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

Deep Neural Networks for Vocal Emotion Analysis Listener Response Prediction

Haiwang Yang[1,*]

[1] School of Educational Sciences, Zhaotong University, Zhaotong 657000, Yunnan Province, China

SUMMARY: *For digital music dissemination and intelligent audio analysis scenarios, this paper constructs a deep neural network model around the problem of audience reaction prediction in vocal emotion analysis. Based on the time-frequency feature extraction and preprocessing of human voice audio, the model uses the fully convolutional network to extract the spatial information in the spectral domain, combines the bidirectional long short-term memory network to capture the time dependence of emotion in phrase progression, and introduces the context attention fusion mechanism to adaptively weight the key frequency band, key frame and cross-segment association information. Thus a computational mapping between vocal expression and listener feedback is established. The experimental results show that the accuracy of the model on the human voice emotion recognition task reaches 91.8%, and the macro-average F1 value is 91.1%. In the listener response prediction task, the mean absolute errors of preference prediction and arousal prediction are reduced to 0.356 and 0.339, respectively. The results show that the proposed model can stably improve the accuracy of emotion analysis and the ability of listener feedback prediction in complex vocal clips.*

KEYWORDS: *Deep neural network; Vocal emotion analysis; Listener response prediction; Spatio-temporal feature fusion*

1 Introduction

With the continuous expansion of digital music platforms, short video transmission scenarios and online education environments, the transmission mode of vocal music works has shifted from one-way playback to high-frequency interaction. Listeners 'stay, thumbs up, comments, re-listening and emotional feedback have gradually become an important basis for measuring the effect of vocal expression. Different from general speech signals, vocal music has linguistic information, melodic lines, rhythm organization and singing skills. Its emotional communication is not only reflected in the semantic level, but also deeply embedded in the pitch fluctuation, resonance position, energy distribution, timing change and timbre control. Because of this, the same song often elicits distinct subjective reactions in different listener groups. Traditional analysis methods that rely on manual annotation and experience judgment are difficult to support stable recognition and prediction in complex scenes.This limitation becomes more prominent in contemporary vocal communication environments, where audience response is generated under the joint influence of platform exposure, fragmented listening behavior, repeated playback and rapid emotional switching. In such settings, the target of analysis is no longer limited to the singer’s expressed affect itself, but extends to the dynamic

*15593555451@163.com
https://doi.org/10.65102/is2026082 Published: 30 April, 2026

correspondence between acoustic presentation and listener-side feedback signals. Once the analytical object changes from static category recognition to continuous response estimation, traditional approaches based on isolated handcrafted descriptors can hardly maintain consistency across heterogeneous clips, diverse singing styles and varying recording conditions. This also explains why a computational framework capable of integrating local spectral evidence, long-range phrase dependency and contextual weighting is needed in vocal emotion analysis.

In recent years, deep learning has promoted the rapid development of speech emotion recognition research. Convolutional neural networks, recurrent neural networks, self-attention networks, and pre-trained acoustic representation models have been able to extract high-level emotion representations from MEL spectrum, acoustic statistical parameters and raw waveforms, and have shown strong performance in category recognition tasks such as anger, joy, sadness, and calm. However, most of the existing studies focus on "what emotion the speaker expresses", but pay insufficient attention to "how the audience will react". The former belongs to source-side emotion discrimination, while the latter is closer to perception-side outcome modeling. Although they are related, they are not equivalent. In vocal music, vitrto, air sound, strong and weak contrast and syntactic extension may enhance the appeal, but may also be offset by different audience experience, contextual cues and aesthetic preferences. It is difficult to completely describe this process based on only a single emotional label.

From the perspective of computer modeling, listener response prediction is essentially a complex task that integrates time-frequency pattern recognition, context modeling and probability mapping. On the one hand, the model needs to capture the trajectory of emotional changes from long-term acoustic signals, and identify key clues such as local high-energy segments, timbre mutation and rhythm reconstruction. On the other hand, the corresponding relationship between these cues and the listener feedback data should be established to complete the mapping from "singing characteristics" to "response results". Without the modeling of context information, the model is easy to stay on the surface acoustic similarity, which leads to the fluctuation of prediction results across works, languages, or scenes.

Based on this, this paper focuses on the "deep neural network vocal emotion analysis audience response prediction", trying to build a set of computing framework for vocal audio: Based on time-frequency feature extraction and data preprocessing, deep neural network is used to complete the emotional spatio-temporal feature modeling, and the contextual feature fusion mechanism is combined to improve the prediction ability of audience response. The core problem to be solved in this paper is not to simply improve the accuracy of emotion recognition, but to construct a computable mapping from singing expression features to audience feedback results, and to provide method support for intelligent music analysis, personalized recommendation and online vocal music teaching evaluation.More specifically, the research attempts to connect three layers that are often discussed separately in existing studies, namely acoustic representation, emotional progression and perceptual outcome. At the representation layer, the model needs to preserve fine-grained cues related to timbre, pitch movement and energy redistribution. At the temporal layer, it must characterize how these cues evolve within and across phrases rather than treating each frame as an isolated unit. At the outcome layer, the model should convert the learned emotional organization into quantitative indicators that can reflect audience preference tendency and arousal change. The significance of this design lies in pushing vocal emotion computing from source-side recognition toward response-oriented prediction.

2 Literature review

Most of the existing research on vocal emotion analysis is extended from the framework of speech emotion recognition, but the structure and expression of vocal signals are more complex than that of general spoken language. Vocal signals not only contain basic acoustic information such as rhythm, intensity, pitch and rhythm, but also superimpose melodic organization, vocal skills and performance styles, which makes the emotional expression show stronger hierarchical and time-varying. Early studies mostly relied on manual features such as Mel-frequency cepstral coefficient, fundamental frequency and short-term energy, and combined with classifiers such as support vector machine and extreme learning machine to complete emotion recognition. Such methods have certain feasibility on small-scale data sets, but they are often difficult to stably depict the emotional evolution process in the face of long-term vocal music clips, emotional transition sections and complex noise environments [1, 2].

With the development of deep learning, convolutional neural network, recurrent neural network and attention mechanism have gradually become the mainstream path of speech emotion computing. Han et al. introduced deep neural network into speech emotion recognition earlier and proved that deep acoustic representation was superior to traditional shallow features [3]. Huang et al., Mirsamadi et al., started from the sub-sentence structure discovery and local attention mechanism respectively, and improved the model's ability to perceive key emotional segments [4, 5]. Zhang et al. proposed an attention-based fully convolutional network, which showed good local pattern extraction ability in time-frequency graph modeling [6]. Tarantino et al. further show that the self-attention mechanism helps to enhance the effect of long-distance dependency modeling [7]. These studies show that deep models are no longer limited to static feature matching, but gradually shift to the joint learning of temporal association and emotional context.

