Machine Intelligence Research
www.mi-research.net

20(4), August 2023, 595-604
DOI: 10.1007/s11633-022-1356-x

Speech Emotion Recognition Using Cascaded Attention Network with Joint Loss for Discrimination of Confusions

Yang Liu* Haoqin Sun* Wenbo Guan Yuqi Xia Zhen Zhao

School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China

Abstract: Due to the complexity of emotional expression, recognizing emotions from the speech is a critical and challenging task. In most of the studies, some specific emotions are easily classified incorrectly. In this paper, we propose a new framework that integrates cascade attention mechanism and joint loss for speech emotion recognition (SER), aiming to solve feature confusions for emotions that are difficult to be classified correctly. First, we extract the mel frequency cepstrum coefficients (MFCCs), deltas, and delta-deltas from MFCCs to form 3-dimensional (3D) features, thus effectively reducing the interference of external factors. Second, we employ spatiotemporal attention to selectively discover target emotion regions from the input features, where self-attention with head fusion captures the long-range dependency of temporal features. Finally, the joint loss function is employed to distinguish emotional embeddings with high similarity to enhance the overall performance. Experiments on interactive emotional dyadic motion capture (IEMOCAP) database indicate that the method achieves a positive improvement of 2.49% and 1.13% in weighted accuracy (WA) and unweighted accuracy (UA), respectively, compared to the state-of-the-art strategies.

Keywords: Speech emotion recognition (SER), 3-dimensional (3D) feature, cascaded attention network (CAN), triplet loss, joint loss.

Citation: Y. Liu, H. Sun, W. Guan, Y. Xia, Z. Zhao. Speech emotion recognition using cascaded attention network with joint loss for discrimination of confusions. *Machine Intelligence Research*, vol.20, no.4, pp.595–604, 2023. http://doi.org/10.1007/s11633-022-1356-x

1 Introduction

As artificial intelligence continues to evolve, the field of affective computing has attracted a lot of interest from researchers. The purpose of emotion recognition, a major branch of affective computing, is to recognize significant information from the data, including face, speech, and text. Speech is a reliable source of emotional information among these data, containing not only textual content as well as paralinguistic elements such as emotions. In recent years, speech emotion recognition (SER) has been extensively utilized in many fields, including distance education, personalized customer service, and medical science. However, SER remains a difficult problem because of the variation in speech and the complication of expressed emotions. Therefore, the works of SER have received much focus in recent years.

Feature extraction is a vital process in SER systems, aiming to produce valid high-level feature representations for various emotions. In terms of acoustic feature extraction, several feature sets based on low-level descriptors (LLDs) and high-level statistic functionals (HSFs), including INTERSPEECH-2010, GeMAPS, AVEC-2013, and ComParE, have been developed for a wide range of applications[1]. However, these hand-crafted feature sets might not be ideal for representing emotions in speech, thus leading to suboptimal performance. With the increase in computing power, deep learning has become mainstream, which provides superior capabilities for high-level feature capturing. For example, Schmidt and Kim[2] employed the deep belief network (DBN) to extract the representation of emotional features from the magnitude spectrum and exhibited improved performance compared to hand-crafted features. Han et al.[3] used the highest energy segments to train a deep neural network (DNN) to extract the mel frequency cepstrum coefficients (MFCCs) and pitch, which contain valid emotional information. Mao et al.[4] first employed a convolutional neural network (CNN) to extract emotional highlighted features for SER, which displayed excellent results on a few common datasets.

Although DNNs have gained enormous success in the field of SER, they still utilize personalized features as input that are sensitive to various speaking styles, speech contents, and environments. Despite that most of the current research for SER is related to personalized emotional features and has gained excellent recognition performance, particularly for specific speakers, it is still challenging to decrease huge differences in individualized features for different speakers and the modes of speaking. Recently, researchers introduced the rate of change to re-

Research Article
Manuscript received on April 26, 2022; accepted on July 8, 2022; published online on June 1, 2023
Recommended by Associate Editor Jian-Hua Tao
Colored figures are available in the online version at https://link.springer.com/journal/11633
* These authors contribute equally to this work

flect the change in emotion during speech as a way to obtain speaker-independent and stable speech emotion features. For example, Chen et al.[5] utilized deltas and delta-deltas of log mel-spectrogram to reduce the differences in personalized features across individuals. However, the deltas and delta-deltas of personalized features reflect the process of emotional changes and retain valid affective information while decreasing the impact of emotional-irrelevant factors. In addition, MFCCs are features that accurately describe the short-time power spectral envelope of speech, which combines the auditory perceptual properties of the human ear with the mechanism of speech production and emphasizes prior knowledge of human beings. Therefore, we compute MFCCs, deltas, and delta-deltas as inputs to the proposed network in this paper.

Since emotions are expressed just in specific moments of utterance, how to efficiently highlight emotional-relevant regions becomes a pivotal point for SER. In recent years, the attention mechanism has attracted widespread focus[6, 7]. Chen and Huang[8] proposed dual attention-based bidirectional long short-term memory (BLSTM) network in order to extract sequence information from MFCCs and spatiotemporal features from log-mels for SER. Nevertheless, for 3-dimensional feature inputs, the attention mechanism treats each channel′s attention equally, while the features of each channel have different contributions to emotion. To address this issue, a recently developed spatiotemporal attention consists of channel attention and spatial attention, which could focus on the meaningful feature detectors and the emotional regions. Spatiotemporal attention is verified to be effective in any computer vision and SER tasks. However, spatiotemporal attention cannot capture information over long distances. Liu et al.[9] introduced self-attention, which disregards the distance between features to directly calculate the dependencies, thus capturing relevant information in the features. However, self-attention focuses on the overall feature rather than the regional information about the feature. Xu et al.[10] proposed self-attention with head fusion to generate multiple subspaces to produce feature points that highlight emotions, allowing the model to focus on different aspects of information. Compared to the general self-attention, the self-attention with head fusion allows the model to focus on each part of the feature representation in detail rather than the overall features at once. It is noted that channel attention locates the targeted emotional detectors, and spatial attention locates the targeted emotional areas from detectors selected by channel attention, while the self-attention with head fusion focuses on the degree of dependency of each part of the feature representation, thus capturing long-distance dependencies. This demonstrates that spatiotemporal attention and self-attention with head fusion may be complementary to each other. Therefore, they are integrated into the cascaded attention network (CAN) to further enhance the SER performance in this paper.

