Knowledge-Based Systems 281 (2023) 111077

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Spatio-temporal representation learning enhanced speech emotion recognition with multi-head attention mechanisms☆

Zengzhao Chen [a,b], Mengting Lin [a], Zhifeng Wang [a,*], Qiuyu Zheng [a], Chuan Liu [a]

[a] Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan, 430079, China
[b] National Intelligent Society Governance Experiment Base(Education), Wuhan, 430079, China

ARTICLE INFO

Keywords:
Speech emotion recognition
Spatio-temporal representation learning
Multi-head attention mechanisms
Deep learning

ABSTRACT

Speech emotion recognition (SER) systems have become essential in various fields, including intelligent healthcare, customer service, call centers, automatic translation systems, and human–computer interaction. However, current approaches predominantly rely on single frame-level or utterance-level features, offering only shallow or deep characterization, and fail to fully exploit the diverse types, levels, and scales of emotion features. The limited ability of single features to capture speech emotion information, along with the ineffective combination of different features' complementary advantages through simple fusion, pose significant challenges. To address these issues, this paper presents a novel spatio-temporal representation learning enhanced speech emotion recognition with multi-head attention mechanisms(STRL-SER). The proposed technique integrates fine-grained frame-level features and coarse-grained utterance-level emotion features, while employing separate modules to extract deep representations at different levels. In the frame-level module, we introduce parallel networks and utilize a bidirectional long short-term memory network (BiLSTM) and an attention-based multi-scale convolutional neural network (CNN) to capture the spatio-temporal representation details of diverse frame-level signals. Consequently, we extract deep representations of utterance-level features to effectively learn global speech emotion features. To leverage the advantages of different feature types, we introduce a multi-head attention mechanism that fuses the deep representations from various levels. This fusion approach retains the distinctive qualities of each feature type. Finally, we employ segment-level multiplexed decision making to generate the ultimate classification results. We evaluate the effectiveness of our proposed method on two widely recognized benchmark datasets: IEMOCAP and RAVDESS. The results demonstrate that our method achieves notable performance improvements compared to previous studies. On the IEMOCAP dataset, our method achieves a weighted accuracy (WA) of 81.60% and an unweighted accuracy (UA) of 79.32%. Similarly, on the RAVDESS dataset, we achieve a WA of 88.88% and a UA of 87.85%. These outcomes confirm the substantial advancements realized by our proposed method.

1. Introduction

Speech emotion recognition (SER) has gained increasing popularity in some fields, e.g., human–computer interaction, public opinion analysis, automatic customer service and quality measurement in voice portals, so it has attracted more attention from industry and academia [1]. Speech emotion recognition (SER) systems typically involve two stages: feature extraction and emotion classification [2]. A key focus of SER research is the identification of effective features that enable accurate emotion recognition. Currently, a range of speech emotion features is available, including short-term energy, pitch period, formant, duration, prosody, zero-crossing rate, linear predictor cepstrum coefficient (LPCC), log-mel frequency spectrum (Log-Mel) [3], and mel frequency cepstrum coefficient (MFCC) [4]. Recent studies have indicated that employing Log-Mel features with deltas and delta-deltas can effectively capture the dynamics of emotions, facilitating reliable emotional information retrieval while minimizing the influence of factors unrelated to emotions, such as speakers, equipment, and recording environments [5]. Moreover, manual

☆ This work was supported by National Key R&D Program of China (2022ZD0117103), the National Natural Science Foundation of China (62077022) and Research Project of National Collaborative Innovation Experimental Base for Teacher Development of Central China Normal University (CCNUTEIII 2021-21).

* Corresponding author.
E-mail addresses: zzchen@ccnu.edu.cn (Z. Chen), linmengting@mails.ccnu.edu.cn (M. Lin), zfwang@ccnu.edu.cn (Z. Wang), qiuyu@mails.ccnu.edu.cn (Q. Zheng), liuchuan@mails.ccnu.edu.cn (C. Liu).

https://doi.org/10.1016/j.knosys.2023.111077
Received 6 July 2023; Received in revised form 19 September 2023; Accepted 13 October 2023
Available online 21 October 2023
0950-7051/© 2023 Elsevier B.V. All rights reserved.

features, such as the Interspeech 2020 dataset [6], the geneva minimalistic acoustic parameter Set (GeMAPs) [7], and the audio/visual emotion challenge (AVEC)-2016 [8], have shown promising results in recognizing emotional states.

The second stage of SER involves performing emotion recognition based on the extracted features. Traditional machine learning algorithms, including hidden Markov models (HMM) [9], Gaussian mixture models (GMM) [10], random forests (RF) [11], support vector machines (SVM) [12], plain Bayesian models (PBM) [13], and k-nearest neighbors (kNN) [14], were initially widely used for emotion recognition. However, in recent years, deep learning has gained prominence due to its remarkable performance. Consequently, neural networks have been employed to learn deep emotional representations from shallow input features and to perform emotion classification in SER. Deep neural networks (DNNs) have exhibited excellent performance in extracting discriminative features for SER, enabling the extraction of personalized, task-specific feature representations from diverse training samples [15]. Recently, recurrent neural networks (RNNs) with local attention structures have been proposed for speech emotion tasks. Local attention allows the network to focus on emotionally salient regions of speech while addressing time-dependent issues in SER tasks. Long short-term memory (LSTM) networks, in particular, excel in handling temporal tasks and can resolve the gradient explosion problem associated with RNNs [16]. Additionally, convolutional neural network (CNN)-based frameworks have been introduced for speech emotion tasks, demonstrating exceptional performance on multiple benchmark datasets [17]. Consequently, LSTM networks and CNNs have been extensively employed in SER tasks due to their respective advantages.

Researchers have recently endeavored to combine manually engineered features and automatically learned features to enhance the comprehensiveness of emotion recognition. For instance, Luo et al. [18] developed a two-channel SER system (HSF-CRNN) that integrates high-level statistical features with CRNN to jointly learn more distinctive emotion-related features, achieving excellent performance on the ComParE2018 and IEMOCAP datasets. Similarly, fusion schemes for different features have been proposed by Kumaran et al. [19] and Guo et al. [20]. However, prior research has primarily focused on simple early fusion of different-scale features, without delving into the full potential of diverse-scale emotional features. To maximize the contributions of various features in SER tasks, this study proposes customized models for different-scale emotion features, generating distinct deep emotion representations. Additionally, a multi-head attention mechanism is introduced to leverage multi-feature fusion, capturing multi-granularity information and integrating emotional information from diverse feature scales. The multi-head attention mechanism effectively combines the deep emotion representations generated by the different models, extracting complementary information from each feature scale and enhancing the performance of SER systems.

The main contributions of this paper are as follows:

i. Addressing the limitations of traditional single features, we propose a methodology that leverages various types, layers, and scales of emotion features. Our approach employs a parallel training approach that independently feeds different layer-level features into distinct network architectures. The frame-level module captures the spatio-temporal representation of different types of frame-level features, while the utterance-level module complements them with global features. Finally, a multi-head attention mechanism fuses multi-scale features, enhancing information density.

ii. We present a CNN with an attention mechanism that incorporates multi-scale feature representation. By integrating multi-scale perceptual fields through parallel convolutional layers, our CNN effectively captures the temporal-frequency information of frame-level features, significantly improving the representation capability of existing CNNs. Furthermore, the attention mechanism concentrates on emotionally significant regions.

iii. To enhance the robustness of classification decisions and mitigate local interference signals, we propose a segment-level speech emotion multiplex decision method. Additionally, segmental training increases the training sample size. Our approach achieves a weighted accuracy (WA) of 81.60% and an unweighted accuracy (UA) of 79.32% on the IEMOCAP dataset, and a WA of 88.88% and a UA of 87.85% on the RAVDESS dataset for SER tasks.

The remainder of this paper is structured as follows: Section 2 is a review of related work in the existing literature. Section 3, we provide a problem definition of this research and summarize the notations that appear in this paper. Section 4 introduces our multi-scale speech emotion recognition method combined with an attention mechanism. Section 5 presents the experimental details and analyzes the results. Finally, Section 6 concludes the paper.

2. Related work

In this section, we provide a comprehensive review of relevant research on speech emotion recognition. The review is divided into three main parts: speech emotion recognition, representation based on deep learning techniques, and the incorporation of attention mechanisms.

2.1. Speech emotion recognition

SER is a burgeoning technology that analyzes speech recordings to classify speakers' emotions and predict their physiological and psychological states. Currently, two main approaches dominate the field: unimodal recognition and multimodal recognition.

In unimodal recognition, the focus is solely on speech features for emotion analysis. The raw speech signal undergoes transformations into acoustic features, such as continuous features, spectral features, and prosodic features. Frame-level Low-Level Descriptors (LLDs) can be computed from signal statistics. Subsequently, one or multiple features are selected as input, and machine learning or deep learning techniques are employed for feature representation and emotion classification. For instance, Krishna et al. [21] introduced various classification algorithms, including support vector machines and multilayer perceptrons, for emotion classification using audio features like MFCC, Mel, Chroma, and Tonnetz. Zhao et al. [22] proposed the CNN-LSTM network to extract local and global deep emotion features from sound and log-mel spectrograms, achieving notable results in speech emotion recognition.

In contrast, multimodal recognition encompasses audio-text [23] and audio-visual [24] categories. While these tasks primarily focus on audio features, they incorporate supplementary information from text or visual cues to enhance emotion classification accuracy. Despite advancements in multimodal recognition, speech information remains the primary foundation for practical applications in emotion recognition.

2.2. Representation based on deep learning techniques

Over the past decade, CNNs have become the mainstream architecture for learning speech emotion representation. Their spatial modeling capabilities and parallel computing prowess make them an ideal choice. For instance, Mao et al. [25] harnessed CNNs to learn critical features for speech emotion recognition, achieving high performance across multiple benchmark datasets. Schluter et al. [26] introduced pitch variation and time stretching as effective spectrogram data augmentation techniques, applying them to 2D CNNs for deep feature learning. Chen et al. [5] mitigated the impact of non-linguistic factors on recognition outcomes by calculating delta and delta-delta from log-mel filterbank coefficients. They also proposed a 3D attention-based convolutional recurrent neural network to better capture spatiotemporal relationships within features.

Recent research has emphasized the significance of temporal information in achieving more accurate and comprehensive predictions and inferences [27,28]. Consequently, RNNs and their variants have gained

widespread adoption. LSTM networks, for instance, excel in learning sequential representations of diverse speech styles [29]. Hu et al. [30] introduced the BiLSTM model, which extracts deep emotion representations by incorporating contextual information. Trigeorgis et al. [31] merged CNNs and LSTM models to directly learn optimal audio signal representations from raw waveform data, addressing the challenge of extracting context-aware emotion features.

The Transformer, initially designed for machine translation tasks, has gained prominence in natural language processing due to its ability to model global relationships and its parallel computing capabilities. Gulati et al. [32] introduced a Transformer variant with convolutional operations, which has emerged as a state-of-the-art model for speech emotion recognition. Tarantino et al. [33] utilized a pure Transformer as the backbone network to extract deep emotion representations from feature sets derived from the openSMILE toolkit.

2.3. Attention mechanisms

In recent years, models incorporating attention mechanisms have made significant strides in emotion recognition. Self-attention, multi-head attention, local attention, and global attention have all found applications in this domain. These attention mechanisms enable the model to focus on key information during training, improving recognition outcomes.

