Joint Enhancement and Classification Constraints for Noisy Speech Emotion Recognition

Linhui SUN
sunlh@njupt.edu.cn

Nanjing University of Posts and Telecommunications

Shun WANG
Nanjing University of Posts and Telecommunications

Shuaitong CHEN
Nanjing University of Posts and Telecommunications

Min ZHAO
Nanjing University of Posts and Telecommunications

Pingan LI
Nanjing University of Posts and Telecommunications

Research Article

Keywords: speech emotion recognition, Convolution neural network, Long short-term memory-attention, Multi-domain statistical feature, Enhancement and classification constraints

Posted Date: May 25th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2965810/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Digital Signal Processing on May 1st, 2024. See the published version at https://doi.org/10.1016/j.dsp.2024.104581.

Joint Enhancement and Classification Constraints for Noisy Speech Emotion Recognition

Linhui SUN[1*], Shun WANG[1†], Shuaitong CHEN[1†], Min ZHAO[1†], Pingan LI[1†]
[1*]Nanjing University of Posts and Telecommunications, Nanjing, 210003, Jiangsu, China.
*Corresponding author(s). E-mail(s): sunlh@njupt.edu.cn;
Contributing authors: 1220013714@njupt.edu.cn; 1222014404@njupt.edu.cn; 1220014002@njupt.edu.cn；lpa@njupt.edu.cn;
†These authors contributed equally to this work.

Abstract

In the natural environment, the received speech signal is often interfered by noise, which reduces the performance of speech emotion recognition (SER) system. To this end, a noisy SER method based on joint constraints, including enhancement constraint and arousal-valence classification constraint (EC-AVCC), is proposed. This method extracts multi-domain statistical feature (MDSF) to input the SER model based on joint EC-AVCC using convolution neural network and long short-term memory-attention (CNN-ALSTM). The model is jointly constrained by speech enhancement (SE) and arousal-valence classification (AVC) to get robust features suitable for SER in noisy environment. Besides, in the auxiliary SE task, a joint loss function simultaneously constrains the error of ideal ratio mask and the error of the corresponding MDSF to obtain more robust features. The proposed method does not need to carry out noise reduction preprocessing. Under the joint constraints, it can obtain robust and discriminative deep emotion features, which can improve the emotion recognition performance in noisy environment. The experimental results on the CASIA and EMO-DB datasets show that compared with the baseline, the proposed method improves the accuracy of SER in white noise and babble noise by 4.7%-9.9%.

Keywords: speech emotion recognition, Convolution neural network, Long short-term memory-attention, Multi-domain statistical feature, Enhancement and classification constraints

1 Introduction

Human emotions can be recognized by different kinds of information sources, such as facial expressions, words, brain signals, and speech [1]. Among the above-mentioned sources, speech is the easiest to obtain. Speech signals can not only effectively convey semantic information, but also timely convey human emotional information, reflecting the potential emotional state of the speaker. Speech emotion recognition (SER) plays a very important role in human-computer intelligent interaction scenarios [2–4]. Most SER researches mainly focus on acoustic features, system models and classifiers in ideal environment. In [5], we constructed a decision tree structure and used different DNNs to extract bottleneck features, the experiment results show that the SER of the proposed method is 6.25% and 2.91% higher than that of the traditional SVM and DNN-SVM classification methods, respectively. Zhou et al. [6] proposed a novel multi-classifier interactive learning method, which improved the classifier's performance on the benchmark corporas. In recent years, many researchers have applied deep learning to SER. Yi et al. [7] proposed an adversarial data augmentation network, and the resulting emotion classifiers are competitive with state-of-the-art SER systems. Mustaqeem et al. [8] proposed a novel SE framework using redial based function network (RBFN) similarity measurement to select a key sequence segment. The robustness and effectiveness of the suggested SER model are proved from the experiments. Fan et al. [9] proposed an individual standardization SER network to alleviate the interindividual emotion confusion problem. The above researchers use discrete emotion types as the final recognition results. However, discrete emotion types cannot comprehensively represent daily human emotions. Many researchers also use the two-dimensional arousal-valence model to classify the emotional results. Abdelwahab et al. [10] used unsupervised domain adaptation on arousal and valence prediction, and the accuracy was improved by 6.6% compared with no adaptation. Perez-Toroet al. [11] focused on modeling users' health states based on the arousal-valence plane in different scenes, and their experiments verified the effectiveness of the arousal-valence plane. Zhu et al. [12] proposed a deep recursive architecture for arousal-valence affective states, and experimental results showed that the proposed architecture is competitive enough for the recognition task. Dahmane et al. [13] proposed a stress level estimation method to confirm that people's emotional states are located in a projection area in a space defined by two dimensions of arousal and valence, and conducted experiments on a one-minute progressive emotional challenge dataset. Barros et al. [14] proposed an unsupervised neural framework, and the experimental results showed that this method can improve the performance of arousal and valence recognition. Millan-Castillo et al. [15] applied some procedures and ranking schemes for selecting the number of variables in the arousal and valence models, and the experimental results showed that the proposed method is very competitive. These studies are aimed at SER performance in the ideal environment. However, the signal received by system often mixed with noise, which will lead to a significant reduction in SER performance in the natural environment.

