Human Speech Emotion Recognition Model Based on FCN-LSTM Model

Mingjie Wang[1], Hexi Wang[2*]
[1]Research Department, China Academy of Social Sciences Educating Think-Tank, Beijing, 100009, China
[2]School of Arts & Communication, Beijing Normal University, Beijing, 100091, Chin
E-mail: hxpersonal2023@163.com
[*]Corresponding author

Keywords: Speech emotion recognition; Fully convolutional networks; Long short-term memory networks; Attention mechanisms; Feature fusion

Received: September 30, 2025

With the popularization of smart devices and the growing demand for mental health monitoring, speech emotion recognition (SER) is becoming increasingly important in intelligent interaction. The study proposes the FLA-SER model, a hybrid architecture for robust SER. The proposed architecture consists of three components: an FCN backbone for spectral-spatial feature extraction, a Bi-LSTM for modeling temporal dependencies, and a Transformer enhanced by a dynamic memory pool to capture global dependencies. Adaptive fusion of spatio-temporal features is realized by a hierarchical attention framework. The experimental results on the RAVDESS dataset revealed that the model achieved 95.3% accuracy for 'anger' emotion recognition, representing a 15% improvement over the traditional LSTM model. On the CMU-MOSI cross-lingual dataset, the average accuracy was 94.2%. The FLA-SER model is a robust solution for SER applications across languages and noisy environments. It demonstrates significant practical value in mental health monitoring and intelligent interaction scenarios.

Povzetek: Model s hierarhično pozornostjo robustno prepoznava čustva iz govora ter doseže 95,3 % pri "jezi" na RAVDESS in 94,2 % na CMU-MOSI, tudi v večjezičnih in šumnih okoljih.

1 Introduction

With the increasing popularity of smart devices and the continuous surge of mental health needs, the application of speech emotion recognition (SER) in the field of intelligent interaction is becoming more and more important [1]. SER can enhance the naturalness of human-computer communication and help provide personalized services. Meanwhile, in mental health monitoring scenarios, it can realize early screening and remote warning of emotional abnormalities, which is of great practical significance [2]. However, traditional single utterance models only model single sentence features in isolation. It is difficult to capture the temporal dependence and dynamic association of contextual emotions in speech sequences, leading to insufficient modeling of emotional coherence and limited mining of long-distance dependent features [3]. In this context, Falahzadeh et al. proposed to transform speech signals into 3D image representations for the input compatibility problem of SER deep convolutional neural networks (CNNs). Moreover, a pre-trained visual geometric group network was used for migration learning, while the model parameters were optimized by combining with the gray wolf optimization algorithm. Experimental results indicated that the research model performed well on relevant datasets and could significantly improve the performance of SER applications [4]. Albadr et al. addressed the problem that most automatic SER ignored the classification link and evaluated a single scenario, and extracted the features by using Mel frequency cepstrum coefficients. The optimization genetic algorithm-extreme learning machine was also used to optimize the classification process. Experimental results demonstrated that the optimization method significantly improved its performance in several test scenarios, with a maximum accuracy of 100%, and was able to recognize emotions efficiently [5]. To address the issue of poor storage and processing efficiency in standard machine learning for processing high dimensional speech emotion features, Chattopadhyay S et al. suggested a hybrid wrapped feature selection technique. It also incorporated linear predictive coding with linear predictive cepstrum coefficients for feature dimensionality reduction. The outcomes revealed that the model achieved a recognition accuracy of up to 98.72% on four benchmark datasets, which significantly outperformed the existing algorithms [6].

Fully convolutional network (FCN) is a CNN that removes the fully connected layers (FCLs) and has a powerful spatio-temporal feature extraction capability. Kapoor S et al. proposed a CNN method that fuses artificially designed features with deep learning features in order to fulfill the need for early automated monitoring of stress and anger. Experimental data showed that the method achieved up to 97.5% classification accuracy in multiple datasets and up to 96.7% accuracy in the validation set, with a significant reduction in loss [7]. As a special kind of recurrent neural network, long short-term memory ((LSTM) network has a unique advantage in time series modeling with its gating mechanism. Time series analysis, natural language processing, and other domains have made extensive use of it [8]. Gupta et al. proposed an acoustic feature blending method that fuses Mel-frequency cepstral coefficient features with visual bag-of-words to address the problem of limited accuracy of single model in SER. Moreover, an integrated multilayer perceptron classifier was used. The experimental results indicated that LSTM with multilayer perceptron classifier based on Mel frequency cepstrum coefficients performed the best. Moreover, the classification accuracy of all six categories of emotions was significantly improved [9]. Tejaswini et al. proposed a hybrid model of fast text CNN and LSTM to address the difficulty of early detection of depression and the lack of accuracy of existing text detection models. Experimental results indicated that the detection accuracy of this model on real datasets was better than that of existing methods, and it could provide an effective solution for the early identification of depression [10]. Yang et al. focused on the difficulty of acquiring reservoir information triggered by the lack of logging data, and proposed a method of fusing convolutional layers and LSTM. The study constructed a model containing an attention mechanism (AM), a cycle skipping mechanism and an autoregressive component to estimate the missing logs. The outcomes based on multiple well data indicated that the stability and robustness of the optimized model were better than that of benchmark models such as recurrent neural networks. Moreover, it was able to accurately generate missing logging curves such as acoustic waves [11].

Table 1 provides a comparative summary of key related works, detailing the datasets, features, models, and performance metrics used in recent SER research. Although significant progress has been achieved, persistent limitations remain in traditional models, specifically: 1) the difficulty in capturing the temporal dependency and dynamic association of contextual emotions in speech sequences, 2) the insufficient mining ability for long-distance dependent features, and 3) the limited focus on robust, cross-domain performance tied to computational efficiency. The development of the spatio-temporal feature modeling approach utilizing a fully convolutional network-long short-term memory network (FCN-LSTM) and the proposed AM to integrate contextual features is motivated by these specific gaps. The goal is to solve the aforementioned issues.

Table 1: Comparative summary of recent speech emotion recognition studies

Study	Model/Architecture	Key feature(s)	Datasets used	Metric (UAR/WA)	Representative performance
Falahzadeh et al.	GWO-CNN	2D spectrogram, GWO Optimization	RAVDESS, SAVEE	WA (%)	88.5%~90.1%
Albadr et al.	GA-ELM	MFCC, GA optimization	EMO-DB, RAVDESS	WA (%)	82.3%~85.0%
Chen et al.	Vesper (Transformer)	Pre-training, Speaker/Emotion separation	RAVDESS, IEMOCAP	WA (%)	91.5%~93.8%
Gupta et al.	LSTM-MLP	MFCC, visual bag-of-words fusion	RAVDESS, SAVEE	WA (%)	78.0%~82.5%
Proposed study	FLA-SER (FCN-LSTM-Trans-Attn)	Spatiotemporal fusion, hierarchical attention	RAVDESS, CMU-MOSI	WA (%)	94.2%~95.3%

Table 1 illustrates the prevailing trend toward deep learning and optimization in SER. However, it also highlights a lack of focus on integrated spatio-temporal modeling and demonstrated robustness. This justifies the current research direction.

Motivated by the persisting gaps in the literature, this study addresses the following research questions: (1) Can integrating the FCN backbone, Bi-LSTM temporal modeling, and Transformer-based AMs effectively enhance SER performance across languages and noisy environments? (2) Can the proposed hybrid architecture minimize inference time to maximize cross-domain generalization capability (measured by single-sample time and domain adaptability score)? (3) Can the architecture's inherent robustness, similar to principles in adaptive control theory, be leveraged to maintain performance under noisy conditions?