Given that not all speech frames contain emotional information, these attention mechanisms allow the model to concentrate on frames that are most relevant to emotions. Chen et al. [6] employed the self-attention mechanism to fuse different feature types and assign weight values, giving higher importance to critical features and enhancing their contribution to final recognition results. Anish et al. [34] utilized a multi-head attention mechanism to capture complex relationships and dependencies in Log Mel-Filter Bank Energies (LFBE) spectrogram features, enabling the model to concentrate on different aspects of emotional cues and improving its ability to recognize subtle emotional patterns. Yoon et al. [35] introduced a multi-hop attention model for SER, extracting hidden information from speech data using bidirectional LSTM and applying the multi-hop attention model to generate final weights for emotion classification. In summary, attention mechanisms have gained widespread adoption in emotion recognition, allowing models to focus on key information, improve recognition of subtle emotional patterns, and enhance prediction accuracy.

Summary: Previous research has primarily relied on single-type features or early feature fusion for recognition tasks. However, single-feature approaches have limitations in capturing emotional information in speech, and simple fusion of diverse features may not effectively exploit their complementary advantages. To address these challenges, this paper introduces a novel approach to speech emotion recognition, leveraging spatio-temporal representation learning enhanced with multi-head attention mechanisms.

3. Preliminaries

In this section, we provide formal definitions of key concepts related to speech emotion recognition and the multi-scale spatio-temporal features used in this paper. Additionally, we describe the proposed multi-attention mechanism. Table 1 summarizes important notations, which will be further explained in the following sections.

3.1. Problem definition

Definition 1 (*Speech Emotion Recognition Problem*)**.** Speech emotion recognition aims to detect the emotion labels (A^m) of a test audio signal based on the emotion labels ($\{A_d^e \mid d = 1, 2, \ldots, D\}$) available in the training data. Mathematically, it can be formulated as:

$$d^* = \arg\max_d \left\{ f\left(A_1^e, A^m; \gamma\right), f\left(A_2^e, A^m; \gamma\right), \ldots, f\left(A_D^e, A^m; \gamma\right) \right\} \quad (1)$$

Table 1
Summary of notations.

Notation	Description
A	Audio signal
δ	ReLU activation function
BN	Batch normalization layer
MP	Maximum pooling layer
Conv	Convolution layer
w,W	Weights
f_C	Segment level classifier
$G = \{X_i, Y_i\}_{i=1}^K$	Speech samples and corresponding labels
$C_i = \left\{ s_i^0, s_i^1, \ldots, s_i^{M_i} \right\}_{i=1}^K$	Audio sample segmentation set
L	Loss function

Here, $f(\bullet)$ represents the trained network model, and the back-end parameter γ helps mitigate the effects of speech diversity issues (e.g., speaker gender and speech content mismatch) on emotional features. A^e and A^m represent the training and testing speech emotion features, respectively. The variable D denotes the label of the training set. If A^m does not exist in the features with d labels, the speech emotion recognition problem is considered a closed-set problem; otherwise, it is an open-set problem. This paper focuses on the closed-set problem. Fig. 1 illustrates the flowchart of speech emotion recognition.

3.2. Definition of multi-scale deep features and multi-head attention mechanism

Definition 2 (*Frame-Level Depth-Time Features*)**.** We utilize the BiLSTM neural network to extract depth-time features from frame-level features. In analyzing time series feature data, it is essential to not only consider information within individual frames but also capture the connections between frames. Deep learning networks, such as RNNs and LSTMs, are well-suited for modeling sequences and extracting deep temporal features from the original features. In this paper, we modify the BiLSTM network to extract deep temporal features at the frame level, providing a reflection of temporal information from both past and future frames.

Definition 3 (*Frame-Level Deep Spatial Features*)**.** To capture deep spatial features from two-dimensional data, we employ a CNN along with multi-scale convolutional kernels. This allows us to learn temporal and frequency information of the LLDs features separately, resulting in an accurate representation of the spatial features in the frame-level feature data. To evaluate the relevance of speech emotion information, we introduce a self-attention network that assigns weights to the deep spatial features. The attention mechanism is designed to learn the emotion relevance of the feature matrix, with optimized structure and parameters to accurately capture the emotional nuances in speech data.

Definition 4 (*Multiscale Features*)**.** This paper presents a multiscale feature extraction technique that incorporates three main aspects. First, we fuse frame-level depth features and utterance-level depth features to enable a detailed analysis of emotion at the frame level while providing a coarser, overall description of the speech signal. Second, we achieve multiscale feature fusion by considering both temporal and spatial perspectives during frame-level depth feature extraction. Finally, we utilize multi-scale convolutional kernel settings to effectively capture time–frequency features in frame-level deep spatial features. By considering these three aspects, we comprehensively learn the features, resulting in complementary advantages.

Definition 5 (*Multi-Attention Mechanism*)**.** This paper proposes a multi-attention mechanism that focuses on learning the distribution of importance within input features using attention blocks. This mechanism enhances the relevance of information for classification by assigning weights to highlight relevant features. By fusing the information from

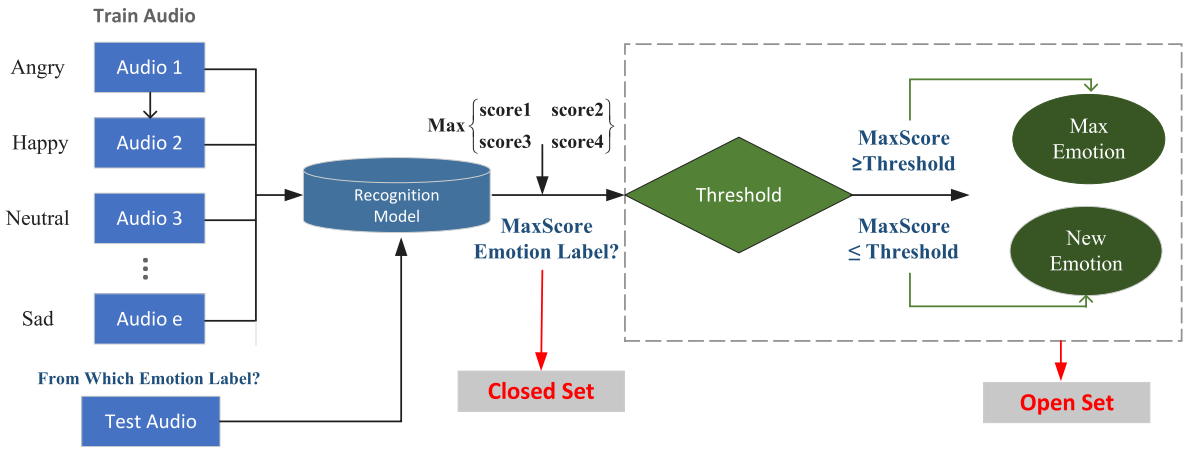

Fig. 1. Flowchart of speech emotion recognition.

the two-way network, the multi-attention mechanism achieves complementary utilization of multi-scale feature information and improves the information density.

Overall, the definitions provided in this section establish the foundation for understanding the concepts of speech emotion recognition, multi-scale deep features, and the multi-attention mechanism proposed in this paper.

4. Methods

To extract varying levels of emotion representation from speech signals, this study develops two modules: the frame-level module and the utterance-level module. The frame-level module, consisting of CNN and BiLSTM sub-modules, captures both spatial and temporal information to analyze fine-grained emotion features. The utterance-level emotion feature model preserves the global feature information of the speech signal and complements the advantages of different emotion feature levels. Additionally, a multi-head attention mechanism integrates features from different scales, emphasizing emotion task-related information. The final emotion classification results are obtained using the segment-level multiplex decision method. The overall framework is illustrated in Fig. 2.

4.1. Feature reconstruction phase

Given the significant variation in duration among different audio samples in the dataset, it is necessary to divide the continuous speech sequence into several segments to meet the input requirements of the network. In this study, the audio in the IEMOCAP and RAVDESS datasets is divided into equal-length segments of 2 s and 1.4 s, respectively, with a sampling rate of 16 000 Hz. To ensure a natural transition between frames, adjacent frames are overlapped. The selected features used in our study are based on The INTERSPEECH 2010 Paralinguistic Challenge feature set.

For frame-level features, we apply a hamming window to frame the speech with a frame length of 128 ms and a frame shift of 32 ms. The last frame is padded to account for the missing time. A total of 76-dimensional features are extracted, including various types of shallow emotion features, as shown in Table 2. These include 34 low-level descriptors (LLDs), 34 LLDs first-order differences, 4 pitch-related features, and 4 pitch-related first-order features. The frame-level feature dimension is set to (B, D).

Utterance-level features consist of 1582 features derived from the base of 34 LLDs. Additionally, 34 corresponding delta coefficients are attached, applying 21 general statistical functions. Moreover, 19 statistical functions are applied to the four tonal LLDs and their four first-order differences, excluding percentile 1.0 and PCTLrange0 – 1 from the set of 21 statistical functions.

Table 2
Interspeech 2020 paralinguistics challenge feature set.

Low-level descriptors	Quantity
pcm_loudness: Loudness and Loudness delta	2
MFCC: Mel-frequency cepstrum coefficients and its delta	30
logMelFreqBand: Mel-frequency cepstrum and its delta	16
1spFreq: Instantaneous frequency and its delta	16
FOfinEnv: Fundamental frequency contour line and its delta	2
voicingFinalUnclipped: Voicing probability and its delta	2
FOfinal: Fundamental frequency and its delta	2
jitterLocal: local frequency perturbation and its delta	2
jitterDDP: Differential inter-frame perturbation and its delta	2
shimmerLocal: local amplitude perturbation and its delta	2

4.1.1. Frame-level deep sentiment feature extraction

This study employs a CNN to extract frame-level deep spatial feature information from the input data. Additionally, an attention mechanism is utilized to reduce the sensitivity of parameters in the subsequent fusion process after completing the spatial feature extraction. CNN networks offer unique advantages in extracting deep features and have shown effectiveness in addressing challenges associated with speech emotion recognition, such as diverse speech emotion signals from different speakers and acquisition devices. These networks exhibit translation invariance and local perception properties in their unique structure and computational methods. The feature extraction process using convolutional kernels is task-specific and converts the two-dimensional data into a feature map, effectively extracting spatial information from the input data. Utilizing CNN networks for feature extraction solves the problem of speech signal diversity and provides robustness in applications dealing with two-dimensional data, specifically locally relevant features. In this study, the feature extraction process is performed by convolutional kernels, resulting in a feature map that enhances the spatial information in the input data.

