Neurocomputing 612 (2025) 128711

Contents lists available at ScienceDirect

Neurocomputing

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

ESERNet: Learning spectrogram structure relationship for effective speech emotion recognition with swin transformer in classroom discourse analysis

Tingting Liu [a,b,1], Minghong Wang [a,1], Bing Yang [b,*], Hai Liu [b,c], Shaoxin Yi [c]

[a] Faculty of Education, The University of Hong Kong, 999077, Hong Kong
[b] School of Education, Hubei University, Wuhan 430062, China
[c] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China

ARTICLE INFO

Keywords:
Speech emotion recognition
Intelligent education
Feature extraction
Swin Transformer
Classroom discourse analysis

ABSTRACT

Speech emotion recognition (SER) has received increased attention due to its extensive applications in many fields, especially in the analysis of teacher-student dialogue in classroom environment. It can help teachers to better learn about students' emotions and thereby adjust teaching activities. However, SER has faced several challenges, such as the intrinsic ambiguity of emotions and the complex task of interpreting emotions from speech in noisy environments. These issues can result in reduced recognition accuracy due to a focus on less relevant or insignificant features. To address these challenges, this paper presents ESERNet, a Transformer-based model designed to effectively extract crucial clues from speech data by capturing both pivotal cues and long-range relationships in speech signal. The major contribution of our approach is a two-pathway SER framework. By leveraging the Transformer architecture, ESERNet captures long-range dependencies within speech mel-spectrograms, enabling a refined understanding of the emotional cues embedded in speech signals. Extensive experiments were conducted on the IEMOCAP and EmoDB datasets, the results show that ESERNet achieves state-of-the-art performance in SER and outperforms existing methods by effectively leveraging critical clues and capturing long-range dependencies in speech data. These results highlight the effectiveness of the model in addressing the complex challenges associated with SER tasks.

1. Introduction

Classroom dialogue is a crucial component of the teaching and learning process. Though classroom interaction, teachers can convey knowledge and students can share their opinion and confusion, promoting knowledge construction and mutual understanding [1,2]. However, the complexity of classroom dialogue extends beyond the exchange of words, it also involves emotional communication, which plays an important role in learning environment and may influence the motivation, engagement, and academic performance of students. Speech emotion recognition (SER) technology offers a novel approach to analyzing and understanding the emotional aspects of classroom dialogue. SER aims at providing high-accuracy emotion classification predictions for human, which can be used in the field of intelligent education [3,4]. The study of SER has gained increasing attention due to the vital role of emotional states in interpersonal communication and its numerous applications. It has garnered significant interest from various down-pathway tasks, including classroom dialogue analysis [5,6], education [1,7], healthcare [8], and intelligent voice assistants [9]. Furthermore, automatic emotion classification technology can assist in monitoring emotions and interpreting data in psychological treatment, such as depression analysis [10], which is useful for taking care of the emotion state of students and effective protection can be offered through targeted treatment if their hidden emotional states are detected. However, SER remains a challenging task due to the ambiguous nature of emotions and the presence of background noise.

To this end, numerous studies [1,11–13] on SER have been conducted over the past decades, elevating the task to a new level. Firstly, machine learning-based methods such as support vector machines (SVM) [14] were predominant in SER. SVMs were also combined with hidden Markov models (HMM) [15] for this purpose. However, with the emergence of deep learning, its exceptional performance in various tasks quickly captured the attention of many SER researchers. Consequently, deep learning-based methods, including convolutional neural networks

* Corresponding author.
E-mail addresses: yangbing.cn88@gmail.com (B. Yang), hailiu0204@mail.ccnu.edu.cn (H. Liu).
[1] These authors (Tingting Liu and Minghong Wang) contributed equally to this work.

https://doi.org/10.1016/j.neucom.2024.128711
Received 12 July 2024; Received in revised form 16 September 2024; Accepted 1 October 2024
Available online 9 October 2024

(CNN) [16], recurrent neural networks (RNN) [17], and long short-term memory networks (LSTM) [18], have been widely utilized in SER, achieving impressive results. Recently, the Transformer model has brought innovation to nearly every field, known for its exceptional ability to learn long-range dependencies. As a result, many Transformer-based methods [19–21] have been proposed, demonstrating satisfying performance [22]. In general, existing methods can be roughly categorized into two families: handicraft feature-based (HFB) and deep learning-based (DLB) methods.

For HFB methods, acoustic features such as pitch [23], intensity [24], and prosody [25] are extracted from speech signals, after which machine learning algorithms are employed for emotion classification. One approach [26] introduces a pre-processing step that clusters speech segments by formant characteristics and uses phoneme occurrence rates as features. Wu et al. [27] proposed utilizing modulation spectral features combined with prosodic features for SER. Some researchers [24] have investigated generating feature representations by combining acoustic features with lexical features for SER. An innovative SER framework [28] utilizes an adversarial joint loss strategy and integrates a cascaded attention network, enabling the extraction of significant emotional attributes. This approach efficiently grasps the extensive temporal dependencies within targeted regions.

When compared to previously discussed HFB methods, DLB methods exhibit greater advancement and promise, attributed to their capacity to leverage global-level signal data and mitigate sentiment analysis inaccuracies arising from feature selection bias. Among DLB methods, CNNs serve as the most widespread backbone, encompassing architectures like ResNet [29], AlexNet [30], and LIGHT-SERNET [31]. Lei et al. [32] introduced an efficacious technique called MsEmoTTS, which delves into the interplay of emotional characteristics across various levels to facilitate multiscale emotional speech synthesis. Li et al. [33] explores the use of unsupervised representation learning for speech SER to address the challenge of limited labeled datasets and archive outstanding concordance correlation coefficient performance on emotion primitives by applying contrastive predictive coding on unlabeled data.

Extensive experimentation has revealed that speech synthesis techniques grounded in speech and text surpass those reliant on audio and text, respectively, in sentiment analysis. A different methodology [34] devised a temporal alignment mechanism utilizing mean-max pooling, offering a straightforward yet potent network design that facilitates precise capture of discourse nuances and enables emotion computation via textual data for cross-excitement analysis. Zhu et al. [35] present a unique global-aware multi-scale neural network tailored for SER, which acquires multi-scale feature depictions and incorporates a global-aware fusion module to seize crucial emotional cues. Ye et al. [36] introduce an innovative technique for amalgamating features across various temporal scales. To tackle the intrinsic challenge of feature depiction in language-driven emotion recognition, a novel approach known as the attention mechanism-based multiscale SER network [37] was introduced. This groundbreaking framework incorporates a parallel network structure that efficiently merges fine-grained, frame-level attributes with coarse-grained, utterance-level depth features. Furthermore, the SER task harnesses the exceptional capabilities of the Transformer encoder [38], which relies on multi-head attention. Zhang et al. [39] proposed an unsupervised pre-training model base on Transformer and verified the validity of proposed model. The approach of fusing features also archives inspiring performance. Zou et al. [40] employ multi-level acoustic information and a co-attention mechanism to fuse features extracted from MFCC, spectrogram and wav2vec2. While DLB methods significantly reduce the labor cost of dataset annotation and outperform traditional methods, numerous challenges remain in SER.

Drawing from its successes in computer vision and natural language processing, the Transformer architecture has been adapted for speech emotion recognition (SER) to derive representations from mel-spectrograms [13]. Dutta et al. [41] presented a novel multi-modal emotion recognition technique that amalgamates audio and text through a learnable audio front-end (LEAF) model and BERT-derived text embeddings, which are then processed using a bidirectional GRU and self-attention mechanisms. Zhang et al. [42] introduced a feature fusion model, Dual-TBNet, comprising two 1D convolutional layers, two Transformer modules, and two BiLSTM modules, preserving speech data and ensuring robust feature extraction. He et al. [44] developed an SER system utilizing a cross-attention transformer (CAT) to integrate three acoustic features: raw waveform, spectrogram, and MFCC. The adoption of the Transformer architecture in this study is attributed to its proficiency in capturing feature correlations.

Unlike prior SER research, which mainly concentrated on utilizing general emotion features, our approach offers a unique viewpoint by elucidating the distinctive attributes of speech. Our motivation is twofold: identifying pivotal indicators of specific emotions and differentiating nuances for classifying similar emotions. To harness these insights, we propose a token-mask strategy that entails concealing linguistic information and subsequently predicting the concealed segments, thereby augmenting our model's feature learning capacity. The main contributions of our work are summaried as three aspects:

1) We proposed a novel ESERNet model to exploit observations in speech emotion signals. To our knowledge, this model is the first to uncover pivotal cues and long-range semantic relationships in speech signals. Furthermore, the Transformer architecture is leveraged to investigate mel-spectrogram relationships.
2) We introduce a two-pathway model to capture both pivotal cues and long-range relationships in speech signals. Specially, the crucial cue pathway learned key features by recovering the masked spectrogram and relationship pathway can get multi-scale features and learn the long-distance dependencies by swin Transformer.
3) Experiments conducted on the IEMOCAP and EmoDB datasets demonstrate that the proposed ESERNet model outperforms several state-of-the-art methods, confirming its effectiveness.

