Dynamic-static Cross Attentional Feature Fusion Method for Speech Emotion Recognition *

Ke Dong[1], Hao Peng[2,3], and Jie Che[1]

[1] Hefei University of Technology, Hefei, China
[2] Dalian University of Technology, Dalian, China
[3] Newcastle University, Newcastle, UK
coliapaston@163.com, ph97135@163.com, 3417410684@qq.com

Abstract. The dynamic-static fusion features play an important role in speech emotion recognition (SER). However, the fusion methods of dynamic features and static features generally are simple addition or serial fusion, which might cause the loss of certain underlying emotional information. To address this issue, we proposed a dynamic-static cross attentional feature fusion method (SD-CAFF) with a cross attentional feature fusion mechanism (Cross AFF) to extract superior deep dynamic-static fusion features. To be specific, the Cross AFF is utilized to parallel fuse the deep features from the CNN/LSTM feature extraction module, which can extract the deep static features and the deep dynamic features from acoustic features (MFCC, Delta, and Delta-delta). In addition to the SD-CAFF framework, we also employed muti-task learning in the training process to further improve the accuracy of emotion recognition. The experimental results on IEMOCAP demonstrated the WA and UA of SD-CAFF are 75.78% and 74.89%, respectively, which outperformed the current SOTAs. Furthermore, SD-CAFF achieved competitive performances (WA: 56.77%; UA: 56.30%) in the comparison experiments of cross-corpus capability on MSP-IMPROV.

Keywords: Speech Emotion Recognition · Attention Mechanism · Feature Fusion · Multi-view Learning · Cross-corpus

1 Introduction

With the development of human-computer interaction (HCI) system, speech emotion recognition (SER) has gradually become a hot topic. The human-computer interaction systems with efficient SER method can provide targeted feedback and support based on the emotional state of the specific speaker.

The main purpose of speech emotion recognition is to provide assistance in recognizing the emotional information in the speech signal and understanding the emotional activities of the human. In order to explore the features in speech signals, a series of feature vectors of speech signals are introduced into

* Supported by National Natural Science Foundation (NNSF) of China (Grant 61867005)

SER, including Mel frequency cepstral coefficients (MFCC), differential coefficient (Delta). In addition, the deep learning techniques can automatically extract the deep information contained in the feature vectors without manual calculation and feature adjustment. Typically, CNN is widely applied to extract static acoustic features of different scales in SER [9], and LSTM is often employed to extract dynamic features (or temporal features) of original data due to the temporal characteristics of its network structure. However, the deep speech emotion recognition model with single-view inputs can not further improve the performance during training process. Sun et al. suggest that the accuracy of the model can further improved by introducing multi-view features [16]. The dynamic-static dual-view fusion feature can be a widely utilized multi-view features in SER [17]. For example, Sun et al. proposed a serial method which puts the static speech data from CNN into LSTMs to mine the potential temporal correlations in the speech signals [15]. However, the serial strategy only connects the existing features which can not mine the underlying information in original features. In comparison with the serial fusion strategy, the parallel fusion can extract the potential information from the original inputs more effectively [19]. Furthermore, we note that the performance of the feature fusion algorithm can be enhanced by utilizing the attention mechanism [18]. For instance, Dai et al. proposed an attentional feature fusion module (AFF), which can utilize the properties of the encoder-decoder to fuse the dual input features of the module [5]. In AFF, however, the global features and local features should be distinguished before input. It thus can be found that AFF is not good at fusing the original inputs with unpredictable relationships or equal importance. To address this problem, we propose an cross attentional feature fusion module (Cross AFF) that can be utilized to not only fuse the dual input features equivalently, but also recognize the importance of each feature in feature fusion process. Moreover, a dynamic-static cross attentional feature fusion method (SD-CAFF) is proposed to obtain the improved multi-view fusion features based on the Cross AFF. The main contributions of this paper are summarized as follows:

- A dynamic and static cross attentional feature fusion method (SD-CAFF) is developed to extract and integrate the complementary information existing in dynamic and static features.
- We propose a CNN-based static feature extraction module to mine the deep static features in mel frequency cepstrum coefficient (MFCC).
- The LSTM-based dynamic feature extraction module is proposed to explore the underlying deep dynamic features in the combination of MFCC and MFCC differential coefficients (Delta and Delta-delta).
- We develop a cross attentional feature fusion mechanism (known as Cross AFF), which can be applied to equivalently fuse the dual-view inputs and automatically ascertain the weight of each feature.

