{Z}_2, \cdots, \mathrm{Z}_\mathrm{V}]$. Here, $Z_1 = F_{11} + F_{21} + \cdots + F_{M1}; Z_2 = F_{11} + F_{12} + \cdots + F_{M2}; Z_V = F_{1V} + F_{2V} + \cdots + \mathrm{F}_{\mathrm{MV}}$. Applying this operation to all the music $\{\alpha_1, \alpha_2, \dots, \alpha_\mathrm{N}\}$ in the chord information matrix G, the final chord vector matrix C can be represented as in Eq. (4), where Z_{ij} corresponds to the i-th weight for the j-th piece of music, and N represents the number of pieces of music.

$$\mathbf{C} = \begin{matrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{matrix} \begin{bmatrix} \mathrm{Z}_{11} & \mathrm{Z}_{21} & \cdots & \mathrm{Z}_{V1} \\ \mathrm{Z}_{12} & \mathrm{Z}_{22} & \cdots & \mathrm{Z}_{V2} \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{Z}_{1N} & \mathrm{Z}_{2N} & \cdots & \mathrm{Z}_{VN} \end{bmatrix} \tag{4}$$

Fig. 1 displays the overall process of extracting chord vectors using Chord2Vec.

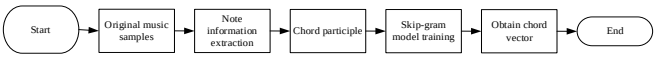

Fig. 1. Chord2Vec process diagram.

IV. MULTIDIMENSIONAL EMOTION FEATURE EXTRACTION BASED ON SPECTROGRAMS

To acquire a more comprehensive set of emotional information, this section introduces a deep fusion model based on neural networks called MultiSpec-DNN. Initially, the model inputs two types of spectrograms: narrowband and wideband spectrograms, corresponding to better frequency resolution and time resolution, respectively. These are extracted from each speech signal by setting frame windows. Given the excellent performance of convolutional neural networks (CNNs) in image processing in recent years, our MultiSpec-DNN model incorporates modules such as CNN, LSTM, and attention mechanisms to fully learn the emotional information within the spectrograms. The MultiSpec-DNN model thoroughly mines the temporal and frequency domain information contained in both types of spectrograms, ultimately obtaining spectrogram features that enhance the performance of speech emotion recognition.

A. MultiSpec-DNN Feature Extraction Model

In this section, we propose a speech emotion feature learning model, MultiSpec-DNN, which takes multidimensional spectrograms as input and integrates modules such as CNN, LSTM, and attention mechanisms. Our model's network structure design draws upon some content from study, and the overall network structure of the MultiSpec-DNN model is shown in Fig. 2.

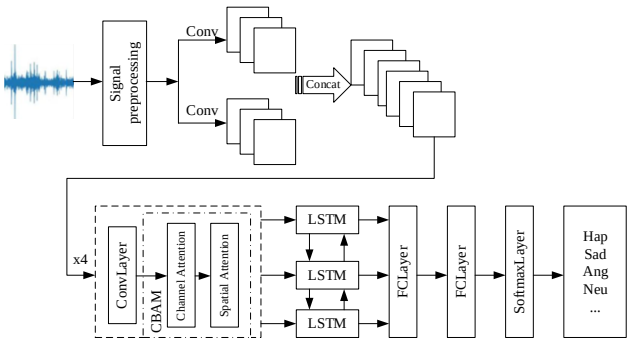

Fig. 2. MultiSpec-DNN network structure.

The MultiSpec-DNN model is based on deep feature learning from two different bandwidth spectrograms. Firstly, the speech signal undergoes preprocessing, which includes pre-emphasis, framing, and windowing, with specific preprocessing steps referenced in the corresponding sections of subsequent experimental chapters. Fourier analysis is performed on the preprocessed speech to obtain two types of spectrograms through different window lengths, namely, the wideband spectrogram (Narrow Band Spectrum) with better time resolution and the narrowband spectrogram (Wide Band Spectrum) with better frequency resolution, which serve as the raw input data for the overall network. The two types of spectrograms are fed into two convolutional layers for convolution operations. The resulting feature maps are concatenated in the channel dimension and then trained in a four-layer convolutional neural network to learn deeper temporal and frequency domain spatial features of the spectrograms. Attention mechanisms are integrated into each convolutional layer to enhance the learning of emotion-related features. To further explore the temporal information within the convolutional feature maps, the output of the last convolutional layer is fed into a bidirectional LSTM network (BLSTM) for learning temporal features. Finally, the output of the BLSTM passes through two fully connected layers before entering the Softmax layer to obtain the emotional classification output.

B. Key Design of the MultiSpec-DNN Model

In this section, we will detail the key step designs within the MultiSpec-DNN model, as well as the specific parameter settings used in subsequent experiments.

