Neurocomputing 555 (2023) 126623

Contents lists available at ScienceDirect

Neurocomputing

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

Speech emotion recognition via multiple fusion under spatial–temporal parallel network

Chenquan Gan [a,b], Kexin Wang [b], Qingyi Zhu [a], Yong Xiang [c], Deepak Kumar Jain [d], Salvador García [e,*]

[a] School of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[b] School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[c] School of Information Technology, Deakin University, Victoria 3125, Australia
[d] Key Laboratory of Intelligent Control and Optimization for Industrial Equipment of Ministry of Education, Dalian University of Technology, Dalian 116024, China
[e] Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence, University of Granada, Granada 18071, Spain

ARTICLE INFO

Communicated by Z. Wang

Keywords:
Speech emotion recognition
Speech spectrum
Spatial–temporal parallel network
Multiple fusion

ABSTRACT

Speech, as a necessary way to express emotions, plays a vital role in human communication. With the continuous deepening of research on emotion recognition in human–computer interaction, speech emotion recognition (SER) has become an essential task to improve the human–computer interaction experience. When performing emotion feature extraction of speech, the method of cutting the speech spectrum will destroy the continuity of speech. Besides, the method of using the cascaded structure without cutting the speech spectrum cannot simultaneously extract speech spectrum information from both temporal and spatial domains. To this end, we propose a spatial–temporal parallel network for speech emotion recognition without cutting the speech spectrum. To further mix the temporal and spatial features, we design a novel fusion method (called multiple fusion) that combines the concatenate fusion and ensemble strategy. Finally, the experimental results on five datasets demonstrate that the proposed method outperforms state-of-the-art methods.

1. Introduction

Speech is the basic, direct, and convenient medium in human daily communication, so speech emotion recognition (SER) is widely used in practice, such as depression auxiliary diagnosis [1], intelligent car [2], child-centered medical research [3] and human–computer interaction [4]. Unfortunately, the uncertain duration and ambiguous emotional characteristics of speech make SER research very difficult [5].

Since the speech spectrum has the ability to express temporal and spatial information, how to automatically extract distinctive emotion features from the spectrum has become a new trend in the development of SER [6]. Neural networks have strong representation ability for samples and can handle large-scale data, enabling learning, recognition, and solving a wide range of complex problems [7]. Hence, speech emotion recognition methods based on neural networks are gradually favored by many scholars [8]. However, the indeterminate duration of speech leads to the inconsistent length of speech spectrum, but neural networks usually require the same input size. Two approaches are normally used to deal with this issue. The first approach cuts the entire speech spectrum into multiple segments of the same length, and the second approach does not cut the speech spectrum.

For the first approach of cutting the speech spectrum, the entire speech spectrum is cut into segments of the same length and each segment is assigned an emotion label of the whole speech [9]. They often adopt cascaded [10] or parallel structure [11] or concatenate fusion [12] to learn emotion features of the speech spectrum. However, since emotion is not evenly distributed throughout the speech, cutting the speech spectrum into the same length makes each segment contain incomplete emotion cues. To avoid this problem, the second approach does not cut the speech spectrum by processing the speech spectrum in two ways.

One way is to take the converted speech spectrum as input, such as extracting acoustic feature sets from the speech spectrum [13] or converting the speech spectrum into a fixed-size picture [14]. But these methods all result in the lack of emotion information to varying degrees. Another way is to take the speech spectrum directly as

* Corresponding author.
E-mail addresses: gcq2010cqu@163.com (C. Gan), s200101155@stu.cqupt.edu.cn (K. Wang), zhuqy@cqupt.edu.cn (Q. Zhu), yong.xiang@deakin.edu.au (Y. Xiang), dkj@ieee.org (D.K. Jain), salvagl@decsai.ugr.es (S. García).

https://doi.org/10.1016/j.neucom.2023.126623
Received 8 June 2023; Received in revised form 17 July 2023; Accepted 24 July 2023
Available online 28 July 2023

input, and generally adopt a cascaded structure [15,16] or concatenate fusion [17]. However, the cascaded structure is composed of different functional modules in series, and the latter module can affect the feature extraction result of the former module, which will cause feature loss and deviation. Besides, different features contain different emotion cues, but the concatenate fusion does not consider each feature separately, which will lead to the omission of emotion cues.

To avoid the loss of spatio-temporal features caused by cutting the speech spectrum, our proposed model will use the complete speech spectrum as input. Additionally, due to strong inter-module dependencies, cascading different feature extraction modules is not conducive to extracting speech emotion features. To address this, we parallelly process the time-domain feature extraction module with the spatial-domain feature extraction module in our approach. Subsequently, we employ a multi-feature fusion mechanism for speech emotion recognition. In summary, our main contributions are as follows:

(1) The proposed spatial–temporal parallel network without cutting the speech spectrum can extract temporal and spatial features from the complete speech spectrum, mitigating feature loss and deviation.
(2) The proposed model extensively leverages the benefits of both temporal and spatial domains. By employing parallel extraction, the model effectively captures and incorporates the temporal and spatial features of speech emotions. To enable a two-way interaction between the temporal and spatial domain features, our model utilizes a concatenate fusion and ensemble strategy.

2. Related work

The work Satt et al. [18] divided the speech spectrum into a set of segments of 3 s and utilized the convolutional neural networks (CNNs) cascaded with long short-term memory (LSTM) to extract emotion cues from the linear scale spectrum. Since then, Li et al. [10] used cascaded CNNs with spatial–temporal attention and large margin learning method, and Li et al. [9] redesigned a cascaded structure of CNNs and attention mechanism. Further, Wu et al. [19] replaced the above attention mechanism with a capsule network. In addition, Mustaqeem et al. [11] and Zhao et al. [12] used a parallel network with cutting speech spectrum and concatenate fusion. However, these studies may introduce incorrect emotion labels or make emotion information missing because they assign emotion labels of the complete speech to the cut segments. Noticing this, our method does not cut the speech spectrum.

To avoid the loss of speech emotion information caused by cutting, the input methods without cutting the speech spectrum bring new inspiration to SER. Among them, some methods take the converted speech spectrum as input, e.g., Daneshfar et al. [20], Yi et al. [21], and Xiao et al. [22] utilized acoustic feature sets as input. Li et al. [23] slice the feature matrix to form new features and Singh et al. [24] proposed new modulation spectrum features based on spectral maps and time modulation. Zhang et al. [14] converted the speech spectrum into pictures. Unfortunately, the acoustic feature sets or generating new features cannot deeply describe the emotion in speech [25], and expanding or compressing speech with inconsistent lengths into fixed-size pictures will lose the temporal domain information of the speech spectrum.