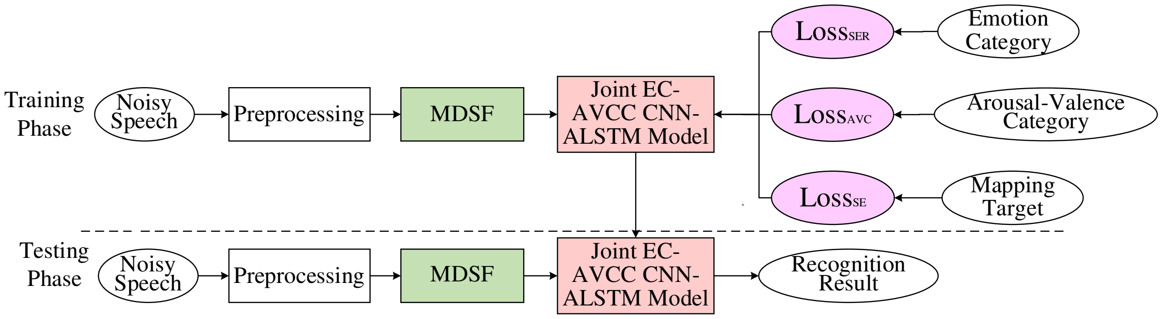

Fig.1 Block diagram of the proposed SER

Taking appropriate methods to reduce the impact of noise will help improve SER accuracy in noisy environments. Firstly, the accuracy of noisy SER can be improved by using robust features. Mansour and Lachiri [16] proposed MFCC-shifted-delta-cepstral (SDC) coefficient features. In [17], Mansour et al. showed MFCC-SDC features are more robust than traditional features. Satt et al. [18] proposed the modified spectrogram for CNN-LSTM in noisy environment. Jing et al. [19] proposed the graph total variation regularization (GTVR) method in noisy environment. Huang et al. [20-21] proposed and adopted the weighted wavelet packet cepstral coefficient (W-WPCC) feature in white Gaussian noise environment. In addition, the selection of a suitable system model is also a key factor in reducing the impact of noise on SER. Zhang et al. [22] proposed an SER method based on binaural representation and deep convolution neural network. Atila et al. [23] proposed a 3DCNN-LSTM emotion recognition model based on attention guidance. In [24], Xu et al. proposed a multi-head attention mechanism fusion method in noisy environment. Finally, the preprocessing of noisy speech can effectively improve the SER performance of the system. Akhtar et al. [25] proposed a modulation spectral feature pool scheme that takes speech enhancement (SE) as the preprocessing of SER. Nam et al. [26] proposed acascade denoising DnCNN-CNN structure based on residual learning. The above approaches either use speech enhancement as a preprocessing operation to get the noise-reduced speech signal, or simply ignore the speech enhancement step.

In this paper, to directly obtain robust and discriminative deep features for emotion classification, a noisy SER method based on joint constraints, including enhancement constraint and arousal-valence classification constraint (EC-AVCC) using convolution neural network and long short-term memory-attention (CNN-ALSTM) in noisy environment is proposed. This method does not need to reconstruct speech in the test phase, which avoids the preprocessing step of speech enhancement and is more convenient and effective. The contribution of the proposed method lies in the following points:

(1) Based on the complementarity between different features, we extract multi-domain statistical features (MDSF). The feature set is a blend of features that contribute more to emotion classification and features that contribute more to SE. The robustness of the features is better than that of the traditional emotion classification feature (TECF) set.