1) Wideband and narrowband spectrograms: Spectrograms provide an intuitive representation of how the vocal frequency spectrum changes over time, containing rich speech information. Digging deep into these features to extract them can help improve the performance of speech emotion recognition. The foundation of the MultiSpec-DNN model is based on two types of spectrograms: the wideband spectrogram (Narrow_band Spectrum) and the narrowband spectrogram (Wide_band Spectrum). Although these spectrogram types only differ due to the size of the Fourier transform window set, they present their own characteristic feature expressions. Previous research indicates that combining wideband and narrowband spectrograms can better reflect the entirety of the speech signal. Therefore, the model innovatively proposes the analysis and extraction of speech features based on these two types of spectrograms for emotion classification training, which helps achieve a more comprehensive and holistic expression of emotions within speech. Specifically, the wideband spectrogram, due to its corresponding short frame window settings, is formed by stacking a large number of short frames, thus providing better time resolution. The narrowband spectrogram corresponds to longer frame window settings, with longer frames stacked, reflecting the distribution of different frequencies over a period of time, and therefore, has a higher frequency resolution. Extracting features based on both types of spectrograms is equivalent to analyzing speech features from both the time and frequency domain perspectives.

Wideband and narrowband spectrograms are typically generated by framing and windowing the speech signal with window widths of approximately 3 ms/25 ms, followed by Fourier transform and stacking the frames to produce the spectrogram. When viewed horizontally, the same types of spectrograms correspond to four different emotional speeches; vertically, they are narrowband and wideband spectrograms extracted from the same speech. Vertical comparison of the same speech reveals that the general trend of both spectrograms is consistent, both reflecting the variation of frequency over time, but a detailed observation reveals clear differences between the two:

The narrowband spectrogram is characterized by its narrow horizontal bands, which appear as narrow, bright yellow lines parallel to the horizontal axis, creating a ripple-like pattern, as shown in the black box. These narrow bands represent the fundamental frequency of vowels and harmonics in the sentence, with their vertical position on the frequency axis corresponding to the pitch frequency value, showing the inflections of pitch over time. The dark blank areas from top to bottom correspond to pauses in speech.

The wideband spectrogram shows wider horizontal bands, also parallel to the time direction, as indicated by the black box in the figure. These wider bands represent the position of the vowel formants in the sentence. Different vowels have different formant frequencies, and different people pronounce the same vowel differently, all of which are reflected in the distribution differences of the wide bands on the frequency axis, so the vowel can be distinguished based on the position of the wide bands. The wideband spectrogram also has evident narrow blank stripes parallel to the frequency axis, representing the plosive sounds in the speech. Larger blank areas, similar to the narrowband, indicate pauses in the sentence.

Based on the analysis of the two types of spectrograms, it is evident that they contain different speech information. Emotion classification is based on refining the emotional expression within speech features, which is also associated with the expression of speech information. Therefore, by delving into the features of the two types of spectrograms for emotional speech, richer emotional information can be obtained from both the time domain and frequency domain perspectives, enhancing the performance of the emotion recognition system.

2) CNN Module design: The spectrogram presents the information contained in the speech signal in the form of an image. Using image analysis methods to extract features from spectrograms can effectively obtain emotional characteristics. Therefore, in the MultiSpec-DNN model, the CNN network commonly used for image feature extraction is adopted for feature extraction of the spectrogram. From the structural diagram of the MultiSpec-DNN model, it can be seen that the entire model can be divided into two CNN structures. The first part conducts preliminary feature extraction on two types of spectrograms, and the second part is the four-layer CNN network designed after concatenating the convolutional features of the two types of spectrograms in the channel dimension, which is used for in-depth mining of emotional information.

a) Preliminary feature extraction of spectrograms: The first part of the CNN network uses two convolutional layers to convolve the wide and narrow spectrograms, respectively. This part is based on the network in the study, but due to the difference in input spectrograms, the specific network parameter settings also vary.

First, unlike the two types of spectrograms proposed in this paper, the spectrogram used as input in the study has only one type. Specifically, in the preprocessing, the window width of each frame is set to 40 ms, and referring to previous work, a high Fourier transform frequency point is set at 1600 (corresponding to 10kHz), which distinguishes the wide and narrow spectrograms by extracting an ultra-narrowband spectrogram with a very high time resolution. Based on the subsequent truncation of the input frequency, the actual corresponding Fourier transform frequency points are equivalent to 640 (truncating 0-4 kHz from 10kHz). For the purpose of fully extracting time and frequency domain features, the paper designs two different rectangular convolutional kernels for the spectrogram. One is a horizontal convolutional kernel that is consistent with the time direction, covering a larger frequency range at the same time point; the other is a vertical convolutional kernel parallel to the frequency direction, which can present the changes in the current frequency range over time. Finally, the feature maps obtained by the two different convolutional kernels are concatenated and used as the input for the subsequent convolutional layers. Unlike the study ^([63]), the MultiSpec-DNN model proposed in this paper obtains two types of spectrograms at the input stage, corresponding to wideband spectrograms with high temporal resolution and narrowband spectrograms with high frequency resolution, naturally expressing more detailed time domain and frequency domain information. Therefore, when conducting preliminary feature extraction on the two types of spectrograms, the CNN convolutional layers did not choose rectangular long convolutional kernel sizes but performed convolution operations with the same convolutional kernel settings on the two types of spectrograms. The convolutional kernel size is set to a regular 3×3, to extract preliminary feature maps from both the time and frequency domain perspectives for the two types of spectrograms, and then concatenated in the channel dimension as the input for subsequent convolutional layers.

b) In-depth mining of emotional features in spectrograms: The second part of the CNN network further mines the concatenated feature maps of the two types of spectrograms to obtain deeper spatial information of both spectrograms. For the purpose of comparing emotional recognition performance, MultiSpec-DNN adopts the four-layer design of the CNN network in the latter part of the study, with the convolutional kernel size also set at 3×3 and the number of convolutional kernels set sequentially at 32, 48, 64, and 80. The specific parameters of the network layers are listed in table form in the subsequent content. Different from the convolutional layers in the study, the MultiSpec-DNN model proposed in this paper also explored the role of convolutional attention mechanisms in in-depth mining of emotion-related features.