The CNN used for deep spatial feature extraction comprises a convolutional layer, pooling layer, and ReLU layer. Convolutional layers extract deep features from shallow features using convolutional kernels. By setting the convolution kernel as (H, R, C), where C is the number of channels and (H, R) is the size of a single convolution kernel, the feature undergoes convolution calculations to generate the data for the convolution layer, as shown in Eq. (2).

$$s(i,j) = f\left(\sum_{h=1}^{H}\sum_{r=1}^{R}\sum_{C=1}^{C'} x_{h,r,c}^{j} w_{h,r,c}^{i} + b^{i}\right) \tag{2}$$

Here, i represents the ith channel of the convolutional layer, $s(i,j)$ indicates the specific value of the jth value of the ith channel, $x_{h,r,c}^{j}$ is the jth input of the input layer (or max pooling layer), $w_{h,r,c}^{i}$ denotes

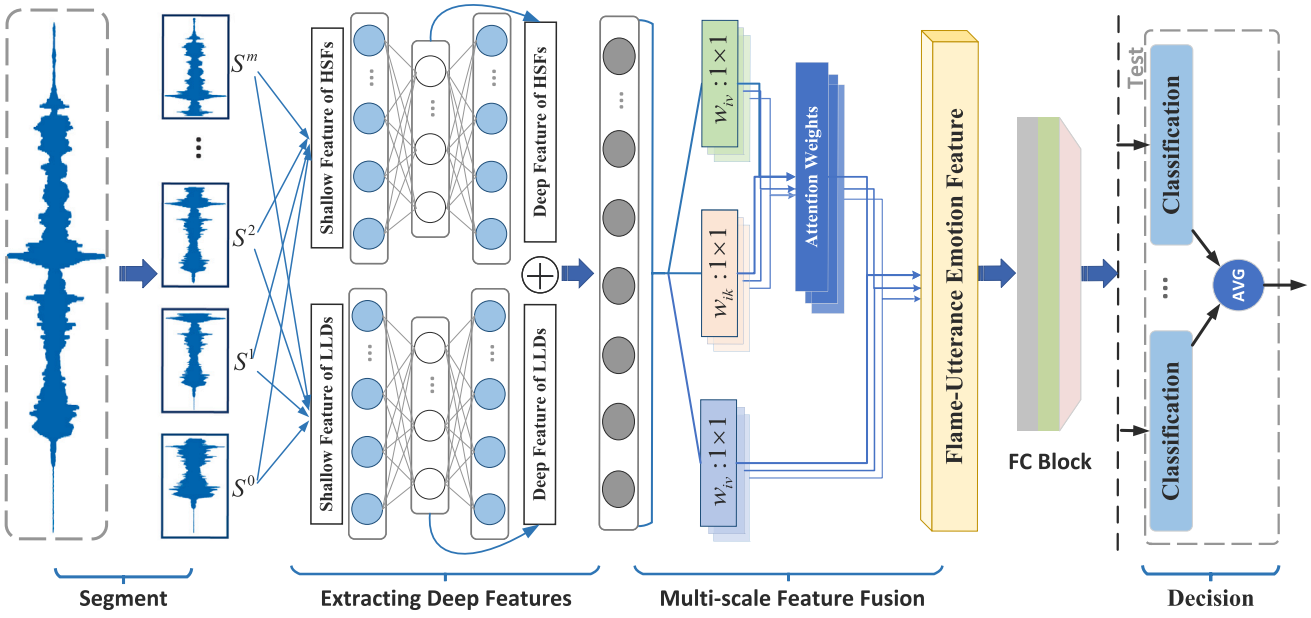

Fig. 2. Spatio-temporal representation learning framework.

the convolutional kernel of the ith channel of the convolutional layer, and $f(\bullet)$ is the ReLU activation function. To capture time–frequency information in the frame-level spatial features, this paper proposes parallel multi-scale convolutional blocks. Based on the dimensional ratios of time and frequency in the frame-level features, frequency convolution kernels (b,1) and time convolution kernels (1,d) with different scales are set up, and the output channels are all 16. Different time and frequency convolutional kernels with 16 output channels are applied at various scales, depending on the dimension of the initial frame-level features. The results are then concatenated and input into 3 convolutional layers with a kernel size of $3 * 3$, generating a representation of 48 channels. After each convolutional layer, a max pooling layer (pooling kernel size set to $2 * 2$) is connected. The input and output of the convolutional neural network are denoted by X and X', respectively, as shown in Eqs. (3)–(5).

$$\hat{X} = \left[Conv_{b*1}(X), Conv_{1*d}(X)\right] \tag{3}$$

$$F(x) = BN\left(MP_{2*2}\left(\delta\left(Conv_{3*3}(x)\right)\right)\right) \tag{4}$$

$$X' = F(F(F(\hat{X}))) \tag{5}$$

Here, δ represents the ReLU activation, BN, MP, and Conv respectively represent the batch normalization layer, max pooling layer, and convolutional layer.

In speech emotion recognition, the density of emotion-related information in different parts of the frame-level deep spatial features extracted by the CNN network is not equal, highlighting the importance of focusing on parts relevant to the emotion classification task. To address this challenge, researchers have developed an attention-based weight allocation method [36]. This method utilizes an attention matrix to assign higher weights to emotion-related information in the feature map generated by the multi-scale convolutional network, thereby enhancing the contribution of emotional components to the final recognition result. In this study, the attention matrix is calculated as follows:

$$\alpha = softmax(w^T X') \tag{6}$$

$$\widetilde{X} = \alpha^T X' \tag{7}$$

Here, w^T represents the transposed trainable parameters, and α represents the weight assigned to the speech emotion. Through the attention mechanism, the model can focus on the connections between different parts of the input, thereby enhancing the accuracy of speech emotion recognition. The convolutional neural network framework for frame-level deep spatial feature extraction is illustrated in Fig. 3.

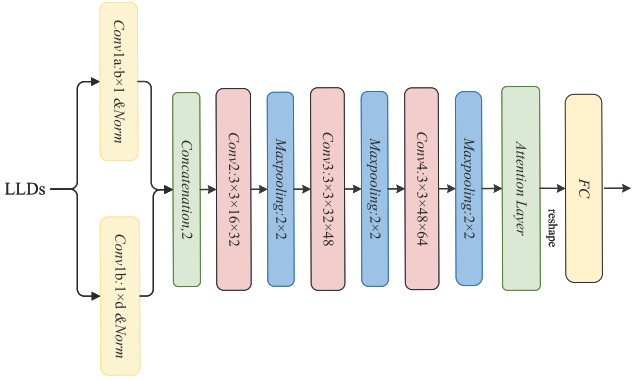

Fig. 3. Frame-level depth space feature extraction network.

As a typical time series signal, contextual information is also crucial for speech emotion recognition. Li et al. [37] discovered that speech emotion-related features exist in both forward and backward speech signals. To better analyze the high-level representation of time-related features in speech emotion through frame-to-frame feature changes, a network model based on the BiLSTM structure is proposed, as shown in Fig. 4. BiLSTM incorporates the inverse phase operation into the LSTM module, enabling it to exploit past and future contextual information, which plays a crucial role in analyzing time-related features in speech emotion. The LSTM layer addresses the long-term dependency problem by introducing a cell state C to the hidden layer in the standard RNN model. For the Cell state C, the LSTM network utilizes three gate control switches: Forget Gate, Input Gate, and Output Gate, to regulate the cell state. The structure and calculations of these gates are described as follows. The forget gate, as shown in Eq. (8), is similar to a single-layer neural network:

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

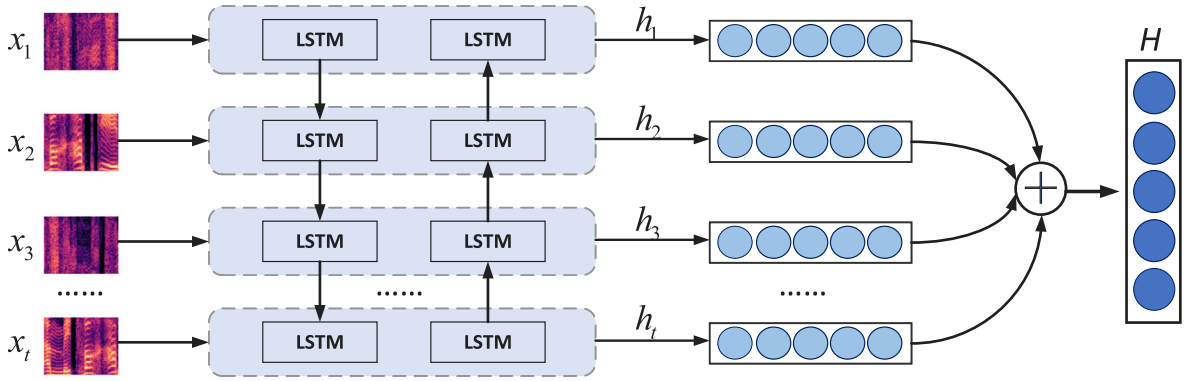

Fig. 4. BiLSTM structure diagram.

Where x_t is the current timestep input, h_{t-1} is the hidden later output at the previous timestep, C_{t-1} represents the memory of the previous sequence LSTM unit, b_f represents the bias term, W represents the weighted vector, and σ represents the sigmoid function and hyperbolic tangent function respectively. The input gate's main function is to selectively update the current input information x_t to the cell state C_t. The formula is expressed as follows:

$$i_t = \sigma\left(W\left[x_t, h_{t-1}, C_{t-1}\right] + b_i\right) \tag{9}$$

$$C_t = f_t C_{t-1} + i_t \bullet tanh\left(\left[x_t, h_{t-1}, C_{t-1}\right] + b_c\right) \tag{10}$$

Finally, the output gate is mainly used to control how much information of the current cell state C_t can be saved to the output h_t. The calculation formula of the output gate is shown in Eqs. (11)–(12).

$$\sigma_t = \sigma\left(W\left[x_t, h_{t-1}, C_{t-1}\right] + b_o\right) \tag{11}$$

$$\overrightarrow{h_t} = o_t \bullet tanh\left(C_t\right) \tag{12}$$

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] \tag{13}$$

The symbol $\overrightarrow{h_t}$ represents the output of the forward LSTM layer, and $\overleftarrow{h_t}$ represents the output of the backward LSTM layer. The final output is obtained by concatenating these two outputs, resulting in $H = \{h_1, h_2, \ldots, h_T\}$. The emotional representation output from the attention layer is combined with this, serving as the frame-level deep emotional representation, as shown in Eq. (14):

$$\mathrm{FL} = \mathrm{Concat}(\widetilde{\boldsymbol{X}}, \boldsymbol{H}) \tag{14}$$

4.1.2. Utterance-level deep emotion feature extraction

The utterance-level deep emotion feature extraction module combines the openSMILE 2010 feature extraction method, which provides statistical features describing the rhythmic, spectral, and speech quality aspects of speech, with the frame-level deep emotion feature module. These two types of features capture emotion information at different levels, combining coarse and fine-grained information to utilize sentiment information more effectively and extract robust features for SER, resulting in improved performance.

In the utterance-level deep emotion feature module, a fully connected layer is used to map the high-dimensional High-Level Semantic Features (HSFs) to a lower-dimensional feature space through a non-linear transformation. This module outputs utterance-level deep emotion feature vectors, as shown in Eq. (15). These vectors are then connected with the frame-level deep sentiment feature vector and serve as the input for the subsequent multi-scale sentiment feature fusion module, as depicted in Fig. 5.

$$UL = FC(\boldsymbol{HSFs}) \tag{15}$$

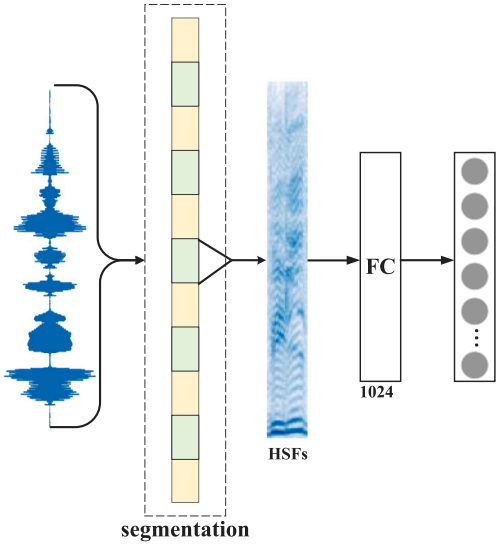

Fig. 5. Utterance-level deep sentiment feature extraction network framework.