In addition, the original length of the speech spectrum remains unchanged and is directly used as input. Ma et al. [15] proposed a method of not cutting speech spectrum with processing redundant zero-padded, and compared with the method of cutting speech spectrum in [18]. Later, Zhang et al. [16] developed a fully CNN for the complete speech spectrum. However, these methods all employ cascaded networks to generate single-type features. It is not conducive to the model to establish a comprehensive emotion cognition. For this issue, Cao et al. [26] used a parallel LSTM to get the dynamic and static features, but the parallel LSTM ignores the spatial features. Tseng et al. [27] enhanced emotion learning through the speech and text modalities, multimodal approaches rely on the contributions of additional modalities. And Li et al. [28] improved the recognition rate of speech emotion by integrating the output of automatic speech recognition into the pipeline of SER, but the speech recognition features that are beneficial to SER still need to be studied. Besides, Wu et al. [17] adopted concatenate fusion to fuse the features generated by a parallel network, but concatenate fusion does not consider the unique information of each feature. For this, our method designs a novel fusion method named multiple fusion. Taking the above factors into consideration, in order to better extract the spatio-temporal features of continuous speech spectrograms, we separately extract the spatio-temporal features of speech emotions using parallel time-domain and spatial-domain feature extraction modules. Finally, we combine a concatenate fusion and ensemble strategy to integrate the spatio-temporal features of speech emotions.

Instead of cutting the speech spectrum into a set of segments and using the cascaded structure, our method keeps the original speech spectrum unchanged and adopts a spatial–temporal parallel structure with multiple fusion. It not only can avoid the loss of emotion information caused by cutting the speech spectrum, but also can learn multi-type emotion features.

3. The proposed method

3.1. Motivation

The speech spectrum includes both spatial and temporal domains, where the spatial domain comprises the relationship between time and frequency, while the temporal domain expresses the contextual relationship of time frames and the emotion richness of each time frame. Inspired by the spatial feature extraction in the image field and the feature extraction of the time context in the natural language processing field, we consider extracting features from the spatial and temporal domain of the speech spectrum. Furthermore, in order to extract more comprehensive emotion features from speech samples, we input the speech without cutting and independently extract the temporal and spatial information of the speech spectrum in parallel.

3.2. Overview

As shown in Fig. 1, the proposed method mainly consists of four parts: preprocessing, temporal module, spatial module, and multiple fusion module. Specifically, the preprocessing module maps the original one-dimensional speech waveform to a two-dimensional spectrum containing both spatial and temporal characteristics. The temporal module utilizes the bidirectional gated recurrent unit (Bi-GRU) and attention mechanism (AM) with zero fill masking to capture temporal emotional changes in the spectrum. The spatial module uses multi-layer CNNs and global average pooling to locate emotion triggering regions in the spatial domain of the spectrum. The multiple fusion module generates spatio-temporal interaction features based on single connection fusion and ensemble strategy, and outputs the probability distribution of sentiment classification.

3.3. Preprocessing

Firstly, Hamming window and short-time Fourier transform are used to map speech into a linear spectrum. Secondly, in order to better match human auditory characteristics, a Mel filter bank is applied to the linear spectrum to obtain the log Mel filter bank energy spectrum (henceforth: spectrum):

$$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_L] \in \mathbb{R}^{f \times L}, \tag{1}$$

where $\mathbf{x}_i$ is the ith frame of spectrum, L is the spectrum length, and f is the number of filters. Finally, for speeding up the efficiency of computation, all spectrums are sorted by length and divided into

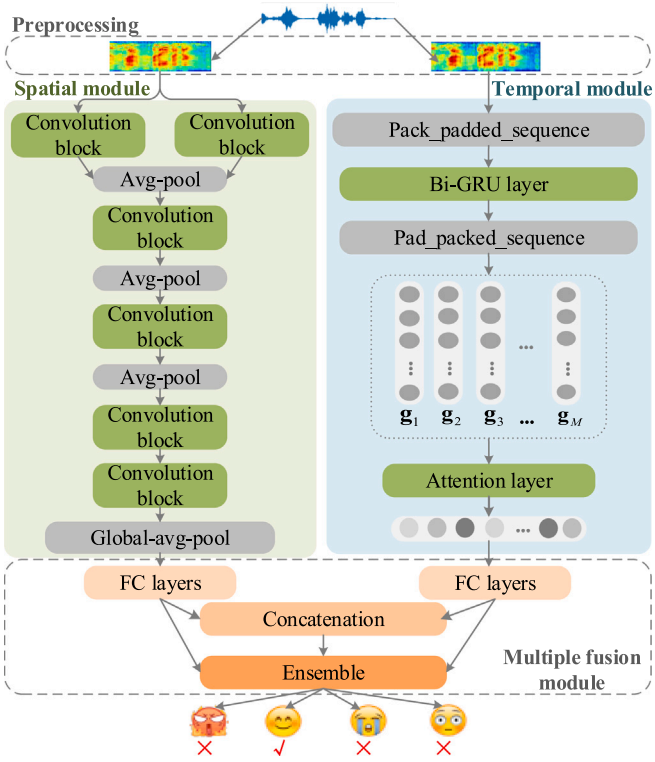


Fig. 1. The overview framework of the proposed method.

batches according to the proximity of their lengths. Meanwhile, each spectrum is padded with zeros to the maximum length of the spectrum in each batch. Given the maximum length M, then the zero-padding spectrum can be represented as:

$$\mathbf{X}_{in} = \left[\mathbf{X}, \mathbf{x}_{L+1}, \ldots, \mathbf{x}_M\right] \in \mathbb{R}^{f \times M}, \tag{2}$$

where $\mathbf{x}_{L+1}, \ldots, \mathbf{x}_M$ are zero-padding frames.

3.4. Temporal module

The time frames of the spectrum have contextual relationships and varying emotional richness. When learning temporal features of the spectrum, the dependencies between consecutive frames and salient emotional moments usually need to be considered [29]. On this basis, the bidirectional gated recurrent unit (Bi-GRU) and attention mechanism (AM) are combined to model temporal properties of the spectrum, as illustrated in Fig. 2.

Firstly, zero-padding frames will waste storage resources and affect the calculation process in Bi-GRU, so we compress them of $\mathbf{X}_{in}$ through the *pack_padded_sequence* and the *pad_packed_sequence* functions in PyTorch [30]. The output of Bi-GRU can be represented as $\mathbf{G} \in \mathbb{R}^{2d_T \times M}$, where d_T is the number of hidden neurons. The zero-padding output $\mathbf{X}_p \in \mathbb{R}^{1 \times dL}$ is represented as:

$$\mathbf{X}_p = pack\_padded\_sequence(\mathbf{X}_{in}). \tag{3}$$

Secondly, in order to capture the temporal context from both forward and backward directions, the input $\mathbf{X}_p$ is fed into a BiGRU with a hidden neuron count of d_T. In order to facilitate subsequent feature extraction operations, the output of the BiGRU is transformed back into a two-dimensional form using the corresponding sequence expansion function. The resulting output $\mathbf{G} = [g_1, g_2, \ldots, g_L, g_{L+1}, \ldots, g_M]$ can be represented as:

$$\mathbf{G} = pad\_packed\_sequence(BiGRU(\mathbf{X}_p)), \tag{4}$$

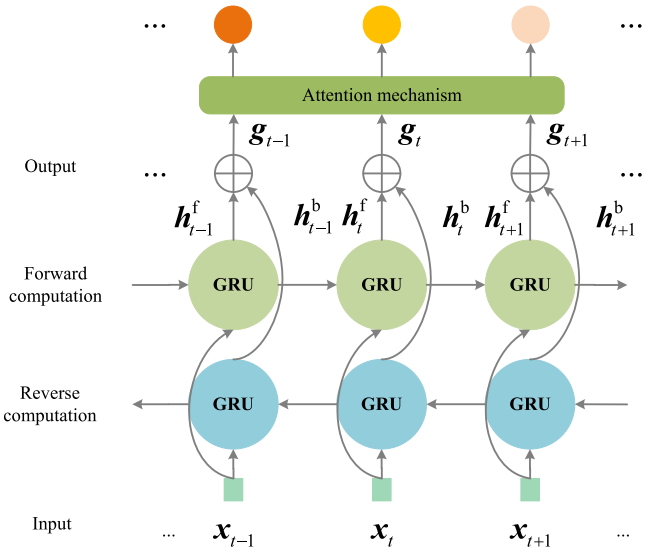


Fig. 2. The structure diagram of temporal module.

where $BiGRU(\cdot)$ represents the computation process of BiGRU in both forward and backward directions. $g_{L+1}, \ldots, g_M$ frames refer to zero vectors. *pad_packed_sequence* represents the sequence expansion function in the PyTorch toolkit, corresponding to the aforementioned sequence compression function *pack_padded_sequence*.

