electronics

MDPI

Article

A Speech Emotion Recognition Model Combining WavLM Pre-Trained Features and Attention Mechanism

Xuhong Huang [1,2], Yinan Wang [1,*], Huiling Huang [3], Chaohui Zhou [1], Xuanyu Zhou [1] and Yongbin Chen [1]

[1] School of Electronic, Electrical Engineering and Physics, Fujian University of Technology, No. 69 Xuefu South Road, Minhou County, Fuzhou 350118, China; huangxuhong@fjut.edu.cn (X.H.); 2241905049@smail.fjut.edu.cn (C.Z.); 2241905033@smail.fjut.edu.cn (X.Z.); 2241908027@smail.fjut.edu.cn (Y.C.)
[2] Institute of Intelligent Ocean and Engineering, Fujian University of Technology, No. 69 Xuefu South Road, Minhou County, Fuzhou 350108, China
[3] Fujian Dongfang Ruizhi Information Technology Group Co., Ltd., Fuzhou 350108, China; 2231905072@smail.fjut.edu.cn
* Correspondence: 2241908028@smail.fjut.edu.cn ; Tel.: +86-188-1444-4164

Abstract

Speech emotion recognition is a key technology for achieving intelligent human–computer interaction. Traditional methods rely on manually designed features and struggle to fully capture the complex emotional information in speech. To address this issue, this paper proposes a novel speech emotion recognition model that integrates the WavLM pre-trained model with a multi-head attention mechanism. The model first utilizes the WavLM self-supervised pre-trained model to extract deep acoustic features. Then, a bidirectional long short-term memory network (BiLSTM) captures the temporal dependencies in the feature sequence. Finally, a multi-head attention mechanism is introduced to adaptively focus on emotionally salient time segments, generating a more discriminative emotional feature representation. Experimental results on two public emotion datasets, RAVDESS and EmoDB, show that the proposed model achieves an unweighted accuracy of 97.22% and a weighted accuracy of 96.88% on RAVDESS, and 96.06% unweighted accuracy and 95.33% weighted accuracy on EmoDB, significantly outperforming various baseline models, thus validating the model's effectiveness and generalization ability.

Keywords: speech emotion recognition; WavLM; bidirectional long short-term memory network; multi-head attention mechanism

Check for updates

Academic Editors: Krzysztof Szklanny and Piotr Wrzeciono

Received: 4 May 2026
Revised: 7 June 2026
Accepted: 10 June 2026
Published: 1 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Speech, as the most natural and efficient communication medium for humans, carries rich semantic and emotional information. Speech emotion recognition aims to automatically analyze speech signals to identify the speaker's emotional state, which is a key technology for achieving intelligent and harmonious human–computer interaction. It has broad application prospects in areas such as intelligent customer service, mental health assistance diagnosis, and in-vehicle emotion interaction [1]. Unlike multimodal emotion recognition, speech emotion recognition relies solely on audio signals, offering advantages such as easy data acquisition, low cost, and flexible deployment. Traditional speech emotion recognition methods mainly depend on hand-crafted acoustic features, such as Mel-frequency cepstral coefficients, fundamental frequency, and energy, combined with machine learning algorithms like support vector machines and hidden Markov models for classification [2]. The performance of such methods heavily relies on the quality of

feature engineering. However, manually designed features are not only cumbersome but also struggle to comprehensively and accurately capture the complex, high-dimensional, and dynamic emotional information embedded in speech signals, thereby limiting the generalization ability and recognition accuracy of the models [3].

In recent years, the emergence of deep learning techniques has brought breakthrough progress to speech emotion recognition. Convolutional neural networks, recurrent neural networks, and their variants (e.g., long short-term memory networks) have been widely adopted for automatically extracting emotional features [4]. In terms of temporal and key feature modeling, Ref. [5] proposed a speech emotion recognition framework that integrates a bidirectional long short-term memory network (BiLSTM) with an attention mechanism, effectively enhancing the model's ability to capture long-range temporal context and local emotional features. However, the performance of deep neural networks heavily depends on large-scale, high-quality annotated data, while the high cost of annotating emotional corpora results in limited data availability, which has become one of the bottlenecks restricting model performance [6]. To address the data scarcity problem, recent studies have conducted systematic explorations from multiple dimensions. A review of speech emotion recognition in noisy environments [7] revealed the significant impact of noise degradation on emotion recognition performance. For speech emotion recognition under limited-data or low-resource scenarios, Ref. [8] provided a systematic review and discussion of current deep learning approaches, offering theoretical guidance for developing efficient models in annotation-scarce scenarios. To more effectively utilize pre-trained features in low-resource scenarios, Ref. [9] systematically evaluated the adaptation performance of various self-supervised speech models on emotion recognition tasks, showing that with proper fine-tuning strategies, pre-trained models can effectively alleviate the overfitting risk caused by data scarcity.

To tackle data scarcity, self-supervised learning has gained rapid traction in the speech processing field. Pre-trained models such as Wav2Vec 2.0 [10], HuBERT [11], and WavLM [12] learn general-purpose speech representations by pre-training on massive amounts of unlabeled speech data. Among them, WavLM introduces a dual objective of joint denoising and masked prediction during pre-training, and innovatively employs a data augmentation strategy simulating overlapping speech, making it particularly effective in downstream speech tasks [12]. Recent studies have shown that, in Spanish speech emotion recognition, WavLM demonstrates significant advantages over other mainstream pre-trained models (e.g., Whisper, Wav2Vec 2.0, HuBERT) across multiple emotion databases, further validating its robust emotional representation in cross-language scenarios [13]. Using such pre-trained models as feature extractors provides a high-quality acoustic representation starting point for speech emotion recognition tasks with relatively scarce data [14]. However, simply feeding pre-trained features into a simple classifier still struggles to fully capture the complex temporal patterns of emotional evolution in speech signals. Emotional expression is often dominated by key speech segments (such as stress, pauses, and intonation shifts), which are unevenly distributed over time and require effective mechanisms to filter and focus on them [15].

To address the above issues, this paper proposes a speech emotion recognition model that integrates WavLM with a multi-head attention mechanism and BiLSTM. The model first employs WavLM to extract deep acoustic features, then uses BiLSTM to capture bidirectional temporal dependencies in the feature sequence, and finally introduces a multi-head attention mechanism [16] to adaptively assign weights to key emotional segments from different subspaces, thereby generating more discriminative emotional feature representations. The main contributions of this paper are as follows:

- Introducing the advanced WavLM self-supervised pre-trained model as a feature extractor to effectively alleviate the performance bottleneck caused by insufficient labeled emotion data;
- Constructing a "WavLM-BiLSTM-Multi-Head Attention" hybrid neural network architecture that synergistically utilizes the powerful representation ability of the pre-trained model, the temporal modeling advantage of BiLSTM, and the key information filtering capability of the attention mechanism;
- Designing a multi-head attention mechanism to weight the hidden states output by BiLSTM from multiple subspaces, enabling the model to adaptively focus on emotionally significant time segments;
- Conducting comparative experiments on the RAVDESS and EmoDB datasets to verify the effectiveness and generalization ability of the proposed model.