Another critical issue in SER is the confusion of features among emotion classes. Feature confusion is a phenomenon in which the clusters of features from different emotion classes might overlap with each other. In most previous studies, some specific emotions have been wrongly classified. From the results shown in the interactive emotional dyadic motion capture (IEMOCAP) dataset[11], it is observed that the utterances with happy and neutral labels are seriously confused. We consider it to be the result of similar levels of happy and neutral activation, and the subtle difference could not be captured by the model. To tackle this issue, Sahu et al.[12] constrained the feature distribution through an adversarial loss. However, there is no expanded decision margin between the various categories. Dai et al.[13] utilized center loss and cross-entropy loss to obtain discriminative features from variable length spectrograms, which significantly enhanced the intra-class compactness. However, in the case of short inter-class distances, the decision margin was not clear. To address this problem, Gao et al.[14] developed a framework for metric learning-based feature representation. They used triplet loss to solve the problem of emotions being grossly misclassified. Features from the same class are pulled closer, and features from different classes are pulled further apart by triple loss, which eliminates the effect of feature confusion and extracts more distinguishable emotion features. Inspired by the advantages of triplet loss, we introduce cross-entropy loss and triplet loss as a joint loss function to ensure improved classification accuracy.

In this paper, we propose an architecture that integrates a cascaded attention network with a joint loss for the SER. First, we extract MFCCs, deltas, and delta-deltas of MFCCs features from the original speech to reduce the interference caused by factors unrelated to emotion such as speaker, content, and environment. Second, to enable the neural network to learn parts of speech with salient emotion, we introduce the cascaded attention network to locate a few targeted emotional areas, where channel attention is used to select the important feature detectors, spatial attention is used to locate the location of emotional regions, and self-attention with head fusion captures long-distance dependencies. Finally, we introduce a novel joint loss strategy consisting of triplet loss and cross-entropy loss, which constitutes the heart of our contribution to solving the feature confusion problem. It explicitly promotes intra-class compactness and inter-class severability among the learned features, which gives rise to larger decision margin to improve the overall classification accuracy. Experimental results on the benchmark dataset IEMOCAP demonstrate that our method achieves 80.34% and 77.91% in weighted accuracy (WA) and unweighted accuracy (UA), respectively.

The remainder of this paper is organized as follows. Section 2 provides the proposed method in detail. Section 3 provides the experimental setup in detail. Section 4 provides the experimental results. Section 5

analyzes the experimental results. Finally, Section 6 presents our conclusions and future work.

2 Proposed method

Fig. 1 depicts an overview of the proposed framework. First, we calculate the MFCCs (static, deltas, and delta-deltas) from the speech signals as input for CAN. Secondly, we briefly describe the architecture of CAN, which integrates CNN and BLSTM with a cascaded attention mechanism, followed by a fully connected layer. Finally, the joint loss function enhances inter-class severability and intra-class compactness.

2.1 Acoustic feature extraction

Given a speech signal, we first reduce the variation among different speakers by utilizing normalization. Then, the speech signal is split into short frames with Hamming windows of 40 ms and a window shift of 10 ms. Finally, spectrograms are mapped to mel-scale by mel filters, and we use the logarithm to calculate the MFCCs and extract the deltas and delta-deltas of MFCCs simultaneously, which are used as the input to CAN as x_i, where i represents the i-th sample.

2.2 Cascaded attention network

We perform feature extraction by CAN. In the proposed framework, firstly, CNN is utilized for extracting spatial features. Next, spatiotemporal attention is proposed to extract emotional features from 3-dimensional (3D) features. Then, BLSTM is used for extracting the time sequence feature. Finally, self-attention with head fusion is used for capturing long-distance dependencies.

CNN. We use four convolution layers normalized after each convolution layer, where the kernel size of the first three convolution layers is set to 3×3, and the padding is set to 1. Next, frequency information is convolved in the last convolution layer to calculate $\boldsymbol{F}_0$, which is fed into the channel attention. Table 1 describes the details of each convolutional layer.

Channel attention. Due to the fact that the channel of the feature maps is regarded as feature detectors, channel attention concentrates on the important feature detectors, when searching for regions related to emotion.

We first compress each channel of $\boldsymbol{F}_0$ to generate the spatial descriptors $\boldsymbol{F}_c^{\max}$ and $\boldsymbol{F}_c^{\text{avg}}$ by max-pooling and average-pooling operations. Max-pooling features collect contributions from all features, and average-pooling features collect important clues about distinctive audio features. These descriptors are then used as input to the shared network to obtain the feature weights. The activation function Relu is used to normalize the feature weights. Finally, after the shared network is applied to each channel, the channel weight and the input feature map are summed using an element-wise summation to generate the feature map $\boldsymbol{F}_c$. We calculate the channel attention:

$$\begin{aligned}\boldsymbol{F}_c &= \sigma(S(\text{Avg}(\boldsymbol{F}_0)) + S(\text{Max}(\boldsymbol{F}_0))) = \\ &\sigma\left(\boldsymbol{W}_1\left(\boldsymbol{W}_0\left(\boldsymbol{F}_c^{\text{avg}} + \boldsymbol{F}_c^{\max}\right)\right)\right)\end{aligned} \tag{1}$$

where σ represents sigmoid function, S represents the shared network, and $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$ represent the shared weight matrices.

Spatial attention. Compared to channel attention, spatial attention focuses on the emotion regions of feature detectors selected by channel attention, which is complementary to channel attention.

