Investigation of Speech Emotion Recognition Techniques Utilizing Bidirectional LSTM and Attention Mechanism

Chengxia Li, Tiantian Liu

School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, Henan, China

Abstract: Conventional speech emotion identification techniques depend on manually crafted acoustic features and superficial classifiers, resulting in restricted feature representation abilities and inadequate model generalization. This research develops a speech emotion recognition model that integrates bidirectional long short-term memory (LSTM) networks with attention mechanisms. The model initially extracts multidimensional acoustic properties from speech data, such as MFCC, Mel-spectrum, and spectral centroid. It subsequently employs a bidirectional LSTM layer to record contextual dependencies inside speech sequences and integrates an attention mechanism to emphasize emotion-critical portions. A multi-task learning framework is established to concurrently identify emotion categories, speech pace, and volume. Experiments indicate that the suggested model attains a validation accuracy of 95.28% across five emotion detection tests, surpassing SVM, LSTM, and Bi-LSTM models. This study presents a feasible approach for speech emotion recognition in complex environments and is instrumental in enhancing the emotional comprehension capabilities of human-computer interaction systems.

Keywords: Speech Emotion Recognition; Bidirectional Long Short-Term Memory Network; Attention Mechanism; Multi-Task Learning; Acoustic Feature Extraction

1. Introduction

Speech functions as a fundamental medium for conveying emotion, and its extensive linguistic data has considerable ramifications for enhancing human-computer connection [1]. Speech emotion recognition utilizes computer models to autonomously assess emotional states in speech, establishing itself as a crucial research avenue in affective computing. It exhibits extensive application potential in contexts such as intelligent customer service, remote psychological evaluation, and customized entertainment systems [2].

Initially, speech emotion recognition research depended on manually crafted acoustic features, including fundamental frequency, energy, speech rate, and various prosodic and spectral features such as MFCC, in conjunction with conventional machine learning algorithms like support vector machines [3]. Although these methodologies yielded certain advancements, their efficacy was significantly contingent upon the caliber of feature engineering. Furthermore, the constrained expressive capacity of manually crafted features led to inadequate generalization abilities across various speakers and contexts [4].

In recent years, deep learning methodologies have yielded significant advancements in voice emotion recognition [5]. Recurrent neural networks and their enhanced variations, long short-term memory (LSTM) networks, have attracted interest for their proficiency in modeling the temporal attributes of speech signals [6]. Standard LSTMs, however, can only process sequence information unidirectionally, so they fail to fully leverage the overall context of speech. Bidirectional LSTMs thoroughly capture temporal patterns in speech by concurrently analyzing input from both preceding and subsequent time steps [7]. Moreover, studies demonstrate that various portions of speech contribute differently to emotional expression. Inspired by attention mechanisms in cognitive science, the ability of models to autonomously concentrate on emotionally salient parts has demonstrated efficacy in improving recognition performance [8].

Despite previous research validating the efficacy of LSTMs and attention mechanisms, opportunities for enhancement persist in several domains: Primarily, the sufficiency of feature extraction, since a singular feature type fails to comprehensively encapsulate intricate emotional data [9]. The modeling capacity for long-term sequence relationships requires enhancement [5]. The model's generalization performance necessitates additional improvement [10]. This study proposes a model for identifying speech emotions that incorporates multidimensional acoustic characteristics, bidirectional LSTMs, and attention processes to tackle these issues.

This study's primary contributions include the invention of a multidimensional feature extraction module that effectively employs characteristics such as MFCC, Mel-spectrum, and spectral centroid to enhance the representation of emotional information. Secondly, an attention mechanism is incorporated into the bidirectional LSTM to significantly augment the model's feature extraction ability, maintaining comprehensive contextual information while emphasizing critical emotional segments. A multi-task learning method is integrated to enhance the model's generalization capability by concurrently learning interrelated tasks, including emotion, speech rate, and volume.

2. Theoretical Framework and Model Foundations

2.1 Bidirectional Long Short-Term Memory Network

In speech emotion recognition tasks, a thorough comprehension of the emotional content of speech signals frequently necessitates contextual information from both prior and subsequent segments [7]. Conventional unidirectional Long Short-Term Memory (LSTM) networks are limited to sequentially capturing historical information from the past to the present, neglecting the incorporation of future contextual information. This shortcoming is a considerable impediment in speech emotion analysis.