Thirdly, we adopt AM to evaluate the contribution of each frame in $\mathbf{G}$. The output of this process is denoted as $\mathbf{c} = [c_1, c_2, \ldots, c_i, \ldots, c_M]$, the sentiment score c_i of the i^{th} frame is calculated as:

$$c_i = \frac{\exp\left(h_i\right)}{\sum_{i=1}^{M} \exp\left(h_i\right)}, \tag{5}$$

where $h_i = \mathbf{V}\tanh\left(\mathbf{g}_i\mathbf{W} + \mathbf{b}\right)$, $\mathbf{g}_i$ is the ith frame of $\mathbf{G}$, $\mathbf{V}$ and $\mathbf{W}$ are trainable weights, $\mathbf{b}$ is the trainable bias, $\exp(\cdot)$ and $\tanh(\cdot)$ are exponential function and hyperbolic tangent activation function, respectively.

To ensure that zero-padding frames have a sentiment score of zero, we reset $\mathbf{c}$ with the mask matrix $Mask(\cdot)$ and then obtain the corresponding $\mathbf{c}' = [c'_1, c'_2, \ldots, c'_i, \ldots, c'_M]$:

$$\mathbf{c}' = \mathbf{c} \odot Mask(\mathbf{X}_{in}), \tag{6}$$

where $\odot$ indicates the Hadamard product.

Finally, according to the summation of the sentiment score of each frame, the temporal feature $\mathbf{U}_T \in \mathbb{R}^{2d_T \times 1}$ can be obtained:

$$\mathbf{U}_T = \sum_{i=1}^{M} c'_i \mathbf{g}_i. \tag{7}$$

3.5. Spatial module

The spectrum is an image-like representation that contains specific emotion information in the spatial domain. Inspired by [9] and verified by experiments, we adopt a CNN-based spatial domain module to extract spatial domain features of the spectrum, and its detailed structure is shown in Table 1. The structure of the spatial domain module is illustrated in Fig. 3.

Firstly, we add a channel dimension for $\mathbf{X}_{in}$ to get the corresponding $\mathbf{X}'_{in} \in \mathbb{R}^{1 \times f \times M}$ as the input, the local spatial features $\mathbf{X}_c$ can be obtained through five convolution blocks, where each convolution block are sequentially stacked by CNN, batch normalization (BN), and rectified linear unit (ReLU). Specifically, the first layer is set up as a set of parallel convolution blocks with average pooling, which are used to focus on the temporal and frequency dimensions of the spectrum, in order to capture the temporal and frequency variation relationships contained in

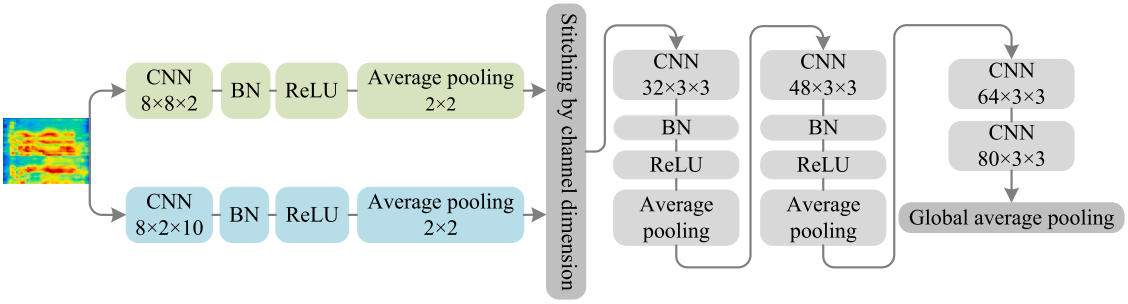


Fig. 3. The structure diagram of spatial module.

Table 1
Architectural details of spatial module.

Layer name	Backbone	Kernel size	Output size
Expand	Expand	–	$1 \times d \times M$
Conv1	Convolution block Avg-pooling	10×2 2×8 2×2	$16 \times d/2 \times M/2$
Conv2	Convolution block Avg-pooling	3×3 2×2	$32 \times d/4 \times M/4$
Conv3	Convolution block Avg-pooling	3×3 2×2	$48 \times d/6 \times M/6$
Conv4	Convolution block	3×3	$64 \times d/6 \times M/6$
Conv5	Convolution block	3×3	$80 \times d/6 \times M/6$
GAP	Global-avg-pool	$d/6 \times M/6$	$80 \times 1 \times 1$
Squeeze	Squeeze	–	80×1

the spatial domain. Then, two convolution blocks with average pooling are sequentially used to mix different feature information generated by the time and frequency dimensions of the spectrum, and redundant invalid information is discarded. Finally, the local spatial feature $\mathbf{X}_c \in \mathbb{R}^{80\times f/8\times M/8}$ is generated by encoding the features through the two stacked convolution blocks. The whole process can be summarized as follows:

$$\mathbf{X}_c = Conv(\mathbf{X}'_{in}), \tag{8}$$

where $Conv(\cdot)$ represents the feature processing of five convolution blocks.

Secondly, the global average pooling (GAP) is used to map $\mathbf{X}_c$ to global spatial feature $\mathbf{U}_S \in \mathbb{R}^{80\times1\times1}$. Notably, $\mathbf{U}_S$ is squeezed to 80×1 for subsequent concatenation with the temporal feature.

$$\mathbf{U}_S = Squeeze(GAP(\mathbf{X}_c)). \tag{9}$$

Finally, it is worth mentioning that this module does not deal with zero-padding, because the work [17] proves that zero-padding has no impact on the feature extraction of CNN.

3.6. Multiple fusion module

To better fuse the above-mentioned temporal and spatial features, we design a multiple fusion that combines the concatenate fusion and ensemble strategy, the structure diagram is shown in Fig. 4.

Firstly, the temporal and spatial sentiment probability distributions $\mathbf{Y}_T \in \mathbb{R}^{N\times1}$ and $\mathbf{Y}_S \in \mathbb{R}^{N\times1}$ are calculated through the temporal classifier and spatial classifier:

$$\begin{aligned} \mathbf{Y}_T &= \mathrm{Softmax}(\mathbf{W}_T\mathbf{U} + \mathbf{b}_T), \\ \mathbf{Y}_S &= \mathrm{Softmax}(\mathbf{W}_S\mathbf{U}_S + \mathbf{b}_S), \end{aligned} \tag{10}$$

where $\mathbf{U} = \mathrm{ReLU}(\mathrm{BN}(\mathbf{W}_t\mathbf{U}_T + \mathbf{b}_t))$, N is the number of emotion categories, $\mathrm{Softmax}(\cdot)$ represents the Softmax function, $\mathbf{W}_T, \mathbf{W}_S, \mathbf{W}_t$ are trainable weights, and $\mathbf{b}_T, \mathbf{b}_S, \mathbf{b}_t$ are trainable biases.