The study proposes a novel spatiotemporal feature modeling approach, named the FLA-SER model, to solve the aforementioned issues. The innovation of the study is twofold: First, it constructs a hybrid architecture by using FCN to capture spectral spatial features (SFs) and LSTM to model temporal dependencies. Second, a hierarchical attention framework and a Transformer enhanced by a dynamic memory pool are introduced to enable the adaptive fusion of contextual features and solve the long-distance dependency problem. The project aims to increase the model's accuracy and computational efficiency for human SER, as well as explore further possible uses of SER in intelligent interaction, mental health monitoring, and other domains.

2 Methods and materials

2.1 Framework for time-frequency feature extraction and preprocessing of speech signals

Time-frequency feature extraction of human speech signals is the modeling basis of the human SER model. The raw speech signal needs to be preprocessed to improve the signal quality and transformed into a form that adapts to the network input. The specific processing flow is shown in Figure 1 [12].

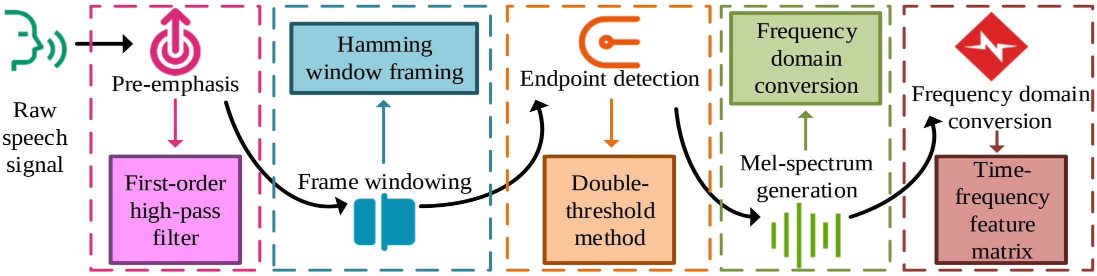


Figure 1: Schematic diagram of the time-frequency feature extraction and preprocessing framework of the original speech signal

In Figure 1, the original speech high frequency energy is attenuated, so it is first pre-emphasized. Meanwhile, the emotion-related spectral features are enhanced [13]. Equation (1) displays the pre-emphasis formula.

$$H(z)=1-\mu z^{-1},\ \ \mu=0.9375 \quad (1)$$

In Equation (1), $H(z)$ is the transfer function of the first-order high-pass filter for boosting the high-frequency energy of speech. μ is the pre-emphasis coefficient to satisfy the enhancement of high frequency emotion sensitive region in SER. z^{-1} is the unit delay operator. The speech signal possesses short-time smooth characteristics, and its spectral characteristics are stable in a short time period, so it can be processed in frames. The frame length of each frame is taken as tens of milliseconds, and overlapping frame splitting is used to maintain continuity. The spectral leakage is suppressed by weighting the window function, and the Hamming window function is used for adding the window function, as shown in Equation (2) [14].

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

In Equation (2), $w(n)$ is the n th sample value of the Hamming window to reduce spectral leakage. N is the quantity of sample points corresponding to the frame length. n is the sample index in the window. Among them, the frame length and frame shift are shown in Equation (3).

$$\begin{cases} a_1=30ms \\ a_2=10ms \end{cases} \quad (3)$$

In Equation (3), a_1 is the frame length, which ensures short-time smoothness. a_2 is the frame shift, which balances the temporal resolution and feature continuity. Speech data endpoint detection is the key technology to accurately recognize the start and end positions of valid speech segments from speech signals. The double threshold method is used to discriminate the features such as short-time energy (STE) and over-zero rate extracted from the previous frames by setting high and low thresholds. The formula for STE is shown in Equation (4) [15].

$$E(n)=\sum_{i=0}^{N-1}\left|x(i)\cdot w(i)\right|^2 \quad (4)$$

In Equation (4), $E(n)$ is the STE of the n th frame, reflecting the speech intensity. $x(i)$ is the value of the i th sampling point. $w(i)$ is the Hamming window function. The formula for the short-time over-zero rate is shown in Equation (5).

$$Z(n)=\frac{1}{2}\sum_{i=1}^{N-1}|\operatorname{sgn}(x(i))-\operatorname{sgn}(x(i-1))| \quad (5)$$

In Equation (5), $Z(n)$ is the short-time over-zero rate of the n th frame, reflecting the spectral complexity. The dual threshold judgment formula is shown in Equation (6).

$$S_n=\begin{cases} 1,(E_n>T_H)\wedge(Z_n>T_H) \\ 0,(E_n<T_L)\wedge(Z_n<T_L) \\ \quad S_{n-1},\text{otherwise} \end{cases} \quad (6)$$

In Equation (6), S_n is the judgment result of the n th frame. 1 is a speech frame and 0 is a non-speech frame used to mark the valid speech segment boundary. T_H is the set high threshold. T_L is the set low threshold for determining the non-voice segment start. S_{n-1} is the judgment result of the $n-1$ frame, which maintains the judgment state continuity to avoid misjudgment. Silent segments, such as ambient noise, vocal stops, etc., can be eliminated by the double threshold method. Then the split-frame signal is fast Fourier transformed to obtain the spectrum. The linear frequency domain is converted to the Mel frequency domain perceived by the human ear by means of a Mel filter bank. The Mel spectrum generation formula is shown in Equation (7) [16].

$$\text{Mel}(f) = 2595 \cdot \log_{10}\left(1+\frac{f}{700}\right) \quad (7)$$

In Equation (7), f is the linear frequency (Hz). $\text{Mel}(f)$ is its corresponding Mel frequency. Among them, the formula for calculating the output of the Mel filter bank is shown in Equation (8).

$$S(m) = \sum_{k=0}^{K-1} \left|X(k)\right|^2 \cdot H_m(k), m = 0,1,\ldots,M-1 \quad (8)$$

In Equation (8), $X(k)$ is the frequency domain signal after short-time Fourier transform. k is the frequency index. $H_m(k)$ is the frequency response of the m th Mel filter, and the number of filters should cover emotionally sensitive frequency bands, such as the angry high frequency region. K is the quantity of fast Fourier transform points to ensure the balance between frequency resolution and computational efficiency. Finally, independent normalization is performed by session to eliminate acoustic differences between different speakers and recording environments. The cross-speaker feature normalization method formula is shown in Equation (9) [17].

$$x_{\text{norm}} = \frac{x-\mu_s}{\sigma_s + \grave{o}}, \grave{o} = 10^{-8} \quad (9)$$

In Equation (9), x is the original Mel spectral eigenvalue. μ_s and σ_s are the eigenmean and variance of the s th session, respectively. $\grave{o}$ is the smoothing term to prevent the denominator from being zero.

2.2 Spatio-temporal feature modeling based on FCN-LSTM

In the SER, the FCN module refines speech spectral-SFs layer by layer using multi-layer convolutions. It enhances emotion-sensitive band weights and extracts emotional state-space patterns. This provides a basis for subsequent joint modeling of spatio-temporal features using the FCN-LSTM. Its specific structure for extracting speech spectral SFs is shown in Figure 2 [18].