The rest of this paper are organized as follows: The specific SD-CAFF structure and related algorithms are discribed in Sec. 2. The related comparative experiments and ablation experiments are conducted and analyzed in Sec. 3. Finally, Sec. 4 concludes this paper.

2 Methodology

This section introduces the main framework of the SD-CAFF model in detail (see Fig. 1).

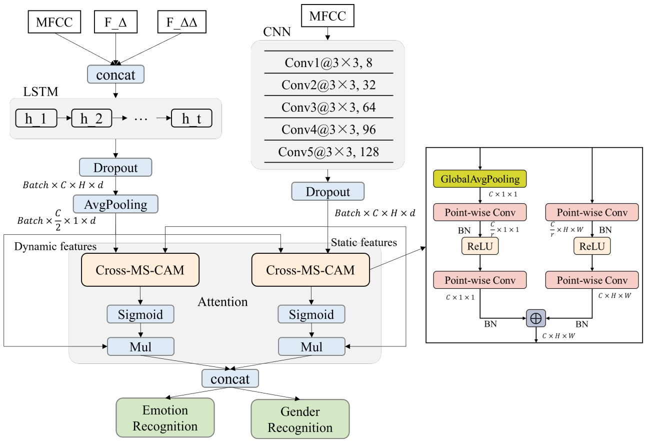

Fig. 1: The framework of the SD-CAFF architecture. The architecture consists of three parts: CNN static feature extraction module (CNN), Lstm dynamic feature extraction module (LSTM) and cross-attention feature fusion module (Cross AFF). In terms of classification, SD-CAFF also utilizes multi-label auxiliary learning based on emotional labels and gender labels.

2.1 Acoustic Feature Extraction

Mel frequency cepstrum coefficient (MFCC) is a widely utilized acoustic feature. In this paper, MFCCs are adopted as the static acoustic feature. Moreover, the first-order and second-order differential coefficients (Delta and Delta-delta) of MFCC are utilized to introduce the dynamic information. The process to obtain the acoustic features is illustrated in Fig. 2. The obtained MFCC features are employed as the input static acoustic features of SD-CAFF and can be utilized to calculate the Delta coefficients (defined in Eq. (1)).

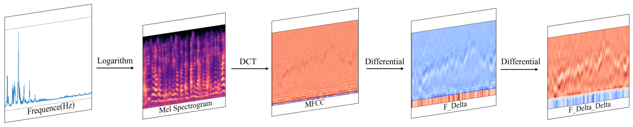

Fig. 2: Extraction of Acoustic Features

$$delta(t) = \frac{\sum_{n=1}^{N} n\,(mfcc_{t+n} - mfcc_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{1}$$

where $mfcc_{t+n}$ and $mfcc_{t-n}$ are MFCC coefficients, and $delta(t)$ is the final delta coefficient. The Delta-delta coefficient is obtained by applying the same algorithm for the delta coefficient, which can be calculated as Eq. (2).

$$delta\text{-}delta(t) = \frac{\sum_{n=1}^{N} n\,(delta_{t+n} - delta_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{2}$$

with $delta\text{-}delta(t)$ being the final Delta-delta coefficients. Therefore, we have obtained the static and dynamic acoustic features.

2.2 Dynamic-static Cross Attentional Feature Fusion Method

The framework of dynamic-static cross attentional feature fusion method (SD-CAFF) consists of three parts: CNN static feature extraction module, LSTM dynamic feature extraction module and Cross attentional feature fusion module (Cross AFF).

CNN Static Feature Extraction Module This paper proposed a CNN static feature extraction module, which is mainly composed of 2D convolution layer, batch normalization layer and activation function layer. Following the principle of lightweight network, this module utilizes DY-ReLU [4] as the activation function for superior performance.

As shown in Fig. 1, the CNN static feature extraction module is structured by five 2D convolutional kernels. For the first and second layers, each convolutional kernel in conjunction with a layer of a layer of BatchNorm2d, and a layer of DY-ReLUB. The other convolutional kernels are jointly with a layer of MaxPool2d, a layer of BatchNorm2d, and a layer of DY-ReLUB.

The issue accomplished by CNN static feature extraction module can be expressed as Eq. (3).

$$X_S = CNN(mfcc(n)) \tag{3}$$

where X_S is the deep static feature and $mfcc(n)$ is MFCC coefficients.