The utterance-level deep sentiment feature extraction network framework shown in Fig. 5 illustrates the flow of information in this module.

Overall, the methods section now provides a clear and organized explanation of the utterance-level deep emotion feature extraction module. The purpose and benefits of combining openSMILE 2010 features with the frame-level deep emotion features are highlighted. The role of the fully connected layer in mapping high-dimensional features to a lower-dimensional space is clarified. The connection between the utterance-level deep emotion features and the frame-level deep sentiment features is described, emphasizing their collaboration in the subsequent sentiment feature fusion module.

4.2. Multi-scale deep emotional feature fusion

In this section, we propose a multi-headed attention mechanism to fuse different types of features. Each feature type has limited capacity to capture speech emotion information. Therefore, we combine temporal–spatial emotion representation based on frame-level information with emotion representation at different granularities, incorporating frame-level and utterance-level information, to obtain multi-scale emotion representation information.

Algorithm 1: The proposed Multi-scale deep emotional representation learning model.

Input: Flame-level feature X: a feature matrix of shape (B,D) a sequence of LLDs vectors $\{x_1, x_2, \cdots, x_T\}^T$;Utterance-level feature U: a feature matrix of shape (1,1582)

Output: Emotion prediction for input sample 1 Extracting frame-level deep emotion spatial features by Attentional Multi-Scale CNN blocks

1 Extracting frame-level deep emotion spatial features by Attentional Multi-Scale CNN blocks:

$$\begin{cases} \hat{X} & = \left[Conv_{b*1}(X), Conv_{1*d}(X)\right] \\ F(x) & = BN\left(MP_{2*2}\left(\delta\left(Conv_{3*3}(x)\right)\right)\right) \\ X' & = F(F(F(\hat{X}))) \\ \alpha & = softmax\left(w^T \mathbf{X}'\right) \\ \widetilde{\mathbf{X}} & = \alpha^T \mathbf{X}' \end{cases}$$

2 Extracting frame-level deep emotion temporal features by BiLSTM blocks:

$$\begin{cases} h_t & = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] \\ H & = \left\{h_1, h_2, \cdots, h_T\right\} \end{cases}$$

3 Concatenate frame-level deep temporal features $\widetilde{\boldsymbol{X}}$ and deep spatial features H to form frame-level depth-emotion features:

$$\mathrm{FL} = \mathrm{Concat}(\widetilde{\boldsymbol{X}}, \boldsymbol{H})$$

4 Extracting utterance-level deep emotion features by FC blocks:

$$UL = FC(\boldsymbol{HSFs})$$

5 Concatenate frame-level deep features FL and utterance-level deep features UL,and assign weights by multi-headed attention mechanism to achieve feature fusion:

$$\begin{cases} f & = Concat(FL, UL) \\ Q_i, \boldsymbol{K}_i, \boldsymbol{V}_i & = w_{iq}(\boldsymbol{f}), w_{ik}(\boldsymbol{f}) w_{iv}(\boldsymbol{f}) \\ O_i & = \mathrm{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \\ \widehat{\boldsymbol{O}} & = \mathrm{Concat}\left(\boldsymbol{O}_1, \boldsymbol{O}_2 \cdots, \boldsymbol{O}_N\right) \end{cases}$$

6 Compute classification loss L:

$$\begin{cases} \hat{y}^h & = \arg\max_{h=h^*} \mathrm{Softmax}(W\widehat{\boldsymbol{O}} + \boldsymbol{b}) \\ L & = -\sum_h^H y^h log\left(\hat{y}^h\right) \end{cases}$$

7 Predict the emotional label h^*

The fusion of these features is achieved by concatenating the frame-level deep sentiment feature (FL) with the utterance-level deep sentiment feature (UL), as shown in Eq. (16).

$$f = Concat(FL, UL) \tag{16}$$

However, it is important to note that different types of features contribute differently to the final speech emotion recognition rate. To address this, our study employs a multi-headed attention mechanism that generates multiple sets of attention weights by transforming inputs. This allows the model to capture information from different spatio-temporal emotion subspaces at various locations. By sequentially modeling the relative dependencies between elements at different locations, the model enhances the weight of information related to emotion features, leading to a greater contribution of these features to the final recognition result. The formulas for the attention mechanism are as follows:

$$Q_i = w_{iq}(f) \tag{17}$$

$$K_i = w_{ik}(f) \tag{18}$$

$$V_i = w_{iv}(f) \tag{19}$$

$$\mathrm{O}_i = \mathrm{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \tag{20}$$

$$\widehat{\boldsymbol{O}} = \mathrm{Concat}\left(\boldsymbol{O}_1, \boldsymbol{O}_2 \cdots, \boldsymbol{O}_N\right) \tag{21}$$

Here, w_{iq}, w_{ik}, and w_{iv} are projection functions implemented as 1×1 convolutions. The dot product between the query vector matrix and the key vector matrix calculates the weight of each dimension of data in the feature. This weight indicates the degree of relevance between the queried task and each input data, effectively highlighting the weights of emotional elements. Finally, the results from each attention head are concatenated to obtain the fused feature data.

The proposed multi-scale deep emotion representation learning model is summarized in Algorithm 1, which presents a detailed breakdown of the algorithm, illustrating the step-by-step process of the model.

4.3. Segment-level speech emotion multiplexing decision

In the speech dataset, a small portion of data exhibits difficulty in distinguishing emotional features. This difficulty arises due to relatively long speech durations, which may contain additional emotional features that interfere with emotion analysis. Moreover, emotional representations of different speech emotions that are closely related, resulting in overlaps that can lead to confusion. To address this issue, this paper proposes a novel segment-level multiplex decision method.

Algorithm 2: Algorithm for model objective function learning.

Input: Input audio signals

Initialization: convolutional layer weights W_{CNN} , Attention layer weights W_{Att}, BiLSTM weights W_{BiLSTM}, Utterance-level FC weights W_{FC}, Multi-headed attention layer weights W_{M-Att},learning rate $lr = 0.001$

for e = 1: E (E = epoch) **do**

  $lr * 0.95$ at each epoch;

  for t = 1: T (T = samples / batch size) **do**

    Calculate the loss:

$$L = -\sum_{h}^{H} y^h log\left(\hat{y}^h\right)$$

    Compute the backpropagation error: $\frac{\partial L(t)}{\partial W(t)}$

    Update the parameters:

$$\begin{cases} W_{CNN}(t+1) & = W_{CNN}(t) - \mathrm{lr}\,\frac{\partial L(t)}{\partial W_{CNN}(t)} \\ W_{Att}(t+1) & = W_{Att}(t) - \mathrm{lr}\,\frac{\partial L(t)}{\partial W_{Att}(t)} \\ W_{BiLSTM}(t+1) & = W_{BiLSTM}(t) - \mathrm{lr}\,\frac{\partial L(t)}{\partial W_{BiLSTM}(t)} \\ W_{FC}(t+1) & = W_{FC}(t) - \mathrm{lr}\,\frac{\partial L(t)}{\partial W_{FC}(t)} \\ W_{M-Att}(t+1) & = W_{M-Att}(t) - \mathrm{lr}\,\frac{\partial L(t)}{\partial W_{M-Att}(t)} \end{cases}$$

  end

end

Let $G = \left\{X_i, Y_i\right\}_{i=1}^{K}$ represent the original speech samples, where $Y_i \in \{0,1\}^H$ denotes the label of speech sample X_i. Here, H and K represent the numbers of classes and training samples, respectively. Each speech sample is divided into multiple equal-length speech segments, with the number of segments determined by the length of the original speech sample. We assume that the segmentation quantity of speech sample X_i is M_i. Each speech segment is assigned the same emotional label as the original speech sample. The segmented fragment set of the speech sample is defined as $C_i = \left\{s_i^1, s_i^2, \ldots, s_i^{M_i}\right\}_{i=1}^{K}$, where s represents the segmented fragment, i represents the speech sample sequence, and m represents the segmentation sequence of the speech sample.

Using this method, we construct a new training set $\tilde{G} = \left\{s_i^{m_i}, Y_i\right\}_{m_i=1,i=1}^{M_i,K}$, where M_i represents the number of speech segments into which the ith speech sample is divided. Based on this, we train a segment-level classifier f_c. The function $f_c\left(s_i^{m_i}\right) \in [0,1]^H$ predicts the probabilities of emotion classes for the speech segment, and the cross-entropy loss function is used for optimization. Here, y^h denotes the probability that sample $s_i^{m_i}$ belongs to category h. The learning algorithm for the objective function, loss, and network weights is shown in Algorithm 2.

$$L\left(f_c\left(s_i^{m_i}\right), y\right) = -\sum_{i}^{K}\sum_{m_i}^{M_i}\sum_{h}^{H} y^h \log\left(f_c\left(s_i^{m_i}\right)^h\right) \tag{22}$$

Meanwhile, we consider all segmented speech fragments of a speech sample in the test set as one set, denoted as $G_{C_i} = \left\{C_i, Y_i\right\}_{i=1}^{K}$. However, the speech fragments generated from a speech sample may predict multiple emotional categories. If all speech segments in a speech sample contribute equally to the labels, the sample-level prediction can be obtained by averaging all segment-level predictions:

$$F\left(C_i\right) = \frac{1}{M_i}\sum_{m_i=1}^{M_i} f_c\left(s_i^{m_i}\right) \tag{23}$$

Here, $num(C_i)$ denotes the number of divisions for the ith speech sample. Additionally, this method can also be modified to use a voting scheme. Based on $f_c\left(s_i^{m_i}\right)$, the emotional category of each segmented speech can be predicted, and then a vote is taken for the emotional category with the highest probability. The emotion class with the highest voting value is selected as the final predicted result for the speech sample. In this paper, the mean value strategy is used. The specific framework is illustrated in Fig. 6.

Furthermore, Algorithm 2 presents the algorithm for learning the objective function of the model.

5. Experimental results and analysis

5.1. Dataset

The IEMOCAP dataset is a multi-modal, multi-speaker emotional database recorded by the University of Southern California. It consists of 5 dialogues, each featuring a male and a female speaker. The dataset includes a total of 10,039 audio samples with a sampling rate of 12 kHz and an average duration of 4.5 s. It covers seven emotions: neutral, fearful, happy, angry, sad, excited, and frustrated. For our experiment, we focused on the utterances labeled as angry, happy, sad, and neutral. We grouped the "excited" category under the "happy" category due to their similarity.

The RAVDESS dataset is another widely used dataset for studying SER tasks. It comprises recordings and videos of 24 professional actors (12 males, 12 females) reciting English sentences. The dataset contains a total of 7356 files, including 1440 audio files. The audio files have a sampling rate of 48 kHz and an average duration of 3 s. It covers eight emotions: sad, happy, angry, calm, fearful, surprised, disgusted, and neutral. For our study, we used all the emotion categories available in the RAVDESS dataset.