2. Analysis of speech signal characteristics

2.1. Speech emotion recognition

SER focuses on analyzing emotional states conveyed through voice, which aids in understanding human emotions more deeply. Recent years have seen numerous SER methods [43] [44]. Peng et al. [45] introduce an effective neural network framework for speech emotion recognition, leveraging multi-scale convolutional layers and statistical pooling to extract and enhance acoustic and lexical features using an attention module. Sun et al. [46] present the Multimodal Cross- and Self-Attention Network, which employs cross- and self-attention mechanisms to model interactions between textual and acoustic information effectively. Xu et al. [47] propose a deep convolutional neural network that utilizes multi-scale area attention to capture emotional features at various granularities, combined with vocal tract length perturbation for data augmentation. Despite the impressive performance of these methods, they have limitations, such as insufficient comprehensiveness in feature extraction. In [40], Zou et al. tackle this challenge by presenting an end-to-end SER system that amalgamates various acoustic features, such as MFCC, spectrogram, and high-level embeddings, utilizing a unique co-attention mechanism. Nevertheless, this methodology neglects the equilibrium among different granularities and the interplay of components within the signal. Shen et al. [48] introduced a novel model, the Emotion Neural Transducer (ENT), tailored for fine-grained SER and employing lattice max pooling to differentiate emotional frames. In order to mitigate these constraints, Chen et al. [49] unveiled the SFormer, a Transformer-based framework designed for the speech emotion recognition task. While these studies have contributed to the progress of SER, little research has delved into learning long-range dependencies in speech. Consequently, our research aims to harness long-range

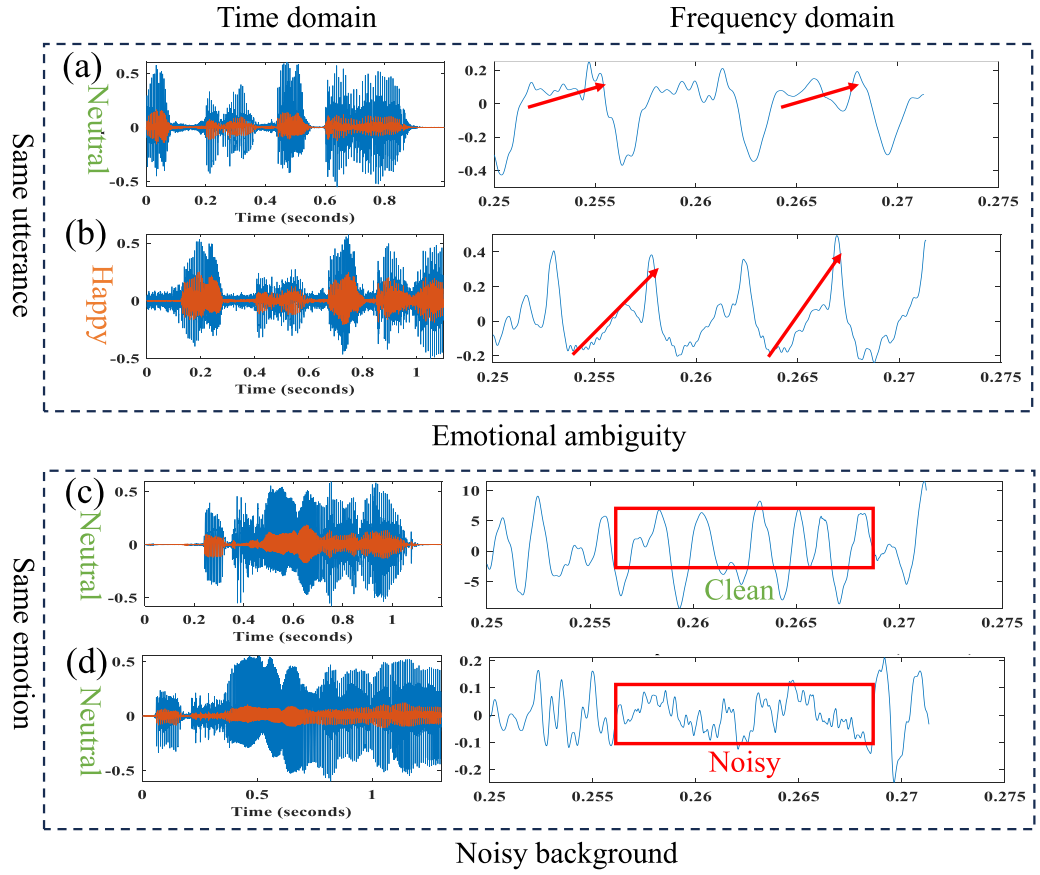

Fig. 1. Two challenges exist in SER. In the left column are time-domain plots, while those in the right column are frequency-domain plots. (a), (b): Emotional ambiguity: The same utterance can be expressed with different emotions. (c), (d): Noisy background: The same speech emotion signal can be interfered by background noises.

dependencies more efficiently to augment SER performance. Despite the efficacy of CNNs in capturing speech signal features and their successful application in various computer vision tasks, two core challenges persist for CNN-based SER.

i) *Ambiguous emotions*: The emotional content in speech can vary significantly with changes in the environment (Fig. 1(a), (b)). Depending on whether the atmosphere is friendly or impersonal, the same utterance can be expressed with different tones and speeds, making it difficult to recognize linguistic emotions in different contexts. Therefore, the ambiguity of speech emotions in different communication environments poses a significant challenge for SER.
ii) *Noisy background*: Communication often takes place in complex and diverse environments. As shown in Fig. 1(c), a neutral emotion can be perceived against a clear background, whereas Fig. 1(d) shows a neutral emotion against a noisy background. In noisy environments, the presence of multiple overlapping sounds makes it difficult to distinguish the target voice from the background, let alone recognize the speaker's particular emotion.

2.2. Observation and motivation

By carefully analyzing the spectrogram of the voice signal, we identify several characteristics that address these challenges. As shown in Fig. 2, speeches exhibit various emotions, such as happiness, sadness, anger, and neutrality. The features of each spectrogram differ from one another, with some emotions sharing similarities in pitch, tone length, tone repetition, and timbre. Although their spectrograms may appear similar, crucial features that can be leveraged for SER are always exist.

Finding I: *Key cues of speech emotions.* Emotions cannot be easily categorized based on surface-level features alone. However, some core features within each class are distinct from others. In Fig. 2(a), we show that the mel-spectrogram is corresponding to an angry speech emotion. While this mel-spectrogram displays various attributes, the analysis of these features could potentially introduce ambiguous data, thereby presenting a challenge for SER. To alleviate this issue, it is crucial to recognize long-range dependencies within spectral components, which are regarded as key cues.

Finding II: *Minuscule discrepancies of different emotions.* In the mel-spectrograms depicting diverse emotions, slight variations are frequently noted. These nuanced distinctions constitute a distinctive attribute in SER, rendering the detection of fine-grained, decisive features imperative. In Fig. 2, it shows that mel-spectrograms of angry emotions display stronger energy levels compared to others when the same statement. Conversely, the arrow indicating a point on the mel-spectrogram representing sadness shows weaker energy. These inconspicuous yet significant features are easily overlooked, highlighting that certain features are more meaningful and discriminative than others.

The findings above can be summarized as the problem of identifying crucial cues and long-range dependencies. We argue that it is essential to recognize the critical cues of the same emotion that remain consistent across different settings, as well as the subtle yet significant long-range dependencies often overlooked by existing approaches. Based on these two findings, our primary motivations are how to focus on critical minority regions and how to uncover the relationships of long-range dependencies. To this end, we proposed a two-pathway model comprising a crucial cue pathway for learning key features and a relationship pathway for capturing long-range information. These two pathways