(2) We present a noisy SER model based on joint EC-AVCC using CNN-ALSTM. There is no need to denoise the noisy speech, the auxiliary SE task can reduce the impact of noise on the deep features, and the auxiliary AVC task can increase the emotional contribution of deep features. The system can extract multiple depth features with robustness and high correlation with emotion classification tags.

(3) The ideal ratio mask (IRM) [27] and MDSF mapping are used to jointly constrain SE subtask in the system model. It not only minimizes the difference between the MDSF estimated by the network and the MDSF of clean speech, but also minimizes the difference between the mask estimated by the MDSF and the real mask of speech.

2 Proposed SER method

The proposed SER system block diagram based on joint EC-AVCC using CNN-ALSTM is shown in Fig.1. As shown in Fig.1, the noisy speech signal is firstly preprocessed and MDSF is extracted, and then the deep features are extracted by CNN-ALSTM under the joint constraint of SE and AVC. Finally, they are input to the softmax function to get the emotion judgment result.

2.1 MDSF

To obtain robust features, speech feature sets with high contributions to emotional classification are extracted. They are from MFCC, MFCC first-order difference, zero crossing rate, energy and pitch frequency feature sets, namely TECF. On the other hand, to obtain more useful features for SE, the amplitude spectrum feature (ASF) is also extracted. Then they are spliced and fused into multi-domain feature sets. In a noisy environment, due to the interference of noise, the emotional contribution of the TECF set decreases a lot, while the ASF is robust to noise. In addition, the ASF also can represent the global information of speech signals, which can enrich the features of classification models. TECF and ASF are fused into a multi-domain feature set. The acoustic features of each sentence are extracted, and then the global feature of these features, namely MDSF, is calculated by the five global statistics including mean, median, variance, minimum and maximum. The global features are better than the local features in emotional classification accuracy.

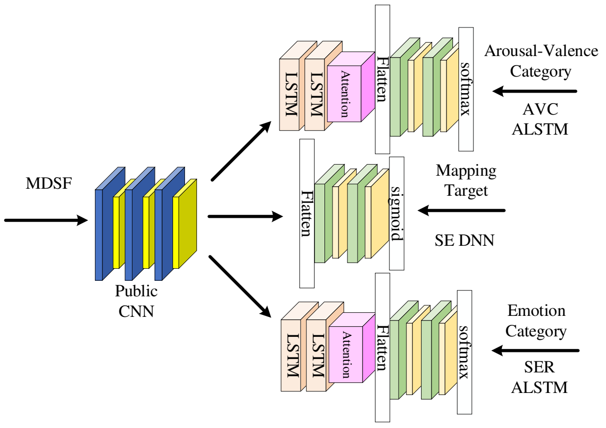

Fig.2 Joint EC-AVCC CNN-ALSTM model structure

2.2 Proposed Network model

2.2.1 Model structure

The proposed SER method based on joint EC-AVCC using CNN-ALSTM does not need to carry out noise reduction preprocessing on noisy speech. It directly inputs its MDSF to the system model. The proposed joint EC-AVCC CNN-ALSTM model structure is shown in Fig. 2. In the model, SER is the main task, and SE and AVC are subtasks. The input feature of the system is MDSF obtained by multiple acoustic features fusion and feature screening. The public network of primary and secondary tasks is a CNN structure with three convolution layers and three maximum pooling layers. In the primary and AVC auxiliary tasks, the public CNN is followed by a two-layer LSTM network and a self-attention model. The attention module is used to determine the emotional contribution of different utterance-level features based on the query vector, to improve the emotion recognition accuracy of the system. It assigns higher weights to the utterance-level features containing more emotional information, and reduces the impact of utterance-level features containing less emotional information. The attention output sequence F_i ($i \in [1, T]$) are obtained by:

$$G(h_t,q_i)=h_t\cdot q_i \qquad (1)$$

$$\theta_{i,t}=soft\max(G(h_t,q_i)) \qquad (2)$$

$$F_i=\sum_{t=1}^{T}(\theta_{i,t}h_t) \qquad (3)$$

where $t \in [1, T]$, h_t represents the output sequence of LSTM, q_i represents the query sequence, $G(h_t,q_i)$ is the scoring function of the attention mechanism, and T represents the length of the utterance-level feature.