5.2. Evaluation metrics

To evaluate the performance of our technique, we compared its results with those of other studies that conducted speech emotion recognition experiments using the IEMOCAP and RAVDESS datasets. We adopted six evaluation metrics: Weighted Accuracy (WA), Unweighted Accuracy (UA), Macro Precision, Weighted Precision, Macro F1, and Weighted F1. In our experiment, we divided the IEMOCAP and RAVDESS datasets into training and validation sets, which accounted for 80% and 20% of the total samples, respectively.

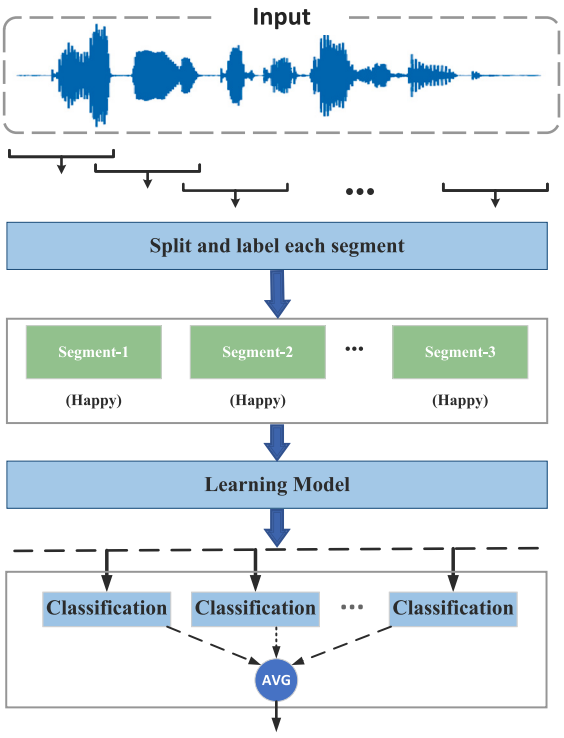

Fig. 6. Multiplex decision framework (happy audio).

The calculation formulas for UA, WA, Macro Precision, Weighted Precision, Macro F1, and Weighted F1 are as follows:

$$UA = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FN_i} \times 100\% \tag{24}$$

$$WA = \frac{\sum_{i=1}^{K} TP_i}{\sum_{i=1}^{K}\left(TP_i + FN_i\right)} \times 100\% \tag{25}$$

$$Macro\ precision = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FP_i} \times 100\% \tag{26}$$

$$Weighted\ precision = \frac{\sum_{i=1}^{K} TP_i}{\sum_{i=1}^{K}\left(TP_i + FP_i\right)} \times 100\% \tag{27}$$

$$Macro\ F1 = \frac{2 \times Macro\ precision \times UA}{Macro\ precision + UA} \tag{28}$$

$$Weighted\ F1 = \frac{2 \times Weighted\ precision \times WA}{Weighted\ precision + WA} \tag{29}$$

UA, Macro Precision, and Macro F1 assign equal weights to each emotion class. Since there is an imbalance in the number of samples across different emotion classes in the IEMOCAP and RAVDESS datasets, WA, Weighted Precision, and Weighted F1 should be calculated by assigning weights to each class based on the number of samples for each emotion classes.

5.3. Baseline

To demonstrate the superior performance of our model, we compared our experimental results with those of other studies. Table 3 presents the specific comparative results.

AMSnet [6]: This paper introduces a parallel network for multi-scale SER called AMSNet. It utilizes a connection attention mechanism to fuse frame-level manual features extracted based on SCNN and utterance-level depth features extracted based on LSTM.

Dual Attention+BiLSTM [38]: This approach addresses the variable lengths between different audio signals by applying a novel dual attention-BiLSTM. It also utilizes a new data preprocessing mechanism using linear interpolation and decimation.

Attention-VTLP [39]: This approach applies multiscale area attention in a deep convolutional neural network to attend to emotional characteristics with different granularity. Additionally, it performs data augmentation with vocal tract length perturbation (VTLP) on the IEMOCAP dataset.

LSTM-Coattention [40]: This method leverages multi-level acoustic information, including MFCC, acoustic spectrograms, and high-level acoustic information embedded in CNN, Bi-LSTM, and wav2vec2, respectively. These extracted features serve as multimodal inputs and are fused using a co-attention mechanism.

FA-UA [41]: This method incorporates a Focus-Attention (FA) mechanism to detect the segment with the largest amplitude. It also proposes a novel Calibration-Attention (CA) mechanism to modulate the information flow and improve the utilization of surrounding contexts.

BAT [42]: This method proposes a self-attentive approach that captures deeper local information by dividing mixed spectrogram features into blocks. It applies a cross-block attention mechanism to capture dependencies between blocks and merge global contextual information in a channel-weighted manner.

DST [43]:This approach utilizes a lightweight decision network to determine the usage of window sizes based on the input speech condition, allowing it to adaptively discover and focus on valuable information embedded in the speech.

Dual Memory-Transformers [44]: This approach introduces a Transformer-based SER architecture that enables the model to understand the relative importance of modalities as they change over time. It leverages Wav2Vec2 and BERT models to extract audio and text features respectively.

Att-Net [45]: This approach introduces a lightweight deep learning-based Self-Attention Module (SAM). It employs a multilayer perceptron (MLP) in channel attention to extract global cues and utilizes a special dilated CNN in spatial attention to extract spatial information from the input tensor.

twine-Shuf-pat [46]: This approach utilizes a shuffle box to generate features and employs iterative neighborhood component analysis to select features.

AAD-Net [47]:This approach introduces a Transformer-based SER (Speech Emotion Recognition) architecture that enables the model to understand the relative importance of modalities as they change over time. It leverages Wav2Vec2 and BERT models to extract audio and text features respectively.

Aural Transformers [48]:The paper proposed an automatic emotion recognition system composed of a SER and a FER. For the SER module, two transfer-learning techniques, namely embedding extraction and fine-tuning, were evaluated on the pre-trained xlsr-Wav2Vec2.0 transformer.

By comparing our results with these state-of-the-art methods, we aim to demonstrate the effectiveness and superiority of our proposed model.

5.4. Comparison with baseline methods

In this section, we present the results of our proposed model compared to several baseline methods. The model was trained for a total of 150 epochs with a batch size of 128. The initial learning rate was set to 0.001, which was multiplied by 0.95 at each epoch. The training process utilized the following experimental hardware configuration: an AMD Ryzen Threadripper PRO 5975WX CPU with 32 cores, an Nvidia GeForce RTX 4090 GPU, and 87.9 GB of memory.

Table 3 provides a comprehensive overview of the experimental results and compares our model with the baseline methods. The evaluation metrics used were WA and UA. The results are presented separately for the IEMOCAP and RAVDESS datasets.

Table 3
Experimental results of comparison with the baseline method.

Dataset	Method	Source	Models	WA (%)	UA (%)
IEMOCAP	Chen et al, 2023 [6]	EXPERT SYST APPL	AMSnet	69.22	70.51
	Chen et al, 2021 [38]	ENG APPL ARTIF INTEL	Dual Attention+BiLSTM	68.73	70.29
	Xu et al, 2021 [39]	ICASSP	Attention-VTLP	79.34	77.54
	Zou et al, 2022 [40]	ICASSP	LSTM-Coattention	71.64	72.70
	Kim et al, 2022 [41]	arXiv	FA-UA	72.01	72.83
	Lei et al, 2022 [42]	Neural Networks 2022	BAT	73.20	75.20
	Chen et al, 2023 [43]	ICASSP	DST	71.80	73.60
	Priyasad et al, 2023 [44]	INTERSPEECH	Dual Memory-Transformers	76.80	77.30
	Proposed method		STRL-SER	**81.60**	**79.32**
RAVDESS	Kwon et al, 2021 [45]	Applied Soft Computing	Att-Net	81.00	83.00
	Tuncer et al, 2021 [46]	Knowledge-Based Systems	twine-shuf-pat	87.43	87.43
	Kim et al, 2022 [41]	arXiv	FA-UA	82.85	82.47
	Mustaqeem et al, 2023 [47]	Knowledge-Based Systems	AAD-Net	88.00	84.00
	Luna-Jiménez et al, 2021 [48]	Applied Sciences	Aural Transformers	86.70	86.42
	Proposed method		STRL-SER	**88.88**	**87.85**

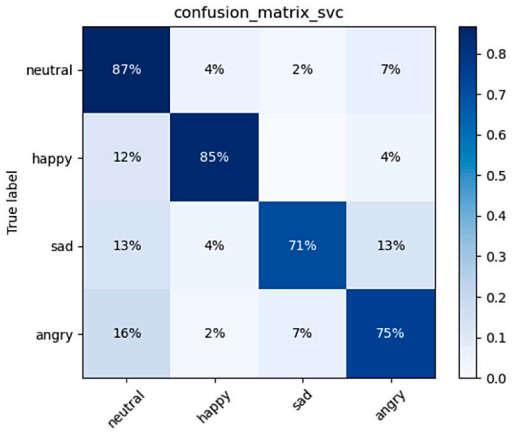

Fig. 7. Confusion matrix of our model on IEMOCAP.

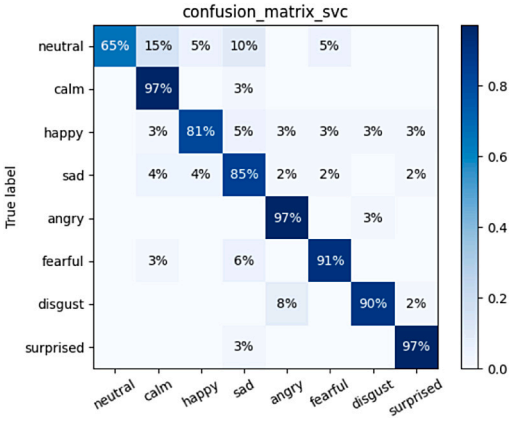

Fig. 8. Confusion matrix of our model on RAVDESS.

Regarding the IEMOCAP dataset, our model outperformed the state-of-the-art approach in terms of WA and UA. Specifically, our model achieved a WA of 81.60% and a UA of 79.32%. These results represent an improvement of 2.26% and 1.78% over the previous best-performing method, respectively.

Similarly, on the RAVDESS dataset, our model exhibited superior performance compared to the baseline methods. It achieved a WA of 88.88% and a UA of 87.85%, surpassing the other models by 0.88% and 0.42% for WA and UA, respectively.

To provide a visual representation of the model's performance, Fig. 7 shows the confusion matrix of our model on the IEMOCAP dataset, while Fig. 8 displays the confusion matrix on the RAVDESS dataset. These figures offer insights into the model's ability to correctly classify different emotions.

Overall, our proposed model demonstrates superior performance compared to the baseline methods, as evidenced by the higher accuracy achieved on both datasets.

Table 4
Experimental results for different numbers of heads on the IEMOCAP (%).

Head	WA	UA	Macro precision	Weighted precision	Macro F1	Weighted F1
1	80.66	78.48	81.52	80.81	79.57	80.44
2	81.36	78.17	80.97	81.23	79.27	81.05
4	79.71	77.12	77.83	79.54	77.40	79.58
8	**81.60**	**79.32**	81.34	81.70	**80.22**	**81.55**
16	79.71	77.49	78.65	79.65	77.98	79.63
32	81.36	77.53	**81.66**	**81.80**	78.65	81.06

Table 5
Experimental results for different numbers of heads on the RAVDESS (%).