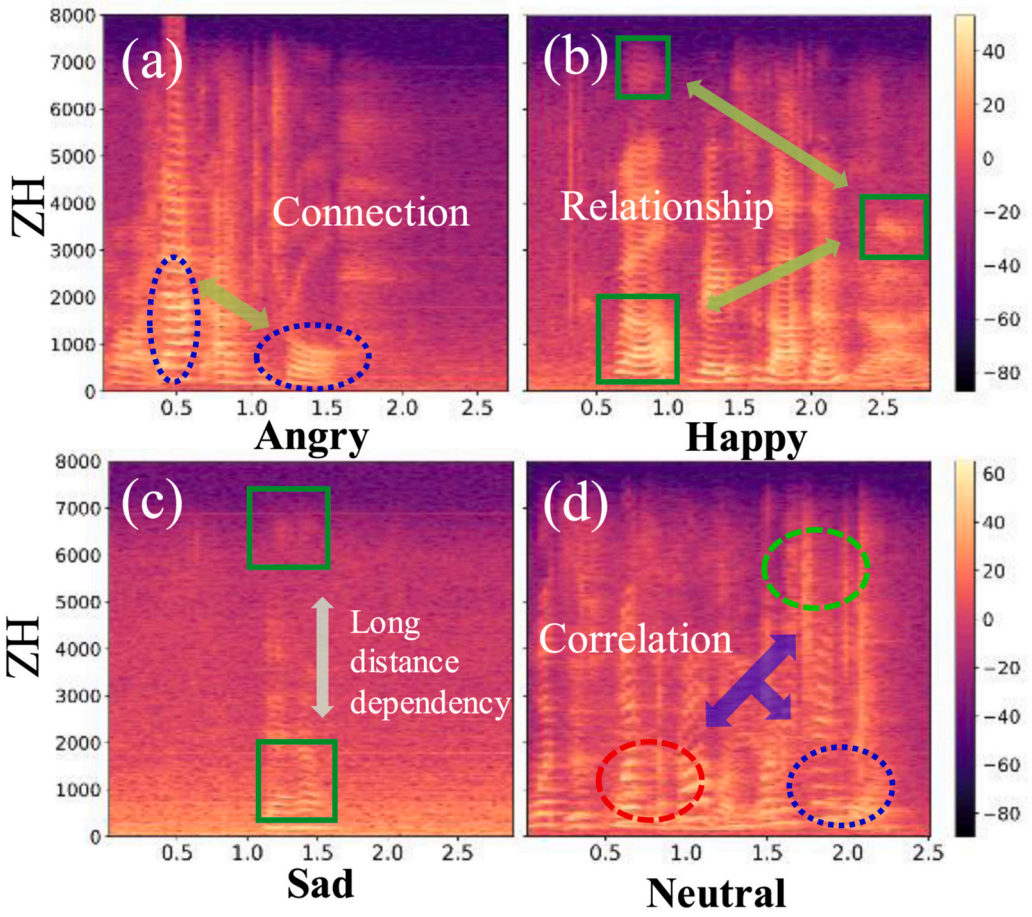

Fig. 2. presents mel-spectrograms depicting various speech emotions. The shade of color reflects variations in energy (dB). The rectangles highlight critical cues and long-range dependencies within the speech signal information.

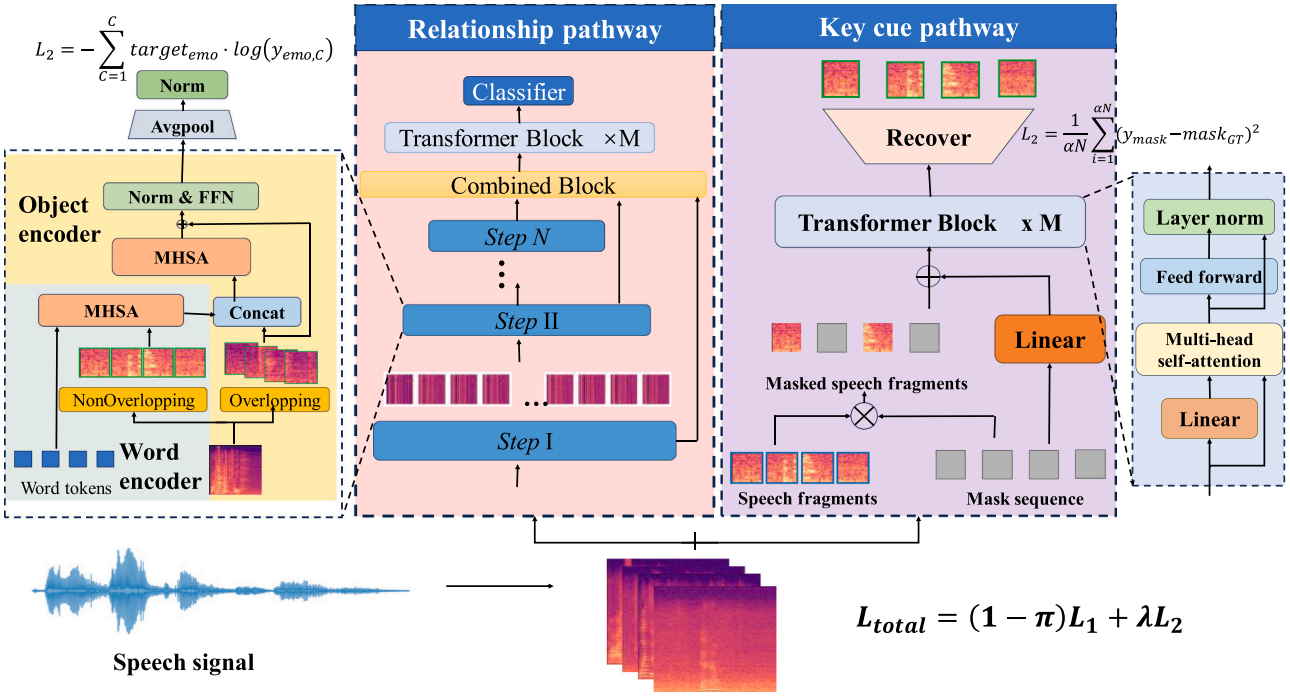

Fig. 3. Two-pathway architecture of the proposed ESERNet model.

work collaboratively to enhance the performance of SER.

3. Proposed ESERNet model

3.1. Overview

We propose a novel two-pathway network for speech emotion recognition with Swin Transformers (called as ESERNet). The framework of ESERNet is illustrated in Fig. 3. In this network, each pathway addresses a different subtask, and the pathways are subsequently fused to solve the final SER task. The relationship pathway processes log-mel features to learn long-range dependencies in log-mel signals. Meanwhile, the crucial cue pathway uses partially masked log-mel features as input to predict missing values. Specifically, for input speech signals, they are first transformed into spectrograms and then fed into the two pathways. In the crucial cue pathway, the input spectrogram is masked by the masked sequence and the Transformer blocks are used to recover the spectrogram and in this progress this pathway model can learn the crucial and discriminative features. In the relationship pathway, spectrograms and word tokens are fed into the word encoder to capture coarse-grained information and then the object encoder is used to capture and encode information about consecutive frames of a speech signal across multiple levels of granularity. After that, a joining block is applied to fuse the two parts of features and the fused is fed into Transformer blocks to find the long-distance dependencies. All components of ESERNet are divisible, allowing our proposed modules to be easily adapted for use in networks designed for other tasks.

3.2. Key cue pathway

The crucial components of a speech signal encompass not only semantics and tones but also convey contextual information and the relationships between semantics and tones. Word-level relationships help identify and locate the connections between words. However, capturing key features is essential for determining the sentiment of a statement when these relationships become unreliable or nonexistent. To tackle this challenge, we propose the ESERNet model, specifically engineered to grasp intricate relationships, thereby bolstering the precision of SER. Our approach incorporates a vital cue pathway, comprising Transformer blocks, which scrutinizes these pivotal cues with the aim of enhancing the precision of masked value recovery.

Within this key cue pathway, the initial input speech segment is denoted as $s = [s_1, s_2, \ldots, s_T]$, where T signifies the sequence's length. Subsequently, a binary mask sequence, sharing the same dimensionality as the input sequence, is generated for the purpose of the mask recovery task. In this mask sequence, a value of 1 signifies that the respective time step ought to be masked, whereas a value of 0 signifies that it should be preserved. Consequently, the mask sequence is articulated as $m = [m_1, m_2, \ldots, m_T]$.

Prior to feeding the input into the model, a partially masked input is created by amalgamating the generated mask sequence with the initial feature sequence. The masked information, derived through the mask processing, can be formulated as $s_m = s \odot (1 - m)$, where $\odot$ signifies element-wise multiplication. Once s_m is obtained, the binary mask is mapped to align with the dimensionality of the original feature in the masking information embedding layer. This mapped projection is subsequently integrated into the input, expressed as,

$$s' = s_m + Linear(m), \tag{1}$$

where $Linear(\bullet)$ denotes a linear layer responsible for embedding occlusion information m into the input. Through the utilization of the multi-head self-attention (*MSAT*) mechanism, the model is capable of dynamically harnessing the relationship between the known and masked segments. This process can be articulated as follows:

$$s_{out} = LNorm(MSAT(s', s', s', m)), \tag{2}$$

where, s_{out} is directed into a feed-forward network (FFN) to extract more advanced features, thus improving the understanding of pivotal attributes. This can be formulated as:

$$s'_{out} = LNorm(s_{out} + FFN(s_{out})). \tag{3}$$

3.3. Relationship pathway

1) Encoding

Firstly, we introduce a network module that integrates a word encoder and an object encoder framework, aiming to comprehensively learn both the macroscopic and microscopic attributes of speech signals. The specifics of this framework will be elaborated in the subsequent subsections.