In order to calculate spatial attention, first, we calcu-

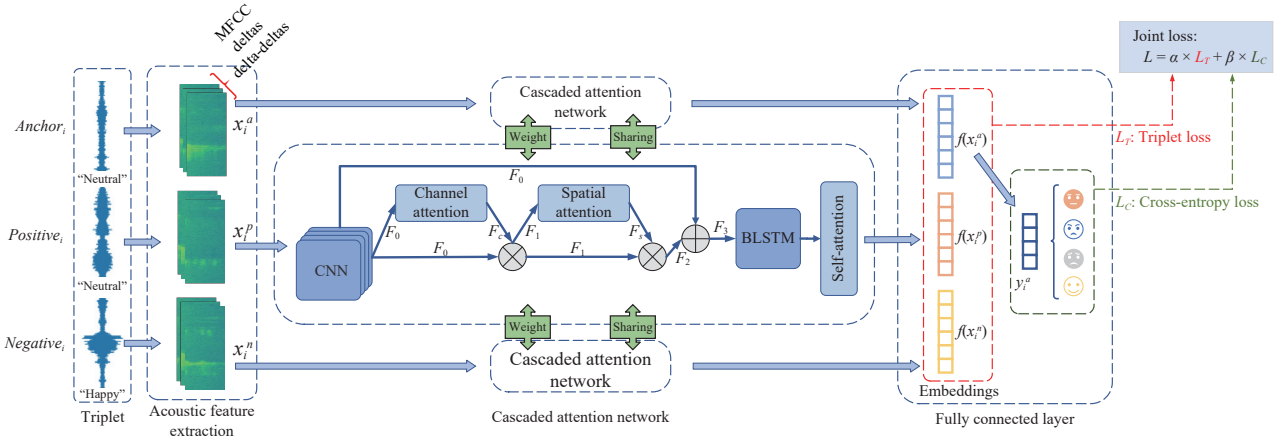


Fig. 1 Details of the proposed method. This network consists of nine phases: 1) Triplet selection. 2) The extracted MFCCs, deltas, and delta-deltas features are used as input acoustic features for CAN. 3) CNN is used to extract spatial features. 4) Channel attention is used to discover the emotion-relevant parts of each channel. 5) Spatial attention is used to complement channel attention and discover emotion-relevant parts of space. 6) BLSTM is used to extract temporal features. 7) Self-attention with head fusion is used to focus on capturing long-distance dependencies. 8) The fully connected layer is utilized to obtain high-level feature representations for better classification accuracy. 9) Joint loss is used to expand the distance between different emotion samples and fuse the label information to learn a better feature representation.

Table 1 Details of four convolution layers

Layer	$kernel\_size$	$channel_{in}$	$channel_{out}$	Padding
Conv1	(3, 3)	3	8	1
Conv2	(3, 3)	8	16	1
Conv3	(3, 3)	16	32	1
Conv4	(26, 1)	32	64	0

late $\boldsymbol{F}_1$ by multiplying $\boldsymbol{F}_0$ and $\boldsymbol{F}_c$. Then, we aggregate the channel information of the feature maps to obtain $\boldsymbol{F}_s$ by using the max-pooled features $\boldsymbol{F}_s^{\max}$ and average-pooled features $\boldsymbol{F}_s^{\text{avg}}$. Then, we calculate the spatial attention:

$$\begin{aligned}\boldsymbol{F}_s(\boldsymbol{F}_1) = \sigma\left(f([\text{Avg}(\boldsymbol{F}_1); \text{Max}(\boldsymbol{F}_1)])\right) = \\ \sigma\left(f\left(\left[\boldsymbol{F}_s^{\text{avg}}; \boldsymbol{F}_s^{\max}\right]\right)\right)\end{aligned} \tag{2}$$

where f represents a convolution operation with a filter size of 7×7, and σ represents the sigmoid function.

BLSTM. We first utilize the sequence features $\boldsymbol{F}_0$ extracted from the four convolutional layers to perform a skip connection via a shortcut and perform an element-wise summation operation between $\boldsymbol{F}_0$ and $\boldsymbol{F}_2$ to obtain $\boldsymbol{F}_3$, where $\boldsymbol{F}_2$ is calculated by multiplying $\boldsymbol{F}_1$ and $\boldsymbol{F}_s$. Then, $\boldsymbol{F}_3$ is activated by the activation function Relu. Finally, the activated values are fed into the BLSTM network with 32 cells per direction.

Self-attention with head fusion. Different from spatiotemporal attention, self-attention intends to capture the long-range dependencies of the BLSTM layers output. We calculate $\boldsymbol{X}_j^{attn}$ by using different parameter sets $\boldsymbol{W}_j^Q$, $\boldsymbol{W}_j^K$, $\boldsymbol{W}_j^V$, $j \in (0, n_{head}]$ as follows:

$$Att(\boldsymbol{W}_j^Q\boldsymbol{Q}, \boldsymbol{W}_j^K\boldsymbol{K}, \boldsymbol{W}_j^Q\boldsymbol{V}) = sft(\boldsymbol{Q}_j{\boldsymbol{K}_j}^T)\boldsymbol{V}_j \tag{3}$$

$$\boldsymbol{X}_j^{attn} = Att(\boldsymbol{Q}_j, \boldsymbol{K}_j, \boldsymbol{V}_j) \tag{4}$$

where each $\boldsymbol{X}_j^{attn}$ is referred to as a head. $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are equal to the output of the BLSTM. sft represents the softmax function. $\boldsymbol{W}$ represents trainable parameters. Then, we superimpose the heads to produce an attention map $\boldsymbol{X}_m$ as follows:

$$\boldsymbol{X}_m = \frac{\sum_0^{n_{head}-1} \boldsymbol{X}_j^{attn}}{n_{head}}. \tag{5}$$

Finally, we generate a feature map for $\boldsymbol{X}_m$ using global average pooling (GAP) as the output of CAN.

2.3 Joint loss

As shown in Fig. 2, the triplet loss training strategy learns an embedding space where the samples from the same class are nearer to others than those from different classes.