Secondly, concatenate the temporal and spatial features to obtain the concatenate fusion feature $\mathbf{U}_F \in \mathbb{R}^{(80+2d_T)\times1} = \mathrm{Concatenate}(\mathbf{U}_T, \mathbf{U}_S)$, and then calculate the corresponding sentiment probability distribution $\mathbf{Y}_F \in \mathbb{R}^{N\times1}$ through the fusion classifier:

$$\mathbf{Y}_F = \mathrm{Softmax}(\mathbf{W}_F\mathbf{U}_F + \mathbf{b}_F), \tag{11}$$

where $\mathbf{W}_F$ and $\mathbf{b}_F$ are trainable weight and bias, respectively.

To further fuse emotion information in the spatial and temporal domains, the probability distributions $\mathbf{Y}_T$, $\mathbf{Y}_S$, $\mathbf{Y}_F$ are integrated, the output $\hat{\mathbf{Y}} \in \mathbb{R}^{N\times1}$ can be represented as:

$$\hat{\mathbf{Y}} = \sum_{j=T,S,F} \mathbf{Y}_j. \tag{12}$$

Finally, the cross-entropy is adopted as the loss function:

$$\mathrm{Loss}(\mathbf{Y}, \hat{\mathbf{Y}}) = -\sum_{k\in N} y_k \ln \hat{y}_k, \tag{13}$$

where $\mathbf{Y} = [y_1, \ldots, y_k, \ldots, y_N], \hat{\mathbf{Y}} = [\hat{y}_1, \ldots, \hat{y}_k, \ldots, \hat{y}_N]$ are real and predicted probability distributions, respectively. $y_k, \hat{y}_k$ represent the real probability and the predicted probability of the k^{th} emotion category, respectively.

4. Experimental setup

4.1. Datasets

To show the performance of our method, this section performs several experiments on five datasets: IEMOCAP [31], MSP-IMPROV [32], EMO-DB [33], eNTERFACE05 [34], SAVEE [35]. The IEMOCAP dataset includes recordings from two scenarios: scripted and improvised, including 9 emotions: anger, happiness, sadness, neutral, excitement, disgust, fear, depression, and surprise. Consistent with most studies on SER, we consider the four emotions of anger, happiness, sadness, and neutral under improvised scenes. Moreover, due to the similar activation and valence states of excitement and happiness in emotional dimension analysis, excitement is divided into happiness. The MSP-IMPROV dataset contains 8348 samples, including 652 samples improvised against the target text, 620 samples read as scripts, 4381 samples improvised by actors, and 2785 samples recorded in natural interactions. However, due to the ambiguity of human emotion perception, different emotion experts may have different emotion cognition for the same speech sample, which has been confirmed in psychological research [36] and statistical models [37]. Therefore, the emotion labels of speech samples may not be unanimously recognized by experts, so only 7798 samples in this dataset have certain emotion labels. The

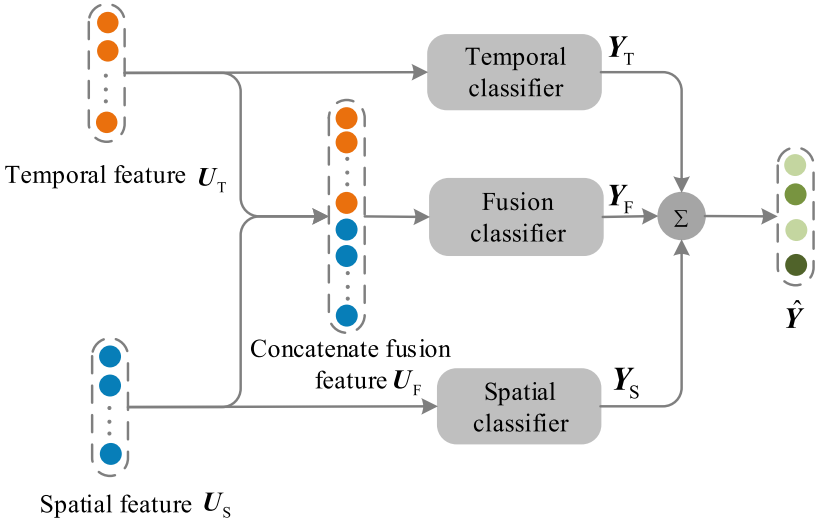


Fig. 4. The structure diagram of multiple fusion module.

Table 2
Details of the five datasets.

Dataset	Speaker	Emotion	Sample number
IEMOCAP	5 female/5 male	happiness, neutral, angry, sadness	2943
MSP-IMPORV	6 female/6 male	happiness, neutral, angry, sadness	7798
EMO-DB	5 female/5 male	happiness, neutral, angry, sadness, fearful, disgusted, bored	535
eINTERFACE05	42 human	happiness, angry, sadness, fearful, disgust, surprise	1257
SAVEE	4 male	happiness, neutral, angry, sadness, fearful, disgusted, surprise	480

Table 3
Sample distribution of five datasets.

Dataset	happiness	neutral	angry	sadness	fearful	disgusted	surprise	bored
IEMOCAP	947	1099	289	608	–	–	–	–
MSP-IMPORV	2644	3477	792	885	–	–	–	–
EMO-DB	71	79	127	62	69	46	–	81
eINTERFACE05	207	–	210	210	210	210	210	–
SAVEE	60	120	60	60	60	60	60	–

Note: Dash (–) indicates the corresponding dataset does not have samples of that emotion category.

current research has discarded samples without labels and selected 7798 samples with labels for training and testing. To facilitate comparison with other methods, we also select 7798 samples with labels for experiments. The EMO-DB and SAVEE datasets are both audio recorded in deductive scenarios. The eNTERFACE05 dataset reflects the corresponding emotion states based on six predefined emotion scenarios. Additional detailed information on these datasets can be found in Tables 2 and 3.

4.2. Performance measures

Weighted accuracy (WA) and unweighted accuracy (UA) are two evaluation indicators commonly used in speech emotion recognition, and the calculation methods are shown in Eqs. (14)–(15).

- **WA**: The ratio of the total number of correctly predicted samples to the total number of samples, representing the classification accuracy of all utterances.
- **UA**: The average of the accuracy of each emotion category, representing the classification accuracy of each emotion category.

$$WA = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} (TP_i + FN_i)}, \tag{14}$$

$$UA = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{(TP_i + FN_i)}, \tag{15}$$

where N represents the number of sentiment categories, TP_i represents the number of correct prediction samples for the ith category, and FN_i represents the number of incorrectly-predicted samples for the ith category. It is worth noting that UA is a more reasonable metric than WA since the uneven distribution of emotion samples has always been an unsolved problem in speech emotion recognition tasks. There are two reasons for this: on the one hand, UA as a class accuracy can more effectively show the accuracy of the model under imbalanced data; on the other hand, the accuracy of WA as a whole cannot reflect the accuracy of each emotion class.