The attention mechanism is a signal processing mechanism discovered by scientists in the 1990s. Its design is based on strategies used by humans and other organisms when processing external data. Specifically, when a vast amount of information floods into the visual range, the human brain will select this information based on its goals, actively ignoring some irrelevant information and focusing on important information, allowing the brain to process more information and quickly find targets. In the field of artificial intelligence, the attention mechanism usually determines the importance of certain features to the target task or strengthens the extracted features with attention, as shown in Fig. 3 for a simple model incorporating an attention mechanism.

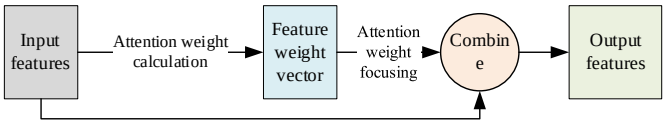

Fig. 3. An example of a simple attention mechanism.

As shown in Fig. 3, the introduction of the attention mechanism into the network starts with calculating the attention weight for each feature value. The weight represents the importance of each feature value relative to the overall feature. Then, the obtained feature weight vector is multiplied by the corresponding position of the original input feature to obtain the output feature enhanced by attention. If the original featureIt seems that your message was cut off before completing your thoughts on CNN module design and attention mechanisms in deep learning. The information you provided indicates an approach to emotion recognition using spectrograms and a CNN architecture tailored to capture both time and frequency domain features.

The above briefly introduced the basic theory of the attention mechanism. For the MultiSpec-DNN model proposed in this paper, after the convolution operation on the spectrogram, a series of feature maps will be obtained. In order to make the network pay more attention to the emotion-related information in the feature maps, the MultiSpec-DNN model introduced a lightweight convolutional attention module, CBAM (Convolutional Block Attention Module), after the convolution operation. The CBAM module enhances the features from the output feature maps of the convolutional

layer in both the channel and spatial dimensions, and its network structure is shown in Fig. 4.

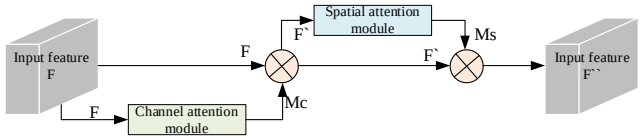

Fig. 4. CBAM network structure diagram.

As shown in the figure, for an output feature map from a certain convolutional layer, the attention mechanism introduced by CBAM is mainly divided into two steps: First, the convolutional feature map F goes through the Channel Attention Module (CAM) to obtain the channel attention weight matrix M_c, which is then element-wise multiplied by the original convolutional feature map to obtain an intermediate feature map F'; Then, F' goes through the Spatial Attention Module (SAM) to obtain the spatial attention weight matrix M_s, which is then element-wise multiplied by F' to obtain the output feature F''.

Fig. 5 shows the internal structure of the channel attention module. The channel attention module aims to spatially compress the convolutional feature map along the channel dimension, that is, to find the spatial weight for each feature map of the corresponding channel. Specifically, assuming the input convolutional feature map has C channels, there are C feature maps. First, each of these feature maps is subjected to Max Pooling (MaxPool) and Average Pooling (AvgPool) operations, focusing on the maximum pixel value and the average state of all pixels, to spatially aggregate and map key and average information of the feature maps, respectively obtaining two pooled feature vectors of size 1×1×C; Then, these two pooled vectors are fed into a shared fully connected layer for channel attention mining, which consists of a double-layer Multilayer Perceptron (MLP) network.

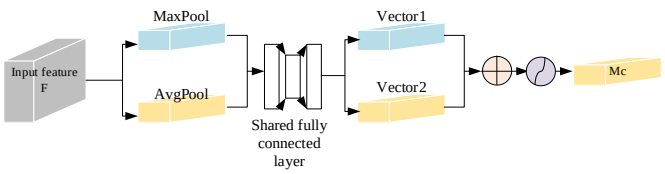

Fig. 5. CBAM's channel attention module.

In this network, two length-C pooled feature vectors are first compressed according to a certain ratio and then restored to C to obtain two intermediate vectors, as shown in Fig. 5's Vector1 and Vector2. These two intermediate vectors are element-wise added and then normalized through the Sigmoid function to obtain the channel weight vector M_c for the original convolutional feature map. This process can be represented by Eq. (5).

$$M_C(F) = \sigma(\mathrm{MLP}(\mathrm{MaxPool}(F)) + \mathrm{MLP}(\mathrm{AvgPool}(F))) = \sigma\left(\boldsymbol{W}_1(\boldsymbol{W}_0(F^C_{\max})) + \boldsymbol{W}_1\left(\boldsymbol{W}_0\left(F^C_{\mathrm{avg}}\right)\right)\right) \quad (5)$$

In the formula, σ represents the Sigmoid function, W_1 and W_0 represent the weight matrices in the shared fully connected layer, and $F^C_{\max}$ and F^C_{avg} represent the pooled feature vectors obtained after Max Pooling and Average Pooling. After multiplying the final channel weight vector $M_C(F)$ with the original convolutional feature map F element-wise, the channel attention feature map F' is obtained, which serves as the input for the spatial channel attention module.