For capturing the macroscopic details from the raw speech signal, we suggest employing a word encoder. More specifically, we initially generate a series of trainable word tokens: $t_1 \in R^{N_x \times W_1}$ for *Step* I, $t_2 \in R^{N_x \times W_2}$ for *Step* II and $t_3 \in R^{N_x \times W_3}$ for *Step* III, where N_x approximates the number of words in the utterance. The tokens for the latter two steps are generated by the combined block. Subsequently, the input variable s_i is partitioned into N_x non-overlapping, uniformly distributed segments, which can be represented as:

$$[n_{i1}, n_{i2}, \ldots, n_{iN_x}] = NOL\left(s_i, \frac{N_i}{N_x}\right), \tag{4}$$

where $NOL(\bullet)$ means the non-overlapping segmentation. n_{ij} is the j-th non-overlapping segment of s_1, and $j \in [1, N_x]$. t_i^j denotes the j-th word token in s_i, and t_i^j is the updated value of t_i^j.

Then, each word token captures coarse-grained characteristics from different segments of the speech signal. This operation is expressed as:

$$n_{ij} = s_i\left[\frac{N_i}{N_x} \times (j-1) : \frac{N_i}{N_x} \times j\right], \tag{5}$$

the token t_i^j is then passed through the object encoder at various steps to effectively leverage the coarse-grained information during the modeling process,

$$t_i^j = MSAT\left(t_i^j, n_{ij}, n_{ij}\right), \tag{6}$$

where $MSAT(\bullet)$ means multi-head self-attention.

2) Object encoder

Within the object encoder, the initial input speech signal undergoes a transformation into acoustic representations, denoted as $a_1 \in R^{N_1 \times W_1}$, where N_1 signifies the number of frames and W_1 represents the dimensionality of each frame embedding. In order to extract information pertaining to sequential frames in *Step* I, we employ an object encoder with a window size of N_{oi} to capture details about contiguous frames within a_i. This procedure can be articulated as:

$$\begin{cases} \left[a_{i_1}, a_{i_2}, \ldots, a_{i_{N_i}}\right] = Overlapping(a_i, N_{oi}), \\ a_{ij} = a_i\left[j - \dfrac{N_{oi}}{2} : j + \dfrac{N_{oi}}{2}\right], \quad j \in [1, N_i], \end{cases} \tag{7}$$

where $Overlapping(\bullet)$ represents the overlapping segmentation, the subscript i denotes *Steps* I–III, and $a_i[x:y]$ consists of the x-th to the y-th tokens of s_i. To allow the object encoder to incorporate coarse-grained information, the learnable tokens t_i are passed to each word encoder at every step. Consequently, the attention mechanism for each segment can be expressed as:

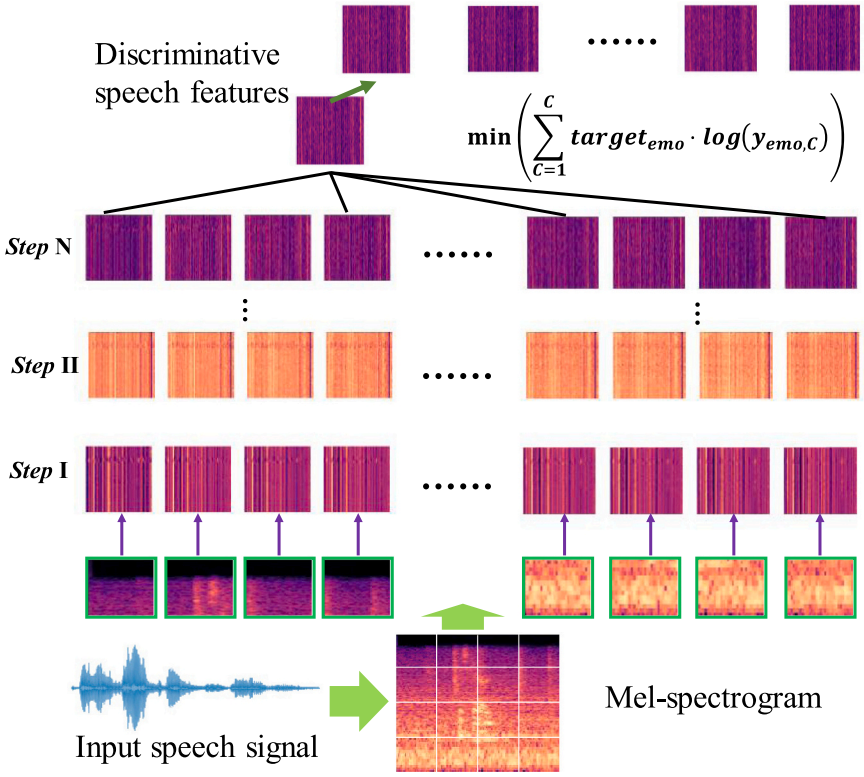

Fig. 4. Training process of the proposed relationship pathway.

Table 1
Experimental results on IEMOCAP by SFormer, AMSNet, ISNet, TIM-Net, CA-MSER, ShiftCNN, ENT, and our ESERNet models.

Models	ACC (%)	WA (%)	UA (%)
SFormer [49]	71.00	70.50	71.5
AMSNet [37]	69.87	69.22	70.51
TIM-Net [36]	72.25	72.50	71.65
CA-MSER[40]	72.17	71.64	72.70
ShiftCNN [52]	73.79	74.80	72.80
ENT [48]	72.94	72.43	73.88
ESERNet	**76.85**	**76.79**	**77.12**

Table 2
Experimental results on EmoDB by GM-TCNet, AMSNet, TIM-Net, SCAR-NET and our ESERNet methods.

Models	UA	WA	ACC (%)
GM-TCNet [53]	90.48	91.39	90.85
AMSNet [37]	88.56	88.34	88.64
TIM-Net [36]	89.19	90.28	89.95
SCAR-NET [51]	-	-	**96.45**
MelTrans [13]	92.50	92.47	92.52
ESERNet	**92.73**	**92.81**	93.85

$$\begin{cases} h_i^{jz} = Concat\left(t_i^{(z)}, a_{ij}\right), \\ g_i^{jz} = MSAT\left(a_i^{(j)}, h_i^{jz}, h_i^{jz}\right), \\ a_i^{(j)} = Norm\left(g_i^{jz} + a_i^{(j)}\right), \end{cases} \tag{8}$$

where $h_i^{jz} \in R^{(1+N_{oi})\times W_1}$ is the enhanced segmention, $z = ceil\left[\frac{j\times N_a}{N_i}\right]$ rounds a number upward to its nearest integer, and $j \in [1, N_i]$. The resultant a is then fed into an FFN that can be expressed as:

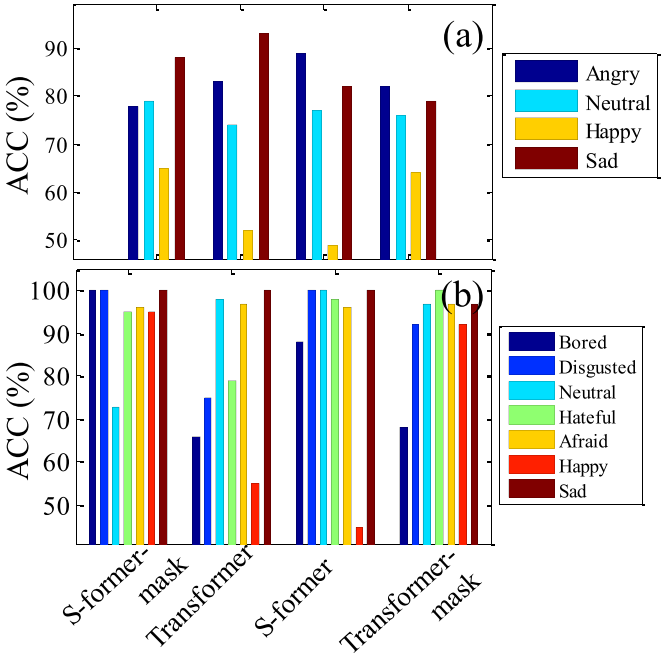

Fig. 5. Performance comparison on two emotion dataset by four different network architectures. (a) IEMOCAP dataset. (b) EmoDB dataset.

$$Step_i = Norm(FFN(a_i) + a_i). \tag{9}$$

The total training process of the relationship pathway is illustrated in Fig. 4.