2. BiLSTM Speech Emotion Recognition Model Integrating WavLM and Multi-Head Attention Mechanism

2.1. Overall Framework

The overall framework of this paper is shown in Figure 1, mainly consisting of three modules: the WavLM feature extraction module, the BiLSTM temporal modeling module, and the multi-head attention emotion classification module. First, the raw speech data passes through the pre-trained WavLM model to extract high-dimensional acoustic features containing emotional information. Subsequently, these acoustic features are fed into a bidirectional long short-term memory network (BiLSTM) to capture the bidirectional temporal dependencies in the feature sequence and extract deep temporal emotional features. Finally, the hidden state sequence output by the BiLSTM is sent to the multi-head attention mechanism, which adaptively calculates importance weights for each time step from multiple representation subspaces, focusing on emotionally salient key segments to generate a more discriminative emotional feature representation. The final emotional category probability distribution is output through fully connected and Softmax layers. This architecture effectively integrates the powerful representation capability of WavLM, the temporal modeling advantage of BiLSTM, and the key information filtering ability of the multi-head attention mechanism, significantly improving the performance of speech emotion recognition.

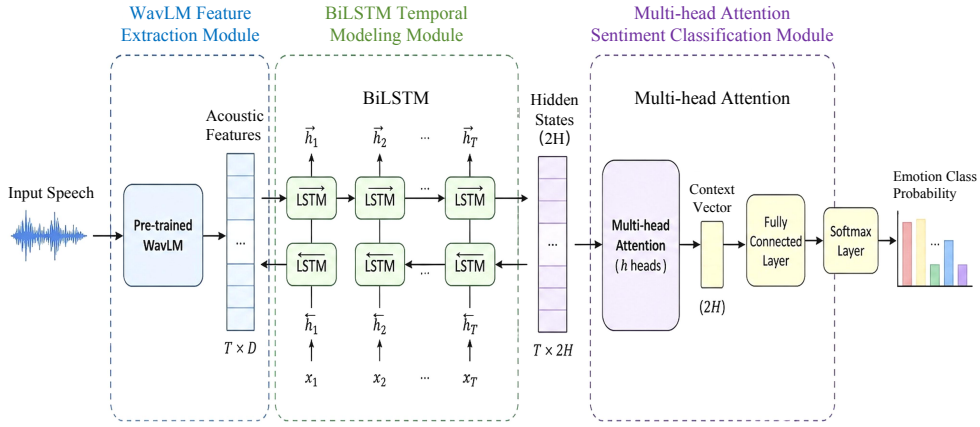


Figure 1. Speech emotion analysis framework.

2.2. WavLM Feature Extractor

This paper adopts the self-supervised pre-trained model WavLM [12] as the speech feature extractor and freezes all its parameters during subsequent training. WavLM is

a large-scale self-supervised speech pre-trained model proposed by Microsoft Research, designed to provide a unified acoustic representation for a wide range of speech processing tasks (e.g., recognition, separation, emotion analysis). The model consists of seven temporal convolutional blocks and L Transformer layers. The convolutional blocks map the raw waveform to frame-level features, while the Transformer layers capture long-range contextual dependencies.

Unlike common self-supervised models for automatic speech recognition (such as wav2vec 2.0 and HuBERT), WavLM adopts a dual optimization strategy of joint denoising and masked prediction during pre-training. Additionally, it innovatively introduces a data augmentation method that simulates overlapping speech. Specifically, during training, the model not only needs to predict masked speech segments but also needs to recover the original signal from noise-disturbed input. This design makes WavLM more robust to complex acoustic scenarios such as background noise and overlapping speech, making it particularly suitable for tasks like emotion recognition that require fine-grained capture of acoustic details.

The feature extraction process based on WavLM is shown in Figure 2. The raw speech input signal is denoted as $X \in \mathbb{R}^{1\times N}$, where N is the number of sampling points. First, the signal passes through a convolutional neural network (CNN) encoder to generate shallow two-dimensional speech encoding features $f_x = \{x_1, x_2, \ldots, x_T\} \in \mathbb{R}^{T\times 768}$, where T is the frame length (related to input duration and convolutional downsampling rate), and 768 is the feature dimension per frame. This paper uses the WavLM Base model (the large model has 1024 dimensions), balancing high performance with computational efficiency.

Subsequently, these low-level speech features f_x are input into a context network consisting of L Transformer layers, which extract high-level feature vectors layer by layer: $f_{h_l} = \{h_{l,1}, h_{l,2}, \ldots, h_{l,T}\} \in \mathbb{R}^{T\times 768}$, where $\{\, l = 1, 2, \ldots, L\}$. Each Transformer layer applies non-linear transformations to the output of the previous layer through a self-attention mechanism, enabling higher-layer features to contain richer global semantic information.

Existing studies have shown that low-level features output by the CNN encoder preserve good local acoustic structures, while high-level embeddings from the context network provide semantic information at different levels of abstraction. Weighted averaging of these two types of features can effectively improve downstream task performance [17], and subsequent studies have further validated the effectiveness of this strategy [18]. Based on this, this paper performs an equally weighted average of the 13 layers of feature vectors output by the context network (i.e., l = 1 to l = L, where L = 12, plus the output of the CNN encoder, totaling 13 layers) to obtain the final acoustic embedding representation:

$$H_X = \{H_1, H_2, \ldots, H_T\} = \frac{1}{13}\sum_{l=0}^{12} f_{h_l} \in \mathbb{R}^{T\times 768} \quad (1)$$

where f_{h_0} denotes the output of the CNN encoder. This cross-layer feature aggregation strategy alleviates the bias that a single top-layer feature might be overly biased toward the pre-training task and enhances the feature's sensitivity to emotion-related acoustic variations. The complete feature extraction process is shown in Figure 2.

It should be noted that during model training, we freeze all parameters of WavLM, i.e., no fine-tuning is performed. This is mainly based on two considerations: First, WavLM has been fully pre-trained on large-scale speech data (e.g., LibriLight, containing 60,000 h of unlabeled speech), and the general acoustic features it extracts already possess strong transferability for emotion recognition tasks. Second, speech emotion datasets are typically small (e.g., RAVDESS has only 1440 samples). Fine-tuning WavLM's huge parameter set (the base model has about 95 million parameters) at the same time would easily lead to overfitting. The freezing strategy allows the subsequent BiLSTM and attention modules

to be optimized on a high-quality fixed representation, achieving better generalization performance while ensuring training efficiency.

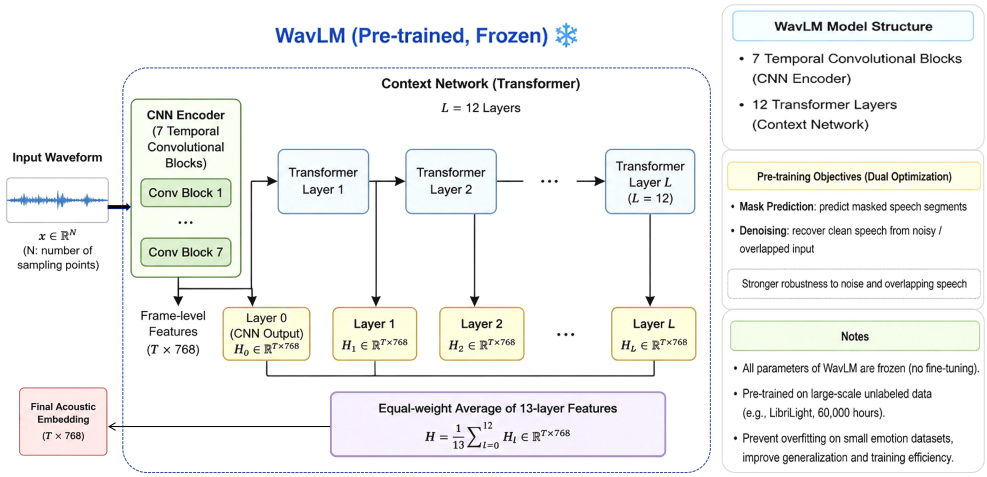


Figure 2. WavLM Feature extraction process.