Another attention module in CBAM is the spatial attention module, whose network structure is shown in Fig. 6. After obtaining the channel attention feature map F', the spatial attention module aims to compress the channel dimension of F' along the spatial plane of the feature map to obtain the spatial attention parameter matrix M_s for the overall feature map. Specifically, first, Max Pooling and Average Pooling are used to compress the channel dimension of the channel attention feature map F', and assuming the original feature map size is H×W×C, two pooled feature matrices of size H×W×1 are obtained after the two types of spatial pooling. Then, these two matrices are concatenated along the channel dimension to form a feature tensor with 2 channels as shown in Fig. 6; To mine spatial attention, the 2-channel feature tensor is fed into a convolutional layer for training, with the kernel size set to 7×7 according to the settings in the literature, and after convolution, it is mapped to an intermediate matrix of size H×W×1, which is then normalized through the Sigmoid function to obtain the spatial attention matrix for the original convolutional feature map F, as shown in Eq. (6).

$$M_S(F) = \sigma(f^{7\times7}([\mathrm{MaxPool}(F); \mathrm{AvgPool}(F)])) = \sigma\left(f^{7\times7}\left(\left[F^S_{\max}; F^S_{\mathrm{avg}}\right]\right)\right) \quad (6)$$

In the formula, σ represents the Sigmoid function; 7×7 indicates the size of the convolution kernel in the module. It has been verified in the original CBAM literature that a convolution kernel of size 7×7 yields better performance than one of 3×3; $F^S_{\max}$ and F^S_{avg} respectively represent the pooled feature matrices obtained after Max Pooling and Average Pooling. Finally, by performing an element-wise multiplication of the channel attention feature map F^' with the spatial attention weight matrix $M_S(F)$, the attention-weighted feature map with respect to the original convolutional feature map F is obtained.

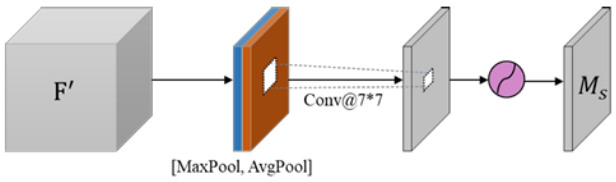

Fig. 6. CBAM's spatial attention module.

3) BLSTM module design: From the introduction of the spectrogram generation and extraction process in the previous text, it can be understood that the spectrogram is actually obtained by performing operations such as Fourier transform on the frame-by-frame speech signal and then stacking them in time order. Therefore, both broadband and narrowband

spectrograms naturally contain temporal information of the speech signal. After the spectrogram goes through the learning of multiple layers of Convolutional Neural Networks (CNNs), a group of spatial feature maps in both the time and frequency domains is obtained. Although these represent higher-dimensional features compared to the original spectrogram, the convolution operation does not change the temporal order of the features, and the output of the convolution layer still retains the temporal sequence of the original spectrogram. On the other hand, from the perspective of emotional expression, the emotional category contained in a sentence is presented through the entire sentence. Learning features both forward and backward in time can obtain richer global emotional information. Based on the above analysis, in order to extract more comprehensive emotional information, the MultiSpec-DNN model inputs the output of the last convolutional layer into the BLSTM in the temporal direction to further enhance the mining of temporal features in the spectrogram. In the experiment, the hidden layer output of the BLSTM is used as the input for the subsequent fully connected layers.

The Bidirectional Long Short-Term Memory network (BLSTM) is built on the foundation of the Bidirectional Recurrent Neural Network (BRNN) and the LSTM], proposed by Graves et al. in 2005. According to the background knowledge, it is understood that RNNs can model sequential data by combining information from the previous moments, and LSTM was designed on this basis to solve the problem of gradient vanishing due to overly long temporal information. However, LSTM can only receive sequence information from before the current moment during training, and the value at a certain moment in the temporal data is often influenced by information from both before and after this moment. Ignoring the sequence information from later moments could lead to prediction errors. Therefore, by training the LSTM with sequences in both forward and backward orders and combining the results from both directions, the BLSTM integrates the information from the entire sequence data, effectively improving the model's performance. The BLSTM network structure is shown in Fig. 7.

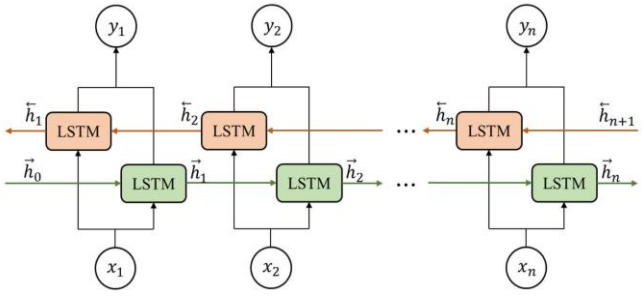

Fig. 7. BLSTM network structure.