While the model structure continues to evolve, pre-trained representation and multi-task learning also promote the expansion of this field. Pepino et al. used wav2vec 2.0 embedding to improve the generalization ability of acoustic representation [8], Pastor et al. used HuBERT self-supervised representation to explore the cross-corpus transfer problem [9], and Akinpelu et al. enhanced the robustness of the model in different data domains from the perspective of transfer learning [10]. Cai et al. proposed a multi-task learning framework to enable emotion recognition to be co-optimized with related auxiliary tasks [11]. Gao et al. further expanded the boundary of traditional acoustic modeling by interactive fusion of multi-layer acoustic information and semantic information [12]. This shows that the current research has extended from a single classification problem to more complex computing paradigms such as representation learning, transfer adaptation and multi-source fusion.

However, most of the existing works focus on "speaker emotion category recognition", and the attention on "audience response prediction" is still obviously insufficient. Although Ando et al have noticed the importance of hearer dependent emotion perception model [13], their research focuses on perception difference modeling and has not yet formed a complete prediction link for listener feedback results. For vocal scene, the listener response is not a simple copy of emotional label, but the output of acoustic cues, performance context, historical preference and perceptual threshold. If only relying on single-layer emotion classification results, it is difficult for the model to explain the response shift of different listeners to the same vocal clip.

In order to present the relevant research path more clearly, Table 1 summarizes the representative work. It can be seen from Table 1 that although the existing research has made continuous progress in feature learning and model performance, there is still room for further

deepening in context fusion, listener side response mapping, and unified optimization of computational efficiency and robustness in the vocal context.

Table 1: Comparison of studies related to speech/vocal emotion computing

Study	Model/Method	Key Features	Research Focus	Main Limitations
Han et al. [3]	DNN + ELM	Deep acoustic features	Improving emotion classification performance	Insufficient modeling of long-term dependencies
Huang et al. [4]	Attention-assisted model	Sub-utterance emotional structure	Enhancing key segment recognition	Limited contextual integration
Mirsamadi et al. [5]	RNN + Local Attention	Temporal emotional cues	Improving local emotion perception	Moderate cross-scenario generalization
Zhang et al. [6]	Attention-FCN	Time-frequency spectrograms	Strengthening spatial feature extraction	Lack of listener-side modeling
Pepino et al. [8]	wav2vec 2.0	Pre-trained acoustic representations	Improving representation generalization	Limited support for specific reaction prediction
Cai et al. [11]	Multi-task Learning	Shared multi-task representations	Optimizing the robustness of emotion recognition	Task objectives still mainly classification-oriented
Ando et al. [13]	Listener-dependent Model	Differences in listener perception	Focusing on individual perceptual bias	No complete reaction prediction framework established
Gao et al. [12]	Multi-layer Acoustic-Semantic Interaction Model	Joint acoustic and semantic features	Improving complex emotion recognition performance	Insufficient adaptation to vocal performance scenarios
This Study	Deep Neural Network Fusion Model	Time-frequency features, temporal dependencies, and contextual feedback	Integrated vocal emotion analysis and listener response prediction	Further experimental validation is still needed

Synthesizing the existing research, it can be seen that vocal emotion computing has entered the stage of deep representation and multi-mechanism fusion from the stage of manual features, but the research focus is still biased towards emotion recognition itself. Concerning the title of this paper, the real key problems are: how to extract high-value cues that can affect the audience's response from the vocal audio, how to establish the "emotional expression-perceptual feedback" mapping relationship in the deep network, and how to control the inference complexity while ensuring the accuracy of the model. Based on this understanding, this paper will design the method from three levels: time-frequency feature extraction, spatio-

temporal feature modeling and context fusion optimization, so as to make up for the shortcomings of existing research in the direction of audience response prediction.

3 Methods and materials

3.1 Framework of human voice audio time-frequency feature extraction and data preprocessing

The time-frequency feature extraction of human voice is the starting point for deep neural network to carry out vocal emotion analysis and audience response prediction. Different from general speech, vocal signals carry multiple information such as lyrics semantics, melody direction, singing strength, resonance changes and emotional rendering at the same time. There are both short-term stationary characteristics and continuous fluctuations across phrases in the signal. If the original waveform is directly sent to the network, it is not only susceptible to the interference of recording equipment, environmental noise, accompaniment residue and silent segments, but also weakens the identifiability of emotion-sensitive segments in time-frequency space. Therefore, this paper constructs a processing flow of "pre-emphasis, framing, window-adding, endpoint detection, spectrum transformation, Mel mapping, statistical normalization" at the input of the model, so as to enhance the acoustic cues more related to emotional expression and audience perception, and provide stable input for subsequent spatio-temporal feature modeling. The overall process is shown in Figure 1.

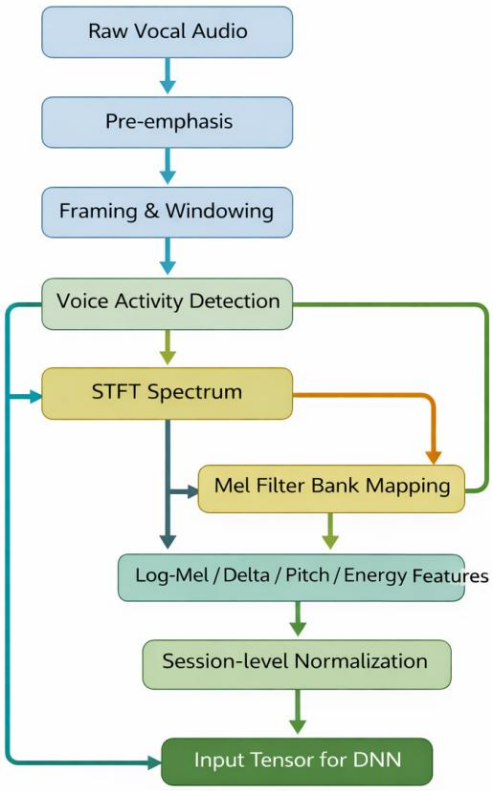

Figure 1: Framework of human voice audio time-frequency feature extraction and data preprocessing

Vocal music signals usually show high frequency energy attenuation after acquisition, and a large number of emotional changes are reflected in the high frequency harmonics, fricative

details and timbre edge regions. In order to weaken the spectrum skew and highlight the high-frequency information, the first-order pre-emphasis processing is used in this paper, and its expression is as follows.

$$y(n) = x(n) - \mu x(n-1), \quad 0.90 \le \mu \le 0.97 \tag{1}$$

where, $x(n)$ is the original sampling sequence, $y(n)$ is the signal after pre-emphasis, and μ is the pre-emphasis coefficient. Combined with the smoothness of the vocal signal and the rhythm of emotional fluctuations, this paper sets the frame length to 25 ms and frame shift to 10 ms, and uses overlapping framing to preserve the continuous relationship between adjacent segments. To suppress the spectral leakage, Hamming window is applied to each frame sample:

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

where N is the number of sampling points per frame and w(n) is the weight of the window function. Windowed short-time segments are more suitable for frequency domain analysis, and are more conducive to convolutional networks to capture local spectral patterns.