4.3. Parameter settings

This model is implemented based on the Pytorch [30] deep learning framework, and the training is done on a 64-bit UBUNTU16.04 system with E5-2598V4CPU and 4 NVIDIA TeslaV100 GPUs. We employ experiments on each dataset with Leave-One-Speaker-Out (LOSO, *i.e.*, keep one speaker for testing and others for training) cross-validation. The Adam with a learning rate of e^{-4} is used as the optimizer, the weight decay is e^{-3} to prevent over-fitting, the batch size is 32, the epoch

Table 4
Comparison results on IEMOCAP and MSP-IMPROV.

Model	Datasets	
	IEMOCAP	MSP-IMPROV
	UA (%)/WA (%)	UA (%)/WA (%)
FCN [16]	63.90/70.40	–
Variable-Length [15]	64.22/71.45	–
ADAN [21]	64.51/66.92	–
CNN+LSTM [12]	68.00/65.20	–
Attention-pool [9]	68.06/71.75	–
LSTM+attention [26]	72.50/70.50	–
Multimodal [27]	–	53.20/–
CGDANN [22]	–	56.20/–
Ours	74.35/72.01	57.70/53.19

Note: Dash (–) indicates the corresponding work does not experiment on the dataset.

Table 5
Comparison results on EMO-DB, eNTERFACE05, and SAVEE.

Model	Datasets		
	EMO-DB	eNTERFACE05	SAVEE
	UA (%)/WA (%)	UA (%)/WA (%)	UA (%)/WA (%)
Cascaded-attention [10]	82.10/83.30	75.60/75.80	54.75/56.50
ADAN [21]	83.31/84.49	–	–
DPTM [14]	86.30/87.31	79.40/79.25	–
PQPSO [20]	–	–	55.00/59.38
Ours	89.02/89.50	88.65/88.64	62.26/66.45

Note: Dash (–) indicates the corresponding work does not experiment on the dataset.

is 100, the hidden neuron d_T of Bi-GRU is 512 to enrich contextual features, the window length of spectrum is 25 ms, and the filter number of spectrum is 40 in the spatial module and 26 in the temporal module.

5. Experimental results and analysis

To ensure the fairness of the comparison results, when comparing our proposed model with other models in terms of performance, we will use the same datasets and evaluation metrics (UA/WA). In exploring the results of ablation experiments, only the variations in filter quantity, sliding window length, fusion and fusion strategies, whether or not to segment different speakers' speech data, and the emotional recognition results of different speakers' speech data in cascading or parallel structures need to be discussed. This will ensure the fairness of the ablation experiment results.

5.1. Comparison with state-of-the-art methods

Tables 4 and 5 show the comparison results of our method with advanced methods. The results of our method are the average results of LOSO cross-validation. On the five datasets, our best results UA/WA are 83.19%/80.59%, 58.58%/60.10%, 94.64%/95.65%, 100%/100%, 80.95%/83.33%, respectively. The corresponding worst results UA/WA are 58.03%/52.05%, 49.90%/44.43%, 84.69%/83.67%, 70%/70%, 40.48%/46.67%, respectively. There is a difference between the best and worst results, but our average result is still better than other methods.

In more detail, since we maintain the integrity of speech and adopt the parallel model to represent emotion from multiple perspectives, our method can express emotion features more comprehensively than the cascaded models using the cut input [9,10]. Compared with the concatenate fusion model using the cut input of a 128-dimensional Mel spectrum [12], our method improves UA and WA by 6.81% and 6.35%, respectively, due to the reason that it is more conducive to feature interaction than concatenate fusion.

Compared with cascaded models without using the cut input [15,16], the performance of our model is more effective. The reason is that parallel models can extract emotion from features through multiple perspectives and express emotion more completely than cascaded models. The advantage of parallel models can also be seen in the parallel LSTM model [26], which takes a 26-dimensional spectrum as input but ignores emotion cues in the spatial domain. In contrast, our method considers the spatial cues, leading to better performance.

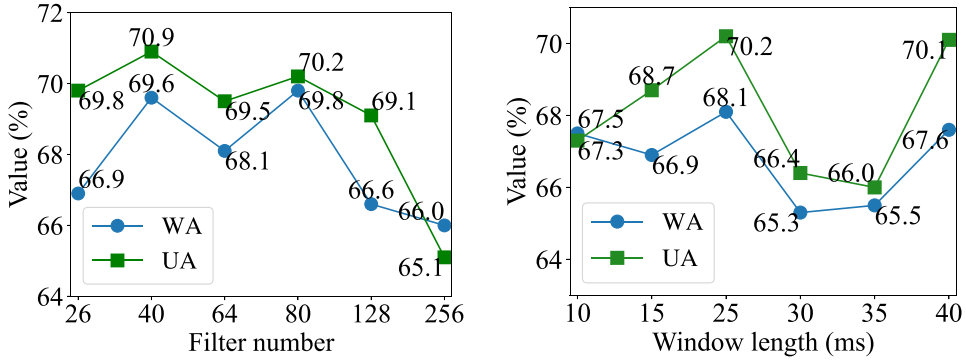
Compared with the models of taking the converted speech spectrum as input (*i.e.*, [14,20–22,27]), our method has advantages on all datasets. This is because we keep the original speech spectrum as input, avoiding the loss of spatial–temporal cues caused by feature set extraction and image compression. Specifically, the work in [20,21], and [22] performs statistical computations on speech to generate feature sets, resulting in the loss of frequency variation information in the spatial domain. The work [14] compresses or expands the spectrum into fixed-size images, causing the loss of context in the temporal domain. In addition, the work [27] takes text and speech as input, requiring additional text-modal assisted speech for emotion recognition.

5.2. Ablation analysis

To further discuss the reasons for our good results, we performed some ablation analysis. Since the ablation results of our method on each dataset are similar, we only show the analysis on the IEMOCAP dataset.

Fig. 5 displays the spatial module performance and temporal module performance of our method under different filter numbers and window lengths. In Fig. 5(a), we explore the influence of filter number on spatial feature extraction while keeping the window length unchanged (*i.e.*, 25 ms). It can be seen that the performance of spatial feature extraction is not proportional to the number of filters. This is because the filter is used to smooth noise components in speech. Too few filters will make noise removal insufficient, while too many filters will excessively remove useful sound components. Similarly, one can see from Fig. 5(b) that, while keeping the number of filters unchanged (*i.e.*, 26), the performance of temporal feature extraction does not improve with the increase of the window length. The main reason is that the window length is related to framing. Too small window length will cause time frame redundancy, whereas too large window length will result in blurred time frame boundaries. According to these experiments, we set the filter number of the spatial module to 40 and the window length of the temporal module to 25 ms.