The BLSTM consists of forward and backward LSTM networks, corresponding to the lower and upper LSTM networks in Fig. 7, respectively. The lower forward LSTM network processes the sequential data in order during training, saving information from before the current moment. The upper backward LSTM processes the sequential data in reverse order, saving information from after the current moment. This means that the output at any moment in the sequence is related to the entire sequence data.

The input data to the BLSTM network is usually a set of vectors corresponding to sequential data. In the MultiSpec-DNN model proposed in this paper, the feature map output from the convolutional layer is used as the input to the BLSTM to learn temporal features. The treatment of input data here refers to the mapping relationship between the convolutional feature map and the BLSTM network in the research of the Connectionist Text Proposal Network (CTPN). CTPN uses the VGG16 network for convolutional training of text images and uses the spatial feature tensor obtained by densely sliding a 3×3 convolution kernel on the last layer of convolutional output as input to the BLSTM. In this model, the convolutional layer has already integrated the CBAM module to strengthen attention in both channel and spatial dimensions. Therefore, when referring to the CTPN network, only the method of converting the three-dimensional feature tensor to BLSTM input is considered. Specifically, assume that the feature map output size from the convolution layer is H×W×C, and the hidden layer output size of LSTM in each direction within the BLSTM is 128, then the hidden layer output dimension of the BLSTM is 256. Since the convolution operation does not affect the original temporal relationship between the frames of the spectrogram, the W dimension from left to right corresponds to the temporal order of the frames. Therefore, with H as the batch size of data for a single time point and W as the maximum time length, such a data stream is input to the BLSTM, learning the sequence temporal features of each row of data in the W dimension, as shown in Fig. 8.

Fig. 8 shows in an intuitive way how to input the convolutional feature map in temporal order into the BLSTM network. After rotating the feature map, the vertical direction corresponds to the temporal sequence, and the batch data stream along the W dimension is transmitted to the BLSTM, resulting in the final output temporal feature map of H×W×256. Finally, the output of the BLSTM network is unfolded into a one-dimensional vector and input into two fully connected layers, and it seems that you are discussing the design and implementation of a Bidirectional Long Short-Term Memory (BLSTM) module for emotion recognition from speech, using a spectrogram as input. This process includes several steps, such as generating the spectrogram, applying convolutional layers to extract spatial features, and then using a BLSTM to capture temporal dependencies in both forward and backward directions to enhance feature learning.

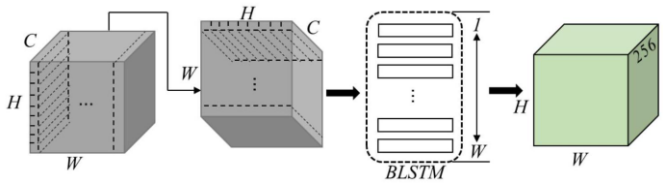

Fig. 8. Convolutional feature map to BLSTM network input method.

V. Case Study

This section will validate the effectiveness of the proposed method using a homemade experimental dataset. The author of

this article confirms that all experiments were conducted in accordance with relevant guidelines and regulations.

A. Experimental Setup

The EMA (Emotion Music Analysis) dataset was collected and produced by referring to the literature, with all categories of emotional music sourced from the internet and uniformly converted to WAV format.

The EMA dataset consists of 4,412 pieces of instrumental music, encompassing four emotional categories: 1,251 pieces of cheerful music, 1,072 pieces of exciting music, 948 pieces of tense music, and 1,141 pieces of joyful music.

The Emotion dataset is composed of 2,978 pieces of MP3 format music, with musical emotions divided into 4 categories: 661 pieces of angry, 739 pieces of happy, 768 pieces of relaxed, and 810 pieces of sad music. The duration of the music ranges from 25 seconds to 55 seconds. For the convenience of the experiment, only the first 25 seconds of each piece of music is used, with zero-padding for those less than 25 seconds.

The 4Q-emotion dataset consists of 1,472 pieces of MP3 format music, where musical emotions are not categorized by emotional words but are classified into four labels: Q1, Q2, Q3, and Q4. There are 442 pieces in Q1, 296 in Q2, 438 in Q3, and 296 in Q4. Only the first 30 seconds of each piece of music are used, with zero-padding for those less than 30 seconds.

For the convenience of processing in the research process, the first 50 seconds of each song were chosen, and zero-padding was performed for those with a duration of less than 50 seconds.

The loss function used in this section's experiment is the Cross Entropy Loss, as shown in Eq. (5), which mainly describes the distance between the actual output (probability) and the expected output (probability); the smaller the value, the closer the two probability distributions are, and the better the model performance, used for multi-label classification tasks. Here, N represents the number of samples i, M represents the number of categories, $y_{ic}$is 1 when the category corresponds to the category of sample i, and 0 otherwise, p_{ic} represents the predicted probability that sample i belongs to category C.

$$L = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -\sum_{c=1}^{M} y_{ic}\log(p_{ic}) \qquad (7)$$

For the simplest binary classification problem, the commonly used evaluation metrics are Accuracy, Precision, Recall, and F-measure. The EMA dataset, Emotion dataset, and 4Q-emotion dataset are randomly allocated into training and test sets in a 9:1 ratio. For the EMA dataset, chord vectors are trained using Chord2Vec and extracted using the Skip_gram model from the Gensim library, with a min_count of 5 and a set chord vector dimension of 500. The vectors of chord combinations that appear in each piece of music are summed up, resulting in a 1×256 dimensional chord vector feature matrix, which serves as the shared chord feature for all three datasets.