Considering that actual vocal recordings often contain breathing, pauses, accompaniment gaps and low-energy trailing segments, only relying on the whole audio will cause invalid frame accumulation. In this paper, endpoint detection is implemented by combining short-time energy and zero-crossing rate to identify valid vocal intervals. The short-term energy is calculated as follows.

$$E_t = \sum_{n=0}^{N-1} |x_t(n)w(n)|^2 \tag{3}$$

where E_t is the short-time energy of frame t and $x_t(n)$ is the NTH sampling point of the frame. By setting the dual threshold, the effective singing segment and the background noise segment can be more stably distinguished, and the silence segment can be avoided to interfere with the emotion modeling results.

After endpoint screening, this paper performs short-time Fourier transform on each frame signal to obtain the time-spectrum representation:

$$X_t(k) = \sum_{n=0}^{N-1} x_t(n)w(n)e^{-j2\pi kn/N} \tag{4}$$

where, $X_t(k)$ is the complex spectrum of the t-th frame at frequency index k. Since listeners' perception of frequency is not linearly distributed, and vocal emotion recognition is more dependent on changes at the perceptual scale, this paper further maps linear frequency to Mel frequency domain:

$$Mel(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right) \tag{5}$$

On this basis, the power spectrum is weighted and accumulated by Mel filter bank to obtain the Log-Mel spectrum. Compared with the single spectrogram, the Log-Mel feature is closer to the human auditory mechanism, and has better expression ability for high and low frequency energy redistribution, formant migration and timbre fluctuation. In order to preserve the change

trajectory of emotion over time, we further concatenate the first-order difference, fundamental frequency and frame-level energy on the basis of the static Mel spectrum to form a multi-channel input tensor, so that the network can perceive the spectrum structure information and the dynamic change of singing at the same time.

There are significant differences in the volume range and spectral distribution between different singers and different recording conditions. Without normalization, the model is easy to misidentify device features as emotional cues. To this end, this paper adopts the session-level normalization method to perform the zero mean and unit variance transformation on each audio feature independently:

$$\hat{x}_i = \frac{x_i - \mu_s}{\sigma_s + \varepsilon} \tag{6}$$

Here, x_i is the original feature value, μ_s and σ_s represent the mean and standard deviation of the current session, respectively, and ε is the smoothing term. After this processing, the dynamic range differences between different samples are compressed, which helps to improve the generalization ability of the model under cross-speaker, cross-work, and cross-scene conditions.At the feature organization stage, this normalization strategy also helps reduce the coupling between speaker-specific recording characteristics and emotion-related acoustic variation. In vocal tasks, the same emotional category may be expressed by different singers through different resonance strategies, breath control patterns and phrase intensities, while similar recording equipment may introduce surface-level consistency unrelated to emotion itself. If these two sources of variation are not effectively separated, the network may learn shortcut cues and produce unstable predictions when the test clips come from unseen singers or performance scenarios. Therefore, the preprocessing scheme in this paper should be understood not merely as signal cleaning, but as an important computational step for constraining feature bias and preserving emotionally meaningful variability.

In general, the time-frequency feature extraction and preprocessing framework constructed in this section does not only complete conventional signal cleaning, but also focus on the core issue of "how to transform vocal emotional expression into computable representation". Through the enhancement of high-frequency emotional information, effective interval screening, perceptual scale mapping and joint encoding of multiple features, the input layer can more fully retain the fine-grained cues related to audience perception, which lays a reliable data foundation for the subsequent deep neural network to carry out spatio-temporal dependence modeling and audience response prediction.

3.2 Modeling spatio-temporal features of human voice emotion based on Deep neural Network

After the time-frequency feature extraction is completed, the model needs to further answer a more central question: how the emotional cues in vocal music are effectively organized in two dimensions of time and frequency, and converted into high-level representations that can be used to predict audience responses. For vocal signals, emotion is not only attached to a single instantaneous spectral surface, but unfolds along pitch advance, intensity change, resonance transfer and rhythm extension. If the classification only relies on static features, the model can only recognize some local acoustic differences, but it is difficult to grasp the emotional progression relationship within the phrase. Therefore, this paper constructs a spatio-temporal modeling framework composed of fully convolutional network (FCN), bidirectional Long Short-Term memory (Bi-LSTM) and frame-level attention mechanism in the feature encoding stage, which is used to jointly characterize the spatial distribution and time evolution of vocal

emotion. Its structure is shown in Figure. 2.