Table 6 shows the ablation results of multiple fusion module. Due to the multiple fusion module being composed of connection fusion and integration strategies, we analyzed the connection fusion effect of the temporal module and spatial module, and also discussed the ensemble effect of the temporal module and spatial module. Finally, we compared them with the multiple fusion of the proposed model. Connection fusion refers to the direct concatenation of temporal and spatial features into an emotion classifier. The ensemble strategy aims to feed the temporal and spatial features into the corresponding emotion classifier, and then add up the outputs of each classifier. From Table 6, it can be observed that under the benchmark of spatial domain and temporal domain recognition performance, connection fusion can only achieve recognition performance similar to that of spatial domain and temporal domain. The reason is that the connection fusion splices the intermediate output features of the temporal module and the spatial module, and cannot fully consider the final emotion decision of the temporal module and spatial module. The connection fusion classifier has only one decision and is a weak classifier. In contrast, the ensemble strategy considers both temporal module and spatial module decisions, which can form a strong classifier and reduce overall misclassification, thus outperforming the temporal module and the spatial module in terms of WA metrics. Although the ensemble strategy has the above capabilities, it lacks intermediate fusion information in the spatial and temporal domain, so the UA indicator does not have an advantage. The multiple fusion module of the proposed model considers both connection fusion and ensemble strategy, which can better achieve spatial and temporal



(a) Spatial module performance with different filter numbers

(b) Temporal module performance with different window lengths

Fig. 5. Performance on IEMOCAP with different filter numbers and window lengths.

Table 6
The ablation results of multiple fusion module.

Model	UA (%)	WA (%)
Only spatial	70.93	69.58
Only temporal	69.78	66.93
Concatenate fusion	70.81	66.47
Ensemble strategy	67.74	71.09
Multiple fusion	74.35	72.01

Table 7
Results of different speakers under different input methods.

Speaker	Input methods	
	Cutting input	Not cutting input
	UA (%)/WA (%)	UA (%)/WA (%)
Session1M	75.53/75.11	80.80/77.60
Session1F	63.56/64.23	69.75/70.11
Session2M	74.19/73.85	82.53/78.21
Session2F	76.23/74.16	83.19/80.59
Session3M	65.54/65.92	74.86/76.08
Session3F	64.05/66.61	73.48/73.11
Session4M	62.87/62.98	73.25/72.93
Session4F	67.17/67.26	75.01/68.66
Session5M	56.22/58.34	58.03/52.06
Session5F	67.44/63.54	72.62/70.77
Average	67.28/67.20	74.35/72.01

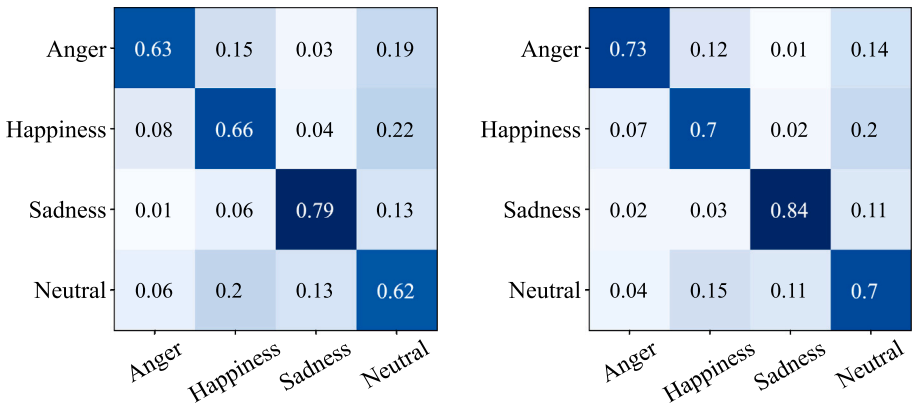
Table 8
Results of different speakers under different structures.

Speaker	Structures	
	Cascaded	Parallel
	UA (%)/WA (%)	UA (%)/WA (%)
Session1M	76.02/74.40	80.80/77.60
Session1F	65.86/64.21	69.75/70.11
Session2M	76.34/71.60	82.53/78.21
Session2F	81.02/78.39	83.19/80.59
Session3M	65.78/65.84	74.86/76.08
Session3F	66.64/69.51	73.48/73.11
Session4M	60.37/62.78	73.25/72.93
Session4F	66.68/62.31	75.01/68.66
Session5M	59.23/54.25	58.03/52.06
Session5F	67.08/63.11	72.62/70.77
Average	68.50/66.64	74.35/72.01

feature interaction. Therefore, the two evaluation indicators of multiple fusion are better than those of connection fusion and ensemble strategy.

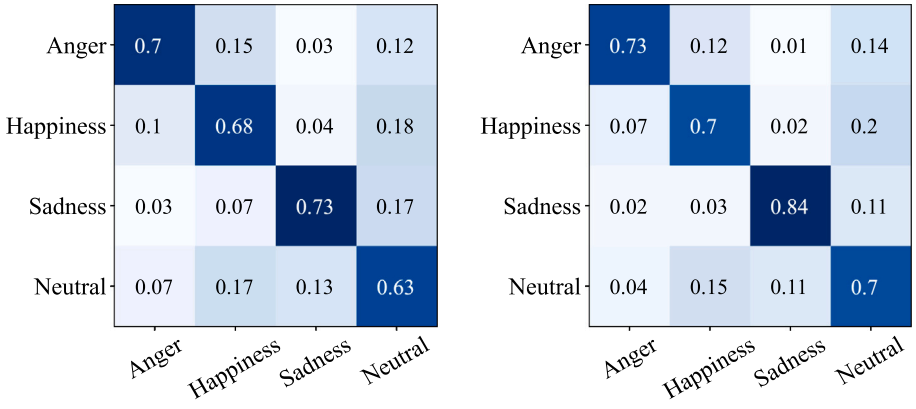
Table 7 shows the results of different speakers under different input methods, to explore the advantages of not cutting input. Specifically, the cutting input setting adopts the most commonly used 3 s segmentation method in [18], where speech longer than 3 s is cut into segments every 3 s, and speech shorter than 3 s is filled with zeros for 3 s. Each segment is assigned an emotion label of the entire speech. It can be found that the effect of not cutting input is better in most speakers, and the average results are higher than cutting input in both WA and UA. The reason is that not cutting input maintains the integrity of the speech, so the extracted emotion is more complete. Further, Fig. 6 shows the recognition rates of various emotions of cut input and not cutting input in the form of a confusion matrix graph, Fig. 6(a) a confusion graph of the cutting input method, and Fig. 6(b) a confusion graph of the not cutting input method. It can be found that the not cutting input method has a higher recognition rate than the cutting input method in all four emotion categories, and has a greater improvement in anger and neutral emotions, narrowing the gap in recognition rates among various emotions. This is because emotions are not evenly distributed throughout the entire speech. Cutting the speech into segments of the same length will result in each segment containing incomplete emotional clues, while the not cutting input method maintains the complete speech as input and processes zero-filled frames to ensure that the extracted emotional clues are more complete.

Table 8 shows the results of different speakers under different structures. In order to explore the advantages of the parallel structure in the proposed model, a comparison was made between the classic cascaded structure network and the spatial–temporal parallel network of the proposed model. Specifically, the cascaded structure network is formed by sequentially connecting the spatial and temporal modules in the proposed model, and the input features are consistent with the input of the spatial module. From the table, it can be observed that in most speakers, parallel structures perform better than cascaded structures, as the multi-type features obtained by parallel structures can more comprehensively express emotions. In addition, Fig. 7 analyzes the change in emotion recognition rate of each category of cascade structure and parallel structure from the confusion matrix to better show the advantages of parallel structure. From the graph, it can be seen that the emotion recognition rate of the four categories of parallel structure is higher than that of cascade structure, especially for the



(a) Cutting input (b) Not cutting input

Fig. 6. The confusion matrix of different input methods.