The extraction of RP features first uses a 16th-order LP to derive the LP residuals, with overlapping of 10 ms between adjacent frames. Then pre-emphasis is applied to the original information to extract LP residuals and identify the maximum value of the Hilbert envelope in each frame, thereby obtaining the required RP features. Next, RP and MFCC are weighted and fused to determine the final MF_RP features. The final feature size is as shown in Table III:

TABLE III. FEATURE EXTRACTION SIZE

Data	Name Size
Music Pre-processing	3895×44100×60
GTF Features	3895×24×44
MF_RP Features	3895×192×44

Fig. 9 shows the time sequence diagram of the 3-frame MF_RP features extracted from the music of 4 emotional types in the EMA dataset.

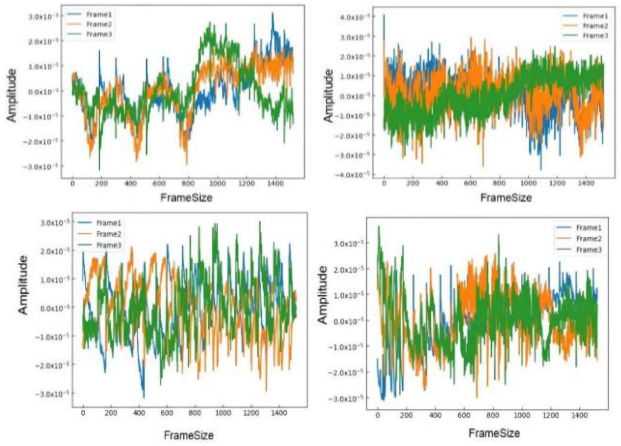

Fig. 9. Music emotion time-series feature graph.

It can be seen that the time-series curves of the MF_RP features for tense and exciting music emotions are significantly different from those of other emotional frames. Cheerful and joyful music have similar features in the mid-high frequency range, but show greater differences in the mid-low frequency range. Therefore, MF_RP features can enhance the extraction capability for emotional information in music signals, effectively capturing the differences even in subjectively similar emotions.

Subsequently, comparative experiments were conducted for the dual-feature filtering channel CNN, as shown in Tables IV, V, and VI. The first two columns represent the GTF feature channel and the MF_RP feature channel, respectively. The four comma-separated numbers in each row represent the number of repetitions, the number of feature mapping layers, the size of the convolutional kernels, and the size of the max-pooling layers, respectively. A stride of 1 is used for all experiments, and zero-padding is performed for each convolutional layer.

TABLE IV. CNN FILTER CHANNEL ARCHITECTURE COMPARISON

GTF Feature Channel	MF_RP Feature Channel	Training Accuracy	Testing Accuracy
3,4,5,4	2,4,6,5	0.75	0.48
2,5,7,4	3,6,8,4	0.88	0.53
2,4,6,4	2,7,9,5	0.91	0.58
3,3,5,5	4,6,8,4	0.87	0.57
2,4,5,4	3,5,7,4	0.96	0.63

TABLE V. DIFFERENT CNN FILTER CHANNEL ARCHITECTURE COMPARISON (EMOTION)

GTF Feature Channel	MF_RP Feature Channel	Training Accuracy	Testing Accuracy
3,5,4,4	2,4,6,5	0.81	0.46
2,4,5,4	3,6,8,4	0.84	0.50
2,5,8,4	2,7,9,5	0.94	0.63
3,3,5,5	3,5,8,4	0.88	0.54
2,4,5,4	3,6,9,5	0.91	0.57

TABLE VI. DIFFERENT CNN FILTER CHANNEL ARCHITECTURE COMPARISON (4Q-EMOTION)

GTF Feature Channel	MF_RP Feature Channel	Training Accuracy	Testing Accuracy
3,4,5,4	2,3,5,4	0.91	0.56
2,3,5,4	2,6,8,4	0.97	0.63
2,4,8,4	2,7,8,4	0.85	0.52
3,4,5,4	3,6,8,4	0.83	0.50
2,3,5,4	3,6,8,4	0.92	0.58

It can be observed that on different datasets, adapting the CNN structure to the type and size of features can improve classification accuracy. The same filter channel structure ignores the feature size and complexity. A too deep structure will extract redundant deep features of GTF, while too shallow structures may result in incomplete deep features of MF_RP. Selecting the appropriate filter channel parameters can better extract both features.

With a batch_size of 128, 3895 pieces of music are processed through Chord2Vec and music preprocessing to extract chord vectors, GTF features, and MF_RP features. The latter two are input into the modified filter channels. The CNN layer, as the filtering channel for MFCC and GTF features, extracts deep information with three feature mapping layers and 2x2 filters for GTF features, and max-pooling layers of 2x2, repeated twice. The MFCC_RP feature's CNN filtering channel contains 6 feature mapping layers and 3x3 filters, with max-pooling layers of 2x2, repeated three times. Both use BN layers to normalize the outputs. Subsequently, a fully connected layer fuses the two features into a 1x256 data matrix, which is then fed into a 1x256 BILSTM layer. The data trained by the BILSTM layer isfurther fused with the chord vectors, and finally, the Broad Learning System (BLS) is used for node enhancement to obtain the final output.