The bidirectional long short-term memory network (Bi-LSTM) ingeniously integrates two independently functioning LSTM layers to proficiently tackle this issue [7]. One LSTM layer analyzes the sequence in a forward direction along the temporal axis, capturing previous dependencies, while the other analyzes the sequence in reverse to obtain future contextual information. This bidirectional processing approach allows the network to acquire a comprehensive contextual representation at every time step.

The essence of Bi-LSTM is in its intricate gating mechanism. Each LSTM unit meticulously manages information using three gate structures: the forget gate ascertains the retention degree of historical data, with its output

$$f_t = \sigma(W_f \cdot [h_{t-1}, \boldsymbol{x}_t] + b_f) \tag{1}$$

The forget gate's output varies between 0 and 1, governing how much of the prior cell state is disregarded; the input gate modulates the intake of fresh information by

$$\dot{\iota}_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

evaluating the extent of information modification, followed by integration.

$$\widetilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{3}$$

The output gate regulates the generation of candidate states.

$$o_t = \sigma(W_o \cdot [h_{t-1}, \boldsymbol{x}_t] + b_o) \tag{4}$$

governs the ultimate output information quantity.

In the Bi-LSTM architecture, the comprehensive hidden state at each time step t is formed by concatenating the forward hidden state $\overrightarrow{\mathrm{h}_\mathrm{t}}$ and the backward hidden state $\overleftarrow{\mathrm{h}_\mathrm{t}}$.

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{5}$$

This bidirectional feature fusion approach allows the model to concurrently analyze both the historical development and future trajectories of speech signals in sentiment analysis, thus enhancing the precision of emotional state progression capture. The Bi-LSTM can identify minor emotional nuances in segments with sudden shifts or slow sentiment changes due to its bidirectional information flow; hence, it considerably improves identification performance.

2.2 Attention Mechanism

Drawing from human visual attention mechanisms, attention techniques in deep learning allow models to autonomously concentrate on task-relevant segments throughout the processing of sequential data. In speech emotion recognition, different speech frames at varying time intervals significantly influence sentiment expression. Attention processes proficiently discern these pivotal frames and allocate greater weights [8].

In practical applications, we analyze the hidden state sequence produced by the Bi-LSTM.

$$H = \{h_1, h_2, \ldots, h_T\} \tag{6}$$

Initially, we calculate

$$e_t = v^T \tanh(Wh_t + b) \quad (7)$$

Compute the attention score at each point, where W, b, and v are parameters subject to training. The scores are subsequently adjusted into a weight distribution using the softmax algorithm.

$$\alpha_t = \frac{\exp(e_t)}{\sum_{j=1}^{T} \exp(e_j)} \quad (8)$$

A weighted sum produces the context vector.

$$c = \sum_{t=1}^{T} \alpha_t h_t \quad (9)$$

This vector embodies the prominent emotional characteristics throughout the entire sequence.

2.3 Multi-Task Learning

Multi-task learning improves model generalization by concurrently training numerous related tasks within a single model, utilizing correlations across tasks [10]. Speech emotion recognition intricately links paralinguistic variables like speech tempo and volume to emotional states. Considering these features as auxiliary tasks promotes the model's acquisition of more distinct shared feature representations.

The primary benefit of this method is its ability to avert model overfitting via cross-task regularization effects. When the model concurrently optimizes various objectives—such as emotion classification, speech rate recognition, and volume estimation—shared layers are necessitated to acquire universal characteristics advantageous for all tasks. This enhances the fundamental emotion recognition task's reliability and precision.

3. Design of the BiLSTM-ATT-Based Speech Emotion Recognition Model

This chapter outlines the architecture of the proposed speech emotion recognition model. Figure 1 illustrates that the model consists of four fundamental components: data preprocessing and feature extraction, bidirectional LSTM temporal modeling, attention mechanism, and multi-task classification output.