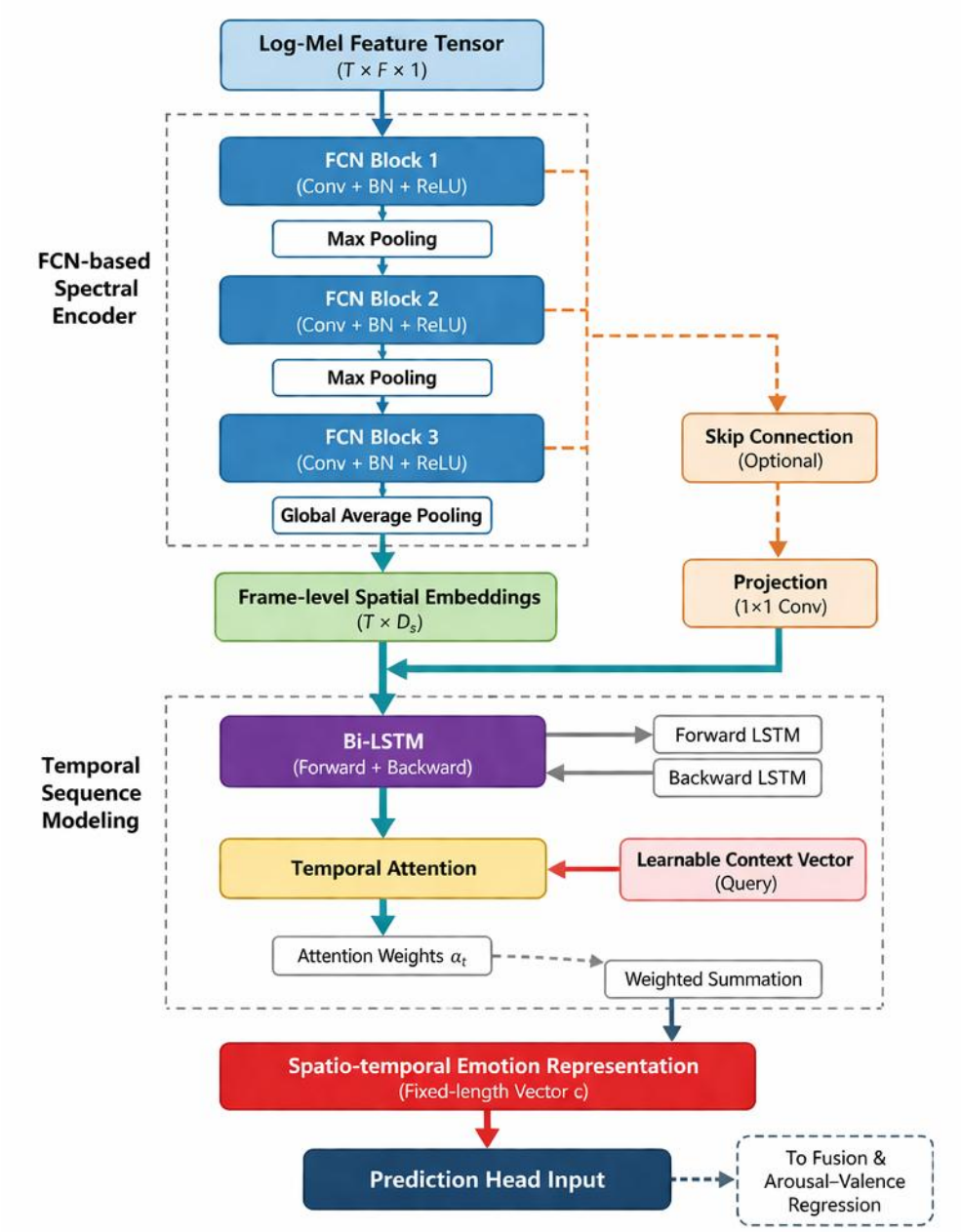

Figure 2: Framework for modeling spatio-temporal features of human voice emotion

In this framework, FCN partially undertakes the task of spectral spatial feature extraction. Log-Mel spectrogram is essentially a two-dimensional time-frequency matrix, and its local texture can correspond to emotion-sensitive information such as timbre roughness, harmonic aggregation, energy mutation and formant shift. In this paper, a three-layer convolutional structure is used to extract the spectrogram features layer by layer, so that the network gradually transitions from the low-level band response to the high-level emotion pattern representation. The convolutional layer output can be expressed as follows.

$$Y_{i,j,k} = \sigma\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\sum_{l=1}^{L} W_{m,n,l,k}\, X_{i+m,j+n,l} + b_k\right) \tag{7}$$

where X represents the input feature map, $Y_{i,j,k}$,k is the response value of position (i,j) in the output feature map on the KTH channel, W is the convolution kernel parameter, b_k is the bias term, and $\sigma(\cdot)$ is the nonlinear activation function. By stacking convolution and pooling alternately, the model can compress redundant information while retaining the local spectral shape structure that is more relevant to emotional expression. In order to avoid too large parameter scale, the global average pooling is introduced at the end of the convolution to compress the high-dimensional feature map into a frame-level spatial embedding vector, which provides a compact input for subsequent time series modeling.

However, spatial features alone are still not sufficient to describe the continuous changes of vocal emotion. Many emotional judgments don't come from a single frame per se, but rather from how the preceding and following segments are organized, such as the sudden increase in intensity after a weak rise, the recovery of timbre in a drawl, or the emotional closure caused by the drop in pitch at the end of a sentence. To this end, in this paper, the FCN output sequence is fed into the Bi-LSTM module to capture the forward and backward dual time dependence. The forgetting gate is calculated as follows.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{8}$$

where x_t is the input vector at the current time, h_{t-1} is the hidden state at the last time, W_f and b_f are the weight matrix and bias term respectively, f_t represents the retention strength of the current state to historical information. With the help of the gating mechanism, the network can screen out the transition information unrelated to emotion in the long time sequence segment, while retaining the dynamic association across phrases. The forward hidden state and backward hidden state of Bi-LSTM output are concatenated to obtain a more complete temporal expression:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{9}$$

This representation is able to incorporate both past and future context, making the model more stable when dealing with delayed expression, emotion inversion, or progressive emotion intensification.

Considering that not all frames in the vocal clip contribute equally to the emotion judgment, this paper adds a frame-level attention mechanism after Bi-LSTM to improve the weight of key emotional moments. The attention distribution is written as follows:

$$\alpha_t = \frac{\exp(v^\top \tanh(W_a h_t + b_a))}{\sum_{r=1}^{T} \exp(v^\top \tanh(W_a h_r + b_a))} \tag{10}$$

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{11}$$

where α_t represents the attention weight of frame t and c is the aggregated spatio-temporal emotion representation. This mechanism enables the model to actively focus on high emotion density segments, such as high pitch bursts, abrupt changes in strength, or significant timbre turning positions, instead of processing all frames equally. The representation thus obtained is no longer just acoustic encoding in the general sense, but is closer to "the outcome of the emotional organization that listeners may perceive".

Taken together, the spatio-temporal feature modeling method constructed in this section essentially forms a computational link from local pattern extraction of spectrum, long-term

temporal dependence memory to selective enhancement of key frames. FCN is responsible for identifying emotion-related spatial textures, Bi-LSTM is responsible for reconstructing the dynamic logic of vocal expression on the time axis, and the attention mechanism further compresses invalid information and highlights emotional core segments. After this process, the model is able to output a more compact and discriminative emotion representation, which provides a stable input for the context fusion optimization for audience response prediction in the next section.

3.3 Context feature fusion optimization mechanism for audience response prediction

After the above joint modeling of FCN and Bi-LSTM, the model has been able to obtain a relatively stable spatio-temporal representation of human voice emotion, but this initial fusion still mainly stays at the acoustic expression level. For the task of listener response prediction, local time-frequency patterns and order dependence alone are not sufficient to fully explain the auditory results. Whether a piece of vocal music can elicit positive feedback often depends not only on the emotional intensity at a certain moment, but also on the cohesion of phrases, the distribution of emotional peaks, the position of paragraphs, the continuous tension and the listener's attention allocation to key segments. Without the context reorganization mechanism, important information in long sequences is easy to decay in the propagation process, and it is difficult to achieve effective alignment between features of different modalities or different levels. Based on this problem, this paper introduces a context fusion optimization module for listener response prediction after spatio-temporal feature modeling, and completes deep feature reconstruction through three steps of "dynamic memory screening, global dependency modeling, and hierarchical attention coupling". The overall structure is shown in Figure 3.

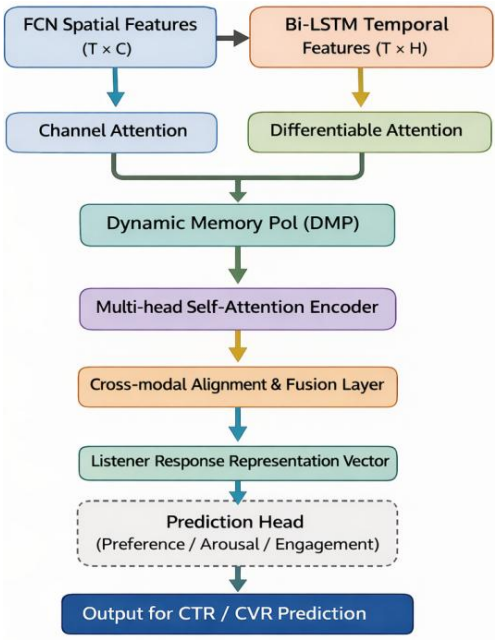

Figure 3: Optimization mechanism of context feature fusion for listener response prediction

In Figure. 3, the FCN branch outputs spectral spatial features, which can reflect emotion-sensitive patterns such as formant migration, energy aggregation and timbral edge changes. The Bi-LSTM branch outputs time series features, which are more suitable for describing the continuous evolution of emotion advancement, strong and weak turning points and phrase

tension. Although both of them serve for emotion modeling, there are obvious differences in statistical distribution and expression granularity, so they cannot be simply spliced. In order to reduce the interference of long sequence redundant information on the prediction results, this paper firstly introduces a dynamic memory pool in the time series branch to screen the high-value emotion frames. The gating calculation is given as follows.

$$g_t = \sigma(W_g h_t + b_g) \tag{12}$$

where, h_t represents the timing feature at the t-th time step, g_t is the gating weight, h_t is the filtered memory unit, and $\odot$ represents element-wise multiplication. This mechanism is able to suppress silent tail segments, low-contribution transition frames and repetitive draw-over segments, so that the emotional segments that are really related to the change of listening perception are retained.