Head	WA	UA	Macro precision	Weighted precision	Macro F1	Weighted F1
1	86.11	86.41	86.41	87.11	85.51	85.80
2	87.15	86.90	87.09	86.99	86.55	86.92
4	86.80	85.44	86.31	86.95	83.54	84.39
8	**88.88**	**87.85**	**89.23**	**89.20**	**88.22**	**88.82**
16	84.02	83.99	84.54	84.86	83.84	84.02
32	84.37	84.51	84.14	85.11	84.08	84.51

5.5. Ablation experiments

5.5.1. Comparative experiments of multi-head attention mechanisms

In this section, we conducted ablation experiments to investigate the impact of the number of heads in the multi-head attention mechanism on the overall performance of our model. We evaluated the recognition accuracy of the model with different numbers of heads, ranging from 1 to 32, on both the IEMOCAP and RAVDESS datasets.

Tables 4–5 presents the experimental results, showing the values of various evaluation metrics for each number of heads. The results indicate the performance of the model under different configurations.

Fig. 9 illustrates the fluctuation of results as the number of heads increases. It is important to note that increasing the number of heads does not necessarily lead to better performance. While more heads allow the model to capture sentiment information from different spatiotemporal subspaces at multiple locations, it also introduces higher computational complexity. Additionally, as the subspaces become finely divided, there is a risk of introducing bias when combining learned weights.

Based on the experimental findings, selecting the optimal number of heads becomes crucial to strike a balance between the amount of information captured and the computational complexity. From Tables 4–5,

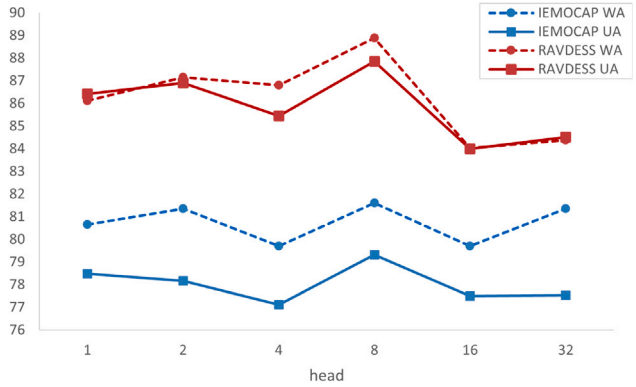

Fig. 9. Experimental results of the number of heads.

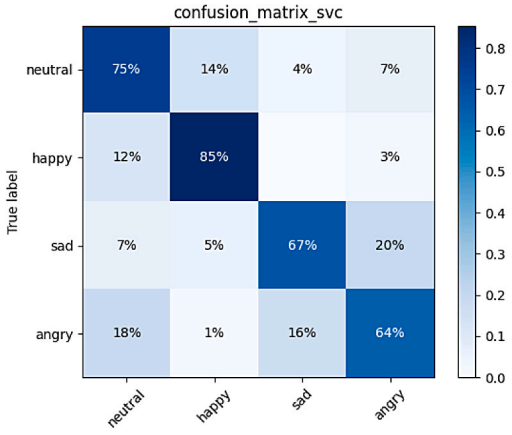

Fig. 10. Results on IEMOCAP using Transformer module.

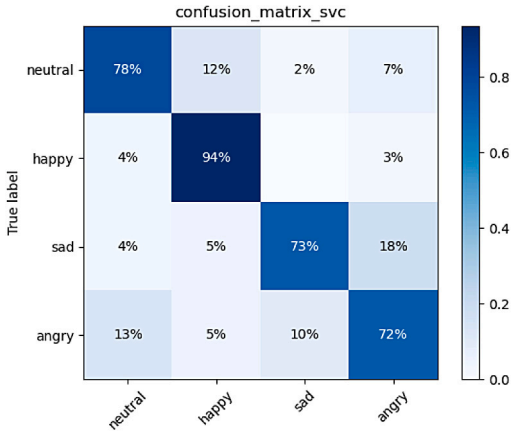

Fig. 11. Results on IEMOCAP using LSTM module.

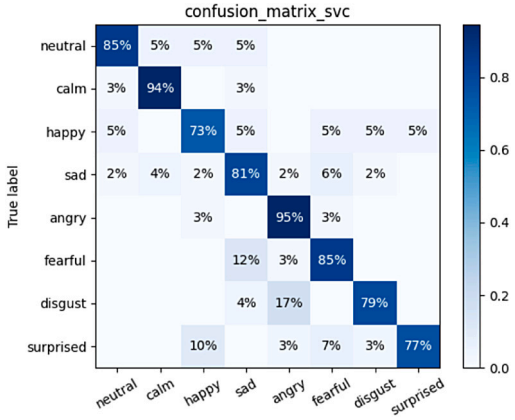

Fig. 12. Results on RAVDESS using Transformer module.

it is evident that the best overall performance is achieved when the number of heads is set to 8.

These results highlight the importance of carefully tuning the number of heads in the multi-head attention mechanism to optimize the model's performance. The selection of an appropriate number of heads ensures the model's ability to effectively capture relevant sentiment information while maintaining a manageable computational load.

5.5.2. BiLSTM, LSTM and transformer module comparison experiment

In this section, we conducted an experiment to compare the effectiveness of the BiLSTM network with the LSTM network, and Transformer network. We replaced the BiLSTM network in the model with either an LSTM or Transformer network and evaluated the performance on both the IEMOCAP and RAVDESS datasets.

Based on the experimental results presented in Tables 6 and 7, the BiLSTM network outperforms the LSTM and Transformer networks across all evaluation metrics in both databases. This can be attributed to the fact that while the Transformer network is known for its exceptional parallel processing capacity, parallel processing is not a crucial requirement for speech emotion recognition. Speech data exhibits strong temporal dependencies between adjacent time steps, which necessitates the extraction of detailed and deep temporal features. Additionally, the BiLSTM network surpasses the LSTM network in capturing contextual temporal information. While the LSTM network only considers past context, the BiLSTM network simultaneously considers both past and future contexts. This bidirectional approach allows the BiLSTM network to leverage a wider range of information and gain a comprehensive understanding of the temporal characteristics of speech data. As a result, the BiLSTM network generates more robust and accurate prediction results.

Table 6
Comparison of Transformer,LSTM and BiLSTM results on IEMOCAP (%).

Modules	WA	UA	Macro precision	Weighted precision	Macro F1	Weighted F1
Transformer	76.41	74.75	74.45	76.30	74.52	76.27
LSTM	80.18	79.13	78.63	80.55	78.64	80.09
BiLSTM	**81.60**	**79.32**	**81.34**	**81.70**	**80.22**	**81.55**

Table 7
Comparison of Transformer,LSTM and BiLSTM results on RAVDESS (%).

Modules	WA	UA	Macro precision	Weighted precision	Macro F1	Weighted F1
Transformer	84.72	84.54	85.10	84.98	84.65	84.69
LSTM	85.76	85.22	85.87	86.07	85.25	85.64
BiLSTM	**88.88**	**87.85**	**89.23**	**89.20**	**84.65**	**84.69**

The confusion matrices in Figs. 10–11 and 12–13 visually represent the performance of the model in emotion classification on the IEMOCAP and RAVDESS datasets using the Transformer and LSTM modules, respectively. These figures support the quantitative results, illustrating the model's ability to classify emotions.

Based on the experimental findings, we can conclude that in the model architecture of this study, the BiLSTM module with its inherent directionality outperforms the LSTM module and Transformer module in capturing temporal dependencies in audio data. By considering both past and future context, the BiLSTM network effectively analyzes feature data and produces more accurate emotion predictions.

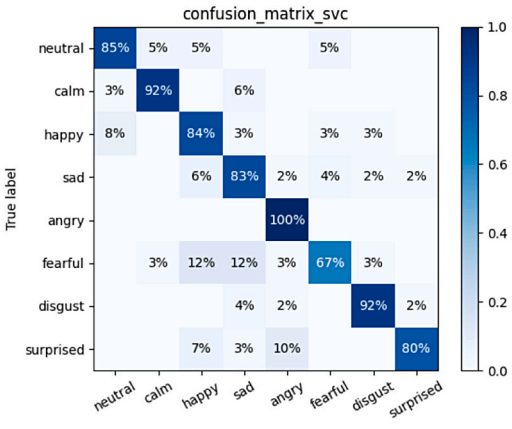

Fig. 13. Results on RAVDESS using LSTM module.

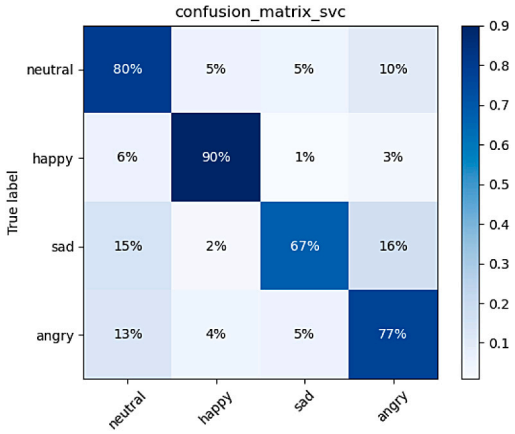

Fig. 14. Results on IEMOCAP using LLDs module.

Table 8
Ablation experiments were performed on HSFs feature on IEMOCAP (%).

Modules	WA	UA	Macro precision	Weighted precision	Macro F1	Weighted F1
LLDs	80.18	78.52	78.57	80.28	78.47	80.17
LLDs+HSFs	**81.60**	**79.32**	**81.34**	**81.70**	**80.22**	**81.55**

Table 9
Ablation experiments were performed on HSFs feature on RAVDESS (%).

Modules	WA	UA	Macro precision	Weighted precision	Macro F1	Weighted F1
LLDs	86.45	86.45	86.73	87.02	86.42	86.41
LLDs+HSFs	**88.88**	**87.85**	**89.23**	**89.20**	**88.22**	**88.82**

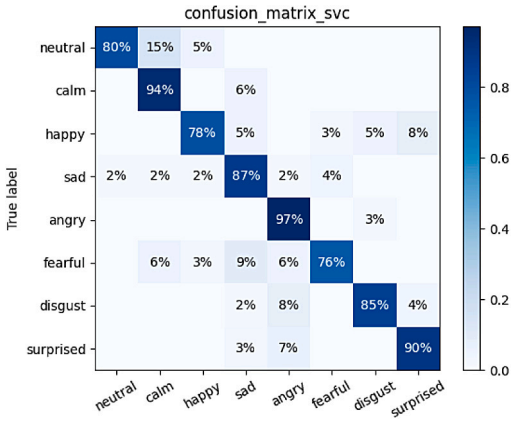

Fig. 15. Results on RAVDESS using LLDs module.

5.5.3. Multi-scale feature ablation experiments

In this section, we conducted ablation experiments to evaluate the effectiveness of multi-feature fusion for the SER task. We compared a frame-level deep emotional representation model with a multi-feature fusion model proposed in this study.