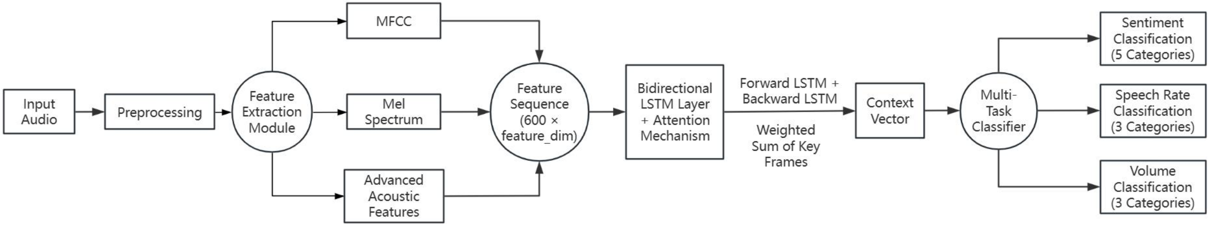
Input Audio
Preprocessing
Feature Extraction Module
MFCC
Mel Spectrum
Advanced Acoustic Features
Feature Sequence (600 × feature_dim)
Bidirectional LSTM Layer + Attention Mechanism
Forward LSTM + Backward LSTM
Weighted Sum of Key Frames
Context Vector
Multi-Task Classifier
Sentiment Classification (5 Categories)
Speech Rate Classification (3 Categories)
Volume Classification (3 Categories)

Figure 1. BiLSTM-ATT-based Speech Emotion Recognition Model

3.1 Feature Extraction and Data Preprocessing

This work develops a multidimensional acoustic feature extraction approach to thoroughly define emotional information in speech. The first step is to prepare the raw voice stream for processing. All audio samples are consistently resampled to a 22.05 kHz sampling rate and converted to mono formats to maintain uniformity in data entry. Framing utilizes a Hamming window with a frame duration of 25 ms and a frame shift of 10 ms to optimize temporal resolution and computational performance [9,10].

During feature extraction, three categories of acoustic features are derived from each speech frame [3,5]: initially, 40-dimensional MFCC coefficients are obtained, followed by the computation of their first- and second-order differences to create a 120-dimensional dynamic feature vector. This feature accurately replicates the properties of human audio perception [3]. Secondly, a 64-dimensional log-Mel spectrum is calculated to delineate the energy distribution of speech within the Mel scale. Furthermore, to delineate the fine structure of the spectrum, the spectral centroid, spectral roll-off point, spectral contrast, and chromaticity features are retrieved. This technique delineates speech features across various dimensions, encompassing spectral energy distribution, fraction of high-frequency components, spectral peak-valley structure, and pitch perception. All features are merged to create a composite feature vector. Z-score normalization is utilized to remove the impact of units, resulting in a fixed-length sequence sample including 600 elements.

3.2 Data Equilibrium and Enhancement

A category merging technique grounded in affective dimension theory was utilized to rectify the unbalanced distribution of emotion categories in the dataset. Samples that were semantically similar to "excited" but less common were put into the "cheerful" category, which made the data distribution more even. During training, various data augmentation techniques were concurrently implemented: random volume scaling (0.6–1.4x) simulated different speaking intensities; moderate Gaussian noise was introduced to improve model resilience against interference; and temporal stretching (0.7–1.3x) and pitch variation (±4 semitones) were utilized to diversify acoustic features, thereby enhancing the model's generalization ability.

3.3 Sequence Modeling and Attention Mechanism

The feature sequence is first input into the bidirectional LSTM network for deep time series modeling. The network adopts a two-layer stacking structure, the hidden layer dimension is set to 128, and each time step outputs a 256-dimensional feature vector spliced by forward and backward hidden states. This approach allows the model to concurrently utilize previous and prospective contextual information, effectively capturing long-range dependencies within speech sequences [7].

Subsequent to the Bi-LSTM layer, the model incorporates an attention mechanism to assign weights to the output hidden state sequence according to significance [8]. Specifically, for all concealed states at each time step produced by the Bi-LSTM

$$H = \{h_1, h_2, \ldots, h_T\} \quad (10)$$

An attention score is calculated for each time step using a trainable weight matrix. A normalized attention weight distribution is produced by the softmax function α_t, and the final context vector, emphasizing emotionally salient passages, is derived through weighted summation.

$$c = \sum_{t=1}^{T} \alpha_t h_t \quad (11)$$

This approach allows the model to autonomously recognize and allocate more weight to speech frames exhibiting substantial sentiment, thereby augmenting the discriminative capacity of feature representations.