After the key memory extraction, this paper further uses multi-head self-attention to carry out global dependence modeling of sequences. Unlike traditional cyclic propagation, self-attention does not rely on a recursive path with a fixed direction, and is more suitable for capturing long-distance emotional echoes, such as the association between the previous segment of the buildup and the subsequent segment of the burst. Its scaled dot product attention is expressed as follows.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{13}$$

Here, Q, K and V are the query matrix, key matrix and value matrix, respectively, and d_k key vector dimension. After multi-head parallel mapping, the model can observe the internal emotional organization of vocal music clips from different subspaces, so as to enhance the ability to identify paragraph level tension and listener's focus.

In addition to the temporal context, the spectrum-spatial features themselves are not uniformly valid. Some frequency bands are more sensitive to the listener's emotional perception. For example, high frequency harmonic enhancement is often associated with tension and agitation, while the smooth distribution of energy in low and middle frequencies is more likely to correspond to lyrical and soothing states. In order to highlight this kind of information, this paper adds the channel attention mechanism in the FCN branch, whose expression is as follows.

$$s_c = \sigma(W_2\, \delta(W_1 z_c)) \tag{14}$$

$$\hat{F}_c = s_c \cdot F_c \tag{15}$$

where F_c is the spatial feature of the CTH channel, z_c is the channel statistics after global average pooling, s_c is the channel weight, $\delta(\cdot)$ is the nonlinear activation function. The function of this module is not to simply enhance the feature amplitude, but to redistribute the weight of the spectral domain according to the auditory sensitivity, so that the model pays more attention to the frequency band region closely related to the listener's perception in the subsequent fusion.

After the internal reinforcement of a single branch is completed, the model also needs to solve the heterogeneous alignment problem between spatial features and temporal features. In this paper, the cross-modal attention fusion strategy is adopted to project the weighted spatial representation F and the global temporal representation H into a unified feature space, and calculate their coupling weights:

$$\alpha = \mathrm{Softmax}(W_f[\hat{F}; \hat{H}] + b_f) \tag{16}$$

$$R = \alpha_1 \hat{F} + \alpha_2 \hat{H} \tag{17}$$

Here, R is the listener response representation vector after fusion, and α_1 and α_2 represent the contribution coefficients of spatial and temporal branches, respectively. In this way, instead of mechanically averaging the two types of information, the model dynamically assigns weights based on the emotional structure of specific samples. For the segments with stronger timbre drive, the spatial branch weight will be relatively improved. For segments that rely on emotional advancement and rhythmic organization, the role of temporal branching is more pronounced.

The fused representation vector is sent to the listener response prediction head, and the output can correspond to the preference tendency, emotional arousal degree, continuous attention probability and other indicators. In terms of method logic, this section does not simply add an attention module, but establishes a closer computational pathway between "vocal expression-contextual organization-auditory response". The dynamic memory pool is responsible for filtering redundancy, the self-attention is responsible for reconstructing global dependencies, and the hierarchical attention is responsible for completing the effective coupling of heterogeneous features. After this optimization, the emotion recognition results obtained by the model are not only in the general sense, but also closer to the high-level representation of the real listener's perception process, which also provides a more reliable model basis for the subsequent analysis of results and practical application verification.

4 Results

4.1 Performance testing of human voice emotion analysis and listener response prediction models

In order to verify the actual performance of the model constructed in this paper in vocal emotion analysis and listener response prediction tasks, it is compared with three representative models such as Bi-LSTM, pure FCN and Transformer-Encoder. Each model was run under the same hardware environment and unified training strategy, the training platform was Python 3.10 and PyTorch deep learning framework, the optimizer was Adam, the initial learning rate was set to 0.001, the batch size was 32 and the training rounds were 50. To avoid information leakage caused by the same singer appearing in the training and testing end at the same time, the ratio of training set, validation set and testing set is 8∶1∶1. Considering the differences in the recording environment and singing conditions in the real scene, the enhancement operations such as noise disturbance, slight pitch modification and time stretching are also introduced in the training stage to improve the robustness of the model.

The experimental data includes two parts: one is the vocal audio samples for human voice emotion recognition, and the other is the corresponding listener feedback annotation data, which mainly includes preference score, emotional arousal score and willingness to pay attention. Based on this setting, this paper tests the classification performance, regression prediction effect and training convergence characteristics. Table 2 shows the comprehensive results of different models on the test set.

Table 2: Performance comparison of different models on the task of human voice emotion analysis and listener response prediction

Model	Emotion Recognition Accuracy / %	Macro-F1 / %	Listener Preference Prediction MAE	Arousal Prediction MAE	Inference Time per Sample / ms
Bi-LSTM	82.6	81.9	0.487	0.451	12.8
Pure FCN	84.1	83.3	0.462	0.438	10.6
Transformer-Encoder	87.4	86.7	0.418	0.401	15.3
Proposed Model	91.8	91.1	0.356	0.339	11.9

It can be seen from Table 2 that the proposed model shows better results on both types of tasks. Compared with Bi-LSTM, the emotion recognition accuracy of the proposed model is increased by 9.2 percentage points, and the macro-average F1 is increased by 9.2 percentage points, indicating that its discrimination of different emotion categories is more balanced. In the prediction of listener preference and arousal, MAE decreases to 0.356 and 0.339 respectively, indicating that the model can not only identify the explicit emotional features in human voice, but also more accurately grasp the deep expression cues that affect auditory feedback. Compared with pure FCN, the advantages of the proposed model are more obvious, which indicates that relying solely on spectral domain spatial features is still not enough to support listener response prediction, and time-dependent modeling and context fusion mechanism play a substantial role in this task. Although Transformer-Encoder outperforms traditional sequence models in global relation capture, its overall performance is still inferior to the proposed method due to the lack of hierarchical modeling for the time-frequency structure of vocal music.

To further investigate the training stability, Figure 4 shows the variation trend of the validation set accuracy of different models during the training process. It can be seen that Bi-LSTM improves rapidly in the first 15 rounds, but tends to be flat after 25 rounds, and the fluctuation is relatively obvious in the later period. The overall rise of pure FCN is relatively stable, but it saturates in advance in complex emotion samples. Transformer-Encoder performs strongly in the middle and late stages, but its convergence speed is slow in the initial stages. In contrast, the proposed model has shown high accuracy after the 10th round, and tends to be stable around the 30th round, with the smallest fluctuation range of the validation set curve. This shows that the time-frequency feature extraction, temporal dependence modeling and context weight allocation form a good synergistic relationship, and the model can effectively converge in a short training period.