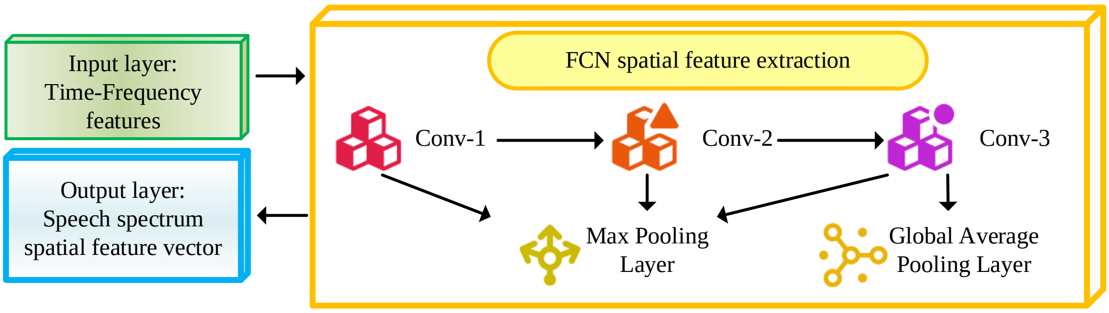

Input layer:
Time-Frequency features
Output layer:
Speech spectrum spatial feature vector
FCN spatial feature extraction
Conv-1
Conv-2
Conv-3
Max Pooling Layer
Global Average Pooling Layer

Figure 2: Flow diagram of FCN spatial feature extraction module

In Figure 2, the raw speech signal will enter the FCN module after preprocessing. FCN SF extraction adopts a 3-layer convolution-pooling structure. Conv-1 captures the underlying frequency patterns, Conv-2 extracts the more complex spectral structure, and Conv-3 focuses on high-dimensional abstract features. The convolutional layer feature extraction formula is shown in Equation (10).

$$Y_{i,j,k} = \sigma\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} W_{m,n,k} \cdot X_{i+m,j+n,l} + b_k\right) \quad (10)$$

In Equation (10), $X_{i+m,j+n,l}$ is the speech spectrum value of coordinates $(i+m, j+n)$, channel l in the input feature map (FM), corresponding to the preprocessed Mayer spectrum time-frequency matrix. $W_{m,n,k}$ is the convolution kernel (CK) weight of size $M \times N \times L$ (M , N is the CK size. L is the quantity of input channels for extracting spectral space features. b_k is the bias term (BT) for the k th output channel to enhance model fitting. σ is the activation function (AF) to highlight emotion-sensitive frequency band features. $Y_{i,j,k}$ is the spatial eigenvalue of coordinate (i, j) and channel k in the output FM, which corresponds to the emotion-sensitive pattern after layer-by-layer refinement. Each convolution layer is connected to a maximum pooling layer. This layer uses a sliding window to identify the maximum value of local area features, achieving local dimensionality reduction while retaining the strongest response features. Global average pooling (GAP) is introduced at the end layer of FCN module. A global averaging operation is done on all spatial dimensions of the FM to compress the 3D FM into one-dimensional vectors, leaving only the global statistical properties of the channel dimensions. Ultimately, the spatial feature vectors (FVs) are output for subsequent temporal feature splicing with LSTM. Aiming at the temporal dependence problem of speech sequences, the study introduces the LSTM module. Through the bidirectional structure and self-AM, it captures the long-distance emotion dynamic association and adaptively focuses on the key emotion frames. Its specific process of temporal feature modeling combined with FCN is shown in Figure 3 [19].

In Figure 3, the preprocessed serialized speech features are fed into the LSTM module. It is first processed by the Bi-LSTM layer, which controls the retention and forgetting of information through a gating mechanism. Meanwhile, the LSTM module constructs a bi-directional network. The forward LSTM parses the sequence along the temporal direction, and the backward LSTM backtracks the information in the reverse temporal direction to collaboratively capture the bidirectional temporal dependence of the speech signal. The formula for the forgetting gate (FG) mechanism is shown in Equation (11).

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (11)$$

In Equation (11), $f_t = 1$ denotes complete retention. $f_t = 0$ denotes completely forgotten. x_t is the input FV of the current moment t, i.e., the subframe Mel spectrum. h_{t-1} is the hidden state of the previous moment $t-1$, which stores the historical timing dependency information. $[h_{t-1}, x_t]$ is the vector that splices the historical hidden state with the current input. W_f is the weight matrix of the FG, which calculates the importance weight of the input information. b_f is the BT of the FG, which regulates the gating activation threshold. σ is the sigmoid AF, which outputs a weight value between 0 and 1. Subsequently, the feature splicing layer maps the 256-dimensional temporal features of the LSTM with the 256-dimensional SFs generated by the FCN module and fuses them into a 480-dimensional vector. Finally, the output layer outputs the spatio-temporal FV of this speech signal, which provides the core input for the subsequent emotion classification task.

For full reproducibility, the complete architectural hyperparameter details for the FCN backbone and the DMP-Transformer module are summarized in Table 2.

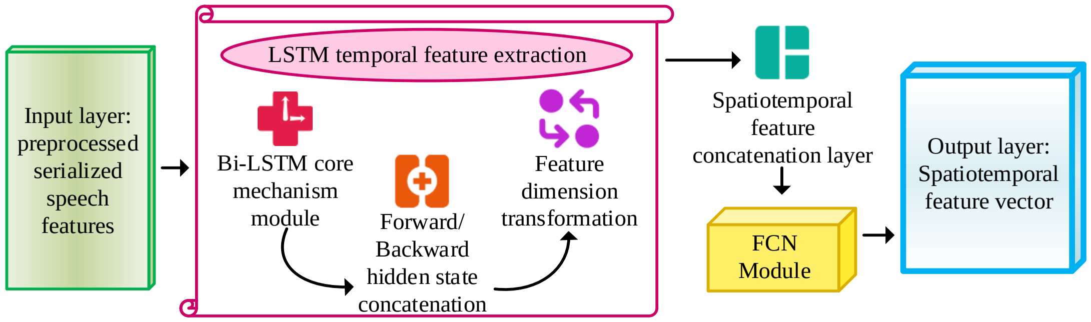

Input layer: preprocessed serialized speech features
LSTM temporal feature extraction
Bi-LSTM core mechanism module
Forward/ Backward hidden state concatenation
Feature dimension transformation
Spatiotemporal feature concatenation layer
FCN Module
Output layer: Spatiotemporal feature vector

Figure 3: Flow diagram of LSTM timing feature extraction module

Table 2: Architectural hyperparameter details of the FLA-SER model

Component	Layer	Kernel size/Head	Stride	Filters/Dim.	Activation	Notes
FCN Backbone	Conv-1	3×3	1	64	ReLU	Captures underlying frequency patterns
	Conv-2	5×5	1	128	ReLU	Extracts complex spectral structure
	Conv-3	3×3	1	256	ReLU	High-dimensional abstract features
	Max pooling	2×2	2	/	/	Dimensionality reduction
DMP-Transformer	Multi-head attention	8 Heads	/	Query/Key: 64	Softmax	Global dependency modeling
	Position-wise FFN	/	/	Hidden: 1024	ReLU	/
	Positional encoding	/	/	/	/	Positional encodings are used

Table 2 provides the details necessary for model replication. The configuration is carefully tuned to balance feature extraction capacity and computational load, particularly the use of 3×3 and 5×5 kernels in the FCN and eight attention heads in the DMP-Transformer.