The attention layer is strategically placed subsequent to the Bi-LSTM layer and prior to the categorization layer. Its purpose is to selectively emphasize the temporal information derived from the Bi-LSTM, underscoring time segments that are more pivotal for sentiment recognition. This strategy efficiently resolves the problem of uneven emotional information distribution in speech sentiment identification, enhancing the model's sensitivity to critical emotional segments.

3.4 Multi-Task Classification Framework

A multi-task learning framework is developed to improve model generalization. The context vector output from the attention layer is concurrently input into three distinct classifiers: the primary task is a sentiment classifier that generates a 5-dimensional probability distribution representing cheerful, friendly, neutral, sad, and unfriendly sentiment states; auxiliary tasks encompass a speech rate classifier (slow, medium, fast) and a volume classifier (soft, medium, loud). This architecture facilitates the acquisition of more discriminative, generalizable characteristics via collaborative training across interconnected challenges.

The model's overall loss function is characterized as the weighted aggregation of task losses:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{emotion}} + \beta \cdot \mathcal{L}_{\text{speed}} + \gamma \cdot \mathcal{L}_{\text{volume}} \quad (12)$$

where $\mathcal{L}_{\text{emotion}}$、$\mathcal{L}_{\text{speed}}$ and $\mathcal{L}_{\text{volume}}$ all employ cross-entropy loss functions. Experimental validation demonstrates that assigning weight factors α =1.0, β =0.3, and γ =0.3 efficiently utilizes the regularization benefits of auxiliary tasks while maintaining the performance of the primary work.

4. Experiments and Results Analysis

4.1 Experimental Configuration

The StyleTalk spoken dialogue dataset, supplied by National Taiwan University, served as the standard. This dataset represents the inaugural spoken dialogue benchmark featuring various speaking styles within a unified conversational setting. The training set consists of 1,878 discourse pairs (1,986 samples), whereas the evaluation set has 486 dialogue pairs (981 samples). The final dataset, following the consolidation of sentiment categories, consists of five sentiment classes: joyful, friendly, neutral, sad, and unfriendly.

The experimental setup utilized an Intel i7 processor and an NVIDIA RTX 4060 GPU, executed with the PyTorch 1.12.1 deep learning framework. The Adam optimizer was used to train the model, starting with a learning rate of 0.001 and using a cosine annealing learning rate scheduling method. The batch size was set to 32, the number of training epochs was set to 50, and an early termination mechanism was used to stop overfitting. The weight coefficients in the multi-task loss function were α =1.0, β =0.3, and γ =0.3.

Accuracy is the principal evaluation metric, augmented by precision, recall, and F1 scores for a thorough study, guaranteeing a comprehensive assessment framework.

4.2 Ablation Studies

Systematic ablation experiments were devised to validate model performance. Table 1 demonstrates that the baseline LSTM model attained a mere 34.56% validation accuracy, suggesting that sequential modeling alone inadequately captures sentiment data. The implementation of a bidirectional structure did not result in any notable enhancement for the Bi-LSTM model, as the validation accuracy persisted at 34.56%, indicating minimal performance benefits from the bidirectional mechanism alone. The incorporation of the attention mechanism led to a significant enhancement in model performance, with the Bi-LSTM+Attention model attaining a validation accuracy of 94.47%. This result illustrates that the attention mechanism may proficiently concentrate on sentiment-related critical speech segments, markedly improving feature extraction capabilities.

The proposed BiLSTM-ATT-MTL model, employing a multi-task learning framework, elevated the validation accuracy to 95.28% while maintaining the accuracy gap between training and validation sets at 4.58%. Experimental findings confirm that multi-task learning significantly improves model generalization by utilizing feature sharing and regularization effects.