The distance between positive samples and anchor samples, which possess the same emotion labels, is minimized by triplet loss, and the distance between negative samples and anchor samples, which possess different emotion label, is maximized by triplet loss. We calculate the triplet loss:

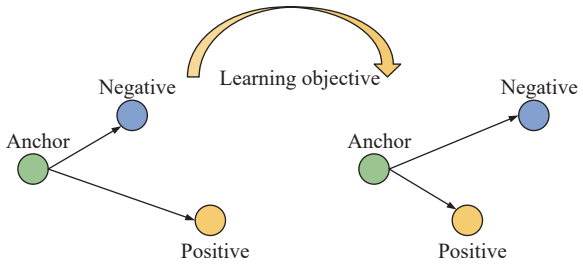


Fig. 2 Learning objective of triplet loss

$$pdist = \|f(x_i^a) - f(x_i^p)\|_2 \tag{6}$$

$$ndist = \|f(x_i^a) - f(x_i^n)\|_2 \tag{7}$$

$$L_T = \max(pdist - ndist + M,\ 0) \tag{8}$$

where x_i^a represents anchor samples, x_i^p represents positive samples, and x_i^n represents negative samples. $f(x_i^a)$, $f(x_i^p)$, and $f(x_i^n)$ represent the embeddings learned by CAN from the anchor, positive, and negative samples, respectively. $\|f(x_i^a) - f(x_i^p)\|_2^2$ and $\|f(x_i^a) - f(x_i^p)\|_2^2$ are the Euclidean distances between the features learned from positive pairs and negative pairs. M represents a minimum distance between two Euclidean distances.

We combine the cross-entropy loss and the triplet loss, assigning different weights α and (1–α), where cross-entropy loss contains emotion label information. The joint loss L is defined as

$$L_C = -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} y_{i,k} \ln p_{i,k} \tag{9}$$

$$L = \alpha \times L_T + (1\text{–}\alpha) \times L_C \tag{10}$$

where L_C represents the cross-entropy loss function. $y_{i,k}$ represents the true probability distribution and the true label of the i-th sample as k. $p_{i,k}$ represents the probability output of the softmax layer and the probability that the i-th sample is predicted to be the k-th label. N represents the number of triplets.

3 Experiments

3.1 Dataset

In order to assess the proposed method, we selected the IEMOCAP corpus, which contains about twelve

hours of audio data from 10 performers. IEMOCAP consists of five sessions, where one female and one male actor perform in improvised and scripted scenes during each session. This corpus includes 10 039 utterances, where emotions such as angry, happy, sad, neutral, surprise, fear, excited, disgust, and frustrated were annotated by at least three expert evaluators. Consistent with prior methods[10, 15], the experiment uses four emotions: happy, angry, neutral, and sad, where the excited class, which replaces the happy class, was selected only from the improvised utterances.

3.2 Experimental setup

We randomly select 80% of the corpus as the training corpus and the other 20% as the test corpus. In the evaluation, we employ the 5-fold cross-validation approach. In our training set, the utterances are split into segments, where the window length is 2 s and the window shift is 1 s. In our testing set, the window shift is set to 1.6 seconds, which is consistent with [10]. We obtain the last prediction by averaging the prediction of all segments of an original utterance in the test process. Adam optimizer is selected with the initial learning rate of 10^{-4} and weight decay of 10^{-6}. We set the batch size to 32. We use weighted accuracy (WA) and unweighted accuracy (UA) as evaluation criteria.

3.3 Triplet selection

Since not all triplets contain information that contributes significantly to training, and some even lead to slower convergence of the model, the selection of triplets is very important to enhance the model performance[16].

In the experiment, we randomly select anchor, positive and negative samples to form triplets. Then, during the training of batches, we choose the hard triplets as follows:

$$\|f(x_i^a) - f(x_i^n)\|_2^2 < \|f(x_i^a) - f(x_i^p)\|_2^2 \qquad (11)$$

to calculate the triplet loss. The choice of hard triplets allows the training to concentrate on undistinguishable samples and decreases computation complexity.

4 Results

4.1 Impact of joint loss weight and triplet margin

Fig. 3 presents the WA and UA assessed at various joint loss weights and triplet margins. We achieve the highest WA and competitive UA when α is set to 0.3 and M is set to 0.2.

As shown in Fig. 3(a), it is difficult for the network to pull the confused samples apart until the weight of the triplet loss α reaches 0.3. With the increase of α from 0.3, the network could not ensure that the divided samples could be classified into the correct class, leading to lower WA and UA.

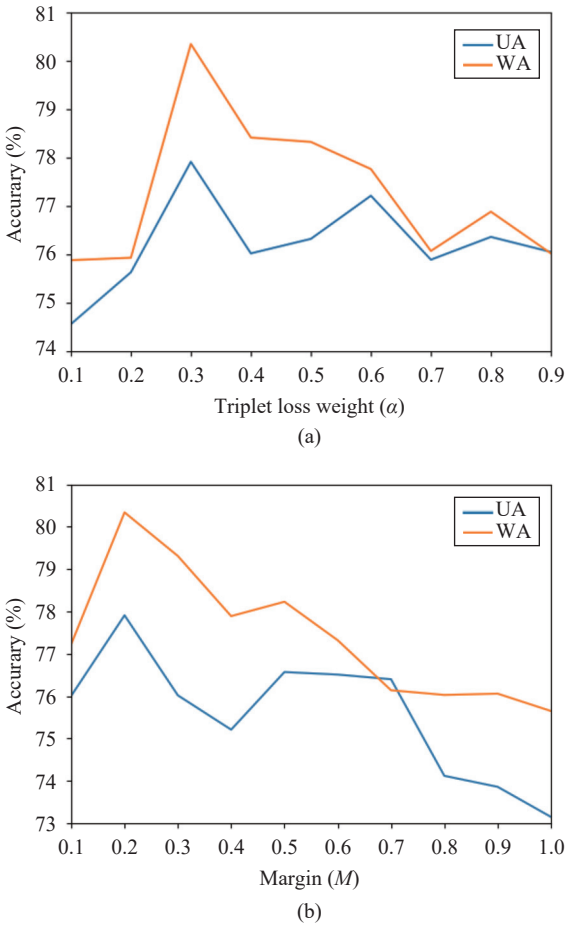


Fig. 3 Impact of different triplet loss weights and margin values. (a) Prediction accuracy of different weights between cross-entropy loss and triplet loss. (1–α) is the cross-entropy loss weight, and α is the triplet loss weight. (b) Prediction accuracy of different margin values.