2.3. BiLSTM Temporal Modeling Module

A key challenge in speech emotion recognition is that emotional information is often reflected through the dynamic changes of acoustic features over time, such as gradual rises in intonation, sudden occurrences of stress, or slow shifts in emotional state. These changes exhibit temporal dependencies of varying lengths that simple feedforward networks or traditional RNNs cannot effectively capture. Long short-term memory networks (LSTMs), through their sophisticated gating mechanisms, can selectively remember or forget historical information, effectively mitigating the vanishing or exploding gradient problem and excelling at handling long-term dependencies in long sequence data [19]. The computation process of a standard LSTM unit at each time step t is as follows: First, the forget gate f_t decides which information to discard from the previous cell state c_{t-1}; the input gate i_t and candidate cell state $\tilde{c}_t$decide which new information to store in the cell state; and the output gate o_t computes the hidden state h_t based on the current cell state c_t. The specific formulas are as follows:

$$\begin{cases} f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \\ i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \\ \tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \\ c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \\ h_t = o_t \odot \tanh(c_t) \end{cases} \quad (2)$$

where x_t is the input feature at time step t (the t-th frame acoustic embedding extracted by WavLM), h_{t-1} is the previous hidden state, W and b are weight matrices and biases, σ is the sigmoid function, and $\odot$ denotes element-wise multiplication. Through this gating mechanism, LSTM can flexibly retain or forget information at different time scales, which is well-suited for capturing the long-term dependencies in the evolution trajectory of emotions.

However, a standard LSTM only utilizes information from the past (unidirectional modeling). In actual speech emotion expression, the emotional state may be influenced not only by past speech segments but also by future changes in intonation, stress positions, or pauses. For example, a strong rising intonation at the end of a sentence might indicate

surprise or anger, information that can only be fully captured through "future" speech segments. Therefore, this paper employs a bidirectional LSTM (BiLSTM), introducing two independent processing channels: a forward LSTM and a backward LSTM.

Specifically, the feature sequence extracted by WavLM $X = \{x_1, x_2, \ldots, x_T\}$ is fed into two LSTM networks in normal and reverse order, respectively. The forward LSTM processes $x_1 \rightarrow x_T$ sequentially, generating the forward hidden state sequence $\overrightarrow{h_t}$; the backward LSTM processes from x_T to x_1, generating the backward hidden state sequence $\overleftarrow{h_t}$. At each time step t, the hidden states from both directions are concatenated to obtain the comprehensive hidden state:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{3}$$

Thus, after BiLSTM modeling, the entire feature sequence outputs the hidden state sequence $H = \{h_1, h_2, \ldots, h_T\}$ as

$$H = \text{BiLSTM}(X) \tag{4}$$

where $h_t \in \mathbb{R}^{2d}$, and d is the hidden dimensionality of the unidirectional LSTM. This bidirectional structure allows the output at each time step to contain both the emotional evolution information before that moment and the emotional trajectory after it, enabling a more comprehensive and accurate representation of emotional dynamics.

In our model configuration, the number of BiLSTM layers is set to 2. The first BiLSTM layer receives the 768-dimensional acoustic feature sequence output by WavLM and maps it to a 256-dimensional hidden space (256 dimensions per direction, concatenated to 512 dimensions); the second BiLSTM layer takes the output of the first layer as input to further extract higher-level temporal emotion patterns, with a final output dimension also of 512. A dropout rate of 0.3 is introduced between the two layers to prevent overfitting. Through this two-layer stacked structure, the model learns temporal dependencies of emotional features at different levels of temporal abstraction, preserving both local acoustic details and capturing emotional evolution trajectories across multiple speech segments.

Compared to unidirectional LSTM or simple RNN, BiLSTM has significant advantages in speech emotion recognition tasks. On one hand, it fully utilizes contextual information, making it more sensitive to key events such as emotional turning points and sudden intonation changes. On the other hand, through the compensatory effect of the backward channel, even if the forward channel is disturbed by noise or information ambiguity at a certain time step, the backward information can still provide effective constraints, enhancing feature robustness. Therefore, feeding the deep acoustic features extracted by WavLM into a BiLSTM network generates emotion representations rich in temporal context, laying a solid foundation for the subsequent multi-head attention mechanism.

2.4. Multi-Head Attention

In the hidden state sequence H output by BiLSTM, different time segments contribute differently to emotion recognition. To enable the model to adaptively focus on emotionally salient time segments, we introduce a multi-head attention mechanism, which computes importance weights for each time step simultaneously from multiple representation subspaces.

For each attention head i ($i = 1, 2, \ldots, n_h$, n_h is the total number of attention heads, set to 8 in this paper), each hidden state h_t is linearly transformed into query Q_i, key K_i, and value V_i:

$$Q_i = HW_i^Q, \quad K_i = HW_i^K, \quad V_i = HW_i^V \tag{5}$$

where W_i^Q, W_i^K, W_i^V are the weight matrices for the i-th head.

The attention output of the i-th head is calculated as scaled dot-product attention:

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^{\mathsf{T}}}{\sqrt{d_k}}\right) V_i \tag{6}$$

where d_k is the dimension of the key vectors, which scales the dot product; the softmax function normalizes the attention weights.

The outputs of multiple heads are concatenated and then passed through a linear transformation to obtain the final multi-head attention output:

$$\text{MultiHead}(H) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_{n_h}) W_{\text{proj}} \tag{7}$$

where W_{proj} is the output linear transformation matrix, and Concat denotes vector concatenation. Denote the multi-head attention output as $A = \text{MultiHead}(H)$.

A is fed into a fully connected layer, and the emotion class probabilities are output via the softmax function:

$$p = \text{softmax}(W_{fc} A + b_{fc}) \tag{8}$$

where W_{fc} and b_{fc} are the weight and bias of the fully connected layer; p is the predicted probability distribution.

The processing flow of the proposed model can be summarized as follows: first, preprocess the raw audio (resampling, silence removal, energy normalization) and extract features using WavLM; second, process the features with BiLSTM to capture long-term dependencies; third, introduce multi-head attention to focus on important temporal features; finally, output the emotion category through fully connected layers and a Softmax activation function.

The main characteristics of this model are:

- End-to-end learning: WavLM directly extracts features from pre-processed audio, avoiding cumbersome manual feature design, allowing the model to automatically learn deep emotion-related patterns from raw signals [20].
- Effective long-term dependency modeling: BiLSTM, through its memory cells and gating mechanisms, effectively handles long-term dependencies in speech signals, overcoming information loss in traditional RNNs, thus accurately capturing the evolution trajectory of emotions.
- Highlighting key emotional information: The multi-head attention mechanism adaptively assigns weights to different time segments, selectively emphasizing the parts most important for emotion recognition, thereby improving accuracy and robustness in complex emotional contexts [21].