During the progression from *Steps* I to III, each step concentrates on a distinct level of detail, advancing sequentially from frames to consonants to words. The input attributes are represented as $I_i \in R^{N_i\times W_i}$,

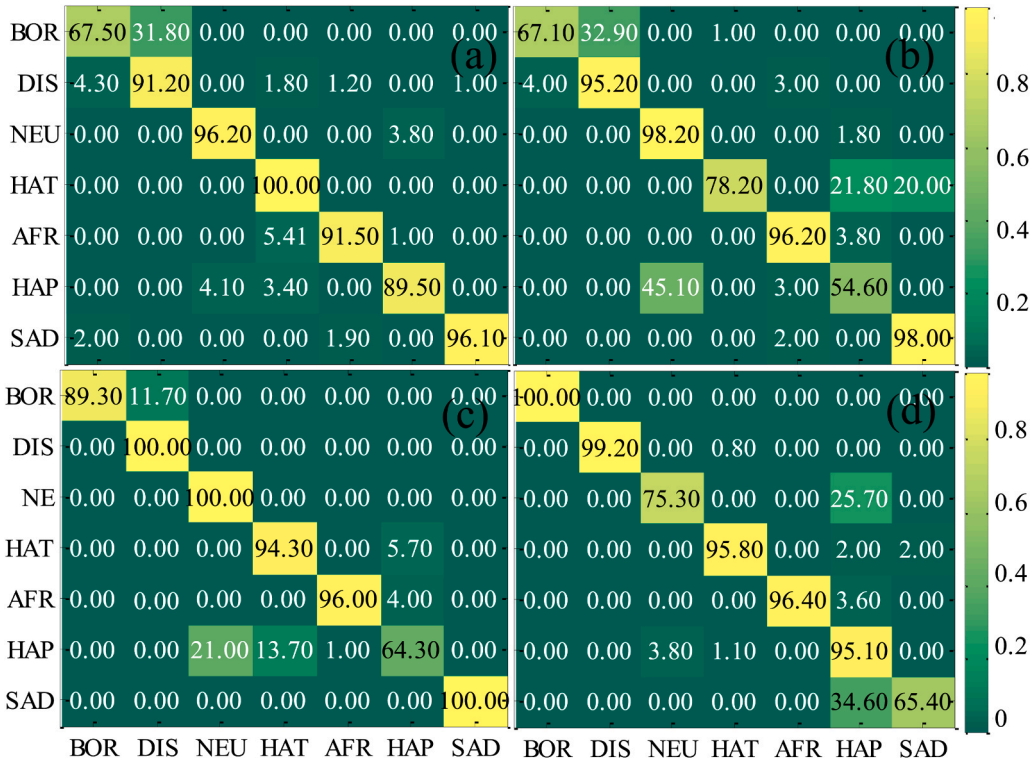

Fig. 6. Confusion matrix analysis of model variants. First row: IEMOCAP dataset. Second row: EmoDB dataset. (a) ST. (b) ST+crucial cue pathway. (c) Relationship pathway. (d) ESERNet.

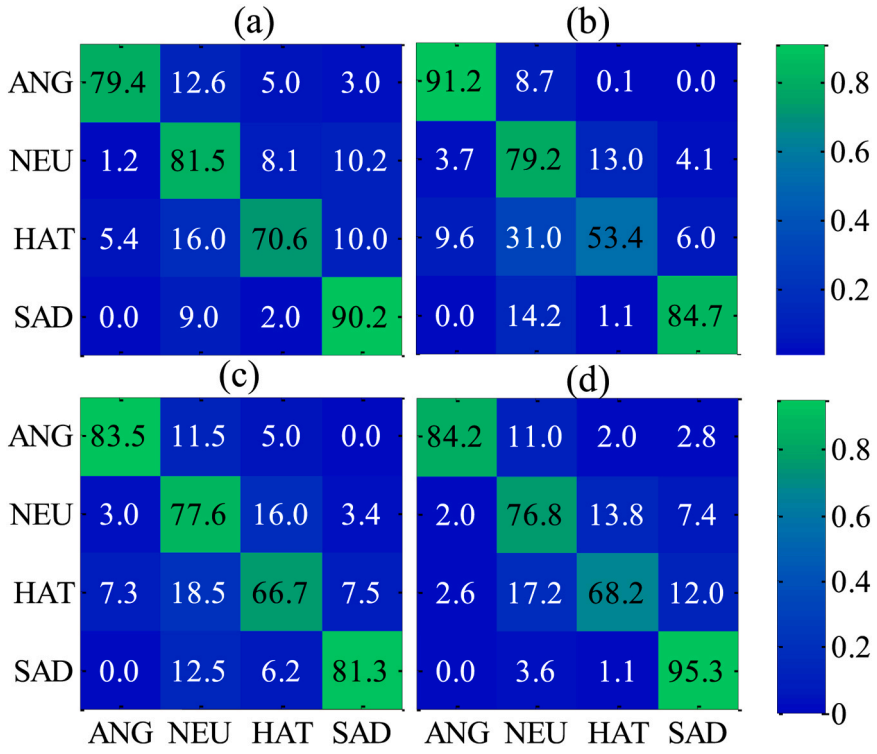

Fig. 7. Confusion matrix analysis of various model variants. (a) ESERNet, (b) ST combined with the crucial cue pathway, (c) ST alone, and (d) Relationship pathway. The first row depicts results from the IEMOCAP dataset, while the second row displays those from the EmoDB dataset.

andi $\in [1, \quad 3]$. N_i signifies the number of tokens with varying levels of granularity, while W_i denotes the associated embedding dimension. Each N_i encapsulates a granular depiction of its respective step, generated by the federated block and transmitted to the subsequent step. Distinct window sizes N_o are utilized for each step to model the interplay between each granularity and its adjacent elements.

Table 3
Performance on the IEMOCAP dataset with different speech emotions, such as the HAP, SAD, ANG, NEU. Here, the 'TRF-mask' denotes the 'Transformer-mask'.

Models	HAP (%)	SAD (%)	ANG (%)	NEU (%)
S-former	52.31	93.62	82.94	74.56
Transformer	55.62	94.53	80.27	72.96
TRF-mask	64.56	84.32	79.83	75.63
ESERNet	**65.48**	**88.76**	**78.34**	**79.15**

3) Combined block

The speech signal advances through *Steps* I to III, with each step highlighting unique feature levels. To efficiently produce pertinent features, we introduce a combined block to be applied between each step pair. This mechanism employs average pooling on the output values of each step and ascertains the combined scale $Q_i, i \in [1, \quad 2]$ (Q_1 of 50 ms and Q_2 of 250 ms), based on the granularity characteristic of each step. Following this, linear projection and layer normalization are executed to derive the granular feature for $Step_i$ (where $i \in \{\mathrm{I}, \mathrm{II}\}$) for the subsequent step, $Step_{i+1}$. This ensures the integration of information from various minimum durations into tokens within s_i, with each token embodying the granular feature of $Step_{i+1}$. Ultimately, the merge scale Q_3 for the union block, which is applied to the output value s_3 of *Step* III, is established at 800 ms, approximating the word count in the utterance sample. Word tags, which represent coarse-grained features within words, are excluded from aggregation. This procedure can be formulated as:

$$\begin{cases} Step_i = Norm(CovPool(Step_i, Q_i)O_i + m_i) \\ cout_i = Norm(cout_i O_i + m_i) \end{cases}, \tag{10}$$

where $CovPool(s, Q)$ signifies a covariance pooling layer applied to s, with both the window size and stride set to Q. The parameters $O_i \in R^{W_i \times W_{i+1}}$ are subject to learning. The outputs of the i-th step are represented by $Step_i$ and $cout_i$, while the inputs for the subsequent step are denoted as x_{i+1} and z_{i+1}, where $i \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}\}$. The output of the concluding combined block is subsequently processed by a series of standard Transformer (ST) encoders to capture the speech signal at a global level. The aggregated output from the utterance step is then merged along the temporal axis and fed into the classifier. The classifier consists of two linear projections separated by an activation function, which produces the final classification result.

3.4. Loss construction of the proposed model

The categorical mean-squared error (MSE) loss is selected as the objective function in the mask pathway. The MSE loss is defined as,

$$L_1 = \frac{1}{\alpha N} \sum_{i=1}^{\alpha N} (y_{mask} - mask_{GT})^2 \tag{11}$$

where the symbol α denotes the mask rate. And y_{mask} and $mask_{GT}$ mean the recovered information of the model's output, and the groundtruth of the cooresponding masked part, respectively.

Then, the categorical cross-entropy (CCE) loss is selected as the objective function in the emotion pathway. The CCE loss is constructed as,

$$L_2 = -\sum_{C=1}^{C} target_{emo} \cdot \log\left(y_{emo,C}\right), \tag{12}$$

where C represents the total number of emotion categories. The ground truth label value for class C is denoted as $target_{emo}$, while $y_{emo,C}$ signifies the model's output probability for class C.

To harmonize the training goals of the speech emotion pathway and the mask pathway, their respective losses are combined with suitable weights. During the training process, the model concurrently learns two tasks pertinent to SER (Speech Emotion Recognition) and mask recovery. Consequently, the overall loss function can be illustrated as follows:

$$L_{total} = (1 - \pi)L_1 + \pi L_2, \tag{13}$$

Table 6
Comparison the performance between the proposed ESERNet method and the SOTA methods.