2.3 Contextual feature fusion optimization based on AM

The constructed FCN-LSTM model has realized the preliminary fusion of spatio-temporal features. However, it still has the problem of gradient vanishing in long-time sequence dependency modeling, and the multimodal feature alignment lacks dynamic adjustment mechanism. As the core technology of current sequence modeling, the AM can dynamically assign weights, effectively model global dependencies, and effectively improve the accuracy and flexibility of feature fusion. Aiming at the problem of long-sequence information decay in FCN-LSTM, the study introduces a hybrid architecture of dynamic memory pool (DMP) and Transformer. The specific optimization architecture is shown in Figure 4 [20].

In Figure 4, the input layer receives the timing FVs output from the Bi-LSTM. The DMP module processes them through a gated filtering mechanism. The mechanism uses a sigmoid AF to calculate the importance weight of each timing frame. The module enables the filtering of redundant information, and the filtered key memory units enter the Transformer timing modeling module. Three mappings are performed first and then with the help of multi-head self-AM. The dot product attention formula is scaled to determine the cross-frame emotional connection weights. Finally, the module stitches the output into 256-dimensional global dependency features. The study designs a hierarchical attention fusion framework around the heterogeneity problem of FCN SFs and LSTM temporal features, as shown in Figure 5.

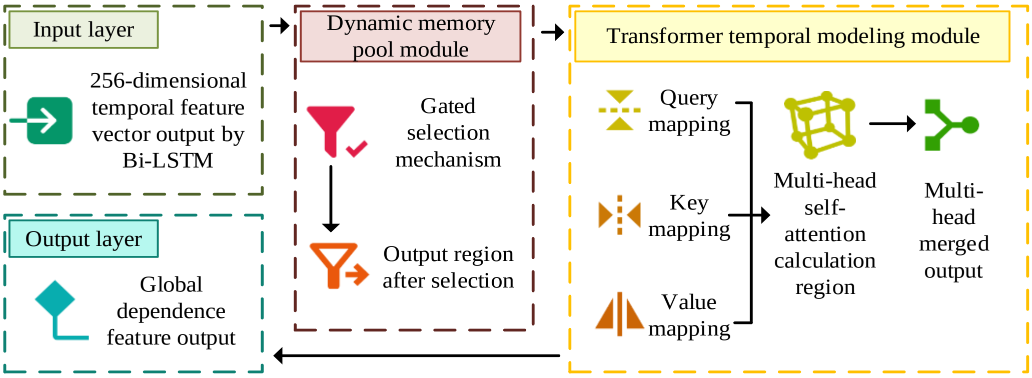

Input layer
256-dimensional temporal feature vector output by Bi-LSTM
Output layer
Global dependence feature output
Dynamic memory pool module
Gated selection mechanism
Output region after selection
Transformer temporal modeling module
Query mapping
Key mapping
Value mapping
Multi-head self-attention calculation region
Multi-head merged output

Figure 4: Schematic diagram of DMP-Transformer fusion mechanism architecture

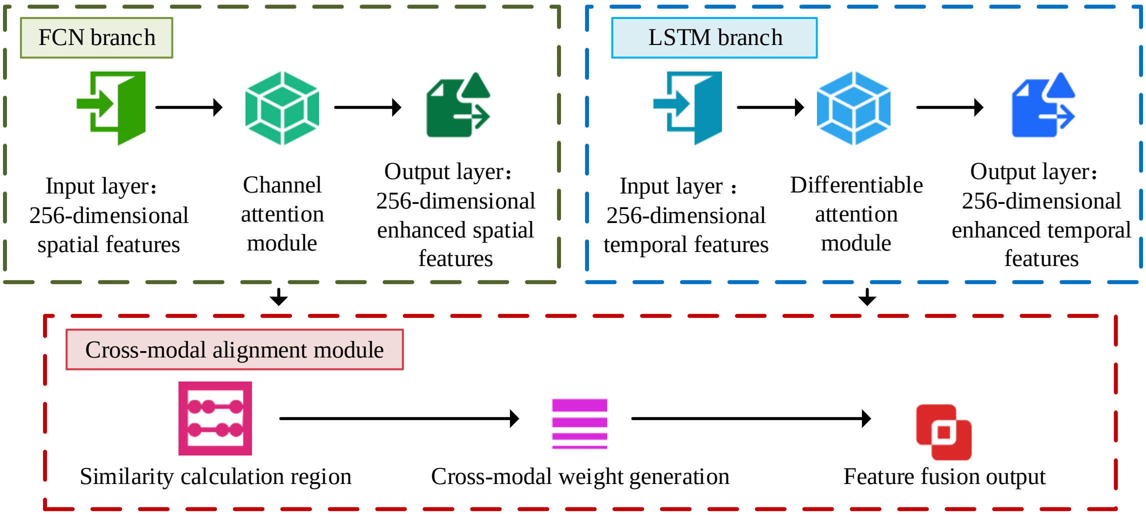
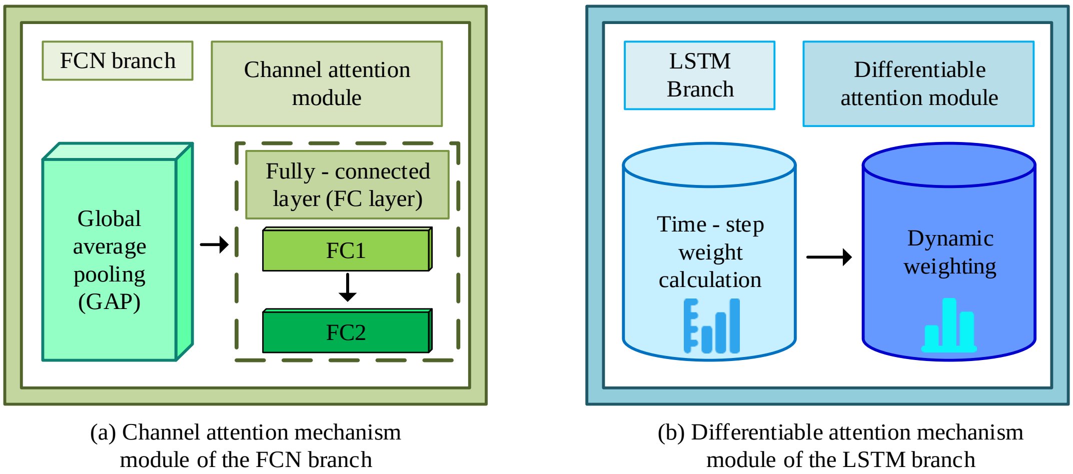

FCN branch
Input layer: 256-dimensional spatial features
Channel attention module
Output layer: 256-dimensional enhanced spatial features
LSTM branch
Input layer : 256-dimensional temporal features
Differentiable attention module
Output layer: 256-dimensional enhanced temporal features
Cross-modal alignment module
Similarity calculation region
Cross-modal weight generation
Feature fusion output

Figure 5: Schematic diagram of the hierarchical attention fusion framework


FCN branch
Channel attention module
Global average pooling (GAP)
Fully - connected layer (FC layer)
FC1
FC2
LSTM Branch
Differentiable attention module
Time - step weight calculation
Dynamic weighting

(a) Channel attention mechanism module of the FCN branch

(b) Differentiable attention mechanism module of the LSTM branch

Figure 6: Structure of the AM in different branches

In Figure 5, the overall framework is divided into two levels. At the intra-modal attention level, the FCN branch introduces the channel AM, which is targeted to strengthen the emotion-sensitive frequency bands in the spectral energy distribution. The LSTM branch adopts a differentiable AM to effectively suppress irrelevant information interference such as silent frames. A dual-tower attention network structure is built at the cross-modal alignment level, and the temporal and SF similarity matrices are calculated independently. The cross-modal weight matrix is generated by Softmax function to realize the dynamic coupling and weight allocation of the two types of features. The FCN branch of the channel AM module and the LSTM branch of the differentiable AM module. The specific structure of the two is shown in Figure 6.