3. Experiments and Result Analysis

3.1. Experimental Datasets

We perform experimental validation on two public speech emotion datasets:

- The RAVDESS dataset [22], recorded by Ryerson University, contains speech and songs performed by 24 professional actors (12 male, 12 female), covering eight emotion categories: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. We use the speech portion, totaling 1440 samples.
- The EmoDB dataset [23], recorded by the Technical University of Berlin, contains German speech performed by 10 actors (5 male, 5 female), covering seven emotion categories: anger, boredom, disgust, fear, happiness, neutral, and sadness, totaling 535 samples.

3.2. Data Preprocessing and Model Configuration

Data Preprocessing: All audio was uniformly resampled to 16 kHz, leading and trailing silence was removed, and energy normalization was applied. To keep input dimensions consistent, audio was truncated or padded to a fixed length (maximum duration 5 s). The dataset was divided into training, validation, and test sets in an 8:1:1 ratio, ensuring that samples from the same speaker did not appear across different sets.

Model Configuration: The WavLM feature extractor uses the WavLM Base model (from HuggingFace transformers library, version 4.30.0) with frozen pre-trained parameters and an output dimension of 768. PyTorch version 1.12.1 was used for implementation. The BiLSTM is set to two layers, with a hidden dimension of 256 and a dropout rate of 0.3. The configuration of stacking two BiLSTM layers follows the mainstream design philosophy in the field of speech emotion recognition; related studies indicate that a two-layer BiLSTM effectively avoids overfitting while capturing bidirectional temporal dependencies [24,25]. The choice of a hidden dimension of 256 has been experimentally validated to strike a good balance between feature representation capability and model complexity [24]. The multi-head attention mechanism uses eight heads, with a key/value dimension of 32. The selection of the number of attention heads refers to the recommendations from the original Transformer paper [16] as well as recent empirical studies in speech emotion recognition, where the eight-head configuration achieved optimal performance in terms of unweighted accuracy [26].

Training Configuration: The initial learning rate was set to 1×10^{-4}, adjusted dynamically using a cosine annealing learning rate scheduler; batch size was set to 16; and training was conducted for 100 epochs with early stopping: if the validation unweighted accuracy did not improve by more than 0.1% for 10 consecutive epochs, training was stopped. The optimizer was AdamW with a weight decay coefficient of 1×10^{-4}. The loss function was cross-entropy loss.

Evaluation Metrics: To comprehensively evaluate model performance, we use weighted accuracy (WA) and unweighted accuracy (UA). WA is the weighted average accuracy across different emotion categories (weights related to the sample size of each category), while UA is the arithmetic mean of accuracies across categories. These two metrics can more objectively evaluate the model's performance in the presence of sample imbalance. The specific calculation methods are

$$Acc_i = \frac{TP_i}{TP_i + FP_i} \tag{9}$$

$$WA = \frac{\sum_{i=1}^{C} N_i \times Acc_i}{\sum_{i=1}^{C} N_i} \tag{10}$$

$$UA = \frac{1}{C} \sum_{i=1}^{C} Acc_i \tag{11}$$

where C is the number of emotion categories, and N_i is the number of samples in category i.

3.3. Confusion Matrix Analysis

To deeply reveal the recognition differences of the proposed model for different emotion categories, we plotted confusion matrices on the RAVDESS and EmoDB test sets (Figures 3 and 4). The unweighted accuracy (UA) and weighted accuracy (WA) calculated from the confusion matrices are as follows: on RAVDESS, UA = 97.22%, WA = 96.88%, and on EmoDB, UA = 96.06%, WA = 95.33%. The UA and WA reported in this paper are obtained from a single optimal model evaluated on the test set, where the model selects the epoch with the best performance on the validation set based on an early stopping

mechanism. To verify the stability of the results, we additionally conduct five independent repeated experiments (each with different random initializations but fixed data splits). On the RAVDESS dataset, the mean unweighted accuracy (UA) is 96.94% with a standard deviation of 0.28%; on the EmoDB dataset, the mean UA is 95.87% with a standard deviation of 0.35%. These results demonstrate that the performance of our proposed model is robust, and the reported single optimal result is not coincidentally due to a favorable data split.

From the RAVDESS confusion matrix, the model accurately recognizes most emotion categories, with "neutral", "calm", and "surprised" achieving near-perfect classification. The main confusions occur between "angry" and "fearful", and between "happy" and "surprised". This is because anger and fear may both exhibit a high pitch and a fast speaking rate, while happiness and surprise often have a high fundamental frequency and rising energy, leading to some acoustic overlap. Nevertheless, due to the multi-head attention mechanism, the model effectively focuses on local time segments where emotional differences are most significant (e.g., stress, sudden rate changes), so these confusions remain at very low levels and do not significantly affect overall accuracy.

On the EmoDB dataset, the confusion matrix also shows excellent class discrimination ability. "Sad", "boredom", and "neutral"—three low-arousal emotions—achieve the highest recognition accuracy, indicating the model can stably capture acoustic patterns with low energy and flat intonation. The main confusions occur between "anger" and "disgust", and between "fear" and "happiness". Analyzing the reasons, anger in EmoDB often has obvious glottalization and strong energy, while disgust sometimes also has a similar rough voice quality; fear and happiness may both have local high pitch variations. However, thanks to BiLSTM's bidirectional modeling of forward and backward temporal dependencies, the model can use long-range emotional context to resolve ambiguities caused by local acoustic similarities, thereby limiting the above confusions to a small range.

To further analyze the sources of misclassification, this paper examines the acoustic features of error cases with relatively high proportions in the confusion matrix. On the RAVDESS dataset, the misclassification between "anger" and "fear" mainly stems from the fact that both emotions exhibit a relatively high mean fundamental frequency, fast speech rate, and possible sudden energy bursts. The confusion between "happiness" and "surprise" is related to both emotions often being accompanied by a high fundamental frequency and energy elevation. On the EmoDB dataset, the confusion between "anger" and "disgust" is often associated with overlapping acoustic features such as glottalization and rough voice quality; the confusion between "fear" and "happiness" may be attributed to locally high pitch variations. Nevertheless, thanks to the robust acoustic representations of WavLM and the bidirectional contextual modeling of BiLSTM, these confusions account for a very small proportion of the test set and do not significantly affect the overall accuracy.

Summarizing the confusion matrix results, we can draw the following conclusions: (1) The proposed model exhibits high discriminative power for different emotion categories, maintaining low confusion even for acoustically similar emotion pairs. (2) The robust acoustic representations provided by WavLM pre-trained features effectively mitigate interference from cross-dataset differences in emotional expression. (3) The synergistic mechanism of multi-head attention and BiLSTM enables the model to ignore incidental acoustic similarities and focus on temporal patterns that truly reflect emotional category differences. This further confirms the superiority of our architecture in fine-grained emotion classification tasks.