Methods	Year	Backbones	Speech Features	Performance
GM-TCNet [53]	2022	CNN	MFCC	★★★★☆
CA-MSER [40]	2022	CNN, LSTM, Transformer	MFCC, Spectrogram, High-level acoustic	★★★☆☆
AMSNet [37]	2023	LSTM + attention	Temporal and spatial feature	★★☆☆☆
SpeechFormer [49]	2023	Transformer	Acoustic feature and text	★★☆☆☆
TIM-Net [36]	2023	CNN	MFCC	★★★★☆
ShiftCNN [52]	2023	CNN	Acoustic feature	★★★★☆
ENT [48]	2024	RNN	Acoustic feature	★★★☆☆
ESERNet	2024	Transformer	Mel-spectrogram	★★★★★

Table 4
Effect of the proposed modules performance on different emotions on the EmoDB dataset.

Models	HAP (%)	AFR (%)	NEU (%)	SAD (%)	DIS (%)	BOR (%)	HAT (%)
S-former	95.78	97.12	74.03	100.00	100.00	100.00	95.67
Transformer	90.47	93.66	97.32	98.61	72.63	67.32	76.56
TRF-mask	50.23	94.09	94.51	97.48	90.24	70.25	96.54
ESERNet	**46.34**	**95.72**	**100.00**	**100.00**	**100.00**	**100.00**	**97.83**

Table 5
Performance comparison on the EmoDB and IEMOCAP datasets by Recall, F1 and Accuracy.

Datasets	Model variants	SpeechFormer	Mask	Recall	F1	Accuracy
IEMOCAP	ST pathway	×	×	0.728	0.743	0.718
	S-former pathway	✓	×	0.741	0.759	0.736
	Transformer-mask	×	✓	0.755	0.757	0.745
	ESERNet	✓	✓	**0.778**	**0.781**	**0.7685**
EmoDB	ST pathway	×	×	0.845	0.865	0.857
	S-former pathway	✓	×	0.899	0.901	0.903
	ST+key cue pathway	×	✓	0.907	0.911	0.908
	ESERNet	✓	✓	**0.931**	**0.932**	**93.85**

where π is a hyperparameter that balances crucial cue pathway loss and relationship pathway loss. Finally, the detailed implementation of crucial cue pathway and speech relationship pathway are presented in Algorithms 1 and 2, respectively.

Algorithm 1. . Crucial cue pathway for the proposed ESERNet model.

Input: Speech fragments $s = [s_1, s_2, \ldots, s_T]$, mask sequence $m = [m_1, m_2, \ldots, m_T]$
1: Mask the input speech fragments: $s_m = s \odot (1 - m)$
2: Embed the mask sequence information into s': $s' = s_m + Linear(m)$
// Transformer block
3: **While** $i < M$:
$s_{out} = LayerNorm(\ MSAT(s', s', s', m)\ )$
$s'_{out} = LayerNorm(FFN(s_{out}) + s_{out})$
end while
4: Recover the masked speech fragments:
$s_r = Recover(s')$
Output: Recovered speech fragments s_r

Algorithm 2. . Speech relationship pathway

Input: Speech fragments s, window size w for different steps, word token t
Set: Batch size s, η and $\square$, *parameters* β_1 and β_2
1: Initialize word token
//Step i
2: **While** $i<N$:
//word encoder
n=Nonoverlapping(s)
$n_{ij} = s_i \left[\frac{N_i}{N_x} \times (j-1) : \frac{N_i}{N_x} \times j\right]$
$t = MHSA(\mathrm{t}, \mathrm{n}, \mathrm{n})$
//object encoder
a=Overlapping(s)
$a_{ij} = a_i \left[j - \frac{N_{oi}}{2} : j + \frac{N_{oi}}{2}\right]$
h=*Concat*(t, a)
a=*Norm*(a + *MSAT*(a, h, h))
End while
3: Combined block
4:M transformer blocks similar to *Algorithm* 1
5: Classification
Output: Predicted labels .

4. Experimental results and discussion

To assess the efficacy of the ESERNet model, its performance across multiple dimensions was evaluated through a series of model comparison experiments.

4.1. General setting

1) Evaluation metrics: In this article, both weighted accuracy (WA) and unweighted accuracy (UA) are employed. Considering that the maximum values of WA and UA may not coincide within the same model, the final evaluation criterion is established by calculating the average of these two metrics [2].

A confusion matrix, serving as a square table, is leveraged to assess the performance of a classification algorithm. It offers a summary and visualization of the algorithm's performance by showcasing the counts of both correct and incorrect predictions. The diagonal entries within the confusion matrix signify accurate predictions, whereas the off-diagonal entries indicate erroneous ones. A substantial proportion of values along the diagonal suggests a high level of algorithm accuracy. Consequently, the confusion matrix reveals the extent to which the algorithm misclassifies instances.

2) Datasets: Two datasets (IEMOCAP and EmoDB) are utilized for training and testing our ESERNet model. **IEMOCAP** [50]: This dataset comprises audiovisual recordings of dyadic conversations involving ten actors (five males and five females), who engage in both scripted and improvised scenarios, totaling approximately 12 hours of content. The dataset includes 10,039 sentences, and our experiments focus on discourses labeled with four emotions: happy (HAP), neutral (NEU), angry (ANG), and sad (SAD). A significant challenge with this dataset is that many sentiments overlap, making it difficult to clearly distinguish between classes. **EmoDB**: It stands as one of the most frequently utilized databases in SER tasks. This dataset comprises 535 sentences, with 302 sentences representing female emotions and 233 sentences depicting male emotions.

3) Compared methods: To validate the proposed model, a variety of methods are selected in the SER task. We introduce several representative methods alongside different varieties of our proposed ESERNet method for the comparative experiments.

CA-MSER [40]: This work introduces an end-to-end SER system that use multi-level acoustic information and a novel co-attention module to understand human emotions conveyed solely through audio information. It also combined kinds of features including MFCC, spectrogram, and high-level features extracted by other models. **ENT** [48]: This work proposes the Emotion Neural Transducer (ENT) for fine-grained speech emotion recognition, which integrates emotion dynamics with automatic speech recognition joint training. By extending neural transducers with an emotion joint network and implementing lattice max pooling, the model effectively distinguishes emotional frames. **TIM-Net** [36]: This paper introduces a novel approach to SER called Temporal-aware bi-direction Multi-scale Network (TIM-Net), which improves human-machine interaction by modeling temporal patterns of emotions at multiple scales. TIM-Net uses temporal-aware blocks to capture affective representations, integrates information from both past and future contexts, and fuses features across different time scales. **SCAR-NET** [51]: The authors introduce an enhanced CNN designed for the extraction of emotional features from speech signals in the context of SER. The model utilizes three parallel pathways to capture spectral, temporal, and spectral-temporal correlation features. Multi-branch deep feature learning is facilitated through the application of split-convolve-aggregate residual blocks.

4.2. Experimental results and analysis

4.2.1. Experimental results on the IEMOCAP

Based on the IEMOCAP dataset, we performed experiments with various contemporary SER methods, and the outcomes are displayed in Table 1. TRIN demonstrates exceptional performance by taking into account the inherent hierarchy of the phonetic structure and its interconnections across different modalities under sequential temporal guidance. ENT effectively differentiates emotional frames by integrating emotion dynamics with joint training in automatic speech recognition. ESERNet, built upon the Transformer architecture, exhibits improvements over cutting-edge methods, highlighting the prowess of our model. In comparison to earlier approaches, our model attains a notable accuracy of 76.50 %, signifying its success in leveraging invariant cues and long-range semantic relationships within speech signals.

4.2.2. Experimental result on the EmoDB

A comparative analysis of the proposed ESERNet method with AMSNet, ENT, and SCAR-NET was also carried out utilizing the EmoDB dataset. The outcomes are summarized in Table 2. Our methodology exhibits a 92.5 % accuracy rate. The exceptional performance of SCAR-NET can be ascribed to its incorporation of split-convolve-aggregate residual blocks for multi-branch deep feature extraction, highlighting the significance of multiscale information in attaining improved SER results. Our approach harnesses information extending over long distances within speech signals, a factor that has been demonstrated to

augment performance. In comparison to LSTM- or CNN-based models, Transformer-based models display superior performance, primarily attributed to the Transformer architecture's inherent capacity to grasp long-range dependent semantic associations among all tokens.