As shown in Fig. 3(b), before M reaches 0.2, the confused samples are pulled apart to a low degree, making it difficult for the network to distinguish between the anchor and negative samples. As M increases from 0.3, model training becomes more challenging, and the model has difficulty converging with poor model fits, leading to lower WA and UA.

4.2 Ablation study

In this section, we analyze the contribution of 3D features, cascaded attention mechanism, and joint loss to the IEMOCAP corpus.

4.2.1 Effects of 3D features

In this section, in order to evaluate the contribution of individual components as the input of CAN, we have ana-

lyzed the following four conditions. The results of this analysis are shown in Fig. 4 and Table 2.

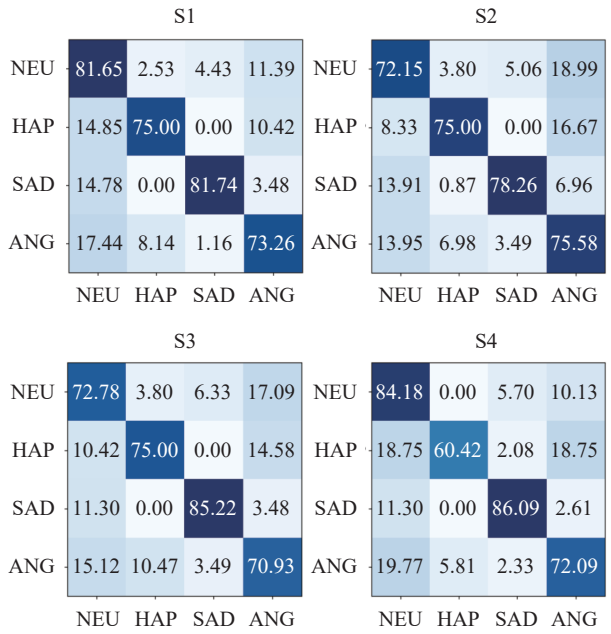


Fig. 4 Confusion matrices for the variants of the proposed approach. M1-1: the proposed approach. M1-2: Proposed without MFCCs. M1-3: Proposed without deltas of MFCCs. M1-4: Proposed without delta-deltas of MFCCs. The diagonal numbers represent the recall rate for each emotion.

Table 2 Ablation study for each component in 3D features. Percentages in bold denote the best-performing method

Model	Methods	WA (%)	UA (%)
M1-1	**Proposed**	**80.34**	**77.91**
M1-2	Proposed without MFCCs	76.17	75.25
M1-3	Proposed without deltas	77.64	75.98
M1-4	Proposed without delta-deltas	79.36	75.70

1) Model 1-1 (M1-1) is the proposed approach.

2) Model 1-2 (M1-2) deletes the MFCCs from M1-1.

3) Model 1-3 (M1-3) deletes the deltas of MFCCs from M1-1.

4) Model 1-4 (M1-4) deletes the delta-deltas of MFCCs from M1-1.

Firstly, to verify the impact of MFCCs, the experimental results in Table 2 indicate that M1-1 has a great degree of improvement over M1-2 and improves by 4.17% on WA and 2.66% on UA in the comparison of M1-1 and M1-2. Compared to M1-2, the method learns the spectral envelope and static characteristics of speech via MFCCs.

Secondly, to verify the impact of deltas and delta-deltas of MFCCs, we have compared the performance of M1-1 with M1-3 and M1-1 with M1-4, respectively. The method improves by 2.7% on WA and 1.93% on UA in the comparison of M1-1 and M1-3. The method improves by 0.98% on WA and 2.21% on UA in the comparison of M1-1 and M1-4. These results show that the deltas and delta-deltas of MFCCs represent non-personalized features, which reflect the dynamic nature of speech and improve the robustness of the network. The approach could improve the classified performance by learning the emotional information included in the non-personalized features. Fig. 4 shows the confusion matrices of these methods, which provide a more visual representation and explicitly reflect the benefits of 3D features.

4.2.2 Effects of cascaded attention mechanism

In this section, in order to evaluate the contribution of individual components in cascaded attention, we have analyzed the following four conditions. The results of this analysis are shown in Fig. 5 and Table 3.

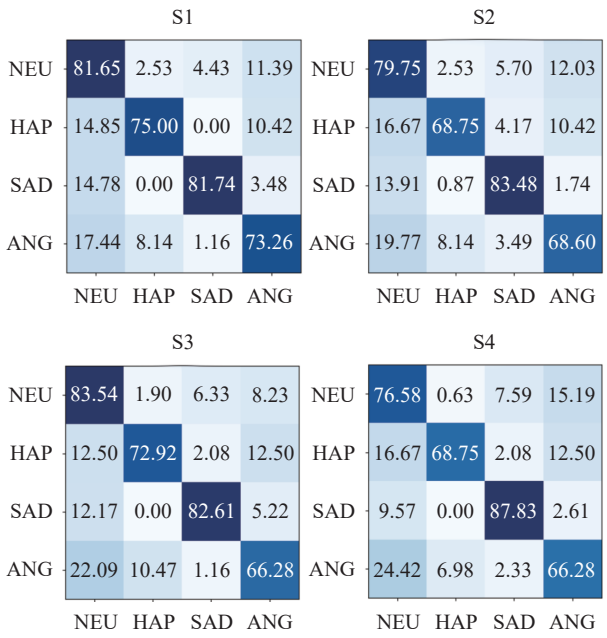


Fig. 5 Confusion matrices for the variants of the proposed approach. M2-1: the proposed approach. M2-2: Proposed without channel attention. M2-3: Proposed without spatial attention. M2-4: Proposed without self-attention. The diagonal numbers represent the recall rate for each emotion.