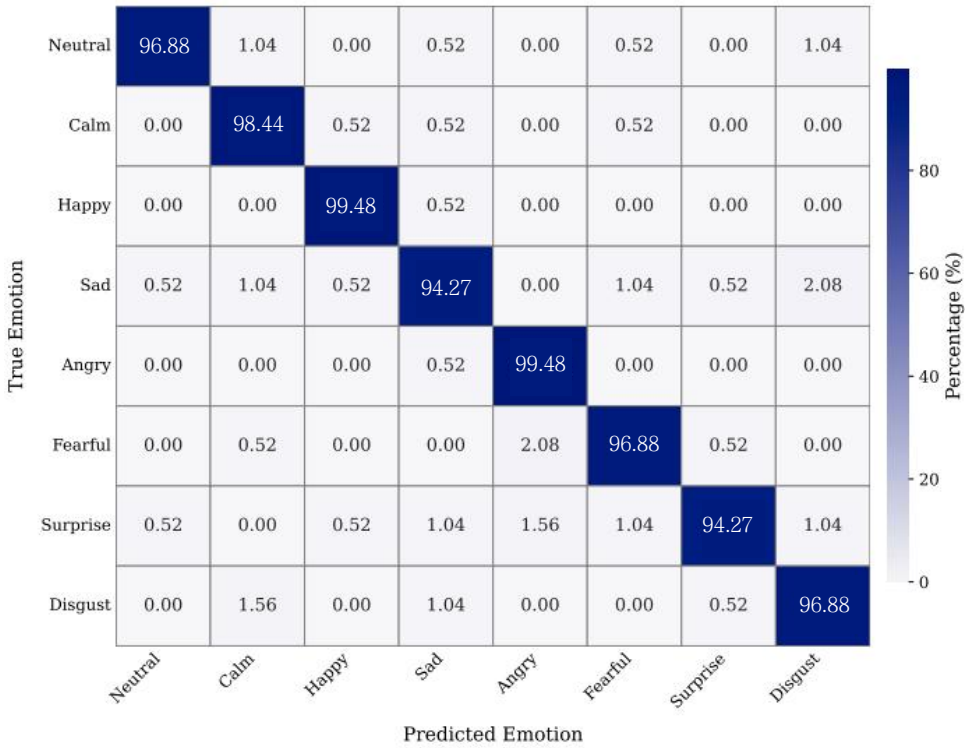


Figure 3. Confusion matrix of the proposed model on the RAVDESS test set.

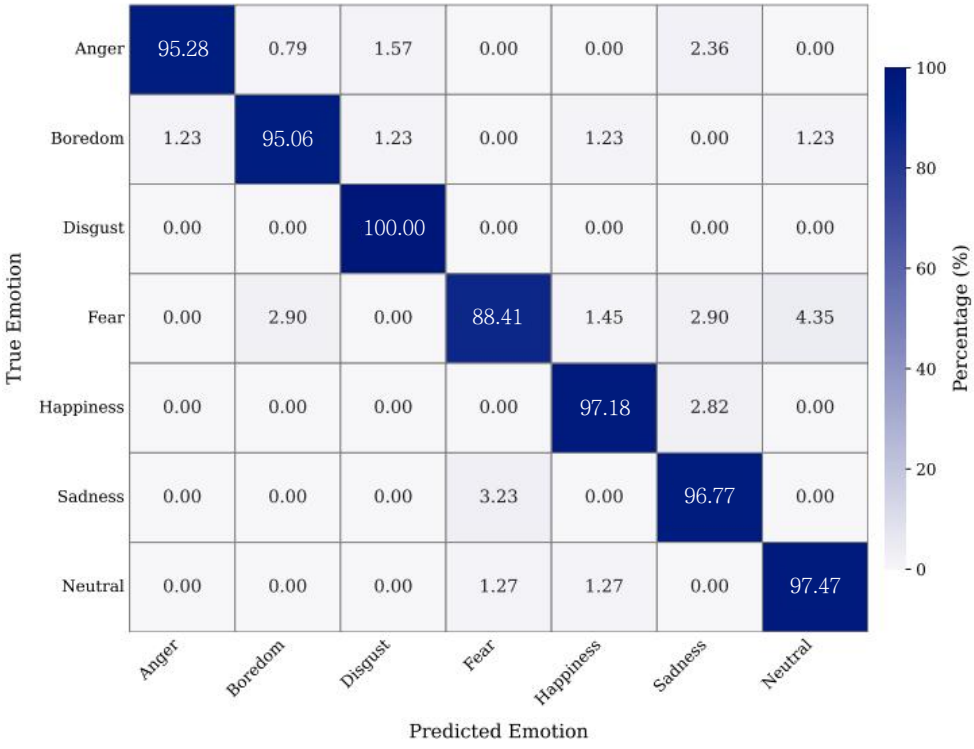


Figure 4. Confusion matrix of the proposed model on the EmoDB test set.

3.4. Ablation Experiments

To further validate the contribution of each component in our proposed model to speech emotion recognition performance, we conduct ablation experiments on the RAVDESS and EmoDB datasets.

3.4.1. Experimental Setup

The ablation experiments follow the same training configuration and evaluation metrics as the main experiments. The evaluation metrics include unweighted average accuracy (UA) and weighted average accuracy (WA). UA reflects the model's average recognition capability across different emotion categories, while WA accounts for the imbalance in sample sizes among categories and is therefore more suitable for imbalanced datasets. Four model variants are compared: (1) the baseline using only the WavLM feature extractor; (2) WavLM followed by a BiLSTM layer (WavLM_BiLSTM); (3) WavLM followed by a multi-head attention layer (WavLM_MHA); and (4) our proposed WavLM_BiLSTM_MHA, which serially fuses both modules.

3.4.2. Result Analysis

Table 1 presents the detailed results of the four model variants on both datasets.

Table 1. Ablation experiment results (%) of different model variants on the RAVDESS and EmoDB datasets.

Model Variant	RAVDESS (UA)	RAVDESS (WA)	EmoDB (UA)	EmoDB (WA)
WavLM (baseline)	76.23	76.39	83.35	83.18
WavLM_BiLSTM	90.54	90.88	89.27	88.03
WavLM_MHA	93.33	92.58	91.35	92.14
WavLM_BiLSTM_MHA(ours)	97.22	96.88	96.06	95.33

The following key observations can be drawn from Table 1:

- Compared with the baseline, any variant that introduces an additional temporal processing module yields a significant performance improvement, with UA increasing by 10–17%.
- On both datasets, using MHA alone consistently outperforms using BiLSTM alone. For example, on RAVDESS, WavLM_MHA achieves a UA that is 2.79% higher than that of WavLM_BiLSTM; on EmoDB, it outperforms WavLM_BiLSTM by 2.08% in UA and by 4.11% in WA.
- The proposed WavLM_BiLSTM_MHA achieves the best results on both datasets, improving UA by 3.89% over the single-MHA variant on RAVDESS and by 4.71% on EmoDB.

3.4.3. Discussion

The above results reveal three important findings.

(1) Critical role of temporal and attention processing for self-supervised features. Although the baseline WavLM extracts rich general acoustic representations, it lacks specialized modeling of temporal emotion dynamics, resulting in limited direct classification performance (UA of only 76.23% on RAVDESS). The introduction of either BiLSTM or MHA boosts performance by more than ten percentage points, demonstrating that self-supervised features must be combined with downstream temporal refinement to fully unleash their potential for emotion recognition.