In Figure 6, the input SFs in the left FCN branch channel attention module are first pooled by GAP. The spatial information of each channel is compressed into a global statistic, and then a channel weight vector is generated via the FCL. This vector is multiplied with the original features after Sigmoid activation to realize the reinforcement of emotionally sensitive spectral energy distribution such as anger high frequency band. The right LSTM branch differentiates the AM and computes the weight coefficients at each time step for the input temporal feature sequence. The weights are generated by the dot product of the query vector and the features of each time step via Softmax. It can dynamically adjust the emotional contribution of different frames and effectively suppress the interference of irrelevant information such as mute frames. Analysis of the attention weights reveals that the highest weights are consistently assigned to frames that coincided with vocal bursts, rapid pitch changes, or areas of high emotional intensity. This finding validates the differentiable attention module's focus on critical time steps and provides interpretability. In summary, the study first preprocesses the speech signal to obtain the Mel spectrum. Then FCN is utilized to extract SFs and LSTM is employed to capture timing dependencies. Finally, the new SER optimization model, named the FLA-SER model (FCN-LSTM network with AM for human SER model), is constructed by fusing features through the DMP-Transformer hybrid architecture and the hierarchical attention framework. The FLA-SER name is used consistently throughout the remaining sections.

3 Results

3.1 Performance testing of the FLA-SER model

To verify the actual performance of the FLA-SER model, it is compared with the traditional LSTM, pure FCN, and 3D-VGG SER models. All comparison models are trained and evaluated under the exact same conditions and hyperparameters to ensure a fair comparison. The experiments are implemented based on Python's sklearn and a deep learning framework. All experiments are conducted using five independent runs to ensure the stability and statistical significance of the results. The average standard deviation across all reported metrics in Figures 7–10 is less than ±1.0%, confirming the robustness of the presented data.

The experimental setup details are critical for replication. The datasets are divided into training, validation, and testing sets using a speaker-independent protocol (80% for training, 10% for validation, and 10% for testing). This ensures that there is no overlap between the training and testing partitions of speakers, which is crucial for evaluating SER generalization. Five-fold cross-validation is performed on the training and validation sets to stabilize model performance. To address data scarcity and mitigate overfitting, data augmentation techniques are applied to the training set. Augmentation includes noise injection (using ambient noise from the ESC-50 dataset at SNR levels between 10dB and 20dB) and pitch shifting (±2 semitones). This strategy effectively increases the training data volume by a factor of four, significantly improving the model's generalization capability and robustness against real-world variations.

The experimental data for the study are derived from RAVDESS, IEMOCAP, and EMO-DB databases, which generally cover a wide range of emotion categories and acoustic scenarios. First, the study preprocesses the raw speech signals from the RAVDESS dataset to obtain Mel spectra of two categories of intense emotions as model inputs. The result of recognition precision for each emotion category after 50 rounds of training for the four models is shown in Figure 7.

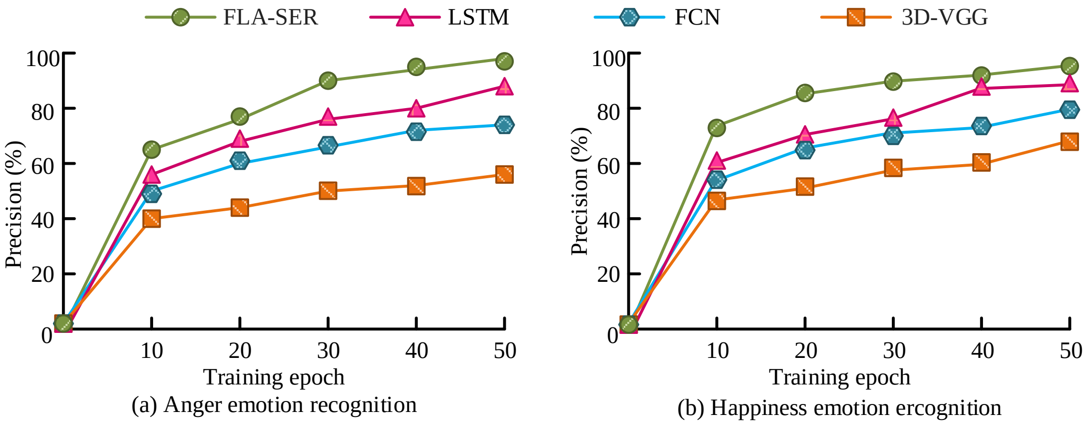
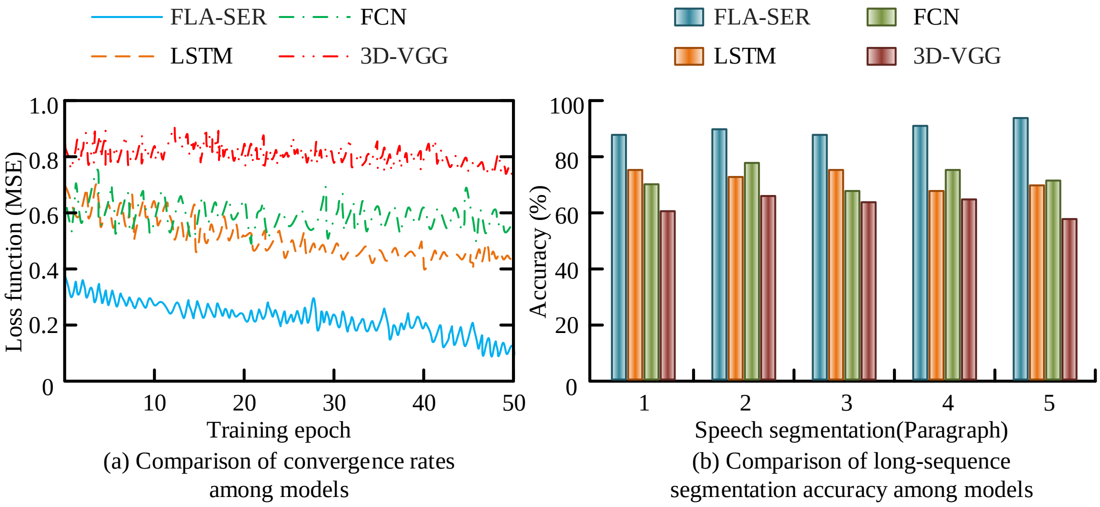


Figure 7: Recognition precision of emotion categories for different recognition models

In Figure 7(a), the recognition precision of the four models for the emotion "anger" increases with the number of training generations. The FLA-SER recognition model has the fastest precision increase, with a final precision of 95.3%. The traditional LSTM recognition model and the pure FCN model do not have much difference in precision, but both of them are lower than the research model. The 3D-VGG recognition model has the lowest improvement and final precision. Figure 7(b) shows the recognition precision of each model for the emotion "happy". Although the four models have similar upward trends in recognition precision for the "happy" emotion, the FLA-SER recognition model achieves a high precision of 75.4% after 10 training sessions. Moreover, it is higher than the other three comparison models throughout the training. This reflects the superiority of the research model's classification accuracy under multiple sentiment categories. The study uses IEMOCAP conversational speech input as a split-frame Mel spectral sequence. The sequence length is minimized to 50 frames to verify the models' ability to capture long-distance emotion dependencies in dialog scenes. The test results are shown in Figure 8.