Table 3 Ablation study for each component in cascaded attention mechanism. Percentages in bold denote the best-performing method

Model	Methods	WA (%)	UA (%)
M2-1	**Proposed**	**80.34**	**77.91**
M2-2	Proposed without channel attention	77.64	75.15
M2-3	Proposed without spatial attention	78.36	76.33
M2-4	Proposed without self-attention	76.76	74.88

1) Model 2-1 (M2-1) is the proposed approach.

2) Model 2-2 (M2-2) deletes channel attention from M2-1.

3) Model 2-3 (M2-3) deletes spatial attention from M2-1.

4) Model 2-4 (M2-4) deletes self-attention from M2-1.

Firstly, in order to validate the impact of channel attention, the experimental results in Table 3 indicate that M2-1 has a great degree of improvement over M2-2 and improves by 2.70% on WA and 2.76% on UA in the comparison of M2-1 and M2-2. The method could improve classification performance by focusing on feature detectors that highlight emotion through channel attention.

Secondly, in order to validate the impact of spatial attention, the method has an improvement of 1.98% on WA and 1.58% on UA in the comparison of M2-1 and M2-3. Spatial attention locates several salient emotional regions of the detectors selected from channel attention, thus improving the classification performance.

Finally, to validate the impact of self-attention with head fusion, M2-1 has a great degree of improvement over M2-4 and improves by 3.58% on WA and 3.03% on UA in the comparison of M2-1 and M2-4. The method captures long-distance dependencies through self-attention with head fusion, which could improve the classification performance. The confusion matrix of these methods is shown in Fig. 5, which provides a more visual representation and explicitly reflects the benefits of the cascaded attention mechanism.

4.2.3 Effects of triplet loss

In this section, to assess the contributions of triplet loss of joint loss, we have analyzed the two conditions below. The results of this analysis are shown in Fig. 6 and Table 4.

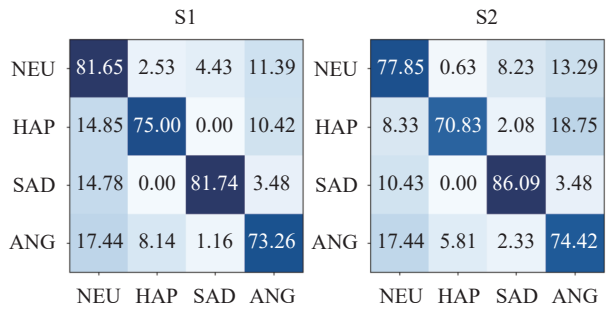


Fig. 6 Confusion matrices for the variants of the proposed approach. M3-1: the proposed approach. M3-2: Proposed without triplet loss. The diagonal numbers represent the recall rate for each emotion.

Table 4 Ablation study for each component in the joint loss. Front of the bold text denotes the best performing method

Model	Methods	WA (%)	UA (%)
M3-1	**Proposed**	**80.34**	**77.91**
M3-2	Proposed without triplet loss	78.05	77.30

1) Model 3-1 (M3-1) is the proposed approach.

2) Model 3-2 (M3-2) deletes triplet loss from M3-1.

To verify the impact of triplet loss, as is shown in Table 4, the model with triplet loss has a great degree of improvement over the model without triplet loss, which achieves a definite improvement of 0.47% on WA and 0.94% on UA in the comparison of the results with and without triplet loss. Similarly, the confusion matrix of these methods is shown in Fig. 6, which provides more visual representation and explicitly reflects the benefits of joint loss. It reveals that the prediction accuracies of angry and sad emotions have been improved by triplet loss. These results demonstrate that triplet loss provides the ability to reduce intra-class distance and expand inter-class distance, thereby enhancing prediction performance.

4.3 Comparison to state-of-the-art approaches

We conducted further experiments in order to validate the effectiveness of our proposed approach. We compared the proposed method with other best methods on the same corpus, where the experimental results are listed in Table 5. The proposed approach shows an absolute improvement of 2.49% on WA and 1.13% on UA compared to state-of-the-art strategies. This is attributed to the excellent ability of the cascaded attention mechanism to focus on salient features and the good discrimination of joint loss. These results strongly prove that the proposed method could produce a positive performance for SER.

Table 5 Classification performance of previous approaches. Percentages in bold denote the best-performing method

Methods	WA(%)	UA(%)
CNN-LSTM [17]	67.30	62.00
3-D ACRNN [5]	–	64.74
CNN-Attn [18]	71.75	68.06
Self-Attn [15]	70.17	70.85
ACNN [10]	76.18	76.36
CNN-LSTM (no augmentation) [16]	77.85	76.78
Proposed	**80.34**	**77.91**

5 Discussions

Compared with all 2D input results, the absolute improvement of WA and UA for the 3D feature combination input is 0.98% and 1.93%, respectively. This shows that the deltas and delta-deltas of MFCCs could preserve valid affective information and reduce the influence of the speaking patterns, speakers, and other factors unrelated to emotion. MFCCs only describe the energy spectrum envelope in a frame of speech. However, the change of MFCC trajectory with time, which is dynamic inform-

ation of speech signal, could be described by these static feature difference spectra. This is crucial to enhance the accuracy of the classification of our model.

To better understand the role of spatiotemporal attention, three samples in a triplet were randomly selected and we plotted their feature maps. For each of the samples, we plotted four kinds of pictures in Fig. 7(a) original feature map of MFCCs; Fig. 7(b) feature map without channel attention; Fig. 7(c) feature map without spatial attention; Fig. 7(d) feature map with spatiotemporal attention.