4.3. Discussion

In this section, a comprehensive examination of the ESERNet model is provided, focusing on analyzing its constituents and documenting the discoveries. Considering that the relationship pathway (S-former) is rooted in the conventional Transformer (ST) framework, a comparative assessment is executed between these two architectures. Specifically, we evaluate the performance of the independent ST model on the SER task to ascertain the individual impacts of the elements within our dual-pathway configuration. Our objective is to enhance the comprehension of ESERNet's operational mechanisms and efficacy in emotion recognition endeavors, thus fostering the ongoing enhancement and refinement of cutting-edge SER models. Our study particularly underscores the ablation of the essential cue pathway and the relationship pathway. The findings, illustrated in Fig. 4, suggest that ESERNet exhibits resilient performance on both IEMOCAP and EmoDB datasets. Notably, our analysis accentuates the significance of the essential cue pathway in bolstering the model's overall performance, uncovering its crucial part within the ESERNet architecture.

4.3.1. Ablation study

In Fig. 5, we show the accuracy of different model versions evaluated on the EmoDB and IEMOCAP datasets. As illustrated in Fig. 5(a), the results on the EmoDB dataset indicate that ESERNet demonstrates commendable performance in most emotion categories, with the exception of neutral emotion classification, where its performance is somewhat lacking. The S-former model, illustrated in Fig. 4(a), encounters difficulty in distinguishing between happy and neutral speech emotions. According to the results displayed in Fig. 4(b) for the IEMOCAP dataset, the recognition accuracy for happy emotions is generally lower compared to other emotion categories. Both the S-former and ESERNet models consistently surpass the ST model in various emotions across both datasets.

4.3.2. Confusion matrix analysis

The confusion matrix shown in Fig. 5 indicates that the S-former model encounters challenges in differentiating between neutral and happy emotions, resulting in potential misclassifications. Upon evaluation on the IEMOCAP dataset, both the ESERNet and S-former models demonstrate enhanced performance in comparison to the ST model, particularly across diverse emotional categories. Further analysis, as depicted in Fig. 5, underscores a reciprocal interference between neutral and happy emotions, indicating that misclassifications between these two emotions are prevalent and frequently impede accurate emotion recognition. Consequently, the indistinct boundary between happy and neutral speech emotions poses a significant challenge, thereby reducing the accuracy of SER. These observations emphasize the complexity involved in identifying nuanced emotional differences.

4.3.3. Effect of the two-pathway strategy

To evaluate the effectiveness of the two pathway strategy, a multi-task model configuration was employed, encompassing both a non-mask model and a masked component for comparative analysis. This design assigned the non-mask model the exclusive task of handling the SER task, independent of any auxiliary analysis by a mask model for specific key features. As shown in Fig. 6, the ESERNet model with the masking strategy consistently exhibits superior performance. Furthermore, the optimal multitask performance is attributed to the S-former submodel and the mask word model.

Fig. 7 illustrates that the multitask model, incorporating a mask, surpasses both the standalone transform and S-former models on the IEMOCAP dataset. According to Table 5, the multitask model demonstrates an approximate 3 % enhancement in performance relative to its single-task counterpart, underscoring the robust capability of the crucial cue pathway to capture vital features, thereby substantially augmenting the overall efficacy of emotion recognition. Table 3 highlights the impressive performance of our ESERNet model in recognizing sad speech emotions within the IEMOCAP dataset. For the EmoDB dataset, Table 4 presents the accuracy of ESERNet and the S-former model for each speech emotion. While the accuracy for emotions such as bored, disgusted, and sad is adequate, it is relatively lower for neutral and happy emotions. The reduced scores for happy emotions can be ascribed to ESERNet's enhanced generalization across other emotions, coupled with a potential imbalance in the representation of angry and neutral emotions. Consequently, ESERNet may be more susceptible to acquiring less discriminative information for these particular emotion categories, resulting in decreased accuracy. Hence, the incorporation of the relationship pathway becomes a pivotal aspect in attaining substantial improvements in SER. Findings from Fig. 5 and Table 3, pertaining to the EmoDB dataset, further accentuate the remarkable effectiveness of the crucial cue pathway.

Furthermore, the comparisons of different methods are provided in Table 6. These columns include the method type, year, backbone and performance, so we can visually observe the difference. In general, ESERNet yields excellent performance among compared methods.

5. Conclusion

In this work, we propose a novel approach for speech emotion recognition by identifying significant features in speech signals. We uncover two key findings: crucial emotional cues and tiny discrepancies between different emotions. Fully leveraging these insights is critical to improving SER performance. To this end, we introduce ESERNet, an efficient method with two-pathway strategy. The critical cue pathway uses a masking strategy to extract essential speech cues, while the relationship pathway aggregates multiscale information from speech signals. Utilizing the Transformer architecture as its foundational structure, our aim is to reveal long-range semantic dependencies within speech signals. The proposed ESERNet model was assessed on two SER datasets. Experimental results show its proficiency in identifying pivotal cues and long-distance dependencies in speech signals. Nonetheless, certain constraints persist, notably the computational demands associated with ESERNet multi-task strategy. In the future, we plan to investigate the development of a streamlined version through the optimization of the framework.

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.

Acknowledgements

This work was supported in part by the Jiangxi Provincial Natural Science Foundation under Grant 20232BAB212026, and Grant 20242BAB25107; in part by the National Natural Science Foundation of Hubei Province under Grant 2022CFB529 and Grant 2022CFB971; in part by the University Teaching Reform Research Project of Jiangxi Province under Grant JXJG-23–27-6; and in part by the Shenzhen Science and Technology Program under Grant JCYJ20230807152900001.

Data Availability

Data will be made available on request.