The results presented in Tables 8 and 9 demonstrate that the multi-feature fusion model outperforms the frame-level feature model in all evaluation metrics on both the IEMOCAP and RAVDESS datasets. This finding highlights the effectiveness of multi-feature fusion in improving the performance of speech emotion recognition. The confusion matrices in Figs. 14 and 15 provide visual representations of the frame-level feature model's performance on the IEMOCAP and RAVDESS datasets, respectively.

Frame-level features primarily analyze emotional feature information based on inter-frame variation, capturing fine-grained temporal dynamics. In contrast, utterance-level feature representation captures global information of speech sequences by removing distracting information such as vocal patterns and irrelevant semantics. The experiments reveal that by combining and complementing different types of features, the performance of speech emotion recognition can be significantly improved.

Tables 8 and 9 specifically focus on the ablation experiments performed on the High-level Statistical Features (HSFs). The comparison between Low-level Descriptors (LLDs) and LLDs+HSFs demonstrates that the inclusion of HSFs leads to improved performance on both datasets.

These findings highlight the importance of multi-feature fusion, which allows for a comprehensive analysis of different aspects of speech signals, capturing both local dynamics and global patterns. The proposed multi-feature fusion model leverages the complementary nature of frame-level and utterance-level features, resulting in enhanced SER performance.

5.6. Sensitivity factor

The effectiveness of segment-level speech emotion multiplexing decision-making in this study was found to depend on the optimal segmentation size. When the segments were too small, it was not possible to capture recognizable speech emotion features. On the other hand, when the segments were too large, they could not effectively make segment-level speech emotion multiplexing decisions. Additionally, changing the size of the segmented segments resulted in changes to the input frame-level feature dimensions. To better capture time–frequency information in frame-level features, parallel convolutional layers were set up in this paper. However, the convolutional kernel settings needed to be adapted for different frame numbers. Thus, several experiments were conducted to determine the optimal values for the multi-scale convolutional neural network. The specific experimental results are presented in Figs. 16–19, and the results for the two datasets are shown separately in Tables 10 and 11 due to their differences in average duration and audio signals.

Table 10 presents the parameter settings and corresponding performance on the IEMOCAP dataset. The segment length refers to the duration of each segment, while Conv Parameter represents the configuration of the parallel convolutional layers. Among the various parameter settings, the highest values for all evaluation metrics were achieved when the segment length was set to 2 s (Frame-level feature dimension is (76,59)) and the parallel convolution kernels were set to $(7 * 1)$ and $(1 * 5)$.

Similarly, Table 11 displays the parameter settings and corresponding performance on the RAVDESS dataset. Again, the segment_length

Table 10
Parameter settings on the IEMOCAP dataset.

Segment_length (LLDs dimensionality)	Conv parameter	WA (%)	UA (%)	Macro precision	Weighted precision	Macro F1	Weighted F1
	(3 * 1) (1 * 3)	79.40	75.98	80.45	79.54	77.53	79.06
	(5 * 1) (1 * 5)	76.85	73.77	76.60	76.73	74.87	76.60
1.6 s	(7 * 1) (1 * 7)	76.64	74.07	77.63	76.78	75.20	76.30
(76,47)	(5 * 1) (1 * 3)	78.12	74.34	75.18	78.10	74.44	77.80
	(7 * 1) (1 * 3)	77.07	74.55	73.36	78.35	74.75	76.79
	(7 * 1) (1 * 5)	79.46	76.55	76.07	79.66	76.01	79.31
	(3 * 1) (1 * 3)	77.00	72.96	77.79	77.16	74.92	76.68
	(5 * 1) (1 * 5)	77.23	72.88	77.11	77.52	74.56	76.89
1.8 s	(7 * 1) (1 * 7)	78.12	73.96	77.15	78.02	75.36	77.92
(76,57)	(5 * 1) (1 * 3)	78.12	74.34	75.18	78.10	74.44	77.80
	(7 * 1) (1 * 3)	75.89	73.49	73.24	76.54	72.40	75.44
	(7 * 1) (1 * 5)	79.46	76.55	76.83	79.59	75.88	79.17
	(3 * 1) (1 * 3)	79.48	78.73	78.20	78.74	78.30	79.43
	(5 * 1) (1 * 5)	79.95	77.59	81.17	80.51	78.66	79.83
2.0 s	(7 * 1) (1 * 7)	78.38	76.82	77.56	78.34	76.91	78.06
(76,59)	(5 * 1) (1 * 3)	79.71	77.26	78.27	79.43	77.68	79.50
	(7 * 1) (1 * 3)	79.24	75.75	78.62	79.41	76.55	78.97
	(7 * 1) (1 * 5)	**81.60**	**79.32**	**81.34**	**81.70**	**80.22**	**81.55**
	(3 * 1) (1 * 3)	74.25	71.99	71.24	74.66	71.26	74.20
	(5 * 1) (1 * 5)	74.75	73.74	69.99	75.71	71.01	74.65
2.2 s	(7 * 1) (1 * 7)	77.00	71.95	73.83	76.55	72.49	76.53
(76,65)	(5 * 1) (1 * 3)	74.50	70.78	71.34	74.30	70.96	74.33
	(7 * 1) (1 * 3)	73.25	71.38	70.15	73.42	70.45	73.05
	(7 * 1) (1 * 5)	77.50	72.68	74.38	77.03	73.02	76.88

Table 11
Parameter settings on the RAVDESS dataset.

Segment_length (LLDs dimensionality)	Conv parameter	WA (%)	UA (%)	Macro precision	Weighted precision	Macro F1	Weighted F1
	(3 * 1) (1 * 3)	84.37	83.09	83.97	84.76	83.37	84.41
	(5 * 1) (1 * 5)	85.06	84.52	85.18	85.27	84.71	85.02
1.2 s	(7 * 1) (1 * 7)	85.76	84.92	86.61	86.14	85.39	85.65
(76,34)	(5 * 1) (1 * 3)	86.80	86.11	87.26	87.46	86.42	86.83
	(7 * 1) (1 * 3)	86.80	86.30	86.92	87.43	86.44	86.76
	(7 * 1) (1 * 5)	87.15	87.01	87.76	86.93	87.24	86.72
	(3 * 1) (1 * 3)	87.84	87.53	87.66	88.24	87.43	87.89
	(5 * 1) (1 * 5)	86.80	85.96	88.22	87.24	86.63	86.69
1.4 s	(7 * 1) (1 * 7)	84.37	84.13	84.34	84.63	84.08	84.36
(76,40)	(5 * 1) (1 * 3)	84.37	83.44	83.88	84.86	83.68	84.66
	(7 * 1) (1 * 3)	**88.88**	**87.85**	**89.23**	**89.20**	**88.22**	**88.82**
	(7 * 1) (1 * 5)	84.37	83.44	83.51	84.29	83.36	84.25
	(3 * 1) (1 * 3)	84.72	84.44	85.59	86.25	84.38	84.78
	(5 * 1) (1 * 5)	86.11	86.55	86.46	86.47	86.24	86.04
1.6 s	(7 * 1) (1 * 7)	85.06	84.78	84.55	85.37	84.44	84.97
(76,47)	(5 * 1) (1 * 3)	82.29	81.74	81.65	82.49	80.99	82.00
	(7 * 1) (1 * 3)	86.80	87.22	86.71	87.34	86.22	86.78
	(7 * 1) (1 * 5)	83.68	83.45	82.86	83.74	83.03	83.56
	(3 * 1) (1 * 3)	84.37	82.53	85.84	85.10	82.52	83.66
	(5 * 1) (1 * 5)	85.76	85.33	85.49	85.81	85.21	85.59
1.8 s	(7 * 1) (1 * 7)	83.33	82.92	83.54	84.00	82.89	83.36
(76,57)	(5 * 1) (1 * 3)	84.37	83.94	83.64	84.61	83.58	84.27
	(7 * 1) (1 * 3)	84.72	83.99	84.84	85.40	84.02	84.67
	(7 * 1) (1 * 5)	85.06	84.50	85.60	85.80	84.52	84.95

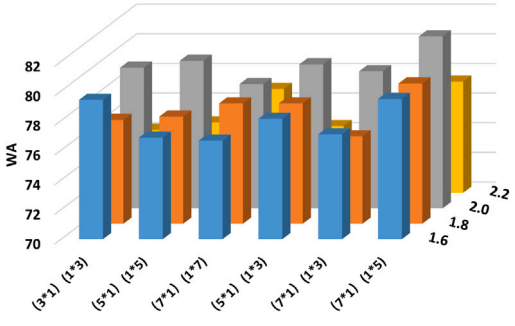

Fig. 16. WA values on IEMOCAP.

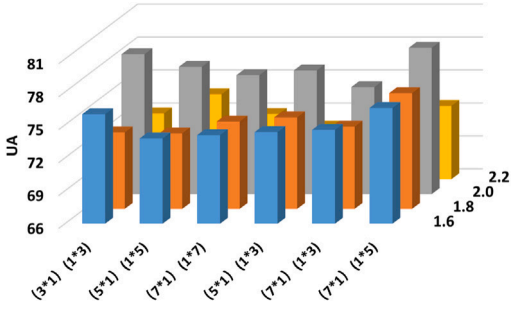

Fig. 17. UA values on IEMOCAP.

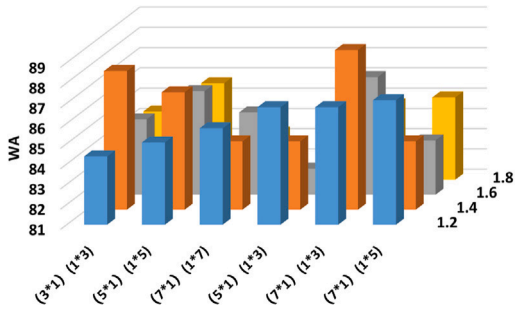

Fig. 18. WA values on RAVDESS.

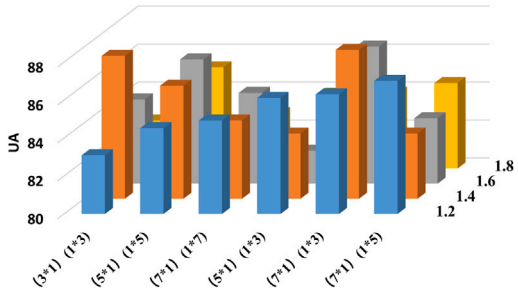

Fig. 19. UA values on RAVDESS.

and Conv Parameter indicate the segment duration and parallel convolutional layer configuration, respectively. For the RAVDESS dataset, the highest values for all evaluation metrics were achieved when the segment length was set to 1.4 s (Frame-level feature dimension is (76,40)) and the parallel convolution kernel is set to ($7 * 1$) and ($1 * 3$).

The experimental results demonstrate that appropriate segmentation of the audio signal followed by multiplexed decision-making can enhance the performance of speech emotion recognition. Moreover, setting different scales of convolutional layers significantly affects the ability to capture time–frequency information of frame-level features.