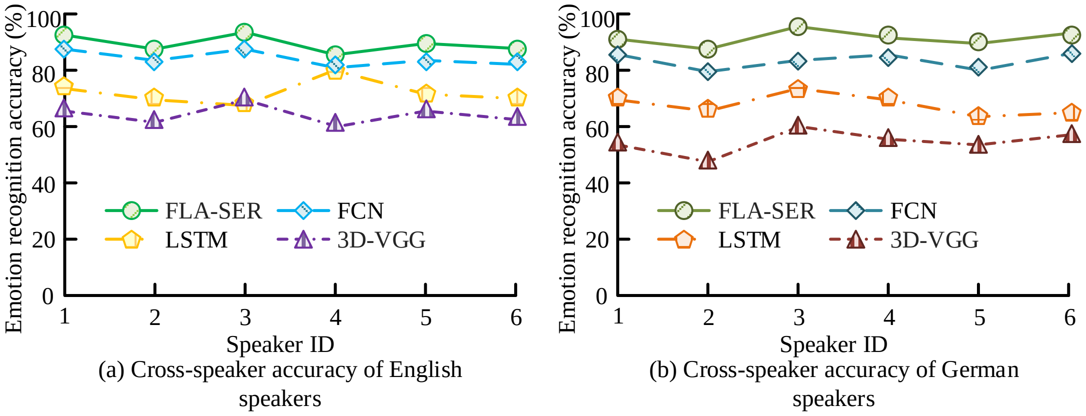
Figure 8(a) shows a comparison of the convergence rates of the four models. The mean squared error (MSE) of the traditional LSTM model decreases faster at the beginning of training, but the MSE decreasing trend tends to level off after 30 generations. The MSE of the pure FCN model decreases slowly and eventually stabilizes above 0.5, and the loss fluctuation is obvious under long-sequences. The 3D-VGG model always has an MSE above 0.6 and converges the slowest. The MSE of the study model drops rapidly to below 0.3 after the 20th generation and stabilizes at around 0.15 by the 50th generation. It is the best model for gradient stability under long-sequences. Figure 8(b) shows the comparison of the long-sequence segmentation accuracy of each model. The accuracy of the first 3 segments of the LSTM model is about 75.3%, but it decreases in the later stages due to the decay of timing information. The accuracy of each segment of the FCN model fluctuates slightly, and the full segmentation accuracy of the 3D-VGG model are all below 65.0%. In contrast, the FCN-LSTM model has excellent accuracy for each long-sequence segmentation, which is over 85.0%. The accuracy of the 5th segment (the end of the long-sequence) is as high as 92.1%. To further validate the cross-linguistic generalization ability of the research models, cross-speaker tests are selected using the RAVDESS (English) dataset and the EMO-DB (German) dataset with inputs of Mayer spectra. The experimental results of each model are shown in Figure 9.



(a) Comparison of convergence rates among models

(b) Comparison of long-sequence segmentation accuracy among models

Figure 8: Comparison of long time-series dependency modeling capabilities of different recognition models



(a) Cross-speaker accuracy of English speakers

(b) Cross-speaker accuracy of German speakers

Figure 9: Comparison of cross-language generalization ability of different recognition models

In Figure 9(a), the FCN-LSTM model fluctuates ≤5% for each speaker, and the accuracy of emoticon recognition is above 90.0% for all of them. The LSTM model and the 3D-VGG model fluctuate more in accuracy, while the FCN model is still lower than the study model, although the accuracy is higher throughout. In Figure 9(b), the average accuracy of FCN-LSTM model is not much different from that of the English group and still significantly higher than the other models. The recognition accuracy of the 3D-VGG model is significantly lower compared to the English group, and significant recognition problems occur due to individual speaker pronunciation differences. In summary, the cross-linguistic generalization ability of the research model is excellent. The fluctuation of motion recognition accuracy is small in the same language cross-speaker recognition. 1000 speech samples from the RAVDESS database are randomly selected as model inputs. The time-consuming results of emotion recognition for the four models are shown in Figure 10.

In Figure 10, the 3D-VGG model shows a significant increase in emotion recognition time with increasing sample size. Its recognition time is up to 32.7 seconds for high sample sizes. The LSTM and FCN models show significant fluctuations in recognition time when recognizing sample sizes from 500 to 1,000, increasing to more than 20 seconds. The overall recognition time of the FCN-LSTM model is significantly reduced compared to the first three models. At high sample sizes, the recognition time is only 14.2 seconds. This demonstrates the filtering effect of the AM on the redundant features and avoids the “sample size-time” linear growth problem of the traditional models.

The contribution of each component within the FLA-SER model is systematically validated through an ablation study. Table 3 summarizes the incremental accuracy improvements on the RAVDESS dataset and demonstrates that the integrated hybrid architecture significantly improves overall performance. The ablation study is performed stepwise, building upon the Bi-LSTM base module.

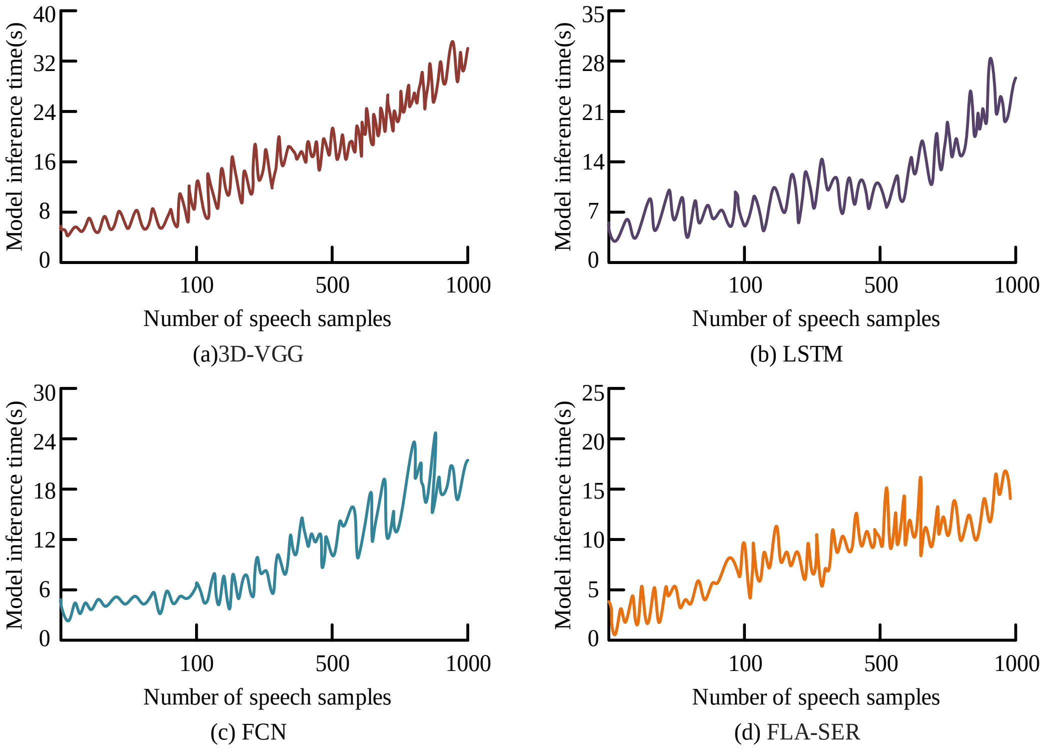


Figure 10: Comparison of sentiment discrimination speed of different recognition models

Table 3: Ablation study of FLA-SER components on RAVDESS (Weighted accuracy %)

Model Architecture	Bi-LSTM Base	+ FCN Backbone	+ DMP Module	+ Hierarchical Attention
Accuracy (%)	80.2±1.5	85.5±1.2	90.1±0.9	95.3±0.8
Improvement (%)	Base	+5.3	+4.6	+5.2

Table 3 clearly demonstrates the incremental value of each proposed module. The FCN backbone provided the largest improvement, confirming the strength of SF modeling. The combination of DMP and hierarchical attention is essential for achieving a final performance of 95.3%.