LSTM Dynamic Feature Extraction Module The recurrent architecture of LSTM can settle the problems in temporal sequence modelling scenarios. Therefore, a BiLSTM module is adopted to extract depth information in dynamic acoustic features. Since all three acoustic features (MFCC, Delta and delta-delta) may contain deep dynamic features, the input of LSTM exists more than one combination. For example, fusion features can be obtained by adding MFCC and Deltas, or combining Deltas and Delta-deltas. This paper utilizes $mfcc \oplus delta$ to represent the combinations of acoustic features. A feature selection experiment is designed to study which feature is the best match for the LSTM module in

SD-CAFF. The specific experimental steps will be demonstrated in Sec. 3. The algorithm of LSTM module can be summarized as Eq. (4).

$$X_D = BiLSTM(mfcc \oplus delta) \tag{4}$$

where X_D is the deep dynamic feature produced by LSTM feature extraction module.

Cross Attention Feature Fusion Mechanism (Cross AFF) Based on the AFF mechanism proposed by Dai et al. [5], this paper proposes a new feature fusion mechanism called cross attention feature fusion (Cross AFF). Different from AFF, Cross AFF can equivalently fuse the dual inputs of the module. Being the most significant part of Cross AFF, Cross MS-CAM (inside the gray dotted frame in Fig. 3) can explore hidden correlation between dual inputs.

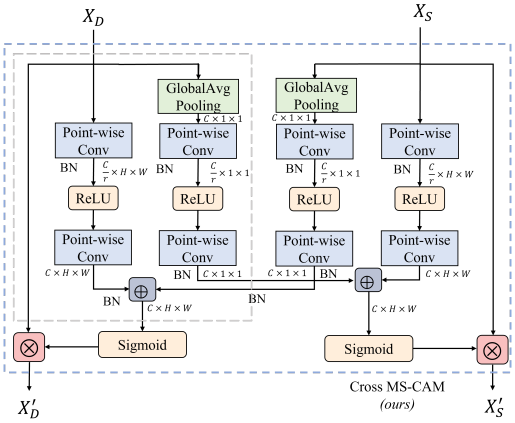

Fig. 3: A cross-attention feature fusion mechanism that can equally integrate dual features. For this module, the input X_S and X_D are equivalent.

Different from the traditional MS-CAM module, utilizing two parallel Cross MS-CAMs can integrate the dual parallel inputs X_S and X_D. Specifically, Cross MS-CAM fuses the global attention and the local attention of the input, and applys the sigmoid function to calculate the weight W_S of the fusion result $X_S \circ X_D$. Then, the attentional feature X_S' can be obtained by multiplying the the weight W_S and the initial static feature X_S. The process of the double Cross MS-CAM can be expressed as Eq. (5).

$$X_S' = X_S \otimes M_C \left(X_S \circ X_D\right) = X_S \otimes \sigma \left(L \left(X_S\right) \oplus g \left(X_D\right)\right) \tag{5}$$

where $M_C(x) \in R^{C \times H \times W}$ represents the attention weight generated by Cross MS-CAM. $\oplus$ refers to broadcast addition, $\otimes$ means element-by-element multiplication, and $\circ$ is the cross-fusion of local and global attention. $\sigma\left(x\right)$ represents the Sigmoid function. Similarly, the solution of attention feature X'_D related to feature X_S is calculated in Eq. (6).

$$X'_D = X_D \otimes M_C \left(X_D \circ X_S\right) = X_D \otimes \sigma \left(L \left(X_D\right) \oplus g \left(X_S\right)\right) \tag{6}$$

In [5], the global and local features should be distinguished before input into the attention feature fusion module. However, if the importance of the dual inputs is equal or unknown, choosing which one is the global feature or the local feature will become a complex issue. Given this, this paper improves the AFF algorithm and proposes a cross-attention feature fusion mechanism (Cross AFF) to fuse the input dual features equivalently. The proposed Cross AFF can better pay attention to the most different information among multiple input features to tackle the feature fusion task effectively [19].

Due to the symmetrical structure of double Cross MS-CAM, the dual inputs of Cross AFF are completely equivalent. As two feature graphs extracted from different views, X and Y are respectively put into the Cross MS-CAM module. To connect the X' and the Y', a concatenation layer (Concat) is applied for obtaining the final fusion feature Z. The process of Cross AFF can be calculated as Eq. (7).

$$Z = X \otimes M_C(X \circ Y) + Y \otimes M_C(Y \circ X) \tag{7}$$

where X'_S and X'_D are inputs of the feature fusion module of SD-CAFF. The Z obtained by fusion algorithm is the final output of the method.