(a) Cascaded structure (b) Parallel structure

Fig. 7. The confusion matrix of different structures.

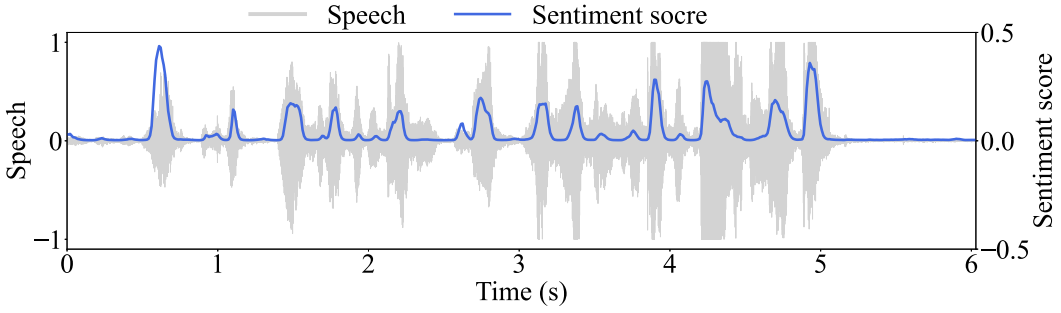


Fig. 8. An example of AM sentiment score on IEMOCAP.

emotional category of sadness, which is about 11% higher than that of cascade structure. This indicates that the cascaded structure of stacking spatial and temporal modules in order will affect each other's feature extraction, while the parallel structure network can simultaneously focus on the emotional information contained in the temporal and spatial domains, capturing discriminative features from the temporal and spatial domains of speech features, and can more comprehensively express emotions.

Fig. 8 shows an example of the AM sentiment score. We randomly selected a speech in the IEMOCAP dataset with an emotion of anger and a duration of 6.03 s. It can be found that the sentiment score has similar fluctuations to the speech, and becomes higher when the speech amplitude value is higher, indicating that AM can observe all frames with sound and thus focus on frames with emotion.

6. Conclusions

In this paper, a spatial–temporal parallel network without cutting speech spectrum has been proposed for SER. Our method not only can avoid the loss of emotion information, but also can extract multi-type emotion features (*i.e.*, temporal and spatial features). To obtain more discriminative speech emotion classification performance, a multiple fusion method has been designed for further mixing the temporal and spatial features. Finally, several experiments on five datasets, including

comparisons with the advanced methods and ablation analysis, have been conducted to discuss further the reasons for our achieved good results.

Despite the advantages mentioned above, there are two problems with the current method. Firstly, speech emotion labels are determined by a few emotion experts, which introduces ambiguity. Different experts may assign different emotions to the same speech samples, making it challenging to find consistent expressions for a single emotion label in real-life situations. Secondly, the annotation of speech samples requires significant manpower and financial resources. Consequently, there is a large number of unlabeled speech samples that remain unused. Deep learning models, which offer strong generalization and performance, typically require extensive training on massive samples, posing challenges for the practical deployment of speech emotion recognition systems. Therefore, future research can focus on addressing the ambiguity of speech emotions and utilizing unlabeled speech data to develop more suitable speech emotion recognition systems for practical applications.

CRediT authorship contribution statement

Chenquan Gan: Conceptualization, Methodology, Validation, Formal analysis, Project administration, Writing – original draft, Supervision. **Kexin Wang:** Conceptualization, Data curation, Methodology, Software, Writing – original draft, Visualization. **Qingyi Zhu:** Supervision, Writing – review & editing, Validation, Investigation. **Yong Xiang:** Supervision, Writing – review & editing, Formal analysis. **Deepak Kumar Jain:** Supervision, Writing – review & editing, Visualization. **Salvador García:** Supervision, Writing – review & editing.

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

The authors do not have permission to share data

Acknowledgments

The authors are grateful to the anonymous reviewers and the editor for their valuable comments and suggestions. This work was supported by the National Natural Science Foundation of China (No. 61702066), the Chongqing Research Program of Basic Research and Frontier Technology, China (No. cstc2021jcyj-msxmX0761) and partially supported by Project PID2020-119478GB-I00 funded by MICINN/AEI/10.13039/501100011033 and by Project A-TIC-434- UGR20 funded by FEDER/Junta de Andalucía Consejería de Transformación Económica, Industria, Conocimiento *y* Universidades.

References

[1] E.W. McGinnis, S.P. Anderau, J. Hruschak, R.D. Gurchiek, N.L. Lopez-Duran, K. Fitzgerald, K.L. Rosenblum, M. Muzik, R.S. McGinnis, Giving voice to vulnerable children: Machine learning analysis of speech detects anxiety and depression in early childhood, IEEE J. Biomed. Health Inform. 23 (6) (2019) 2294–2301.

[2] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, R.W. Picard, Driver emotion recognition for intelligent vehicles: A survey, ACM Comput. Surv. 53 (3) (2020) 1–30.

[3] E. Vaaras, S. Ahlqvist-Björkroth, K. Drossos, L. Lehtonen, O. Räsänen, Development of a speech emotion recognizer for large-scale child-centered audio recordings from a hospital environment, Speech Commun. 148 (2023) 9–22.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J.G. Taylor, Emotion recognition in human-computer interaction, IEEE Signal Process. Mag. 18 (1) (2001) 32–80.

[5] Mustaqeem, S. Kwon, Att-Net: Enhanced emotion recognition system using lightweight self-attention module, Appl. Soft Comput. 102 (2021) 107101.

[6] A.R. Avila, Z. Akhtar, J.F. Santos, D. O'Shaughnessy, T.H. Falk, Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild, IEEE Trans. Affect. Comput. 12 (1) (2018) 177–188.

[7] M. Kaveh, M.S. Mesgari, Application of meta-heuristic algorithms for training neural networks and deep learning architectures: A comprehensive review, Neural Process. Lett. (2022) 1–104.

[8] J. de Lope, M. Graña, An ongoing review of speech emotion recognition, Neurocomputing (2023).

[9] P. Li, Y. Song, I. McLoughlin, W. Guo, L. Dai, An attention pooling based representation learning method for speech emotion recognition, in: Interspeech 2018, ISCA, 2018, pp. 3087–3091.

[10] S. Li, X. Xing, W. Fan, B. Cai, P. Fordson, X. Xu, Spatiotemporal and frequential cascaded attention networks for speech emotion recognition, Neurocomputing 448 (2021) 238–248.

[11] Mustaqeem, S. Kwon, MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach, Expert Syst. Appl. 167 (2021) 114177.

[12] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, B. Schuller, Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition, IEEE Access 7 (2019) 97515–97525.

[13] M.S. Fahad, A. Ranjan, J. Yadav, A. Deepak, A survey of speech emotion recognition in natural environment, Digit. Signal Process. 110 (2021) 102951.

[14] S. Zhang, S. Zhang, T. Huang, W. Gao, Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, IEEE Trans. Multimed. 20 (6) (2017) 1576–1590.

[15] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, L. Cai, Emotion recognition from variable-length speech segments using deep learning on spectrograms, in: Interspeech 2018, ISCA, 2018, pp. 3683–3687.