(2) Superiority of multi-head attention over traditional BiLSTM.BiLSTM is constrained by its recurrent structure, which tends to cause long-range feature smoothing or gradient dilution, making it difficult to accurately highlight local emotional burst points. In contrast, multi-head attention can directly associate any two frames globally, thereby efficiently identifying key frames with high emotional discriminability. The experimental results consistently validate the superiority of MHA.

(3) Optimal performance through synergistic fusion of both modules BiLSTM and MHA are not simply redundant but exhibit strong complementarity and synergy. Specifically, BiLSTM first performs bidirectional temporal smoothing to capture locally continuous speech evolution features; MHA then performs global weighting on this basis to adaptively select the most emotionally salient frames. This dual-layer architecture—"local temporal smoothing followed by global key-frame focusing"—achieves optimal performance on both datasets, confirming the effectiveness of the serial design.

3.5. Attention Mechanism Visualization Analysis

To intuitively demonstrate the selection capability of the multi-head attention mechanism in temporal context modeling within our proposed model, we extract the attention weight distributions for different emotion categories from the model on the EmoDB and RAVDESS test sets. These weights are then mapped onto a physical time axis (0–5 s, consisting of 248 frames). The visualization results are shown in Figure 5 and Figure 6, respectively.

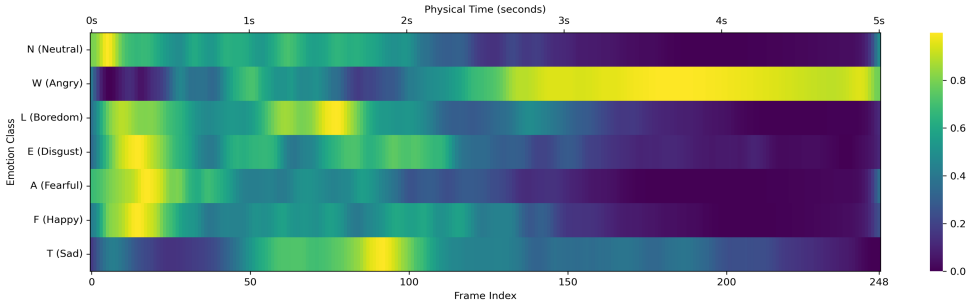


Figure 5. Attention distribution of seven emotion categories on the EmoDB dataset.

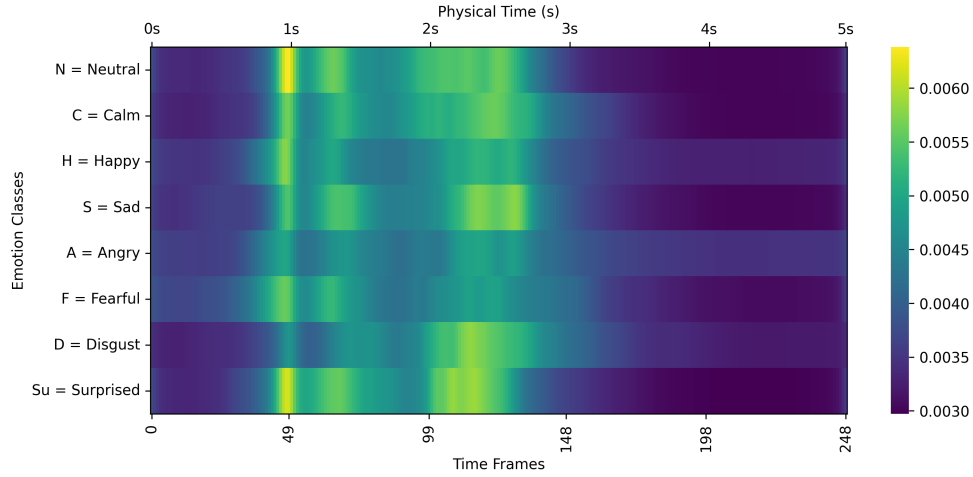


Figure 6. Attention distribution of eight emotion categories on the RAVDESS dataset.

3.5.1. Attention Distribution Analysis on the EmoDB Dataset

Figure 5 shows the attention distributions for the seven emotion categories on the EmoDB dataset. It can be observed that the multi-head attention mechanism exhibits significant differentiated focusing characteristics for different emotions:

- **High-arousal emotions (e.g., anger):** The attention shows extremely strong sustained focusing in the latter half of the time axis (2.5–5 s), with weights close to 1.0. This indicates that when recognizing anger, the model heavily relies on the segments where emotional bursts and sustained acoustic energy accumulation occur in the later part of the audio.

- **Low-arousal emotions (e.g., boredom, sadness):** Their attention peaks precisely lock onto the middle segment of the audio (approximately 1.5–2.0 s), which is highly consistent with the temporal characteristics of these emotions, such as slower delivery, a reduced speech rate, and gradual pitch decay in the middle part.
- **Neutral emotion:** Attention is highly concentrated at the very beginning of the audio (0.1–0.3 s), suggesting that the model tends to quickly establish baseline acoustic features at the onset of speech to determine the neutral state.

3.5.2. Attention Distribution Analysis on the RAVDESS Dataset

Figure 6 presents the attention distributions for the eight emotion categories on the RAVDESS dataset. Compared with EmoDB, the attention distributions on RAVDESS exhibit strong temporal alignment consistency:

- For all emotion categories, an extremely narrow and highly intense bright-yellow peak appears around 1.0 s (approximately frame 49), followed by a secondary high wide-peak region between 2.0 and 2.5 s.
- This phenomenon is closely related to the recording protocol of the RAVDESS dataset, which uses standardized fixed texts (e.g., "Kids are talking by the door" or "Dogs are sitting by the door"). The strong attention focus at 1.0 s precisely corresponds to the onset of the core verb (e.g., "talking", "sitting") in the sentence, while the secondary high focus at 2.0–2.5 s corresponds to the prosodic change point of the sentence-final noun. This strongly demonstrates that the model not only captures acoustic cues of emotion but also automatically learns and aligns key semantic and prosodic turning points in the standardized text, thereby extracting the most discriminative emotional expressions from highly scripted utterances.

3.5.3. Summary

In summary, the visualization results indicate that our proposed "WavLM–BiLSTM–Multi-Head Attention" architecture can effectively overcome the shortcomings of traditional temporal models (e.g., single LSTM or GRU), such as easy forgetting and lack of prioritization. By adaptively allocating temporal weights across multiple subspaces, the multi-head attention mechanism successfully achieves the original design goal of "suppressing irrelevant background while focusing on emotionally salient frames." This constitutes the underlying mechanism for the leading performance of our method on both public datasets.

3.6. Comparison with State-of-the-Art Methods

To verify the effectiveness of the proposed model, we compared it with various baseline models on the RAVDESS and EmoDB datasets. The results are shown in Table 2.

Table 2. Recognition accuracy of different methods on RAVDESS and EmoDB datasets (unit: %).