Table 1. Ablation Experiment Results

Model Configuration	Validation Set Accuracy (%)	Training Set Accuracy (%)
LSTM	34.56	39.88
Bi-LSTM	34.56	39.88
Bi-LSTM+Attention	94.47	98.87
BiLSTM-ATT-MTL	95.28	99.86

4.3 Comparative Experiments

This study assessed various model types with the same dataset to further validate performance. Comprehensive results are shown in Table 2. The conventional SVM technique attained an accuracy of 92.48%, surpassing the baseline LSTM model, although it demonstrated constraints in extracting intricate sentiment features. The Bi-LSTM model attained merely 39.89% accuracy with 4.36 million parameters, indicating inadequate performance. The attention-enhanced LSTM and Bi-LSTM models attained accuracies of 92.02% and 94.74%, respectively, with 1.77 million and 1.85 million parameters.

Table 2. Model Performance Comparison

Model	Accuracy (%)	Parameters (M)
SVM	92.48	-
Bi-LSTM	39.89	4.36
LSTM+Attention	92.02	1.77
Bi-LSTM+Attention	94.74	1.85
BiLSTM-ATT-MTL (the model in this paper)	95.10	0.81

The suggested BiLSTM-ATT-MTL model attains an accuracy of 95.10% with merely 0.81 million parameters, considerably fewer than other deep learning approaches. This outcome illustrates that the proposed model sustains elevated recognition accuracy while providing enhanced parameter efficiency and practical utility.

4.4 Result Analysis

To comprehensively assess the model's recognition efficacy across many sentiment categories, we performed a detailed analysis. Table 3 illustrates that the model exhibits varied identification abilities across the five sentiment tests. The classification performance detailed in Table 3 indicates that the model performs exceptionally well in the sad and neutral categories, attaining F1 values of 97.38% and 96.55%, respectively. The recognition performance for the happy and friendly categories is commendable, with F1 ratings of 95.59% and 95.09%, respectively. The recognition performance for the hostile category is marginally inferior to that of other categories, achieving an F1 score of 90.99%. The analysis suggests that this decrease is mostly attributable to the comparatively limited quantity of training data in this category and certain acoustic features that overlap with other categories, such as friendly. The precision rate of 96.65% signifies that the model's predictions for this category are exceptionally dependable.

Figure 2 illustrates the training procedure. Both the training and validation loss functions demonstrate a consistent decline, accompanied by a corresponding enhancement in accuracy. This signifies that the model exhibits strong convergence and generalization ability.

Table 3. Recognition Performance of the Best Model on the Validation Set

Sentiment Category	Precision (%)	Recall (%)	F1 Score (%)	Number of Samples
Cheerful	94.31	96.91	95.59	291
Friendly	93.62	96.61	95.09	501
Neutral	96.31	96.80	96.55	593
Sad	97.89	96.88	97.38	96
unfriendly	96.65	85.96	90.99	235
Macro Average	95.76	94.63	95.12	1716
Weighted Average	95.32	95.28	95.25	1716