2.3 Multi-label Auxiliary Learning

The current studies regard gender as a factor influencing the results of emotional recognition [7]. In [11], liu et al. proposed a multi-label center loss function, also known as joint loss function. The joint loss function has the advantages of both center loss and multi-label auxiliary learning simultaneously, which can be formulated as Eq. (8).

$$Loss = \mu \left(Loss_0^{\varepsilon} + \lambda \cdot Loss_0^{\varepsilon}\right) + \left(1 - \mu\right) \left(Loss_0^{g} + \lambda \cdot Loss_0^{g}\right) \tag{8}$$

where λ and μ are two hyperparameters to control the loss ratio. In order to eliminate the influence of gender on emotional label classification task, this paper utilized this multi-label center loss as the loss function in the SD-CAFF training process.

3 Experiment

In this section, this paper evaluates the performance of SD-CAFF on two benchmark datasets (IEMOCAP and MSP-IMPROV). To be specific, comparative experiments are first conducted to demonstrate the accuracy advantage of SD-CAFF compared to current SOTAs (state-of-the-art). In addition, ablation experiments are conducted to evaluate the effectiveness of SD-CAFF.

3.1 Datasets

In this work, a series of experiments are conducted on two widely utilized benchmark datasets (IEMOCAP and MSP-IMPROV) to validate the emotion recognition accuracy of SD-CAFF. The illustration of IEMOCAP and MSP-IMOROV are listed as follows.

- **IEMOCAP:** The speech from IEMOCAP is divided into small utterances, each of which is basically 3-15 seconds in length. Note that we only utilized the utterance in the improvised scenario to ensure emotional authenticity. The utterances are classified into ten categories of expert-evaluated emotional labels. Due to the activation and valence domain of Excited and Happy being close, the utterances labeled Excited are merged into the Happy category for dataset augmentation [1]. In this paper, we employed four widely utilized emotional labels (Neutral, Sad, Happy, and Angry) [20] to compare and analyze the performance of SD-CAFF. In addition, we also adopt the gender labels for the multi-label auxiliary learning.
- **MSP-IMPROV:** In this experiment, only utterances marked as improvisation were utilized similar to IEMOCAP, and gender labels and four standard emotional labels are retained [2].

3.2 Experiment Setup

Implementation Details In this study, the performances on IEMOCAP and MSP-IMPROV datasets are regarded as the evaluation criteria of the SD-CAFF. This paper randomly divided the utterances in each dataset into five clusters for 5-fold cross-validation.

Three acoustic features (MFCC, Delta, and Delta-delta) are extracted from the utterances segments employing the librosa library. After setting the parameters in [11], MFCC tensors in the shape of $(mfcc \times time)$ is obtained, where $mfcc$ equals 60 representing the MFCC coefficients, and $time$ equals 251 which is the number of frames. The Delta and Delta-delta coefficients are calculated in the MFCC dimension $(axis = -2)$ of the tensor, and the feature obtained is consistent with the shape of the MFCCs. Finally, 12402 IEMOCAP data and 21895 MSP-IMPROV data are obtained after speech signal preprocessing. Each data comprises an MFCC feature, a Delta feature, a Delta-delta feature, and the corresponding gender and emotion labels.

In order to implement the joint loss function, two sets of hyperparameters are introduced: $center\_rate : lr\_cent = 0.15 : 0.1$; $alpha : beta = 7 : 3$, where $center\_rate$ is the center loss ratio, lr_cent is the center loss learning rate, and $alpha : beta$ is the proportion of emotion and gender labels. The Adam optimizer with a learning rate of 5×10^{-3} calculates the gradient of the method, and the number of epochs and batch size respectively are 30 and 32. In addition, the entire training process was implemented on a GeForce RTX 3090 with 24 GB memory, and the CPU was a 15-core AMD EPYC 7543 32-core processor. The code was written by Python 3.8 according to the framework PyTorch 1.10.0 and released on https://github.com/AdriaKD/SD-CAFF.git.