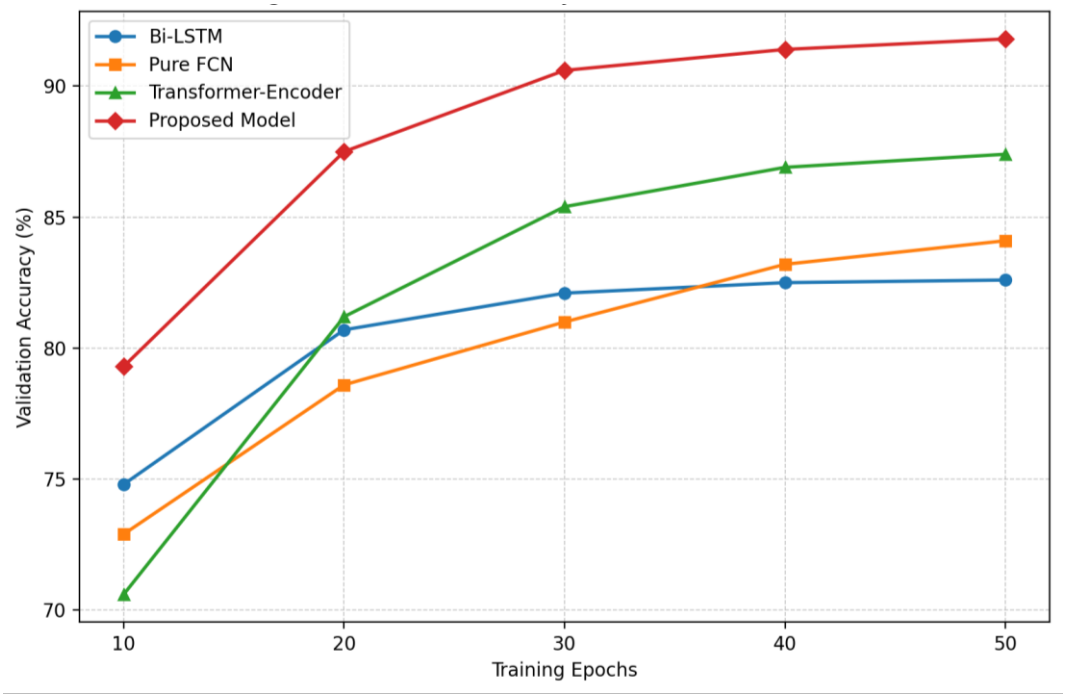

Figure 4: Accuracy variation curves of different models on the validation set

It can also be found from Figure 4 that the proposed model does not show obvious signs of overfitting in the middle and late stages, indicating that the dynamic memory screening and hierarchical attention fusion mechanism constructed in the previous section has a strong inhibitory effect on redundant features. This is particularly important for high-dimensional and long sequence tasks such as vocal emotion analysis, because many samples contain a large number of low-contribution frames in the drawl, stop connection and weak start segments, which may easily lead to the decrease of gradient propagation efficiency or blurred discrimination boundaries if not handled properly.

In order to verify the contribution of each component module, an ablation test is also performed, and the results are shown in Table 3. Taking the Bi-LSTM basic model as the starting point, after adding the FCN spatial branch, the accuracy was improved from 82.6% to 86.3%. On this basis, the accuracy is further improved to 89.7% by adding the context fusion module. When the dynamic memory pool and the hierarchical attention mechanism are enabled at the same time, the final result reaches 91.8%. This result shows that the performance improvement of the proposed model does not come from a single structure stacking, but is based on the continuous coordination of "spatial representation, time dependence and context optimization".

Table 3: Results of ablation experiments for the proposed model

Model Structure	Emotion Recognition Accuracy / %	Listener Preference Prediction MAE
Bi-LSTM Baseline	82.6	0.487
Bi-LSTM + FCN	86.3	0.431
Bi-LSTM + FCN + Context Fusion	89.7	0.381
Full Model	91.8	0.356

In summary, the proposed model has achieved ideal results in human voice emotion

recognition accuracy, listener response prediction error control and training stability. Table 2, Figure 4 and Table 3 jointly show that the joint modeling of the time-frequency spatial characteristics of vocal audio, the long-term emotional evolution information and the context weighting mechanism can effectively improve the model's ability to understand complex singing samples, and also provide a reliable basis for the performance analysis in subsequent practical application scenarios.

4.2 Prediction effect analysis of the model in different application scenarios

In order to evaluate the transfer performance of the model in different application situations, this paper further selects three task scenarios of online music recommendation, vocal music teaching feedback and short video and audio screening for testing, corresponding to three tasks of "audience preference prediction", "singing appeal evaluation" and "high response clip identification" respectively. The test input is standardized human voice audio clips, and the output is mapped to the audience's liking tendency, emotional arousal score and sustained attention probability. As shown in Figure 5, the average prediction accuracy of the proposed model in the three scenes reaches 91.6%, 90.8% and 89.7%, respectively, and the corresponding preference prediction errors are controlled within 0.36, 0.39 and 0.41, indicating that the model can not only recognize the emotional expression in vocal music, but also estimate the feedback results of the listener's side more stably.

In the online music recommendation scenario, the model has the best recognition effect on high acceptance clips, and the inference time of a single sample is maintained at about 12 ms, which can meet the needs of plat-side quick sorting. This indicates that time-frequency features and contextual memory mechanisms have strong explanatory power for audience preference. The overall accuracy in the feedback scenario of vocal music teaching is slightly lower than that in the music recommendation scenario, but the arousal prediction is more stable, indicating that the model has a better perception ability for singing details such as intensity changes, timbral control and emotional promotion, which can provide an auxiliary basis for classroom evaluation and training correction. In the scene of short video and audio screening, the accuracy of the model is decreased due to the influence of background noise, segment duration compression and accompaniment mixing, but it still remains at a high level, indicating that the context fusion structure constructed has certain adaptability to complex propagation environments.

From the application results, the proposed model does not stay at the simple emotion classification, but further realizes the computational mapping of "singing expression-listening reaction". The differences reflected in Figure 5 also show that the requirements for the model are not consistent in different scenarios: the recommendation system emphasizes more on real-time and preference resolution, the teaching feedback pays more attention to fine-grained emotional changes, and the short video screening relies more on robust recognition under noisy conditions. In general, the proposed method shows good prediction ability and transfer adaptability in multi-scenario tests, indicating that it has certain application potential in digital music recommendation, teaching assistant evaluation and content screening tasks.