Dataset	Method	Publication Year	Feature Type	UA	WA
RAVDESS	GM-TCNet [27]	2022	MFCC (hand-crafted)	90.48	90.28
	WavLM-Transformer [28]	2024	WavLM (SSL)	92.45	91.80
	AFT-Net [29]	2024	Hybrid features (SSL)	94.10	93.60
	RNN Temporal Contrastive Model [30]	2023	Wav2vec 2.0 (SSL)	88.50	87.90
	Ours	**This work**	**WavLM (SSL)**	**97.22**	**96.88**

Table 2. *Cont.*

Dataset	Method	Publication Year	Feature Type	UA	WA
EmoDB	GM-TCNet [27]	2022	MFCC (hand-crafted)	95.33	95.33
	WavLM-Transformer [28]	2024	WavLM (SSL)	95.50	95.12
	AFT-Net [29]	2024	Hybrid features (SSL)	95.80	95.20
	RNN Temporal Contrastive Model [30]	2023	Wav2vec 2.0 (SSL)	94.80	94.50
	Ours	**This work**	**WavLM (SSL)**	**96.06**	**95.33**

To comprehensively evaluate the performance of our proposed model—the BiLSTM architecture integrating WavLM and multi-head attention (WavLM–BiLSTM–Multi-Head Attention)—we systematically compare it with a range of existing state-of-the-art methods on two public datasets, RAVDESS and EmoDB, using unweighted accuracy (UA) and weighted accuracy (WA) as evaluation metrics. The results are presented in Table 2. The compared methods include classical temporal models based on hand-crafted features (GM-TCNet [27]), different architectures based on self-supervised learning (SSL) features (WavLM-Transformer [28], AFT-Net [29]), and recent advanced works (RNN temporal contrastive model [30]).

3.6.1. Comparison on the RAVDESS Dataset

The upper part of Table 2 shows the comparison results on the RAVDESS dataset. Our proposed method achieves 97.22% in UA and 96.88% in WA, outperforming all other advanced methods listed in the table. The detailed analysis is as follows:

- **Compared with traditional hand-crafted feature models:** Compared with GM-TCNet [27], which uses traditional hand-crafted MFCC features, our method yields substantial improvements of 6.74 percentage points in UA and 6.60 percentage points in WA. This indicates that self-supervised WavLM features provide deeper and more robust acoustic representations than traditional hand-crafted features, thereby significantly enhancing the model's ability to capture complex emotional signals.
- **Compared with similar pre-trained feature models:** Compared with WavLM-Transformer [28], which also employs WavLM features, our model achieves improvements of 4.77 percentage points in UA and 5.08 percentage points in WA. This comparison strongly demonstrates that, after extracting powerful self-supervised features, our designed "BiLSTM + multi-head attention" architecture is superior to a plain Transformer structure in capturing bidirectional temporal context dependencies and adaptively focusing on emotionally salient frames.
- **Compared with hybrid features and recent temporal models:**
  - Compared with AFT-Net [29], which uses hybrid self-supervised features, our method outperforms it by 3.12 percentage points in UA and 3.28 percentage points in WA.
  - Compared with the RNN temporal contrastive model based on Wav2vec 2.0 features proposed in 2023 [30], our method shows a clear advantage, achieving 8.72 and 8.98 percentage points higher in UA and WA, respectively. These results fully confirm the outstanding performance of the combination of WavLM features and our refined temporal attention architecture on the RAVDESS multimodal acted dataset.

3.6.2. Comparison on the EmoDB Dataset

The lower part of Table 2 presents the comparison results on the EmoDB dataset. Our method achieves 96.06% UA and 95.33% WA. The detailed analysis is as follows:

- **Compared with traditional hand-crafted feature models:** Compared with GM-TCNet [27], our method improves UA by 0.73 percentage points and achieves the same WA (95.33%). This indicates that on the smaller, more focused German EmoDB dataset, our method still maintains a very high recognition upper bound.
- **Compared with similar pre-trained feature models:** Compared with WavLM-Transformer [28], which is also based on WavLM, our method yields improvements of 0.56 percentage points in UA and 0.21 percentage points in WA, further validating the synergistic gain of combining BiLSTM and multi-head attention for downstream sequence modeling.
- **Compared with hybrid features and recent temporal models:**
  - Compared with AFT-Net [29], our method improves UA by 0.26 percentage points and WA by 0.13 percentage points.
  - Compared with the RNN temporal contrastive model based on Wav2vec 2.0 features proposed in 2023 [30], our method achieves a notable improvement of 1.26 percentage points in UA and 0.83 percentage points in WA. These results demonstrate that even on the small EmoDB dataset, which is prone to overfitting, our proposed method exhibits strong robustness and generalization ability.

3.6.3. Summary

In summary, on both the English RAVDESS dataset and the German EmoDB dataset, our proposed WavLM–BiLSTM–multi-head attention architecture achieves state-of-the-art (SOTA) comprehensive performance compared with several recent advanced methods. This not only verifies the excellence of WavLM as a self-supervised feature extractor but also further confirms that bidirectional temporal modeling combined with multi-head attention key-frame selection can maximize the emotional expression potential embedded in self-supervised features.

4. Conclusions

Aiming at the problems in speech emotion recognition—the limited representational capacity of traditional hand-crafted features, the overfitting of deep models on small-scale emotional corpora, and the difficulty of accurately locating key emotional information—this paper proposes integrating the WavLM pre-trained model with a multi-head attention mechanism in a BiLSTM network. The framework fully leverages the general acoustic knowledge obtained by WavLM through pre-training on massive unlabeled speech data, effectively alleviating the constraints caused by scarce annotated emotion samples. It combines BiLSTM's advantage in long-range bidirectional temporal context modeling to accurately characterize the dynamic evolution of emotional states over time. Furthermore, the multi-head attention mechanism adaptively focuses on key speech segments decisive for emotion discrimination from multiple representation subspaces, significantly improving feature discriminability and robustness.

Systematic experimental results on the public RAVDESS and EmoDB datasets demonstrate that our proposed model achieves remarkable recognition performance. On the RAVDESS dataset, it attains an unweighted accuracy (UA) of 97.22% and a weighted accuracy (WA) of 96.88%; on the EmoDB dataset, it achieves a UA of 96.06% and a WA of 95.33%. The ablation study (Table 1) verifies the effectiveness of each module: using only the WavLM baseline yields a UA of only 76.23% (on RAVDESS); introducing BiLSTM increases this to 90.54%; and further incorporating multi-head attention boosts it to 97.22%, demonstrating the synergistic gain between BiLSTM and multi-head attention.

The comparison experiments (Table 2) show that our method achieves state-of-the-art results against recent SOTA methods including WavLM-Transformer (2024), AFT-Net (2024),

and the RNN temporal contrastive model [30] (2023), fully validating the advancement and generalization ability of the proposed architecture. It should be noted that experimental protocols may differ across studies, so absolute numerical comparisons across papers have certain limitations. In our comparative analysis, we prioritize works that use identical or similar protocols and emphasize the consistent performance advantage across multiple datasets. Future work will conduct fair comparisons with more methods under a unified benchmark framework.

Our method confirms the great potential of combining large-scale self-supervised pre-training techniques with refined temporal attention architectures for speech emotion recognition. Subsequent research will pursue the following directions: (1) exploring lightweight model designs to reduce computational cost for edge-side real-time deployment; (2) incorporating multimodal information such as text and facial expressions to build cross-modal emotion perception models; (3) investigating adaptive aggregation strategies for WavLM layers based on learnable weights; (4) evaluating model generalization on natural interaction scenario datasets such as IEMOCAP and MSP-Podcast; (5) studying the application of lightweight adaptation methods such as Adapter and LoRA to WavLM to enhance domain adaptability while maintaining generalization; and (6) introducing statistical significance tests (e.g., bootstrap confidence intervals or paired *t*-tests) to more rigorously assess performance differences between the model and baselines.