6. Conclusions

Existing sentiment analysis methods often rely on single-frame or sentence-level features, which limits their ability to effectively leverage the diverse range of sentiment features available at different types, levels, and scales. In response to this limitation, this paper introduces a novel network architecture. The proposed approach addresses the issue by establishing separate modules for learning deep sentiment features at the frame level and the utterance level. This design enables the acquisition of spatio-temporal representations for frame-level features and global representations for utterance-level features. Additionally, the paper presents a feature fusion method that utilizes a multi-head attention mechanism to combine the advantages of feature fusion at different levels. Furthermore, the paper incorporates a segment-level multiplex decision-making process to enhance the robustness of speech emotion recognition. To validate the effectiveness of the proposed model, extensive experiments were conducted on two datasets: IEMOCAP and RAVDESS. The experimental results demonstrate the superior performance of the proposed model. In conclusion, the proposed network architecture overcomes the limitations of existing methods by effectively integrating various sentiment features and leveraging their full potential. The experimental evaluations on IEMOCAP and RAVDESS datasets provide strong evidence of the model's effectiveness and superior performance.

Abbreviations

The following abbreviations are used in this manuscript:

AVEC	Audio/Visual Emotion Challenge
BiLSTM	Bidirectional Long Short-Term Memory Network
CNN	Convolutional Neural Network
CA	Calibration-Attention
DNNs	Deep Neural Networks
FA	Focus-Attention
FL	Frame-Level Deep Sentiment Feature
GeMAPs	Geneva Minimalistic Acoustic Parameter Set
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
HSFs	High-Level Semantic Features
kNN	K-Nearest Neighbors
LPCC	Linear Predictor Cepstrum Coefficient
Log-Mel	Log-Mel Frequency Spectrum
LSTM	Long Short-Term Memory
LLDs	Deep Neural Networks
LFBE	Log Mel-Filter Bank Energies
MLP	Multilayer Perceptron
MFCC	Mel Frequency Cepstrum Coefficient
PBM	Plain Bayesian Models
RF	Random Forests
RNNs	Recurrent Neural Networks
SER	Speech Emotion Recognition
SVM	Support Vector Machine
SAM	Self-Attention Module
STLR-SER	Spatio-temporal Representation Learning Enhanced Speech Emotion Recognition with Multi-head Attention Mechanisms
UA	Unweighted Accuracy
UL	Utterance-Level Deep Sentiment Feature
VTLP	Vocal Tract Length Perturbation
WA	Weighted Accuracy

CRediT authorship contribution statement

Zengzhao Chen: Validation, Methodology, Funding acquisition, Conceptualization. **Mengting Lin:** Writing – original draft, Conceptualization. **Zhifeng Wang:** Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Qiuyu Zheng:** Investigation, Conceptualization. **Chuan Liu:** Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] I. Shahin, O.A. Alomari, A.B. Nassif, I. Afyouni, I.A. Hashem, A. Elnagar, An efficient feature selection method for arabic and english speech emotion recognition using grey wolf optimizer, Appl. Acoust. 205 (2023) 109279.

[2] J. Liu, Z. Liu, L. Wang, L. Guo, J. Dang, Speech emotion recognition with local-global aware deep representation learning, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 7174–7178.

[3] Y. Zhong, Y. Hu, H. Huang, W. Silamu, A lightweight model based on separable convolution for speech emotion recognition, in: INTERSPEECH, Vol. 11, 2020, pp. 3331–3335.

[4] J. Liu, H. Wang, A speech emotion recognition framework for better discrimination of confusions, in: Interspeech, 2021, pp. 4483–4487.

[5] M. Chen, X. He, J. Yang, H. Zhang, 3-D convolutional recurrent neural networks with attention model for speech emotion recognition, IEEE Signal Process. Lett. 25 (10) (2018) 1440–1444.

[6] Z. Chen, J. Li, H. Liu, X. Wang, H. Wang, Q. Zheng, Learning multi-scale features for speech emotion recognition with connection attention mechanism, Expert Syst. Appl. 214 (2023) 118943, http://dx.doi.org/10.1016/j.eswa.2022.118943.

[7] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. André, C. Busso, L.Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan, K.P. Truong, The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, IEEE Trans. Affect. Comput. 7 (2) (2016) 190–202, http://dx.doi.org/10.1109/TAFFC.2015.2457417.

[8] M.F. Valstar, J. Gratch, B.W. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic, AVEC 2016 - depression, mood, and emotion recognition workshop and challenge, 2016, CoRR abs/1605.01600. arXiv:1605.01600. URL: http://arxiv.org/abs/1605.01600.

[9] T.L. Nwe, S.W. Foo, L.C. De Silva, Speech emotion recognition using hidden Markov models, Speech Commun. 41 (4) (2003) 603–623.

[10] Y. Attabi, M.J. Alam, P. Dumouchel, P. Kenny, D. O'Shaughnessy, Multiple windowed spectral features for emotion recognition, in: IEEE International Conference on Acoustics, 2013.

[11] L. Chen, W. Su, Y. Feng, M. Wu, J. She, K. Hirota, Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction, Inform. Sci. 509 (2020) 150–163, http://dx.doi.org/10.1016/j.ins.2019.09.005.

[12] Y.-L. Lin, G. Wei, Speech emotion recognition based on HMM and SVM, in: 2005 International Conference on Machine Learning and Cybernetics, Vol. 8, 2005, pp. 4898–4901, http://dx.doi.org/10.1109/ICMLC.2005.1527805, Vol. 8.

[13] K. Wang, N. An, B.N. Li, Y. Zhang, L. Li, Speech emotion recognition using Fourier parameters, IEEE Trans. Affect. Comput. 6 (1) (2015) 69–75, http://dx.doi.org/10.1109/taffc.2015.2392101.

[14] R.B. Lanjewar, S. Mathurkar, N. Patel, Implementation and comparison of speech emotion recognition system using Gaussian mixture model (GMM) and K- nearest neighbor (K-NN) techniques, Procedia Comput. Sci. 49 (2015) 50–57.

[15] K. Han, Y. Dong, I. Tashev, Speech emotion recognition using deep neural network and extreme learning machine, in: Interspeech, 2014.

[16] S. Mirsamadi, E. Barsoum, C. Zhang, Automatic speech emotion recognition using recurrent neural networks with local attention, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 2227–2231.

[17] M. Farooq, F. Hussain, N.K. Baloch, F.R. Raja, H. Yu, Y.B. Zikria, Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network, Sensors 20 (21) (2020) 6008.

[18] D. Luo, Y. Zou, D. Huang, Investigation on joint representation learning for robust feature extraction in speech emotion recognition, in: Interspeech, 2018, pp. 152–156.

[19] U. Kumaran, S. Radha Rammohan, S.M. Nagarajan, A. Prathik, Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN, Int. J. Speech Technol. 24 (2021) 303–314.

[20] L. Guo, L. Wang, J. Dang, Z. Liu, H. Guan, Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine, IEEE Access 7 (2019) 75798–75809.

[21] K.V. Krishna, N. Sainath, A.M. Posonia, Speech emotion recognition using machine learning, in: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), 2022, pp. 1014–1018, http://dx.doi.org/10.1109/ICCMC53470.2022.9753976.

[22] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomed. Signal Process. Control 47 (2019) 312–323.

[23] W.Y. Choi, K.Y. Song, C.W. Lee, Convolutional attention networks for multimodal emotion recognition from speech and text data, in: Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), 2018, pp. 28–34.

[24] T. Afouras, J.S. Chung, A. Zisserman, The conversation: Deep audio-visual speech enhancement, 2018, arXiv preprint arXiv:1804.04121.

[25] Q. Mao, M. Dong, Z. Huang, Y. Zhan, Learning salient features for speech emotion recognition using convolutional neural networks, IEEE Trans. Multimedia 16 (8) (2014) 2203–2213.

[26] J. Schluter, T. Grill, Exploring data augmentation for improved singing voice detection with neural networks, 2015, pp. 121–126.

[27] B. Zhou, K. Richardson, Q. Ning, T. Khot, A. Sabharwal, D. Roth, Temporal reasoning on implicit events from distant supervision, 2020, arXiv preprint arXiv:2010.12753.

[28] Z. Yang, X. Du, A. Rush, C. Cardie, Improving event duration prediction via time-aware pre-training, 2020, arXiv preprint arXiv:2011.02610.

[29] F. Chen, LSTM fully convolutional networks for time series classification, arXiv preprint arXiv:1709.05206.

[30] D. Hu, X. Hu, X. Xu, Multiple enhancements to LSTM for learning emotion-salient features in speech emotion recognition, in: Proc. Interspeech 2022, 2022, pp. 4720–4724.

[31] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, S. Zafeiriou, Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5200–5204, URL: https://api.semanticscholar.org/CorpusID:206742471.

[32] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al., Conformer: Convolution-augmented transformer for speech recognition, 2020, arXiv preprint arXiv:2005.08100.

[33] L. Tarantino, P.N. Garner, A. Lazaridis, et al., Self-attention for speech emotion recognition, in: Interspeech, 2019, pp. 2578–2582.

[34] A. Nediyanchath, P. Paramasivam, P. Yenigalla, Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7179–7183, http://dx.doi.org/10.1109/ICASSP40776.2020.9054073.

[35] S. Yoon, S. Byun, S. Dey, K. Jung, Speech emotion recognition using multi-hop attention mechanism, 2019, pp. 2822–2826.

[36] Q. Zheng, Z. Chen, H. Liu, Y. Lu, J. Li, T. Liu, MSRANet: Learning discriminative embeddings for speaker verification via channel and spatial attention mechanism in alterable scenarios, Expert Syst. Appl. 217 (C) (2023) http://dx.doi.org/10.1016/j.eswa.2023.119511.

[37] D. Li, J. Liu, Z. Yang, L. Sun, Z. Wang, Speech emotion recognition using recurrent neural networks with directional self-attention, Expert Syst. Appl. 173 (2021) 114683.

[38] Q. Chen, G. Huang, A novel dual attention-based BLSTM with hybrid features in speech emotion recognition, Eng. Appl. Artif. Intell. 102 (2021) 104277.

[39] M. Xu, F. Zhang, X. Cui, W. Zhang, Speech emotion recognition with multiscale area attention and data augmentation, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 6319–6323.

[40] H. Zou, Y. Si, C. Chen, D. Rajan, E.S. Chng, Speech emotion recognition with co-attention based multi-level acoustic information, 2022, pp. 7367–7371.

[41] J. Kim, Y. An, J. Kim, Improving speech emotion recognition through focus and calibration attention mechanisms, 2022, arXiv preprint arXiv:2208.10491.

[42] J. Lei, X. Zhu, Y. Wang, BAT: Block and token self-attention for speech emotion recognition, Neural Netw. 156 (2022) 67–80.

[43] W. Chen, X. Xing, X. Xu, J. Pang, L. Du, DST: Deformable speech transformer for emotion recognition, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[44] D. Priyasad, T. Fernando, S. Sridharan, S. Denman, C. Fookes, Dual memory fusion for multimodal speech emotion recognition, in: Proc. INTERSPEECH 2023, 2023, pp. 4543–4547.

[45] S. Kwon, et al., Att-net: Enhanced emotion recognition system using lightweight self-attention module, Appl. Soft Comput. 102 (2021) 107101.

[46] T. Tuncer, S. Dogan, U.R. Acharya, Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques, Knowl.-Based Syst. 211 (2021) 106547.

[47] K. Mustaqeem, A. El Saddik, F.S. Alotaibi, N.T. Pham, AAD-net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network, Knowl.-Based Syst. 270 (2023) 110525.

[48] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J.M. Montero, F. Fernández-Martínez, A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset, Appl. Sci. 12 (1) (2021) 327.