B. Experimental Results and Analysis

This section verifies the improvement in emotional classification accuracy of the MULTISPEC-DNN model based on the Chord2Vec chord vector representation, within the same experimental environment. The MULTISPEC-DNN model is also compared with existing mainstream classification models to evaluate whether the proposed model can improve the accuracy and overall efficiency of music emotion classification. Experiments were conducted on the EMA dataset, Emotion dataset, and 4Q dataset, with each dataset divided into three different data partitions: training set, validation set, and test set. Ten-fold cross-validation was employed to ensure that these three partitions do not overlap, thus maximizing the accuracy of the experiments.

1) Experimental scheme: The effectiveness of the chord vector feature is first verified by comparing the accuracy of the MULTISPEC-DNN model that only uses weighted fusion features of GTF and MF_RP with the MULTISPEC-DNN model that adds chord vector features. Subsequently, performance comparisons of the overall model structure are conducted with five mainstream models selected for comparison, introduced as follows:

a) MCCLSTM and MCCBL: Both models start with CNN filtering channels with convolutional kernels of three different sizes to extract music information such as pitch and interval. The former concatenates the output of each CNN channel and uses it as the input for the LSTM layer, while the latter enhances the nodes using a BLS layer and finally trains to obtain the classification results for emotions.

b) RCNNLSTM and RCNNBL: These two models contain two layers of CNN as the filtering channels for input features, where each CNN layer has a fixed convolution kernel size but a random number of kernels. The former uses the output of the final CNN layer as the input for LSTM, while the latter uses BLS layer for enhancement to obtain the final emotion classification results.

c) LSTM_BLS: This model directly uses the extracted features as input for a multi-layer LSTM and as feature nodes for BLS. The latter connects the final output to the enhancement nodes during processing and combines both to obtain the final classification results.

In this experiment, the MCCLSTM, MCCBL, and RCNNBL models refer to the literature for parameter settings. The LSTM_BLS model sets the number of LSTM layers to 2, with memory cell counts of 1024 and 512, respectively, and uses the MF_RP feature as the input for this model. The input and detailed parameters for the models in this section are a batch size of 64, dropout of 0.2, and the optimizer is Adam.

2) Experimental analysis: Tables VII to IX show the results obtained by each model when recognizing emotions on the EMA dataset, Emotion dataset, and 4Q dataset, respectively. A detailed analysis of these tables reveals that the emotional classification accuracy of the model in this section reached 61.8% on the EMA dataset, which is 7.6% higher than MCCLSTM, 4.4% higher than RCNNBL, and 1.5% higher than LSTM_BLS; in the Emotion dataset, the model's classification accuracy reached 63.8%, which is 4.6% higher than MCCLSTM, 2.4% higher than RCNNBL, and 5% higher than LSTM_BLS.

TABLE VII. MODEL CLASSIFICATION COMPARISON (EMA)

Model	Accuracy	Precision	Recall	F1	Traning time(s)
MCCLSTM	0.557	0.615	0.557	0.585	520.43
MCCBL	0.548	0.590	0.548	0.568	95.72
RCNNLSTM	0.562	0.608	0.562	0.584	1105.12
RCNNBL	0.581	0.599	0.581	0.590	120.49
LSTM_BLS	0.610	0.633	0.610	0.621	335.78
MULTISPEC-DNN	0.624	0.645	0.624	0.634	470.94

TABLE VIII. MODEL CLASSIFICATION COMPARISON (EMOTION)

Model	Accuracy	Precision	Recall	F1	Training time (s)
MCCLSTM	0.591	0.642	0.553	0.596	280.45
MCCBL	0.578	0.601	0.525	0.560	42.18
RCNNLSTM	0.573	0.612	0.508	0.556	630.12
RCNNBL	0.589	0.603	0.618	0.610	58.37
LSTM_BLS	0.641	0.655	0.612	0.633	175.49
MULTISPEC-DNN	0.647	0.670	0.625	0.648	223.74

TABLE IX. MODEL CLASSIFICATION COMPARISON (4Q)

Model	Accuracy	Precision	Recall	**F1**	Training time (s)
MCCLSTM	0.581	0.598	0.569	0.583	198.34
MCCBL	0.589	0.624	0.571	0.596	35.21
RCNNLSTM	0.593	0.635	0.574	0.603	550.42
RCNNBL	0.622	0.633	0.624	0.629	50.81
LSTM_BLS	0.661	0.674	0.671	0.673	132.47
MULTISPEC-DNN	0.635	0.658	0.629	0.643	191.54

It is evident that on different datasets, models based on BLS have a much higher training efficiency than those based on LSTM. This is because the model depth of BLS is much shallower compared to LSTM, significantly reducing the complexity of the model, while the accuracy difference between the MCCBL model and the MCCLSTM model is only around 2%. The random number of CNNs can to some extent compensate for the lack of deep information extraction by BLS, therefore the RCNNBL model outperforms the RCNNLSTM model in both accuracy and training efficiency. The LSTM_BLS model further demonstrates that LSTM can extract the temporal relationships of music, thereby maximizing the preservation of musical emotion features. Although the training efficiency is not high when combining BLS with LSTM, the classification accuracy is greatly improved.