Author Contributions: Methodology, Y.W.; Software, Y.W.; Resources, H.H.; Writing—original draft, Y.W.; Writing—review & editing, Y.W.; Supervision, X.H. C.Z., X.Z. and Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the project "Key Technology Research of Intelligent Wheelchair Robot for Elderly Care" (grant number M820726005). The APC was funded by the authors.

Data Availability Statement: The RAVDESS and EmoDB datasets are publicly available. The code used in this study is available from the corresponding author upon reasonable request.

Conflicts of Interest: Author Huiling Huang was employed by the company Fujian Dongfang Ruizhi Information Technology Group Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. George, S.M.; Ilyas, P.M. A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. *Neurocomputing* **2024**, *568*, 127015.
2. Tao, J.H.; Chen, J.J.; Li, Y.W. A review of speech emotion recognition. *Signal Process.* **2023**, *39*, 571–587. (In Chinese)
3. Li, W.K.; Zhang, X.F.; Jia, X.B.; Wang, J.; Liu, Y.; Zhang, S.; Chen, L.H.; Zhao, Y.; Singh, A.K.; Wang, H. ; et al. A survey on deep learning for speech emotion recognition: Models, databases, and trends. *IEEE Trans. Affect. Comput.* **2024**, *15*, 1129–1148.
4. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Qadir, J.; Schuller, B.W. Speech emotion recognition using deep neural networks: A systematic review. *ACM Comput. Surv.* **2023**, *56*, 69 .
5. Zhang, W.; Li, H.; Yang, H. Tibetan-speech-emotion recognition under low-resource conditions. *J. Signal Process.* **2025**, *41*, 1558–1569. [CrossRef]
6. Zou, H.; Si, Y.; Chen, C.; Rajan, D.; Chng, E.S. Speech emotion recognition with co-attention based multi-level acoustic information. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Singapore, 2022; pp. 7367–7371.
7. Mary, D.S.D.; Krishna, M.S.; Kumari, A.B.S. A systematic review on noisy speech emotion recognition: Approaches, databases, and challenges. *Artif. Intell. Rev.* **2024**, *57*, 52.
8. Latif, S.; Alvi, A.; Khalid, U.; Al-Asam, R.; Schuller, B.W. Deep learning for speech emotion recognition in low-resource languages: A systematic review. *Comput. Speech Lang.* **2023**, *80*, 101489.
9. Wang, Y.; Chen, S.; Liu, S.; Chen, Z.; Yu, K. Systematic evaluation of self-supervised speech models for emotion recognition. In *Proceedings of Interspeech 2023*; ISCA: Dublin, Ireland, 2023; pp. 3122–3126.

10. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*; Curran Associates, Inc.: Vancouver, BC, Canada, 2020; pp. 12449–12460.
11. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460.
12. Chen, S.; Wang, C.; Chen, Z.; Wang, Y.; Zhang, S.; Huang, G.; Li, J. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518.
13. Mares, A.; Diaz-Arango, G.; Perez-Jacome-Friscione, J.; Vazquez-Leal, H.; Hernandez-Martinez, L.; Huerta-Chua, J.; Jaramillo-Alvarado, A.F.; Dominguez-Chavez, A. Advancing Spanish speech emotion recognition: A comprehensive benchmark of pre-trained models. *Appl. Sci.* **2025**, *15*, 4340. [CrossRef]
14. Liu, T.; Zhang, L.; Wang, H.; Li, J.; Liu, Y. Feature fusion strategies of WavLM for cross-corpus speech emotion recognition. In *Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Hyderabad, India, 2025.
15. Ye, J.; Wen, X.C.; Wei, Y.; Xu, Y.; Liu, K.; Shan, H. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Rhodes Island, Greece, 2023; pp. 1–5.
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*; Curran Associates, Inc.: Long Beach, CA, USA, 2017; pp. 5998–6008.
17. Pepino, L.; Riera, P.; Ferrer, L. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv* **2021**, arXiv:2104.03502.
18. Lim, J.; Kim, K. Wav2vec-VC: Voice conversion via hidden representations of wav2vec 2.0. In *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Seoul, Republic of Korea, 2024; pp. 10326–10330.
19. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
20. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Shanghai, China, 2016; pp. 5200–5204.
21. Yang, L.; Zhao, H.D.; Yu, K.K. End-to-end speech emotion recognition based on multi-head attention mechanism. *J. Comput. Appl.* **2022**, *42*, 1869–1875. (In Chinese) [CrossRef]
22. Livingstone, S.R.; Russo, F.A. The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef] [PubMed]
23. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; Weiss, B. A database of German emotional speech. In *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2005)*; ISCA: Lisbon, Portugal, 2005; pp. 1517–1520.
24. Adib, M.R.H.; Ali, M.; Mamun, K.A.; Wang, J.; Liu, Y.; Zhang, S.; Chen, L.H.; Zhao, Y.; Singh, A.K.; Wang, H.; et al. Cluster-based evolutionary optimization of deep learning architectures for speech emotion recognition. *Eng. Appl. Artif. Intell.* **2024**, *133*, 107934.
25. Gao, P.Q.; Huang, H.M.; Fan, Y.H. Fusion of coordinate and multi-head attention mechanisms for interactive speech emotion recognition. *J. Comput. Appl.* **2024**, *44*, 2400–2406. (In Chinese) [CrossRef]
26. Wu, Z.; Wang, Y.; Liu, W.; Li, M. BAT: Block attention transformer for speech emotion recognition. *arXiv* **2024**, arXiv:2403.01124.
27. Ye, J.-X.; Wen, X.-C.; Wang, X.-Z.; Xu, Y.; Luo, Y.; Wu, C.-L.; Chen, L.-Y.; Liu, K.-H.; Zhang, H.; Li, M. GM-TCNet: Gated multi-scale temporal convolutional network using emotion causality for speech emotion recognition. *Speech Commun.* **2022**, *145*, 21–35.
28. Wang, X.H.; Zhang, X.F.; Li, W.K.; Wang, J.; Liu, Y.; Zhang, S.; Chen, L.H.; Zhao, Y.; Singh, A.K.; Wang, H.; et al. An efficient WavLM-Transformer framework for speech emotion recognition. *Speech Commun.* **2024**, *158*, 103045.
29. Zhang, L.P.; Zhang, J.C.; Wang, J.; Liu, Y.; Zhang, S.; Chen, L.H.; Zhao, Y.; Singh, A.K.; Wang, H.; Li, M.; et al. AFT-Net: Audio feature temporal network for speech emotion recognition. In *Proceedings of Interspeech 2024*; ISCA: Kos, Greece, 2024; pp. 1120–1124.
30. Wagner, J.; Triantafyllopoulos, A.; Wierstorf, H.; Schmitt, M.; Burkhardt, F.; Eyben, F.; Schuller, B.W. Dawn of the transformer era in speech emotion recognition: Closing the valence gap between wav2vec 2.0 and recurrent neural networks. In *Proceedings of Interspeech 2023*; ISCA: Dublin, Ireland, 2023; pp. 5316–5320.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.