Evaluation Criteria Two kinds (WA and UA) of accuracy are utilized as the criteria to measure the performance of the proposed method and other comparison methods. The weighted accuracy (WA) can be calculated by Eq. (9).

$$WA = \frac{\sum_{\mathrm{n=1}}^{C} \left(\frac{N_C}{N_T} * ACC_{\mathrm{class}} \right)}{C} \tag{9}$$

where N_C represents the total number of samples in specific class, N_T represents the total number of samples, ACC_{class} represents the accuracy of the class, and C represents the total number of classes. In addition, the unweighted accuracy (UA) can be obtained by Eq. (10).

$$UA = \frac{\sum_{\mathrm{n=1}}^{C} ACC_{\mathrm{class}}}{C} \tag{10}$$

where ACC_{class} is the accuracy of the class, and C indicates the total number of classes. UA pays more attention to the average performance of the model on each category compared with WA. Therefore, UA is more advanced in verifying the recognition performance of method between different classes.

3.3 Experimental Results

Feature Selection

In this paper, a feature selection experiment is designed to select the best dynamic input of SD-CAFF. To be specific, we compare the WA and UA of SD-CAFF with different acoustic features combinations on IEMOCAP. We apply $\oplus$ to represent the concatenation of different features, i.e., $mfcc \oplus delta$ refers to the concatenation of MFCC and Delta.

Table 1: The dynamic feature selection experimental results (%) of SD-CAFF on IEMOCAP

The dynamic feature	Overall Acc (WA)	Class Acc (UA)
MFCC	75.06	74.05
Delta	73.80	72.07
Delta-delta	73.18	71.94
Delta + Delta-delta	74.39	73.36
MFCC + Delta	75.77	74.89
MFCC + Delta-delta	75.37	73.96
MFCC + Delta + Delta-delta	74.96	73.99

in Tab. 1, $mfcc \oplus delta$ achieves the best performance (WA: 75.78%; UA: 74.89%) compared to other combinations. Therefore, $mfcc \oplus delta$ is selected as the final dynamic acoustic feature combination and adopted as the input of LSTM feature extraction module.

Comparison with State-of-the-Art Networks In order to further validate the performance of SD-CAFF in speech emotion recognition, we compare SD-CAFF with a series of current state-of-the-art methods (SOTAs).

Table 2: The Overall Acc (WA) (%) and Class Acc (UA) (%) of SD-CAFF and SOTAs on IEMOCAP and MSP-IMPROV

Methods	IEMOCAP		MSP	
	WA	UA	WA	UA
Jiaxing L. (TFCNN+DenseCap+ELM) [10]	70.34	70.78	-	-
Anish N. (MHA+PE+MTL) [12]	76.40	70.10	-	-
Huilian L. (CNN-BLSTM) [6]	74.14	65.62	-	-
Qi C. (HNSD) [3]	72.50	70.50	-	-
S. Latif (Semi-supervised AAE) [8]	-	68.80	-	63.60
Amir S. (CycleGCN) [13]	65.29	62.27	57.82	55.42
Bo-Hao S. (GA-GRU) [14]	62.27	63.80	56.21	57.47
SD-CAFF (our model)	**75.78**	**74.89**	**74.89**	**74.89**

As shown in Tab. 2, SD-CAFF achieves the UA of 74.89% in the IEMOCAP, which is significantly higher than other SOTAs. Also, SD-CAFF obtains the best class balance among all models as its UA is 4.11% higher than the best UA in the control group. Although the best WA in the control group is 76.4% achieved by MHA+PE+MTL, which is silightly higher than SD-CAFF (WA: 75.78%), the UA (70.1%) of MHA+PE+MTL is 4.79% lower than the UA (74.89%) of SD-CAFF. Furthermore, this paper validate SD-CAFF on MSP-IMPROV dataset to evaluate the model performance in the cross-corpus issue. Although the performance of the SD-CAFF on the MSP-IMPROV is slightly lower than some SOTAs specially designed for cross-corpus, it still achieves competitive performance (WA: 75.78%; UA: 74.89%). Compared to the CycleGCN model, SD-CAFF is 0.8% higher in UA. Simultaneously, the weighted accuracy of SD-CAFF is 0.56% higher than that of GA-GRU. Note that both the WA and UA of SD-CAFF on IEMOCAP datasets are much higher than CycleGCN (WA: 65.29%; UA: 62.27%) and GA-GRU (WA: 62.27%; UA: 63.8%). It thus can be found that the performance on speech emotion recognition and the cross-corpus competence can be enhanced by the proposed SD-CAFF.