3.2 Effectiveness of the practical application of the FLA-SER model

The open-source cross-language datasets CMU-MOSI (English), SEMAINE (English, French, German) and EMOVO (Italian) are selected for the study to validate the recognition ability of the FLA-SER model under different languages and emotion types. Table 4 displays the test results.

In Table 5, the FLA-SER model has an average accuracy of 94.2% in the English CMU-MOSI dataset, with an accuracy of more than 92% in the 5th segment of the long-sequence. The F1 values for “angry” and “happy” emotions are 95.6% and 91.4%, respectively. In the face of the multilingual SEMAINE dataset, the model demonstrates its ability to capture multilingual emotion signals with a 90.1% depression F1 value and a 3.5 cross-lingual generalization index. The Italian EMOVO dataset has 89.3% and 86.7% anger and happiness F1 values, respectively. The long-sequence dependency modeling accuracy is over 87%. To further validate the recognition performance of the research model in multi-domain application scenarios, the study uses the MELD dataset, the CSD Chinese customer service dataset, and the DriveTalk dataset as the test data. This is used to simulate the model's EMOTION RECOGNITION ability in mental health monitoring, intelligent customer service and in-vehicle voice interaction scenarios, respectively. Table 5 displays the test results.

Table 4: Recognition ability of the model in different languages and emotion types

Result parameter	CMU-MOSI	SEMAINE	EMOVO
Language	English	English/French/German	Italian
Emotional category	7 categories	Depression	6 categories
Average accuracy (%)	94.2	90.5	90.9
Single sample time	14.5	16.2	15.8
Accuracy of 5th segment in long-sequence (%)	92.3	89.7	87.6
F1 value for anger (%)	95.6	/	89.3
F1 value for happiness (%)	91.4	/	86.7
F1 value for depression (%)	/	90.1	/
Cross-lingual generalization index	2.1	3.5	4.2

Table 5: Recognition performance of the model in multi-domain application scenarios

Result parameter	MELD	CSD	DriveTalk
Application domain	Mental health monitoring	Intelligent customer service	In-vehicle voice interaction
Key task	Early screening of depressive emotions	Classification of customer emotional satisfaction	Real-time recognition of driving emotions
Average accuracy (%)	90.7	91.5	88.3
Single sample time (ms)	17.8	15.6	19.2
Accuracy of 5th segment in long-sequence (%)	89.3	90.2	83.5
Accuracy at 10dB signal-to-noise ratio (%)	86.4	89.7	83.7
Accuracy at 0dB signal-to-noise ratio (%)	78.5	81.2	75.4
Domain adaptability score (1-5)	4.5	4.8	4.2

In Table 5, the FLA-SER model has an average accuracy of 90.7% for early screening of depressive mood in the field of mental health monitoring. The accuracy of the 5th segment of the long-sequence reaches 89.3%. In a noisy environment (0 dB signal-to-noise ratio), it still maintains an accuracy of 78.5%. It is specified that the noise injected for the 0dB SNR tests consisted of Gaussian white noise and background chatter at the signal level, simulating common real-world deployment conditions. In the intelligent customer service scenario, the research model performs customer emotional satisfaction classification. The average accuracy rate reaches 91.5%, and the combination of Chinese tone features makes the domain adaptability score as high as 4.8. In the field of in-vehicle voice interaction, the model's average accuracy rate performs well in the face of a noisy environment. Although the accuracy decreases with the reduction of signal-to-noise ratio, it still reaches 75.4% at 0dB. This reflects its good robustness in noisy environments.

Table 6: Computational complexity and efficiency comparison

Model	WA accuracy (%)	Inference time (ms/sample)	Parameter count (M)	FLOPs (G)
Traditional LSTM	80.3	24.1	5.1	1.2
Pure FCN	82.5	18.9	3.5	2.5
3D-VGG	75.8	32.7	12.8	4.5
FLA-SER	95.3	14.2	7.2	3.1

3.3 Computational complexity analysis

The efficiency of the FLA-SER model is substantiated by its low latency (14.2ms/sample). For reproducibility, all inference timings are conducted on an NVIDIA GeForce RTX 4090 GPU with 24GB of memory and an Intel Core i9-13900K CPU. Table 6 provides a comparative analysis of computational complexity and confirms that the proposed architecture achieves a superior balance between recognition accuracy and efficiency.

Table 6 confirms the superior efficiency of the FLA-SER model. Although it has a moderate number of parameters compared to 3D-VGG, the simplified FCN backbone and efficient AM result in the shortest inference time (14.2 ms/sample) of all the models being compared. This provides a strong basis for real-time deployment.

3.4 Comparison with state-of-the-art (SOTA)

To rigorously contextualize the model's novelty and performance advantage, a quantitative comparison with other State-of-the-Art methods, such as Vesper, GWO-CNN, and GA-ELM, is presented in Table 7.

Table 7: Quantitative comparison of FLA-SER with State-of-the-Art Models (RAVDESS Dataset)

Model	Year	Architecture	WA Accuracy (%)	UAR Accuracy (%)	Reference
GWO-CNN	2023	CNN + GWO	90.1	88.5	[4]
GA-ELM	2022	Feature + ELM + GA	85.0	82.3	[5]
Vesper	2024	Transformer-based	93.8	92.5	[1]
FLA-SER (Proposed)	N/A	FCN-LSTM-Trans-Attn	95.3	94.1	This Study

Table 7 provides the quantitative evidence of the FLA-SER model's leading performance. The FLA-SER model surpasses all contemporary SOTA methods by achieving a weighted accuracy of 95.3% on the RAVDESS dataset. This level of performance justifies the integrated design approach that combines an FCN, a Bi-LSTM, and an advanced attention framework.

4 Discussion

The FLA-SER model outperformed established baseline models, including traditional LSTMs, pure FCNs, and 3D-VGGs. It achieved an 'anger' recognition accuracy of 95.3% on RAVDESS. These results quantitatively validated the effectiveness of the hybrid FCN-LSTM architecture and the hierarchical attention framework, and the incremental accuracy shown in the ablation study (Table 3) further corroborates this effectiveness. The model's superior performance, as demonstrated by the SOTA comparison in Table 7, was primarily due to its hybrid AM and deeper FCN layers. These features effectively facilitated the fusion of spectral, spatial, and temporal features while mitigating the long-dependency issue. The model's robustness was demonstrated by its ability to maintain over 75% accuracy at 0 dB SNR across multiple domains (Table 5). This was consistent with the principles of robust control theory, which emphasized stability under system uncertainty.