In the SER main task and the AVC auxiliary task, the flatten layer connects CNN-ALSTM and DNN. In SE auxiliary task, the flatten layer connects the public CNN and DNN. For primary and secondary tasks, three DNNs are designed as the full connection layer of the system model network. The SER main task and the AVC auxiliary task of the system adopt a five-layer DNN structure. The activation function of the output layer is softmax to achieve the classification tasks. The SE subtask also adopts a five-layer DNN structure. The activation function of the output layer selects sigmoid to achieve the purpose of the regression task. In this way, the joint EC-AVCC CNN-ALSTM model is built.

2.2.2 Loss function

The loss function plays an important role in the joint EC-AVCC CNN-ALSTM model, which mainly restricts the training of the whole model. The main task for emotion recognition selects the multiclassification cross entropy as the loss function, and its specific expression is:

$$L_{SER}=-\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N}q(x_{ij})\log p(x_{ij}) \qquad (4)$$

where M represents the number of samples, N represents the number of emotional categories, and $p(x_{ij})$ represents the prediction probability that the observed sample i belongs to the category j. When sample i does not belong to the category j, $q(x_{ij})$ is equal to 0. When sample i belongs to the category j, $q(x_{ij})$ is equal to 1.

The AVC auxiliary task also uses the multiclassification cross entropy as the loss function, and its specific expression is:

$$L_{AVC}=-\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{Q}q(x_{ik})\log p(x_{ik}) \qquad (5)$$

where Q represents the number of arousal-valence categories, namely low arousal-low valence, low arousal-high valence, high arousal-low valence and high arousal-high valence. $p(x_{ik})$ represents the prediction probability that the observed sample i belong to the category k. When sample i does not belong to the category k, $q(x_{ik})$ is equal to 0. When i belongs to the category m, $q(x_{ik})$ is equal to 1.

Since the SE task is a linear regression deep learning prediction process, the mean square error is generally used as its loss function with the following specific expression (feature mapping(FM)):

$$L_{SE}=\frac{1}{M}\sum_{i=1}^{M}(Y_m-\hat{Y}_m)^2 \qquad (6)$$

where Y_m is the true value of the MDSF vector of the pure speech signal, and $\hat{Y}_m$ is its predicted value which is output by the SE network model. Under the constraint, the training process for the SE task is the

Table 1 SER algorithm flow based on joint EC-AVCC using CNN-ALSTM

Training phase

Input: The noisy mixed speech training set, clean MDSF , emotional labels and arouse-valence labels.

Output: The trained joint EC-AVCC CNN-ALSTM model.

Step 1: Preprocessing operations are performed on noisy and clean speech signals respectively. And then TECF sets are extracted. ASF commonly used in speech enhancement are extracted. The ASF and TECF set are fused to generate multi-domain features. Calculate the statistical features of multi-domain features to obtain the MDSF of noisy and clean speech.

Step 2: The MDSF of noisy speech is used as the input of joint EC-AVCC CNN-ALSTM. The emotional labels, arouse-valence labels and the MDSF of clean speech are used to calculate the loss function of the output layer of the system model.

Step 3: In the forward propagation stage, the weight value and offset value of each layer of neurons in the network are initialized randomly. In the back-propagation stage, the loss function value of the joint constraints of primary and auxiliary tasks are minimized by optimizing the function, and various network parameters are iteratively adjusted, including the weight value and offset value of each layer of neurons.

Step 4: Save the network parameters when the iteration termination condition is reached. The training of the joint EC-AVCC CNN-ALSTM model is completed.

Testing phase

Input: The noisy mixed speech test set and the trained system model.

Output: Emotion category.

mapping process of the MDSF from noisy speech to clean speech.