[16] Y. Zhang, J. Du, Z. Wang, J. Zhang, Y. Tu, Attention based fully convolutional network for speech emotion recognition, in: 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC, IEEE, 2018, pp. 1771–1775.

[17] S. Wu, F. Li, P. Zhang, Weighted feature fusion based emotional recognition for variable-length speech using DNN, in: 2019 15th International Wireless Communications & Mobile Computing Conference, IWCMC, IEEE, 2019, pp. 674–679.

[18] A. Satt, S. Rozenberg, R. Hoory, Efficient emotion recognition from speech using deep learning on spectrograms, in: Interspeech 2017, ISCA, 2017, pp. 1089–1093.

[19] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu, et al., Speech emotion recognition using capsule networks, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2019, pp. 6695–6699.

[20] F. Daneshfar, S.J. Kabudian, A. Neekabadi, Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier, Appl. Acoust. 166 (2020) 107360.

[21] L. Yi, M. Mak, Improving speech emotion recognition with adversarial data augmentation network, IEEE Trans. Neural Netw. Learn. Syst. 33 (1) (2022) 172–184.

[22] Y. Xiao, H. Zhao, T. Li, Learning class-aligned and generalized domain-invariant representations for speech emotion recognition, IEEE Trans. Emerg. Top. Comput. Intell. 4 (4) (2020) 480–489.

[23] D. Li, L. Sun, X. Xu, Z. Wang, J. Zhang, W. Du, BLSTM and CNN stacking architecture for speech emotion recognition, Neural Process. Lett. 53 (6) (2021) 4097–4115.

[24] P. Singh, M. Sahidullah, G. Saha, Modulation spectral features for speech emotion recognition using deep neural networks, Speech Commun. 146 (2023) 53–69.

[25] E. Lieskovská, M. Jakubec, R. Jarina, M. Chmulík, A review on speech emotion recognition using deep learning and attention mechanism, Electronics 10 (10) (2021) 1163.

[26] Q. Cao, M. Hou, B. Chen, Z. Zhang, G. Lu, Hierarchical network based on the fusion of static and dynamic features for speech emotion recognition, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 6334–6338.

[27] S. Tseng, S. Narayanan, P.G. Georgiou, Multimodal embeddings from language models for emotion recognition in the wild, IEEE Signal Process. Lett. 28 (2021) 608–612.

[28] Y. Li, P. Bell, C. Lai, Fusing asr outputs in joint training for speech emotion recognition, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 7362–7366.

[29] S.P. Yadav, S. Zaidi, A. Mishra, V. Yadav, Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN), Arch. Comput. Methods Eng. 29 (3) (2022) 1753–1770.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019) 8026–8037.

[31] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Resourc. Eval. 42 (4) (2008) 335–359.

[32] C. Busso, S. Parthasarathy, A. Burmania, M. Abdel-Wahab, N. Sadoughi, E.M. Provost, MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception, IEEE Trans. Affect. Comput. 8 (1) (2017) 67–80.
[33] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of German emotional speech, in: Interspeech 2005, ISCA, 2005, pp. 1517–1520.
[34] O. Martin, I. Kotsia, B. Macq, I. Pitas, The eNTERFACE'05 audio-visual emotion database, in: 22nd International Conference on Data Engineering Workshops, ICDEW'06, IEEE, 2006, p. 8.
[35] S. Haq, P.J.B. Jackson, J.D. Edge, Audio-visual feature selection and reduction for emotion classification, in: International Conference on Auditory-Visual Speech Processing 2008, ISCA, 2008, pp. 185–190.
[36] M.A. Arbib, Book review: Andrew Ortony, Gerald L. Clore and Allan Collins, the cognitive structure of emotions, Artificial Intelligence 54 (1) (1992) 229–240.
[37] J. Tao, A. Li, S. Pan, A multiple perception model on emotional speech, in: Affective Computing and Intelligent Interaction, Third International Conference and Workshops, ACII 2009, Amsterdam, the Netherlands, September 10-12, 2009, Proceedings, IEEE Computer Society, 2009, pp. 1–6.

Chenquan Gan received the Ph.D. degree from the Department of Computer Science, Chongqing University, Chongqing, China, in 2015. He is currently an Associate Professor with the Chongqing University of Post and Telecommunications (CQUPT), Chongqing. His research interests include sentiment analysis, cybersecurity dynamics, and blockchain.

Kexin Wang is currently pursuing the master's degree with the Chongqing University of Posts and Telecommunications. Her research interests include deep learning and speech emotion recognition.

Qingyi Zhu (Member, IEEE) received the Ph.D. degree in computer science and technology from the College of Computer Science, Chongqing University, Chongqing, China, in 2014. He is currently an Associate Professor with the Chongqing University of Posts and Telecommunications, Chongqing. He has published more than 30 academic articles in peer-reviewed international journals. His current research interests include cybersecurity dynamics, complex systems, and blockchain. He has also served as an invited reviewer for various international journals and conferences.

Yong Xiang received the Ph.D. degree in Electrical and Electronic Engineering from The University of Melbourne, Australia. He is currently a Professor at the School of Information Technology, Deakin University, Australia. His research interests include information security and privacy, data analytics and machine learning, Internet of Things, and blockchain. He has published 6 monographs, over 190 refereed journal articles, and numerous conference papers in these areas. Professor Xiang is the Senior Area Editor of IEEE Signal Processing Letters and the Associate Editor of IEEE Communications Surveys and Tutorials. He was the Associate Editor of IEEE Signal Processing Letters and IEEE Access, and the Guest Editor of IEEE Transactions on Industrial Informatics and IEEE Multimedia. He has served as Honorary Chair, General Chair, Program Chair, Technical Program Committee Chair, Symposium Chair and Track Chair for many conferences, and was invited to give keynotes at several international conferences.

Deepak Kumar Jain received the Bachelor of Engineering degree from Rajiv Gandhi Proudyogiki Vishwavidyalaya, India, in 2010, the Master of Technology degree from the Jaypee University of Engineering and Technology, India, in 2012, and the Ph.D. degree from the Institute of Automation, University of Chinese Academy of Sciences, Beijing, China. He was an awardee of CAS-TWAS Presidential fellowship from 2014-2018. He was invited as "Foreign Experts" by Shandong Taian Administration of foreign Expert Affairs. He has presented several papers in peer-reviewed conferences and has published numerous studies in science cited journals. His research interests include deep learning, machine learning, pattern recognition, and computer vision.

Salvador García received the B.S. and Ph.D. degrees in Computer Science from the University of Granada, Granada, Spain, in 2004 and 2008, respectively. He is currently a Full Professor in the Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain. Dr. García has published more than 110 papers in international journals (more than 85 in Q1), h-index 62. As editor activities, he is editor in chief of "Information Fusion" (Elsevier), and an associate editor of "Swarm and Evolutionary Computation" (Elsevier), "AI Communications" (IOS Press) and "Machine Learning" (Springer) journals. He is a co-author of the books entitled "Data Preprocessing in Data Mining", "Learning from Imbalanced Data Sets" and "Big Data Preprocessing: Enabling Smart Data" published by Springer. His research interests include data science, data preprocessing, Big Data, evolutionary learning, Deep Learning, metaheuristics and biometrics. He belonged to the list of the Highly Cited Researchers in the area of Computer Sciences (2014-2020): http://highlycited.com/ (Clarivate Analytics).