3.4 Ablation Study

In the ablation study, three baselines (CNN baseline, LSTM baseline, and Concat baseline) are designed to analyze the impact of the dynamic-static feature fusion mechanism. Single-view features are directly employed in CNN/LSTM baselines, while the multi-view features are fused by Concat baseline applying the simple operation. The structure of these baselines are shown in Fig. 4.

In order to eliminate irrelevant effects on experiments, the other training parameters of the baselines were consistent with SD-CAFF. In addition, the

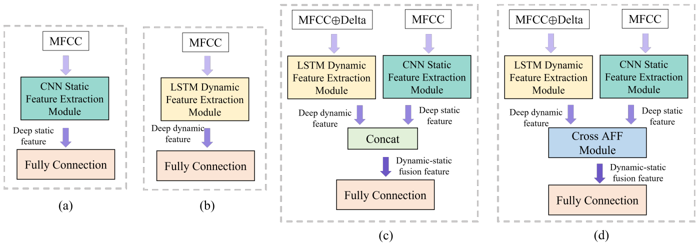
MFCC
CNN Static Feature Extraction Module
Deep static feature
Fully Connection
(a)
MFCC
LSTM Dynamic Feature Extraction Module
Deep dynamic feature
Fully Connection
(b)
MFCC⊕Delta
MFCC
LSTM Dynamic Feature Extraction Module
CNN Static Feature Extraction Module
Deep dynamic feature
Deep static feature
Concat
Dynamic-static fusion feature
Fully Connection
(c)
MFCC⊕Delta
MFCC
LSTM Dynamic Feature Extraction Module
CNN Static Feature Extraction Module
Deep dynamic feature
Deep static feature
Cross AFF Module
Dynamic-static fusion feature
Fully Connection
(d)

Fig. 4: (a). The structure of CNN Baseline; (b). The structure of LSTM Baseline; (c). The structure of Concat Baseline; (d). The structure of SD-CAFF

entire ablation experiment are conducted on IEMOCAP. In Tab. 3, the WA and UA of Concat baseline is higher than that of single-view baselines, and the overall performance of SD-CAFF is superior to Concat baseline. Therefore, we can conclude that combining multi-view deep features is beneficial to extract complementary information in the dynamic-static feature spaces. In addition, the Cross AFF module can better mine the complementary information from the multi-view deep features in comparision with concatenation.

Table 3: Comparative experimental results (%) of baseline methods and SD-CAFF on IEMOCAP

Methods	Neu	Sad	Ang	Hap	WA	UA
CNN Baseline	80.70	77.14	66.09	65.91	73.81	73.06
LSTM Baseline	77.25	76.64	58.48	56.41	69.18	67.81
Concat Baseline	82.80	77.14	62.28	66.97	74.07	73.17
SD-CAFF(final)	**83.99**	**79.61**	**66.44**	**66.82**	**75.78**	**74.89**

4 Conclusion

This paper proposes the dynamic-static cross attentional feature fusion method (SD-CAFF) to improve speech emotion recognition accuracy. SD-CAFF is an attentional feature fusion method, which can parallel fuse the muti-view features effectively by the cross attentional feature fusion module (Cross AFF). Cross AFF with symmetric structure can equivalently fuse the static and dynamic features from the feature extraction modules. In this paper, CNN and LSTM are utilized as feature extraction modules to extract deep features from acoustic features. The CNN static feature extraction module is first utilized to recognize the deep information in the static acoustic feature (MFCC). Then, LSTM is implemented to extract the underlying associations in the combination of static acoustic features (MFCC) and dynamic acoustic features (Delta and Delta-delta). In addition, the training process utilizes a multi-label auxiliary learning loss to enhance the performance of the proposed model. The WA and the UA are applied to measure the performance of SD-CAFF. The experimental results on the benchmark datasets demonstrate that SD-CAFF (WA: 75.78%; UA: 74.89%) achieved superior performance in comparation with the current SOTAs in SER. Furthermore, the rigor of SD-CAFF structure and the necessity of each module are verified by ablation experiment.