To improve the denoising performance of auxiliary SE subtasks and strengthen the noise robustness of deep features, it is necessary to mine the relationship between IRM and MDSF as much as possible. In SE, the loss function usually considers the minimum difference between the estimated mask and the actual mask. Therefore, the error of IRM and the error of the MDSF is constrained jointly in the paper. The joint constraint relationship is taken as the loss function of the subtask of our model, which not only minimizes the error between the MDSF estimated by the network and that extracted from the clean speech, but also minimizes the error between the estimated mask and the real mask of speech. The loss function of the joint constraint in the subtask can be defined as (FM+IRM):

$$L_{SE}=\frac{1}{M}\sum_{i=1}^{M}[(Y_m-\hat{Y}_m)^2+(R_m-\frac{\hat{Y}_m}{Z_m})^2] \qquad (7)$$

where R_m is the actual mask value and Z_m is the noisy MDSF. Equation (7) makes full use of the relationship between MSDF and IRM, which increases the constraints, improves the denoising ability of the model, and further reduces the interference of noise on the deep features extracted by the public network.

Combining equation (4), equation (5) and equation (7), the loss function in the joint EC-AVCC CNN-ALSTM system model is as follows:

$$L=\alpha L_{SER}+\beta L_{SE}+(1-\alpha-\beta)L_{AVC} \qquad (8)$$

where α is the weight of the main task for emotion recognition, β is the weight of the SE auxiliary task, and their value will be determined according to the enumeration method. The proposed joint constraint loss function can constrain emotion classification, SE and AVC at the same time. The mixed speech signal is constrained by the SE module and the AVC module, so that the noise robustness of the deep features related to emotion information is improved and the emotional contribution of deep features is also increased.

2.3 Proposed SER algorithm

The proposed SER framework based on joint EC-AVCC using CNN-ALSTM includes training stage and testing stage in Fig. 1. In the training stage, the main task for emotion classification aims at emotion category labels, the subtask for SE aims at the MDSF of clean speech and the real IRM, and the subtask for AVC aims at arousal-valence category labels. When the MDSF set of noisy speech is input into the CNN-ALSTM with joint constraints, the main task network for emotion classification and the subtask networks for SE and AVC are jointly trained. The weights and offsets of each layer of neurons are randomly initialized. After unsupervised forward training, the parameters are fine-tuned by back-propagation under the supervision of the loss function. The loss function value decreases during the optimization search. Thus, various network parameters, including the weights and biases of each layer of neurons, are iteratively adjusted. When the joint loss function value reaches the minimum, the loss function converges. The training of the CNN-ALSTM system model is completed, and the network parameters are stored. In the test phase, the MDSF of noisy speech signals in the test set is extracted, and then input to the trained CNN-ALSTM system model to extract the deep features. The diversified depth features extracted by the front-end network CNN not only contain information highly related to emotion tags and more differential and discriminative deep features, but also are robust to environmental noise. Such features are used in emotion recognition for the main task of our model, which can effectively improve the performance of SER in noisy environment. Finally, the prediction result of emotion recognition through the classifier is got. The SER algorithm flow based on joint EC-AVCC using CNN-ALSTM is shown in Table 1.

3 Experiment

3.1 Experiment preparation

CASIA speech emotion database and EMO-DB speech emotion database are selected for the experiment. The CASIA chinese database contains six emotions, namely, sad, surprised, happy, fear, angry and neutral, with 200 speech items for each emotion, and a total of 1200 speech items. EMO-DB germany database contains seven emotions, namely, happy, fear, neutral, sad, angry, bored and disgusted, with a total of 535 speech items. In the experiment, the ratio of training samples to test samples is 4:1, and the training samples are further divided into training and validation sets according to the ratio of 4:1. A five-fold cross-validation speaker-dependent method is used in the experiments to obtain more accurate recognition results against the corpus. The sampling rate of each speech is 16KHz. The window function uses a Hamming window with a window length of 512 and a frame shift of 1/4. For CNN, the number of neurons in the convolution layer is set to 256, and the step size is 1. The size of the maximum pooling layer is set to 3*1, and the step size is 3. For LSTM, the number of neurons in both the input layer and the output layer is set to 512. For DNN, the number of neurons in both the input layer and the hidden layer is set to 1024, and the number of neurons in the output layer is consistent with the number of emotional categories classified. In the experiment, the value of the weight of the main task of the total loss function in the system model will be determined according to the enumeration method. By many different signal-to-noise ratio (SNR) experiments, the value of α is set to 0.7 and the value of β is set to 0.2.