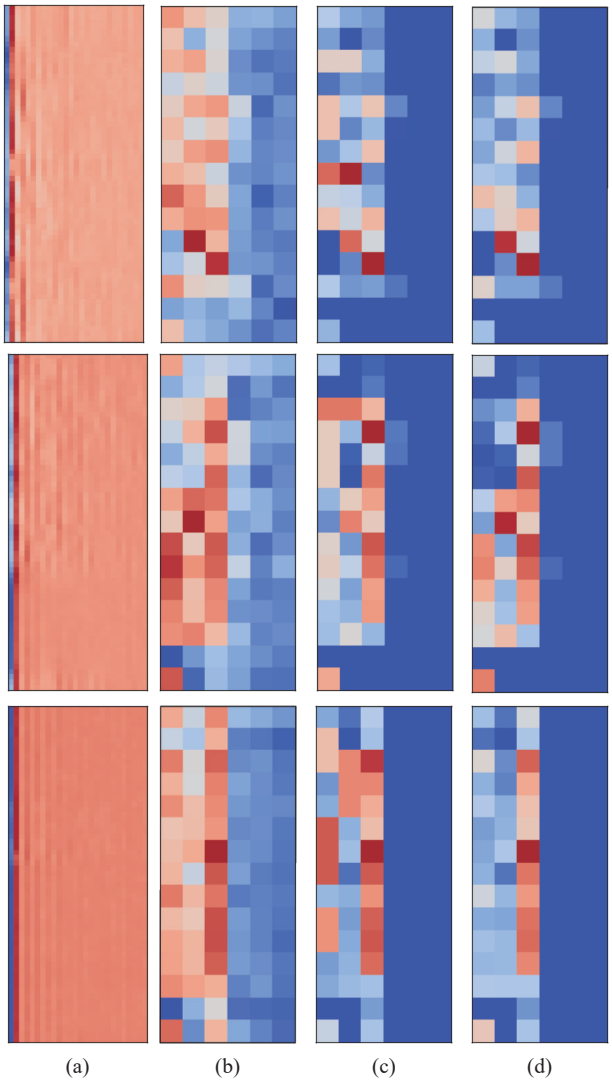


Fig. 7 Feature maps of three samples under different attention mechanisms. (a) Original MFCCs feature maps. (b) Feature maps without channel attention. (c) Feature maps without spatial attention. (d) Feature maps with spatiotemporal attention.

Comparing Fig. 7(d) with their respective feature maps of the original MFCCs, we found that spatiotemporal attention could effectively highlight the speech part of hidden emotional information and significantly suppress the areas unrelated to emotions at the same time, including the silent part of the audio segment. To validate the impact of channel attention, we compared the feature maps in Figs. 7(b) and 7(d). There are significantly more points when channel attention is removed. This indicates that spatial attention plays a suboptimal role in suppressing regions unrelated to emotions. In contrast, in Fig. 7(d), emotionally irrelevant regions are greatly suppressed, and emotionally relevant features are more obvious. To validate the impact of spatial attention, we compare the feature maps in Figs. 7(c) and 7(d). We found that there is no significant difference between both feature maps. However, in Fig. 7(d), certain feature points are more salient in the presence of spatiotemporal attention, and it also suppresses a part of the feature points that are not related to emotions. In conclusion, channel attention plays the main role in spatiotemporal attention, and spatial attention is complementary. Their combination could provide better results to improve the classification performance of the framework.

In order to assess the impact of triplet loss learning feature space embeddings, we used t-distributed stochastic neighbor embedding (t-SNE) to visualize the feature distribution of the proposed model. Fig. 8(a) represents the feature distribution of the model without triplet loss, where all classes could be roughly separated, and the feature distribution is relatively scattered. Fig. 8(b) represents the feature distribution in the model with triplet loss, where all classes could be well gathered, and different classes could be pushed apart with the triplet loss function. It demonstrates that our method of learning feature space embeddings through triplet loss has the ability to enhance the convergence of the feature distribution in the feature space.

6 Conclusions

In this paper, we trained a deep learning framework that combines a cascaded attention network with the joint loss for SER. First, to eliminate the effect of emotionally irrelevant factors, we stacked the MFCCs, deltas, and delta-deltas of the MFCCs along with the channel direction as input. Then, a cascaded attention network was utilized, which integrates CNN and BLSTM with a cascaded attention mechanism, to extract spatial and temporal salient features. Finally, the joint loss strategy was utilized to reduce intra-class distance and expand inter-class distance, which forms a great decision margin and improves classification performance. The validity of the proposed approach was verified in a series of ablation studies and comparative experiments with the IEMOCAP corpus. The absolute increment of WA was more than 2.49%, and UA was more than 1.13% compared to state-of-the-art methods.

For future work, we will apply comprehensive data augmentation techniques to obtain more emotional utter-

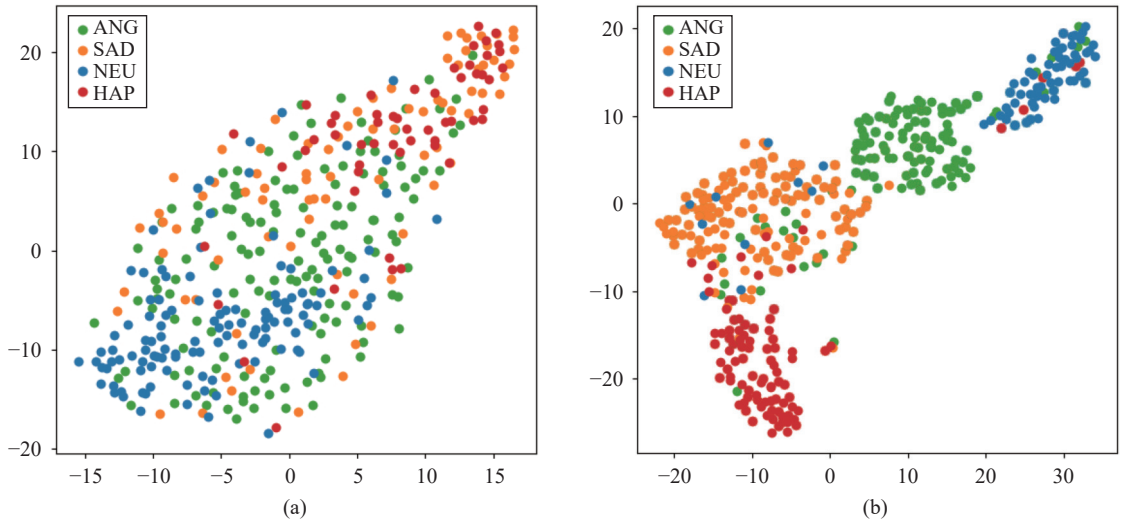


Fig. 8 2D feature distribution in the testing set. (a) Plotted by the proposed model without triplet loss. (b) Plotted by the proposed model. The classes are marked in blue (neutral), green (angry), red (happy), and orange (sad), respectively.

ances for model training as a candidate to enhance the classification performance of the model.