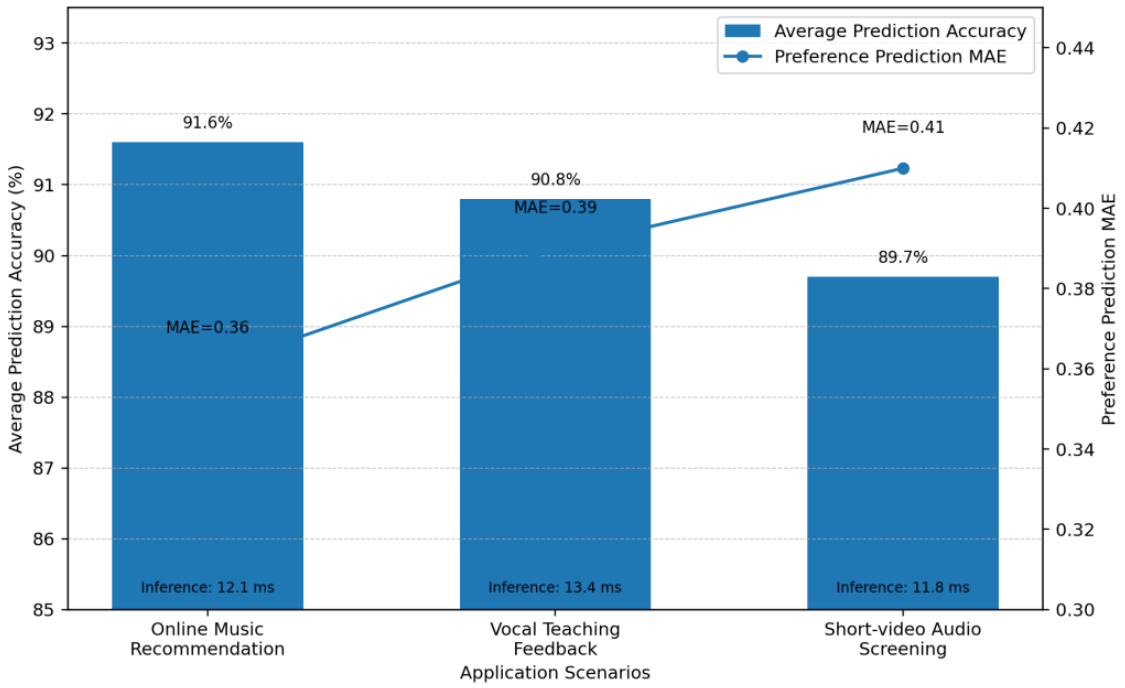

Figure 5: Comparison of model prediction effects under different practical application scenarios

4.3 Analysis of computational complexity of the model

In the task of human voice emotion analysis and listener response prediction, the model performance is not only based on the recognition accuracy, but also needs to consider the parameter scale, floating point computation and single sample inference delay. If the model structure is too complex, even if the offline test results are good, it is difficult to adapt to the application scenarios that are sensitive to response speed, such as online music recommendation, instant feedback in class or short video rapid screening. Based on this, this paper compares the computational complexity of Bi-LSTM, pure FCN, Transformer-Encoder and the proposed model in a unified hardware environment. The experimental platform is Python 3.10 and PyTorch 2.1, the GPU is NVIDIA RTX 4090, the video memory is 24 GB, and the CPU is Intel Core i9-13900K. The inference time statistics are based on the single-sample average of the test set to ensure the repeatability of the comparison between different models.

Table 4 shows the results of each model in terms of emotion recognition accuracy, single-sample inference time, number of parameters and FLOPs. The number of parameters of Bi-LSTM is relatively moderate, but its inference time is not dominant due to the strong recursive calculation of the sequence. Pure FCN performs better in parameter scale control and has lower delay than Bi-LSTM, but its recognition performance is limited due to the lack of long-distance time dependence modeling. Although Transformer-Encoder has advantages in global relation modeling, multi-head self-attention calculation will significantly increase the parameter call and floating point operation burden, so the inference time is relatively long. In contrast, the proposed model achieves a good structural balance between FCN local spectral domain extraction, Bi-LSTM temporal modeling, and attention mechanism weight screening. Although the number of parameters is higher than pure FCN, the overall inference speed is still faster than Bi-LSTM and Transformer-Encoder, showing a good efficiency advantage.

Table 4: Comparison of computational complexity and operating efficiency of different models

Model	Emotion Recognition Accuracy / %	Inference Time / ms·sample^{-1}	Parameters / M	FLOPs / G
Bi-LSTM	82.6	12.8	5.4	1.6
Pure FCN	84.1	10.6	3.8	2.3
Transformer-Encoder	87.4	15.3	8.6	3.8
Proposed Model	91.8	11.9	7.1	3.0

Combined with Table 4, we can see that the proposed model does not suffer from the problem of out-of-control computational cost due to the introduction of multi-module fusion. The number of parameters is 7.1M, which is lower than that of Transformer-Encoder, and the FLOPs is controlled at 3.0G, indicating that the model maintains a relatively compact computing structure while ensuring high expression ability. More importantly, the inference time of the proposed model is only 11.9 ms per sample, which is close to the speed level of pure FCN, but significantly better than its prediction performance. This indicates that the context fusion and dynamic weight allocation mechanism introduced in the previous section does not bring excessive burden, but improves the efficiency of feature utilization by inhibiting redundant feature propagation. In general, the proposed model achieves a balance between accuracy and complexity, which provides a feasible basis for its deployment and application in real-time and interactive scenarios.

4.4 Comparative analysis with advanced methods

In order to further evaluate the relative performance of the proposed model in vocal emotion recognition tasks, this paper selects several representative methods in recent years for horizontal comparison, including GWO-CNN based on optimized convolution structure, GA-ELM based on feature engineering and evolutionary optimization, and wav2vec 2.0 Embeddings based on pre-trained representation. And BAT model based on block-level self-attention mechanism. The comparative experiments were conducted on the RAVDESS vocal subset, and the same data division and evaluation criteria were used to ensure the comparability of the results. Table 5 shows the results of different methods on weighted accuracy, unweighted accuracy, and macro-average F1. Since most of the existing public methods do not report metrics related to listener feedback prediction, this section only makes a horizontal comparison of emotion recognition performance.

Table 5: Performance comparison with state of the art methods

Model	Year	Core Architecture	WA / %	UAR / %	Macro-F1 / %
GWO-CNN	2023	CNN + Grey Wolf Optimization	90.1	88.5	88.9
GA-ELM	2022	Feature Engineering + ELM + Genetic Optimization	85.0	82.3	83.1
wav2vec 2.0 Embeddings	2021	Pre-trained Acoustic Representations + Classification Head	91.3	89.8	90.2
BAT	2022	Block-wise and Token-wise Self-Attention Network	92.0	90.6	90.9
Proposed Model	—	FCN + Bi-LSTM + Contextual Attention Fusion	92.8	91.5	91.7

It can be seen from Table 5 that the proposed model achieves the optimal results on the three indicators. Among them, GA-ELM is strongly dependent on low-dimensional features, and although it has certain computational simplicity, it is obviously insufficient in expression ability when facing the continuous changing emotional texture in the vocal signal. GWO-CNN improves the performance of convolutional network by optimizing search, but its modeling focus is still biased towards local spectral domain response, and the processing of cross-phrase emotional extension and listener perception shift is not sufficient. wav2vec 2.0 Embeddings and BAT have been significantly improved in representation learning and global dependency modeling. However, the former emphasizes more on general acoustic representation, while the latter relies more on the self-attention mechanism to capture sequence relationships, and there is still a gap in the fine-grained context coupling for listener responses.