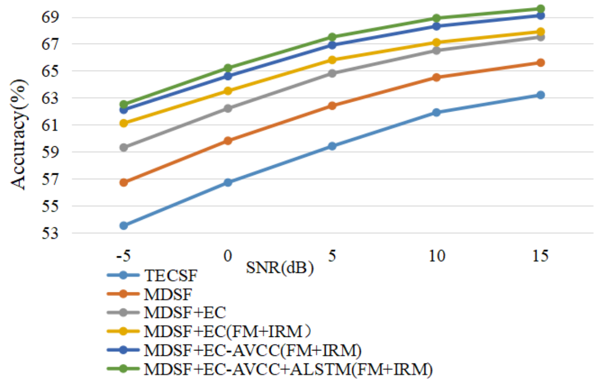

Fig.3 Performance of proposed methods (CASIA+white)

3.2 Ablation experiment comparison

(1) Different features: The MDSF (ours1) and traditional emotion classification statistical feature (TECSF) are used as the input of CNN to carry out the emotional classification comparison experiment. The TECSF is a 255-dimensional feature vector, and MDSF is 1540-dimensional feature vector. The other settings of the two systems are the same. As can be seen in Fig. 3 to Fig. 6, the performance of SER using MDSF is better than that of the TECSF for both white noise and babble noise. This is mainly because the MDSF is dominated by features that contribute more to emotion recognition and complemented by features that contribute more to SE, taking into account the characteristics of two different types of tasks for noisy environments. This makes the robustness of MDSF better than TECSF, and when the noise interference is greater, the emotional recognition performance of the model using MDSF improves more.

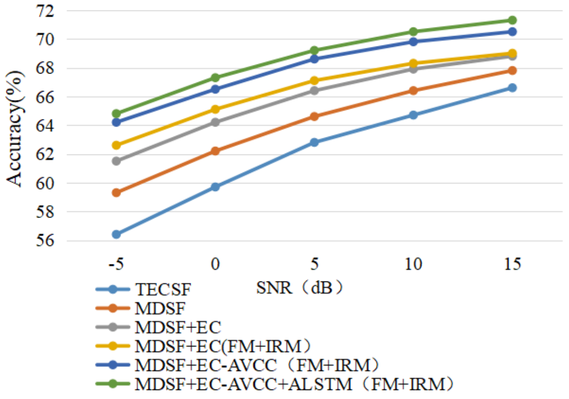

Fig.4 Performance of proposed methods (CASIA+babble)

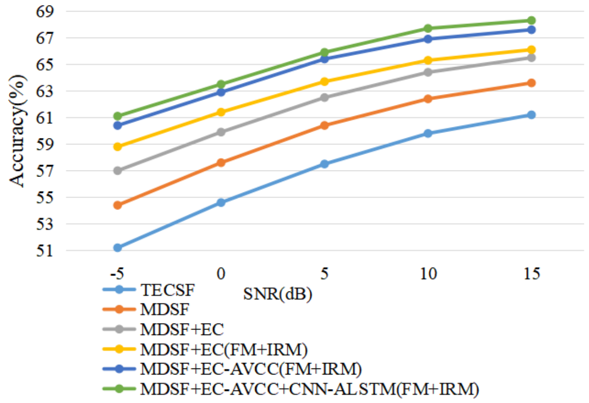

Fig.5 Performance of proposed methods (EMO-DB+white)

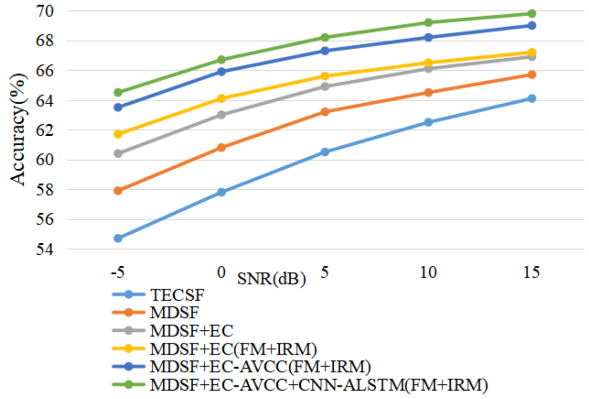

Fig.6 Performance of proposed methods (EMO-DB+babble)

(2) Multi-task and single-task: We compare the recognition accuracy of the proposed MDSF+EC (ours2) method with the MDSF method. The two network models both use MDSF as the input of CNN.