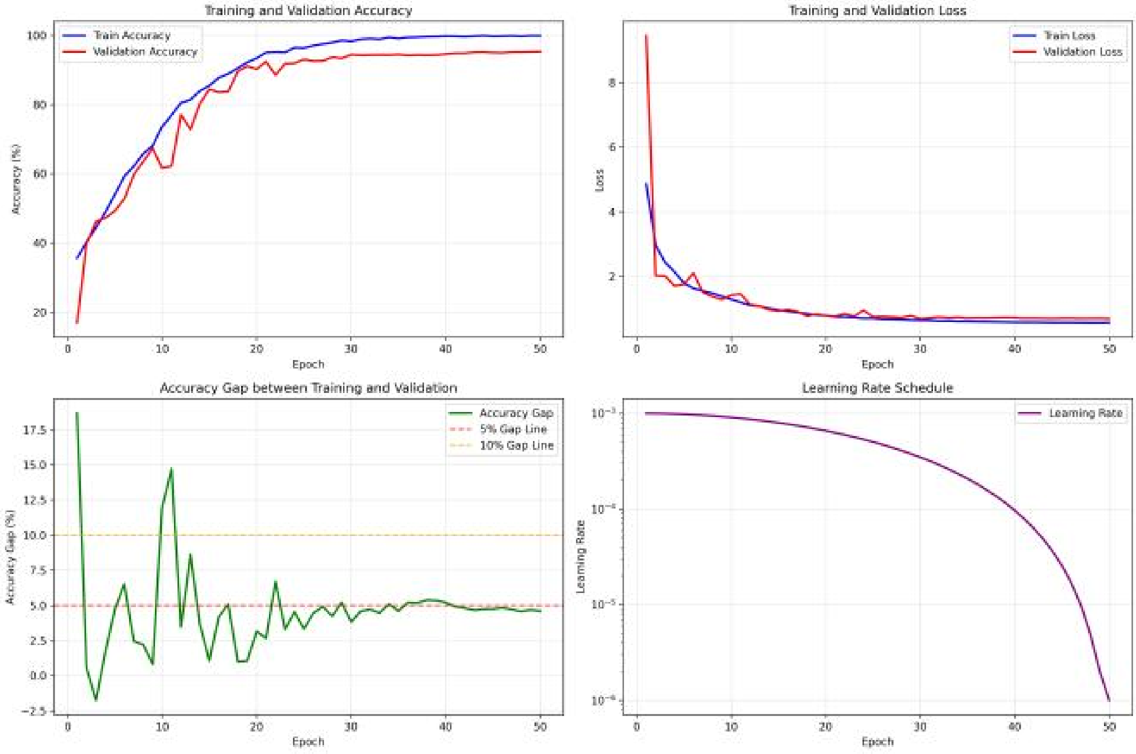

Figure 2. Training Curve

5. Conclusion and Prospective Research

This study presents an approach for identifying voice sentiment that incorporates multidimensional acoustic characteristics, bidirectional LSTM, and attention processes. The method initially extracts multidimensional acoustic features, such as MFCC, Mel-spectrum, and spectral statistics, utilizing a feature fusion algorithm. It subsequently utilizes a bidirectional LSTM network to simulate long-term contextual dependencies within voice sequences. Building on this foundation, an attention mechanism is implemented to allow the model to concentrate on critical speech portions that convey sentiment. Ultimately, multi-task learning improves the model's generalization capacity. Experimental findings indicate that our approach attains an accuracy of 95.28% in a five-class emotion detection task, markedly surpassing models including SVM, LSTM, and BiLSTM.

While our technique attains elevated recognition accuracy, there exists potential for additional optimization. Future research will concentrate on the following avenues: first, examining end-to-end network architectures to directly derive affective feature representations from raw speech signals, thereby diminishing dependence on manually crafted features; second, probing the utilization of higher-order attention mechanisms in speech affect recognition to augment model expressiveness by capturing more intricate feature interactions; furthermore, the generalization capacity of the method will be assessed across a broader range of cross-domain datasets, and training strategies such as curriculum learning will be contemplated to enhance model adaptability in complex scenarios.

References

[1] Ayadi E M, Kamel S M, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 2010, 44(3): 572-587.

[2] Hodgson J P E. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. Computing reviews, 2020(8): 61.

[3] Han K, Yu D, Tashev I. Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine// Interspeech. 2014.

[4] Chen Gang, Zhang Shiqing, Zhao Xiaoming. Natural Speech Emotion Recognition Using CNN+LSTM Combined with Data Balancing and Attention Mechanism. Computer Systems Applications, 2021, 30(05): 269-275.

[5] Shiqing Z, Xiaoming Z, Qi T. Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM. IEEE Transactions on Affective Computing, 2019, 1-1.

[6] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation, 1997, 9(8): 1735-1780.

[7] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks. Acoustics, Speech, and Signal Processing, 1988. ICASSP-88. 1988 International Conference on, 2013, 38.

[8] Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention// 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[9] Zhao Yan, Zhao Li, Lu Cheng, et al. Speech emotion recognition method based on multi-head attention long short-term memory model. Journal of Southeast University (English Edition), 2022, 38 (02): 103-109.

[10] Shen Yan, Li Hongyan, Meng Zhihong, et al. Speech Emotion Recognition Model Integrating Dual-Path CNN-LSTM and Attention Mechanism. Electronic Design Engineering, 2024, 32(18): 6-12.