5 Discussion

The advantages of the proposed model are mainly reflected in the collaborative modeling of three types of information: local acoustic structure, long-term sequential emotion promotion and audience perception context. The experimental results show that the FCN branch can better capture fine-grained patterns such as harmonic aggregation, energy transition and formant shift in the spectrum, Bi-LSTM makes up for the continuous expression of emotional evolution within and across phrases, and the hierarchical attention and context fusion module further reduces the redundant frame interference, so that the model no longer stops at static emotion label discrimination. It is closer to the real listener's perception process of singing clips. Because of this, the full model is superior to the comparison methods in terms of recognition accuracy, prediction error and training stability. From the perspective of computer modeling, the effectiveness of the proposed method does not only mean that the network is deeper or the structure is more complex, but also that the weight distribution method has certain adaptive characteristics. In essence, the attention mechanism assumes the function of dynamic screening, which can adjust the contribution of key frequency bands and key moments according to the emotion density of input segments, which is similar to the idea of improving stability through feedback regulation in complex systems. Especially in the scene with noise disturbance, segment compression and style difference, this dynamic focusing mechanism helps to alleviate the error accumulation caused by long sequence information attenuation and local feature drift. There is still room for further improvement of the research in this paper. Current models still focus on behavioral feedback and rating mapping, and the modeling of individual aesthetic preference, cultural background and repeated listening effect is not sufficient. At the real-time deployment level, although the inference delay has been controlled within an acceptable range, lightweight optimization for edge devices and mobile terminals is still necessary. Further progress can be made in the directions of multimodal collaboration, individual preference modeling and online incremental learning, so as to improve the adaptability and interpretation depth of the model in complex music interactive systems.In addition, interpretability remains an important issue for the practical use of this kind of model. In recommendation or teaching scenarios, users often care not only about the predicted score itself, but also about which phrase, timbre change or emotional turning point contributes most to the final result. This means that future research can further combine attention visualization, segment-level attribution and response-path analysis to make the prediction process more transparent. Once the model is able to provide both quantitative results and structurally meaningful evidence, its value will no longer be limited to automatic estimation, but will extend to teaching diagnosis, content optimization and fine-grained human-computer interaction in digital music systems.

6 Conclusions

Aiming at the problems that vocal emotion expression has strong continuity, obvious time-frequency coupling, and it is difficult to directly characterize listener feedback, this paper constructs a deep neural network model for listener response prediction. Based on the time-frequency feature extraction and preprocessing of human audio, FCN is used to extract spatial information in the spectral domain, Bi-LSTM is used to capture the time dependence of emotion in phrase progression, and the contextual attention fusion mechanism is used to enhance the coupling ability between key frequency bands, key frames and audience-related cues. Experimental results show that the emotion recognition accuracy of the proposed model reaches 91.8%, the macro-average F1 reaches 91.1%, and the MAE of audience preference and arousal prediction are 0.356 and 0.339, respectively. The overall performance of the proposed model is better than the comparison methods. At the same time, there are still some limitations in this study, such as the modeling of individual differences of listeners is not sufficient, and the coverage of multilingual vocal music samples is still limited. In the future, multi-modal information such as text, expression or physiological feedback can be further introduced, and the lightweight deployment strategy can be combined to improve the generalization ability and practical application value of the model.

Funding

General Project of National Social Science Foundation, Survey and Research on "Jia Li" of Miao Nationality in Leigong Mountain, Guizhou Province, 19BMZ093

References

[1] Schuller B, Steidl S, Batliner A, et al. Paralinguistics in speech and language— state-of-the-art and the challenge[J]. Computer Speech & Language, 2013, 27(1): 4-39.

[2] Jahangir R, Teh Y W, Hanif F, et al. Deep learning approaches for speech emotion recognition: state of the art and research challenges[J]. Multimedia Tools and Applications, 2021, 80(16): 23745-23812.

[3] Han K, Yu D, Tashev I. Speech emotion recognition using deep neural network and extreme learning machine[C]//Interspeech 2014. 2014.

[4] Huang C W, Narayanan S S. Attention assisted discovery of sub-utterance structure in speech emotion recognition[C]//Interspeech. 2016: 1387-1391.

[5] Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention[C]//2017 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017: 2227-2231.

[6] Zhang Y, Du J, Wang Z, et al. Attention based fully convolutional network for speech emotion recognition[C]//2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2018: 1771-1775.

[7] Tarantino L, Garner P N, Lazaridis A. Self-attention for speech emotion recognition[C]// Interspeech. 2019: 2578-2582.

[8] Pepino L, Riera P, Ferrer L. Emotion recognition from speech using wav2vec 2.0 embeddings[J]. arXiv preprint arXiv:2104.03502, 2021.

[9] Pastor M A, Ribas D, Ortega A, et al. Cross-corpus speech emotion recognition with hubert self-supervised representation[C]//IberSPEECH 2022. ISCA, 2022: 76-80.

[10] Akinpelu S, Viriri S. A robust deep transfer learning model for accurate speech emotion classification[C]//International Symposium on Visual Computing. Cham: Springer Nature Switzerland, 2022: 419-430.

[11] Cai X, Yuan J, Zheng R, et al. Speech emotion recognition with multi-task learning[C]//Interspeech. 2021, 2021: 4508-4512.

[12] Gao Y, Shi H, Chu C, et al. Speech Emotion Recognition with Multi-level Acoustic and Semantic Information Extraction and Interaction[C]//Interspeech. 2024.

[13] Ando A, Mori T, Kobashikawa S, et al. Speech emotion recognition based on listener-dependent emotion perception models[J]. APSIPA Transactions on Signal and Information Processing, 2021, 10: e6.

[14] Feng H, Ueno S, Kawahara T. End-to-End Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model[C]//Interspeech. 2020: 501-505.

[15] Trinh Van L, Dao Thi Le T, Le Xuan T, et al. Emotional speech recognition using deep neural networks[J]. Sensors, 2022, 22(4): 1414.

[16] Lu C, Zheng W, Lian H, et al. Speech emotion recognition via an attentive time–frequency neural network[J]. IEEE Transactions on Computational Social Systems, 2022, 10(6): 3159-3168.

[17] Lei J, Zhu X, Wang Y. BAT: Block and token self-attention for speech emotion recognition[J]. Neural Networks, 2022, 156: 67-80.

[18] Ramesh R, Prahaladhan V B, Nithish P, et al. Speech emotion recognition using the novel SwinEmoNet (shifted window transformer emotion network)[J]. International Journal of Speech Technology, 2024, 27(3): 551-568.

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

[20] de Lope J, Graña M. An ongoing review of speech emotion recognition[J]. Neurocomputing, 2023, 528: 1-11.

[21] George S M, Ilyas P M. A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise[J]. Neurocomputing, 2024, 568: 127015.

[22] Lausen A, Hammerschmidt K. Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters[J]. Humanities and Social Sciences Communications, 2020, 7(1): 2.