The main difference is that ours2 includes SE auxiliary subtasks, while MDSF is a single task structure. It can be seen from Fig. 3 to Fig. 6 that the performance of the proposed ours2 is better than MDSF. The purpose of the SE auxiliary task is to allow the mapping of noisy speech features to pure speech features. When each parameter of the system model is determined, the deep features extracted by the public network CNN will be less affected by the environmental noise, thus improving recognition performance. In addition, it can be seen that when the SNR is lower, the improvement of the emotional recognition rate is higher. This is because the greater the noise, the greater the constraint of SE subtasks on the public network CNN.

(3) Subtask loss functions: We compare the recognition accuracy of the proposed ours2 with MDSF+EC(FM+IRM) (ours3) method. Both methods use MDSF as the input of the jointly constrained CNN-DNN network model, and the difference mainly lies in the loss function of subtasks. As can be seen in Fig. 3 to Fig. 6, the performance of ours3 is better than that of ours2. A joint constraint on the error of IRM and the one of MDSF is applied and this constraint relation is used as the loss function for the system model auxiliary task. It not only minimizes the difference between the MDSF estimated by the network and the MDSF of clean speech, but also minimizes the difference between the IRM estimated by the MDSF and the true IRM of speech. The joint FM+IRM constrained loss function is able to explore the relationship between IRM and MDSF as much as possible, further making the input noisy speech features mapped to real noiseless speech features, thus improving recognition performance. When the SNR is lower, the improvement of emotional recognition accuracy using FM+IRM is higher than that using FM. While in the case of higher SNR, the improvement of the emotional recognition rate is smaller. This is because the greater the interference of noise on pure speech, the more the joint loss function as well as the SE subtask can play its role.

(4) Different auxiliary tasks: We compare the recognition accuracy of the proposed MDSF+EC-AVCC(FM+IRM) (ours4) method with ours3 method. The two network models both use MDSF as the input of CNN and FM+IRM as the loss function of the SE subtask. The main difference is that ours4 includes SE auxiliary subtask and AVC auxiliary subtask, while ours3 includes SE auxiliary subtask. As can be seen in Fig. 3 to Fig. 6, the performance of ours4 is better than that of ours3. In conclusion, the multiple auxiliary subtasks are effective in improving the performance of the model. This is because they not only improve the robustness of the deep features extracted by the public network through the speech enhancement auxiliary task and reduce the influence of deep features by noise, but also in the shared network layer of the network model, the arousal-valence classification auxiliary task can further provide valid emotional information for the emotion classification task. In the experimental results, the improvement of SER accuracy under the interference of different SNRs is similar, which is because the arousal-valence classification subtask improves the emotion classification performance of the system mainly from the perspective of increasing the emotion information and enhancing the emotional contribution of deep features, which is different from the perspective of the SE auxiliary task to constrain the noise and reduce its influence on deep features.

(5) CNN-ALSTM and CNN-DNN: We compare the recognition accuracy of the proposed MDSF+EC-AVCC+ALSTM (FM+IRM) (ours5) method with ours4 method. The two network models both use MDSF as the input of the model and FM+IRM as the loss function of the SE subtask. They also contain SE subtask and AVC subtask. The main difference is that ours5 includes CNN-ALSTM, while ours4 includes CNN-DNN. As can be seen in Fig. 3 to Fig. 6, the performance of ours5 is better than that of ours4. In short, in the shared network layer of the network model, the LSTM module can extract deep features containing more temporal information, and the Attention module task can multiply each utterance-level feature by a weight to get deep features with higher emotional contribution. This allows the ALSTM module to further provide effective emotional information for the emotion classification task, which enables the CNN-ALSTM network to effectively improve the performance of the model. Similar to the results in the previous subsection, the improvement of SER accuracy is basically the same under the interference of different SNRs in this section. Again, this is because the ALSTM module improves the emotion classification performance of the system mainly from the perspective of increasing the emotional information and enhancing the emotional contribution of deep features, which is different from the perspective of constraining the noise and reducing its influence on deep features mentioned in the previous section.