References

[1] Q. Zheng, Z. Chen, et al., Automated multimode teaching behavior analysis: a pipeline-based event segmentation and description, IEEE Trans. Learn. Technol. 17 (2024) 1717–1733.
[2] Z. Chen, J. Li, et al., Learning multi-scale features for speech emotion recognition with connection attention mechanism, Expert Syst. Appl. 214 (2023) 118943.
[3] L. Chen, K. Wang, et al., K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human–robot interaction, IEEE Trans. Ind. Electron. 70 (2023) 1016–1024.
[4] X. Che, H. Yang, C. Meinel, Automatic online lecture highlighting based on multimedia analysis, IEEE Trans. Learn. Technol. 11 (2018) 27–40.
[5] T.M. Wani, T.S. Gunawan, et al., A comprehensive review of speech emotion recognition systems, IEEE Access 9 (2021) 47795–47814.
[6] Z. Liu, S.X. Yin, et al., Scaffolding language learning via multi-modal tutoring systems with pedagogical instructions, IEEE Conf. Artif. Intell. (CAI)2024 (2024) 1258–1265.
[7] K. Kaur, P. Singh, Trends in speech emotion recognition: a comprehensive survey, Multimed. Tools Appl. 82 (2023) 29307–29351.
[8] C. Zucco, B. Calabrese, M. Cannataro, Sentiment analysis and affective computing for depression monitoring, IEEE Int. Conf. Bioinforma. Biomed. (BIBM) (2017) 1988–1995.
[9] S.A.S. Subhash, et al., A and Santhosh, B, Artificial Intelligence-base Voice Assistant, 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), (2020) 593-596.
[10] M. Deshpande, V. Rao, Depression detection using emotion artificial intelligence, 2017 Int. Conf. Intell. Sustain. Syst. (ICISS) (2017) 858–862.
[11] A. Milton, et al., SVM scheme for speech emotion recognition using MFCC feature, Int. J. Comput. Appl. 69 (2013) 125–137.
[12] J. Zhao, X. Mao, et al., Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomed. Signal Process. Control 47 (2019) 312–323.
[13] H. Li, J. Li, et al., MelTrans: mel-spectrogram relationship-learning for speech emotion recognition via transformers, Sensors 24 (2024) 5506.
[14] P. Shen, et al., Automatic speech emotion recognition using support vector machine, Proc. 2011 Int. Conf. Electron. Mech. Eng. Inf. Technol. (2011) 621–625.
[15] M. Jain, S. Narayan, et al., Speech Emot. Recognit. Using Support Vector Mach. (2020).
[16] Z. Huang, M. Dong, Q. Mao, Y. Zhan, Speech emotion recognition using CNN, Proc. 22nd ACM Int. Conf. Multimed. (2014) 801–804.
[17] J. Lee, I. Tashev, High-level feature representation using recurrent neural network for speech emotion recognition. High-level feature representation using recurrent neural network for speech emotion recognition, Interspeech 2015, 2015.
[18] F. Tao, G. Liu, Advanced LSTM: a study about better time dependency modeling in emotion recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 2906–2910.
[19] J. Wagner, et al., Dawn of the transformer era in speech emotion recognition: closing the valence gap, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 10745–10759.
[20] F. Andayani, et al., Hybrid LSTM-transformer model for emotion recognition from speech audio files, IEEE Access 10 (2022) 36018–36027.
[21] Z. Lian, B. Liu, et al., CTNet: conversational transformer network for emotion recognition, IEEE/ACM Trans. Audio, Speech, Lang. Process. 29 (2021) 985–1000.
[22] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, Y. Ma, SwinFusion: cross-domain long-range learning for general image fusion via swin transformer, IEEE/CAA J. Autom. Sin. 9 (2022) 1200–1217.
[23] S. Chebbi, S.B. Jebara, On the use of pitch-based features for fear emotion detection from speech, 2018 4th Int. Conf. Adv. Technol. Signal Image Process. (ATSIP) (2018) 1–6.
[24] Q. Jin, C. Li, S. Chen, H. Wu, Speech emotion recognition with acoustic and lexical features. 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2015, pp. 4749–4753.
[25] D. Shome, A. Etemad, Speech emotion recognition with distilled prosodic and linguistic affect representations, ICASSP 2024 - 2024 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP) (2024) 11976–11980.
[26] Z.-T. Liu, A. Rehman, M. Wu, W.-H. Cao, M. Hao, Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence, Inf. Sci. 563 (2021) 309–325.
[27] S. Wu, T.H. Falk, et al., Automatic speech emotion recognition using modulation spectral features, Speech Commun. 53 (2011) 768–785.
[28] Y. Liu, H. Sun, et al., A discriminative feature representation method based on cascaded attention network with adversarial strategy for speech emotion recognition, IEEE/ACM Trans. Audio, Speech, Lang. Process. 31 (2023) 1063–1074.
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, 2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2016) 770–778.
[30] Y. Zhang, J. Du, et al., Attention based fully convolutional network for speech emotion recognition, 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), (IEEE2018), pp. 1771-1775.
[31] A. Aftab, A. Morsali, S. Ghaemmaghami, B. Champagne, LIGHT-SERNET: A Lightweight Fully Convolutional Neural Network for Speech Emotion Recognition, 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6912-6916.
[32] Y. Lei, S. Yang, X. Wang, L. Xie, MsEmoTTS: multi-scale emotion transfer, prediction, and control for emotional speech synthesis, IEEE/ACM Trans. Audio, Speech, Lang. Process. 30 (2022) 853–864.
[33] M. Li, B. Yang, et al., Contrastive Unsupervised Learning for Speech Emotion Recognition, 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6329-6333.
[34] H. Li, W. Ding, Z. Wu, Z. Liu, Learning fine-grained cross modality excitement for speech emotion recognition. Learning Fine-Grained Cross Modality Excitement for Speech Emotion Recognition, Interspeech, 2021, pp. 3375–3379.
[35] W. Zhu, X. Li, Speech Emotion Recognition with Global-Aware Fusion on Multi-Scale Feature Representation, 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6437-6441.
[36] J. Ye, X.-C. Wen, et al., Temporal modeling matters: a novel temporal emotional modeling approach for speech emotion recognition. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.
[37] Z. Chen, J. Li, H. Liu, et al., Learning multi-scale features for speech emotion recognition with connection attention mechanism, Expert Syst. Appl. 214 (2023).
[38] A. J. Wagner, et al., Dawn of the transformer era in speech emotion recognition: closing the valence gap, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 10745–10759.
[39] R. Zhang, H. Wu, et al., Transformer Based Unsupervised Pre-Training for Acoustic Representation Learning, 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6933-6937.
[40] H. Zou, Y. Si, et al., Speech Emotion Recognition with Co-Attention Based Multi-Level Acoustic Information, 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7367-7371.
[41] S. Dutta, S. Ganapathy, Multimodal Transformer with Learnable Frontend and Self Attention for Emotion Recognition, 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6917-6921.
[42] Z. Liu, X. Kang, F. Ren, Dual-TBNet: improving the robustness of speech features via dual-transformer-BiLSTM for speech emotion recognition, IEEE/ACM Trans. Audio, Speech, Lang. Process. 31 (2023) 2193–2203.
[43] J. Gideon, M.G. McInnis, E.M. Provost, Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG), IEEE Trans. Affect Comput. 12 (2021) 1055–1068.
[44] Y. Khurana, et al., RobinNet: a multimodal speech emotion recognition system with speaker recognition for social interactions, IEEE Trans. Comput. Soc. Syst. 11 (2024) 478–487.
[45] Z. Peng, Y. Lu, S. Pan, Y. Liu, Efficient speech emotion recognition using multi-scale CNN and attention, ICASSP 2021 - 2021 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP) (2021) 3020–3024.
[46] L. Sun, et al., Multimodal cross- and self-attention network for speech emotion recognition, ICASSP 2021 - 2021 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP) (2021) 4275–4279.
[47] M. Xu, F. Zhang, et al., Speech Emotion Recognition with Multiscale Area Attention and Data Augmentation, 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6319-6323.
[48] S. Shen, Y. Gao, et al., Emotion Neural Transducer for Fine-Grained Speech Emotion Recognition, 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 10111-10115.
[49] W. Chen, X. Xing, et al., SpeechFormer++: a hierarchical efficient framework for paralinguistic speech processing, IEEE/ACM Trans. Audio, Speech, Lang. Process. 31 (2023) 775–788.
[50] C. Busso, M. Bulut, et al., IEMOCAP: interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (2008) 335–359.
[51] K. Mao, Y. Wang, et al., Multi-branch feature learning based speech emotion recognition using SCAR-NET, Connect. Sci. 35 (2023).
[52] S. Shen, F. Liu, A. Zhou, Mingling or misalignment? temporal shift for speech emotion recognition with pre-trained representations, ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (IEEE2023), pp. 1-5.
[53] 2022, J.-X. Ye, X.-C. Wen, et al., GM-TCNet: Gated multi-scale temporal convolutional network using emotion causality for speech emotion recognition, 145 (2022) 21-35..

Tingting LIU (S′18-M′20) obtained her M.S. degree in natural language processing from Huazhong University of Science and Technology, Wuhan, China, in 2014. From September 2024 to August 2028, she is pursuing a PhD in educational technology at the Faculty of Education, The University of Hong Kong, Hong Kong, China. In 2020, she joined Hubei University, Wuhan, as an Assistant Professor in the School of Education. From 2022–2024, she served as a Visiting Scholar at the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, under the supervision of Professor Youfu Li. Her research focuses on learning behavior analysis, human pose estimation, label distribution learning, and graph neural networks. Dr. Liu frequently acts as a Reviewer for several international journals, including IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Neural Networks and Learning Systems, and IEEE Transactions on Multimedia. Since 2020, she has also been a Communication Evaluation Expert for the National Natural Science Foundation of China.

Minghong WANG received the B.S. degree from Nanjing University (NJU) in 1984, Ph.D degree in education information technology from City University of Hong Kong, (CityU), HongKong, in 1998. She is Professor and Director of the Laboratory for Knowledge Management & E-Learning, Faculty of Education, The University of Hong Kong. She is a member of the Advisory Group on Academic Reviews of HKU. She is also Kuang-piu Chair Professor at Zhejiang University, and Eastern Scholar Chair Professor at East China Normal University. She is the Editor-in-Chief of Knowledge Management & ELearning (indexed in Scopus & ESCI). Her research focus is on learning technologies for cognitive development, creative thinking, complex problem solving, knowledge management and visualization, and artificial intelligence applications. She has published 129 journal articles (78 in SSCI/SCI indexed journals; 29 articles in top 10 journals in multiple disciplines). She is recognized as ESI Top 1 % Scholar in (a) Social Sciences, General, and (b) Economics & Business.

Bing YANG currently holds the position of Full Professor of Computing at the School of Education, Hubei University, Wuhan, China. His PhD was awarded by Huazhong University of Science and Technology, located in Wuhan, Hubei, China. Professor Yang's research encompasses computer networking, educational technology, e-commerce, and data mining. He has extensively published his research in computer science journals, the Chinese Journal of Computers, and various international conferences and workshops.

Hai LIU (S'12–M'14) was awarded the M.S. degree in applied mathematics from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2010, followed by the Ph.D. degree in pattern recognition and artificial intelligence from HUST in 2014. Since June 2017, he has served as an Assistant Professor at the National Engineering Research Center for E-Learning, Central China Normal University, Wuhan. Prior to this, he spent two years as a "Hong Kong Scholar" postdoctoral fellow in the Department of Mechanical Engineering at City University of Hong Kong, Kowloon, Hong Kong, under the guidance of Professor You-Fu Li, until March 2020. He has authored over 60 peer-reviewed articles in international journals spanning various domains, including pattern recognition and image processing. His current research focuses on big data processing, artificial intelligence, spectral analysis, optical data processing, and pattern recognition. Dr. Liu regularly reviews for several international journals, such as *Computer & Education, British Journal of Educational Technology, IEEE Translations on Learning Technology, IEEE Translations on Education*. Additionally, he serves as a Communication Evaluation Expert for the National Natural Science Foundation of China.

Shaoxin YI is currently workong toward the B.E. degree in artificial intelligence from Central China Normal University (CCNU), Wuhan, China. His research interests include deep learning, parttern recognition and fine-grained classification.