References

1. Busso, C., Bulut, M., C, L.C., Kazemzadeh, A., Mower, E., Kim, S., ..., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, **42**(4), pp. 335–359 (2008), `https://doi.org/10.1007/s10579-008-9076-6`
2. Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., Provost, E.M.: Msp-improv: An acted corpus of dyadic interactions to study emotion perception. IEEE Transactions on Affective Computing, **8**(1), pp. 67–80 (2016), `https://doi.org/10.1109/TAFFC.2016.2515617`
3. Cao, Q., Hou, M., Chen, B., Zhang, Z., Lu, G.: 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). pp. 6334–6338. IEEE ((2021)), `https://doi.org/10.1109/icassp39728.2021.9414540`
4. Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic relu. In: European Conference on Computer Vision,. pp. 351–367. Springer ((2020)), `https://doi.org/10.1007/978-3-030-58529-7_21`
5. Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., Barnard, K.: Attentional feature fusion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3560–3569 ((2021)), `https://doi.org/10.1109/WACV48630.2021.00360`
6. Huilian, L., Weiping, H., Yan, W.: Speech emotion recognition based on blstm and cnn feature fusion. In: Proceedings of the 2020 4th International Conference on Digital Signal Processing. pp. 169–172 ((2020)), `https://doi.org/10.1145/3408127.3408192`
7. Lambrecht, L., Kreifelts, B., Wildgruber, D.: Gender differences in emotion recognition: Impact of sensory modality and emotional category. Cognition & Emotion, **28**(3), pp. 452–469 ((2014)), `https://doi.org/10.1080/02699931.2013.837378`
8. Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J., W., S.B.: Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. IEEE Transactions on Affective Computing ((2020)), `https://doi.org/10.1109/taffc.2020.2983669`
9. Li, Y., Baidoo, C., Cai, T., Kusi, G.A.: Speech emotion recognition using 1d cnn with no attention. In: 2019 23rd International Computer Science and Engineering Conference (ICSEC). pp. 351–356. IEEE ((2019)), `https://doi.org/10.1109/ICSEC47112.2019.8974716`
10. Liu, J., Liu, Z., Wang, L., Guo, L., Dang, J.: Speech emotion recognition with local-global aware deep representation learning. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7174–7178. IEEE ((2020)), `https://doi.org/10.1109/icassp40776.2020.9053192`

11. Liu, L.Y., Liu, W.Z., Zhou, J., Deng, H.Y., Feng, L.: Atda: Attentional temporal dynamic activation for speech emotion recognition. Knowledge-Based Systems, **243**, pp. 108472–108472 (2022), https://doi.org/110.1016/j.knosys.2022.108472
12. Nediyanchath, A., Paramasivam, P., Yenigalla, P.: Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7179–7183. IEEE ((2020)), https://doi.org/10.1109/icassp40776.2020.9054073
13. Shirian, A., Guha, T.: Compact graph architecture for speech emotion recognition. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6284–6288. IEEE ((2021)), https://doi.org/10.1109/icassp39728.2021.9413876
14. Su, B.H., Chang, C.M., Lin, Y.S., Lee, C.C.: Improving speech emotion recognition using graph attentive bi-directional gated recurrent unit network. In: Interspeech. pp. 506–510 ((2020)), https://doi.org/10.21437/interspeech.2020-1733
15. Sun, B., Wei, Q., Li, L., Xu, Q., He, J., Yu, L.: Lstm for dynamic emotion and group emotion recognition in the wild. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. pp. 451–457 ((2016)), https://doi.org/10.1145/2993148.2997640
16. Sun, S.: A survey of multi-view machine learning. Neural Computing and applications, **23**(7), pp. 2031–2038 (2013), https://doi.org/10.1007/s00521-013-1362-6
17. Ullah, A., Muhammad, K., Del Ser, J., Baik, S.W., de Albuquerque, V.H.C.: Activity recognition using temporal optical flow convolutional features and multilayer lstm. IEEE Transactions on Industrial Electronics, **66**(12), pp. 9692–9702 (2018), https://doi.org/10.1109/TIE.2018.2881943
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ..., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems, **30** (2017), https://doi.org/10.5555/3295222.3295349
19. Yang, J., Yang, J.Y., Zhang, D., Lu, J.F.: Feature fusion: parallel strategy vs. serial strategy. Pattern Recognition, **36**(6), pp. 1369–1381 (2003), https://doi.org/10.1016/S0031-3203(02)00262-5
20. Yoon, S., Byun, S., Jung, K.: Multimodal speech emotion recognition using audio and text. In: 2018 IEEE Spoken Language Technology Workshop (SLT). pp. 112–118. IEEE ((2018)), https://doi.org/10.1109/SLT.2018.8639583