The MULTISPEC-DNN model introduced in this section, which combines dual-channel CNN layer filtering and the novel chord vector features, achieved the best results on both the EMA dataset and the Emotion dataset. Since the BILSTM model itself is more complex than LSTM and CNN, its training efficiency is lower than the MCCBL model, the RCNNBL model, and the LSTM_BLS model. For the 4Q dataset, whether in terms of training efficiency or model classification accuracy, the MULTISPEC-DNN model is not as good as the LSTM_BLS model, indicating that 1286 pieces of music are not sufficient for the MULTISPEC-DNN model to learn enough information, leading to overfitting and ultimately resulting in mediocre classification accuracy.

VI. CONCLUSION

In this paper, we have conducted in-depth discussions and research on the extraction and optimization of musical emotion features within the field of music emotion recognition and analysis. The proposed MultiSpec-DNN model integrates spectral features of different resolutions, using an attention mechanism enhanced CNN and BLSTM networks, to deeply mine the emotional information in the music signals across time, frequency, and temporal dimensions. The emotion recognition rate on the EmoDB dataset is 91.24%, and on the IEMOCAP dataset, it is 71.88%, both demonstrating excellent recognition capabilities. The comparative experiments in this paper further analyze the performance differences between composite features and single features in the task of music emotion recognition, concluding that composite features can significantly improve the accuracy of emotion recognition. In summary, the feature optimization selection algorithm and the MultiSpec-DNN model proposed in this paper have shown significant effectiveness in the field of music emotion recognition. These research findings are of great importance for improving the accuracy and practical application value of music emotion recognition. Future work can be extended on the existing foundation to achieve more accurate and natural music emotion recognition, enhancing people's auditory experience and emotional communication.

ACKNOWLEDGMENT

Supported by the Scientific Research Project of the Department of Education of Jilin Province - Project Name: Application and Promotion of Eight-line Digital Notation in Music Teaching in Higher Education Institutions. Project Number: JJKH20250841SK.

REFERENCES

[1] Liu S, Zheng P, Bao J. Digital Twin-based manufacturing system: a survey based on a novel reference model[J]. Journal of Intelligent Manufacturing, 2023: 1-30.

[2] Liu S, Zheng P, Xia L, et al. A dynamic updating method of digital twin knowledge model based on fused memorizing-forgetting model[J]. Advanced Engineering Informatics, 2023, 57: 102115.

[3] Zheng H, Liu S, Zhang H, et al. Visual-triggered contextual guidance for lithium battery disassembly: a multi-modal event knowledge graph approach[J]. Journal of Engineering Design, 2024: 1-26.

[4] Fu T, Li P, Liu S. An imbalanced small sample slab defect recognition method based on image generation[J]. Journal of Manufacturing Processes, 2024, 118: 376-388.

[5] Sordo M, Celma O, Bogdanov D. MIREX 2011: Audio tag classification using weighted-vote nearest neighbor classification[C]// Music Information Retrieval Evaluation Exchange. 2011.

[6] Yang Y H, Hu X. Cross-cultural Music Mood Classification: A Comparison on English and Chinese Songs[C]//ISMIR. 2012: 19-24.

[7] K Markov, M Iwata, T Matsui. Music emotion recognition using Gaussian Processes. 2014.

[8] Weninger F, Eyben F, Schuller B. On-line continuous-time music mood regression with deep recurrent neural networks[C]// ICASSP 2014 - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.

[9] Chin Y H, Lin P C, Tai T C, et al. Genre based emotion annotation for music in noisy environment[C]// 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2015.

[10] Panda R, Malheiro R, Paiva R P. Novel audio features for music emotion recognition[J]. IEEE Transactions on Affective Computing, 2018, 11(4): 614-626.

[11] Yang Y H, Liu C C, Chen H H. Music emotion classification: a fuzzy approach[C]// Acm International Conference on Multimedia. ACM, 2006.

[12] Schmidt E M, Turnbull D, Kim Y E. Feature selection for content-based, time-varying musical emotion regression[C]// Proceedings of the 11th ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2010, Philadelphia, Pennsylvania, USA, March 29-31, 2010. ACM, 2010.

[13] Wang J C, Yang Y H, Wang H M, et al. The Acoustic Emotion Gaussians Model for Emotion-based Music Annotation and Retrieval[C]// ACM Multimedia. ACM, 2012.

[14] Wang J C, Wang H M, Lanckriet G. A histogram density modeling approach to music emotion recognition[C]// IEEE International Conference on Acoustics. IEEE, 2015.

[15] Li X, Xianyu H, Tian J, et al. A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction[C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[16] Wang Y, Wang H. Multilingual convolutional, long short-term memory, deep neural networks for low resource speech recognition[J]. Procedia Computer Science, 2017, 107: 842-847.

[17] Luz, Santamaria-Granados, Mario, et al. Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS)[J]. IEEE Access, 2018.

[18] Chen X, Wang L, Pan A, et al. Channel-wise Attention Mechanism in Convolutional Neural Networks for Music Emotion Recognition[J]. 2021.