The design principle underlying the combination of FCN-LSTM and hierarchical attention with the DMP-Transformer is analogous to the concept of adaptive and robust control in complex dynamical systems. More specifically, the AM operates as an adaptive gain scheduler, which is similar to the architecture employed in robust neural adaptive control systems for addressing uncertainties [21]. This mechanism enables the model to dynamically prioritize critical emotional frames, or high-value data points, thereby enhancing robustness and adaptability when applied to noisy or cross-domain speech signals. This is similar to adaptive backstepping control for uncertain nonlinear systems [22].

The model’s computational efficiency (Table 6) partially addresses practical deployment issues, such as performance on low-resource edge devices. Further discussion is warranted regarding real-time adaptation to emotional changes, especially the potential for improving the model’s online robustness by leveraging adaptive mechanisms inspired by control theory.

5 Conclusion

Aiming at the difficulties of long-distance temporal dependency mining and cross-modal feature fusion in human SER, the FLA-SER model was proposed. The architecture leveraged FCN for SF extraction, Bi-LSTM for temporal dependency modeling, and an AM for optimized feature fusion. The model outperformed comparable models, achieving 95.3% accuracy in recognizing anger on RAVDESS and an excellent long-sequence end segmentation accuracy of 92.1%. The robustness and efficiency were confirmed by maintaining over 75% accuracy at 0 dB SNR and achieving a processing time of only 14.2 ms for a single sample. Despite the strong performance, a limitation of the current work is the limited coverage of Asian language datasets and the absence of joint modeling with multimodal features. Future work will focus on increasing the linguistic diversity of the cross-language dataset and developing a robust multimodal fusion framework that incorporates facial and textual cues. This will enhance the model's overall generalization and practical value.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: All authors have read and approved the final version of the manuscript and consent to its publication.

Availability of data and material: The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests: All authors declare that they have no conflicts of interest.

Funding: None.

Authors' contributions: All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Mingjie Wang and Hexi Wang. The first draft of the manuscript was written by Mingjie Wang and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgements: None.

References

[1] Chen W, Xing X, Chen P, Xu X. Vesper: A compact and effective pretrained model for speech emotion recognition. IEEE Transactions on Affective Computing, 2024. https://doi.org/10.1109/TAFFC.2024.3369726

[2] Becker D, Braach L, Clasmeier L, Kaufmann T, Ong O, Ahrens K, Wermter S. Influence of Robots' Voice Naturalness on Trust and Compliance. ACM Transactions on Human-Robot Interaction, 2025, 14(2): 1-25. https://doi.org/10.1145/3706066

[3] Lin W C, Busso C. Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling. IEEE Transactions on Affective Computing, 2021, 14(2): 1215-1227. https://doi.org/10.1109/TAFFC.2021.3083821

[4] Falahzadeh M R, Farokhi F, Harimi A, Sabbaghi-Nadooshan R. Deep convolutional neural network and gray wolf optimization algorithm for speech emotion recognition. Circuits, Systems, and Signal Processing, 2023, 42(1): 449-492. https://doi.org/10.1007/s00034-022-02130-3

[5] Albadr M A A, Tiun S, Ayob M, AL-Dhief F T, Omar K, Maen M K. Speech emotion recognition using optimized genetic algorithm-extreme learning machine. Multimedia Tools and Applications, 2022, 81(17): 23963-23989. https://doi.org/10.1007/s11042-022-12747-w

[6] Chattopadhyay S, Dey A, Singh P K, et al. A feature selection model for speech emotion recognition using clustering-based population generation with hybrid of equilibrium optimizer and atom search optimization algorithm. Multimedia Tools and Applications, 2023, 82(7): 9693-9726. https://doi.org/10.1007/s11042-021-11839-3

[7] Kapoor S, Kumar T. Fusing traditionally extracted features with deep learned features from the speech spectrogram for anger and stress detection using convolution neural network. Multimedia Tools and Applications, 2022, 81(21): 31107-31128. https://doi.org/10.1007/s11042-022-12886-0

[8] Lui C F, Liu Y, Xie M. A supervised bidirectional long short-term memory network for data-driven dynamic soft sensor modeling. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-13. https://doi.org/10.1109/TIM.2022.3152856

[9] Gupta V, Juyal S, Hu Y C. Understanding human emotions through speech spectrograms using deep neural network. the Journal of Supercomputing, 2022, 78(5): 6944-6973. https://doi.org/10.1007/s11227-021-04124-5

[10] Tejaswini V, Sathya Babu K, Sahoo B. Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Transactions on Asian and Low-Resource Language Information Processing, 2024, 23(1): 1-20. https://doi.org/10.1145/3569580

[11] Yang L, Wang S, Chen X, Chen W, Saad O M, Chen Y. Deep-learning missing well-log prediction via long short-term memory network with attention-period mechanism. Geophysics, 2023, 88(1): D31-D48. https://doi.org/10.1190/geo2020-0749.1

[12] Warule P, Mishra S P, Deb S. Time-frequency analysis of speech signal using Chirplet transform for automatic diagnosis of Parkinson's disease. Biomedical Engineering Letters, 2023, 13(4): 613-623. https://doi.org/10.1007/s13534-023-00283-x

[13] Deb S, Dandapat S. Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification. IEEE transactions on cybernetics, 2018, 49(3): 802-815. https://doi.org/10.1109/TCYB.2017.2787717

[14] Mathews V, Youn D. Spectral leakage suppression properties of linear and quadratic windowing. IEEE transactions on acoustics, speech, and signal processing, 2003, 32(5): 1092-1095. https://doi.org/10.1109/TASSP.1984.1164418

[15] Ahmed G, Lawaye A A. CNN-based speech segments endpoints detection framework using short-time signal energy features. International Journal of Information Technology, 2023, 15(8): 4179-4191. https://doi.org/10.1007/s41870-023-01466-6

[16] Saxena D G, Farooqui A N, Ali S. Extricate features utilizing Mel frequency cepstral coefficient in automatic speech recognition system. Int. J. Eng. Manuf, 2022, 12(6): 14-21. https://doi.org/10.5815/ijem.2022.06.02

[17] Yang Y, Wang L, Gao S, Yu Z, Dong L. Cross-lingual speaker transfer for Cambodian based on feature disentangler and time-frequency attention adaptive normalization. International Journal of Web Information Systems, 2024, 20(2): 113-128. https://doi.org/10.1108/IJWIS-09-2023-0162

[18] Gupta A, Purwar A. Speech refinement using Bi-LSTM and improved spectral clustering in speaker diarization. Multimedia Tools and Applications, 2024, 83(18): 54433-54448. https://doi.org/10.1007/s11042-023-17017-x

[19] Liu C, Zhang Y, Sun J, Cui Z, Wang K. Stacked bidirectional LSTM RNN to evaluate the remaining useful life of supercapacitor. International Journal of Energy Research, 2022, 46(3): 3034-3043. https://doi.org/10.1002/er.7360

[20] Ren J, Xu D, Yang S, Zhao J, Li Z, Navasca C, Li D. Enabling large dynamic neural network training with learning-based memory management//2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA). IEEE, 2024: 788-802. https://doi.org/10.1109/HPCA57654.2024.00066

[21] Zouari F, Saad K B, Benrejeb M. Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems. International Review on Modelling and Simulations, 2012, 5(5): 2075-2103.

[22] Zouari F, Saad K B, Benrejeb M. Adaptive

backstepping control for a class of uncertain single input single output nonlinear systems[C]//10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13). IEEE, 2013: 1-6. https://doi.org/10.1109/SSD.2013.6564134