Acknowledgements

This work was supported by Natural Science Foundation of Shandong Province, China (No. ZR2020QF007).

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

References

[1] J. H. Tao, J. Huang, Y. Li, Z. Lian, M. Y. Niu. Correction to: Semi-supervised ladder networks for speech emotion recognition. *International Journal of Automation and Computing*, vol. 18, no. 4, Article number 680, 2021. DOI: 10.1007/s11633-019-1215-6.

[2] E. M. Schmidt, Y. E. Kim. Learning emotion-based acoustic features with deep belief networks. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, pp. 65–68, 2011. DOI: 10.1109/ASPAA.2011.6082328.

[3] K. Han, D. Yu, I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association*, Singapore, pp. 223–227, 2014.

[4] Q. Mao, M. Dong, Z. W. Huang, Y. Z. Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014. DOI: 10.1109/TMM.2014.2360798.

[5] M. Y. Chen, X. J. He, J. Yang, H. Zhang. 3D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018. DOI: 10.1109/LSP.2018.2860246.

[6] Y. Liu, H. Q. Sun, W. B. Guan, Y. Q. Xia, Z. Zhao. Discriminative feature representation based on cascaded attention network with adversarial joint loss for speech emotion recognition. In *Proceedings of Interspeech*, pp. 4750–4754, 2022.

[7] M. Seyedmahdad, E. Barsoum, C. Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, Los Angeles, USA, pp. 2227–2231, 2017.

[8] Q. P. Chen, G. M. Huang. A novel dual attention-based BLSTM with hybrid features in speech emotion recognition. *Engineering Applications of Artificial Intelligence*, vol. 102, Article number 104277, 2021. DOI: 10.1016/j.engappai.2021.104277.

[9] Y. Liu, H. Q. Sun, W. B. Guan, Y. Q. Xia, Z. Zhao. Multimodal speech emotion recognition using self-attention mechanism and multi-scale fusion framework. *Speech Communication*, vol. 139, pp. 1–9, 2022. DOI: 10.1016/j.specom.2022.02.006.

[10] M. K. Xu, F. Zhang, S. U. Khan. Improve accuracy of speech emotion recognition with attention head fusion. In *Proceedings of the 10th Annual Computing and Communication Workshop and Conference*, IEEE, Las Vegas, USA, pp. 1058–1064, 2020. DOI: 10.1109/CCWC47524.2020.9031207.

[11] 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. *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359. DOI: 10.1007/s10579-008-9076-6.

[12] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, C. Y. Espy-Wilson. Adversarial auto-encoders for speech based emotion recognition. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, pp. 1243–1247, 2017.

[13] D. Y. Dai, Z. Y. Wu, R. N. Li, X. X. Wu, J. Jia, H. Meng. Learning discriminative features from spectrograms using center loss for speech emotion recognition. In *Proceedings*

of ICASSP/IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Brighton, UK, pp. 7405–7409, 2019. DOI: 10.1109/ICASSP.2019.8683765.

[14] Y. Gao, J. X. Liu, L. B. Wang, J. W. Dang. Metric learning based feature representation with gated fusion model for speech emotion recognition. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, pp. 4503–4507, 2021.

[15] L. Tarantino, P. N. Garner, A. Lazaridis. Self-attention for speech emotion recognition. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, pp. 2578–2582, 2019.

[16] J. W. Liu, H. X. Wang. A speech emotion recognition framework for better discrimination of confusions. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, pp. 4483–4487, 2021.

[17] A. Satt, S. Rozenberg, R. Hoory. Efficient emotion recognition from speech using deep learning on spectrograms. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, pp. 1089–1093, 2017.

[18] P. C. Li, Y. Song, I. V. McLoughlin, W. Guo, L. R. Dai. An attention pooling based representation learning method for speech emotion recognition. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, pp. 3087–3091, 2018.

Yang Liu received the B. Eng. and M. Eng. degrees in computer science and technology from Tianjin University, China in 2010 and 2012, respectively, and the Ph. D. degree in information science from Japan Advanced Institute of Science and Technology, Japan in 2016. Currently, he is a lecturer with Department of Information Science and Technology, Qingdao University of Science and Technology, China.

His research interests include speech signal processing, life prediction of mechanical equipment and robotic theory.

E-mail: yangliu_qust@foxmail.com
ORCID iD: 0000-0002-9976-8671

Haoqin Sun received the B. Eng. degree in international digital media from Qingdao University, China in 2020. Currently, he is a master student in software engineering at Department of Software Engineering, Qingdao University of Science and Technology, China.

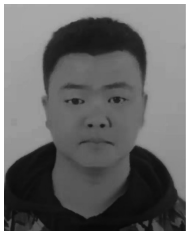
His research interest is speech emotion recognition.

E-mail: 12shq12@163.com
ORCID iD: 00000-0002-8554-8969

Wenbo Guan received the B. Eng. degree in computer science and technology from Jiangsu University of Science and Technology, China in 2019. Currently, he is a master student in electronic information at Department of Electronic Information, Qingdao University of Science and Technology, China.

His research interest is speech separation.

E-mail: g1912913565@163.com



Yuqi Xia received the B. Eng. degree in computer science and technology from Shenyang Normal University, China in 2018. Currently, he is a master student in electronic information at Department of Electronic Information, Qingdao University of Science and Technology, China.

His research interest is speech emotion recognition.

E-mail: 2954200746@qq.com

Zhen Zhao received the Ph. D. degree in systems engineering from Tongji University, China in 2011. Currently, he is an associate professor with Department of Information Science and Technology, Qingdao University of Science and Technology, China.

His research interests include speech emotion recognition, artificial intelligence and edge computing.

E-mail: zzqust@126.com (Corresponding author)
ORCID iD: 0000-0002-7898-8974