Table 2 SER ablation experiments (%) (CASIA+white)

Method	SNR(dB)				
	-5	0	5	10	15
CNN(baseline)	53.5	56.7	59.4	61.9	63.2
Ours1	56.7	59.8	62.4	64.5	65.6
Ours2	59.3	62.2	64.8	66.5	67.5
Ours3	61.1	63.5	65.8	67.1	67.9
Ours4	62.1	64.6	66.9	68.3	69.1
Ours5	**62.5**	**65.2**	**67.5**	**68.9**	**69.6**

3.3 Performance comparison

Table 2 to Table 5 show the performance comparison of ablation experiments. It can be seen that compared with the CNN baseline method, the SER accuracy of the proposed method has been effectively improved in noisy environment. In addition, Table 6 to Table 9 show the performance comparison of the proposed method with other methods under the environment of white noise and babble noise respectively. Literature [28] adopts the decision tree and improved support vector machine (SVM) model. Literature [29] adopts CNN with attention head fusion. Literature [30] adopts CNN with area attention. Compared with Literature [28], [29] and [30], the effect of the proposed method is better in noisy environments with lower SNRs. This is because the joint enhancement constraint is more effective at lower SNRs. Another advantage of the proposed method is that there is no need to denoise and reconstruct clean speech in advance, which is of great significance in practical application.

Table 3 SER ablation experiments (%) (CASIA+babble)

Method	SNR(dB)				
	-5	0	5	10	15
CNN(baseline)	56.4	59.7	62.8	64.7	66.6
Ours1	59.3	62.2	64.6	66.4	67.8
Ours2	61.5	64.2	66.4	67.9	68.8
Ours3	62.6	65.1	67.1	68.3	69
Ours4	64.2	66.5	68.6	69.8	70.5
Ours5	**64.8**	**67.3**	**69.2**	**70.5**	**71.3**

Table 4 SER ablation experiments (%) (EMO-DB+white)

Method	SNR(dB)				
	-5	0	5	10	15
CNN(baseline)	51.2	54.6	57.5	59.8	61.2
Ours1	54.4	57.6	60.4	62.4	63.6
Ours2	57	59.9	62.5	64.4	65.5
Ours3	58.8	61.4	63.7	65.3	66.1
Ours4	60.4	62.9	65.4	66.9	67.6
Ours5	**61.1**	**63.5**	**65.9**	**67.7**	**68.3**

Table 5 SER ablation experiments (%) (EMO-DB+babble)

Method	SNR(dB)				
	-5	0	5	10	15
CNN(baseline)	54.7	57.8	60.5	62.5	64.1
Ours1	57.9	60.8	63.2	64.5	65.7
Ours2	60.4	63	64.9	66.1	66.9
Ours3	61.7	64.1	65.6	66.5	67.2
Ours4	63.5	65.9	67.3	68.2	69.0
Ours5	**64.5**	**66.7**	**68.2**	**69.2**	**69.8**

4 Conclusion

A noisy SER method based on joint EC-AVCC using CNN-ALSTM in noisy environment is proposed. The CNN-ALSTM SER model is jointly constrained by the subtask for SE and the subtask for AVC to get features suitable for SER in noisy environment. In the auxiliary SE task, a joint constraint loss function is used, which can simultaneously constrain the error between the predicted value and the real value masked by the ideal IRM and the error of the corresponding MDSF, so as to obtain more robust features. In the auxiliary AVC task, the different arousal-valence emotion categories provide additional emotional information for the main task of emotion categorization, allowing the public network in the multi-task system model to extract deep features that are more differentiated and discriminative, thus improving the system’s SER accuracy. CNN-ALSTM can extract deep features containing more temporal information, multiply a weight to each utterance-level feature, and thus filter out the deep features with the higher emotional contribution. Experimental results show that the proposed method can effectively improve the SER accuracy compared with the baseline method.

Table 6 SER performance comparison (%) (CASIA+ white)

Method	SNR(dB)		
	-10	-5	0
Document[28]	-	-	47.53
Document[29]	53.2	58.9	63.8
Document[30]	56.3	61.7	**66.6**
Ours	**59.1**	**62.5**	65.2

Table 7 SER performance comparison (%) (CASIA+babble)

Method	SNR(dB)		
	-10	-5	0
Document[29]	57.1	62.1	66.4
Document[30]	58.9	64.5	**69.4**
Ours	**61.7**	**64.8**	67.3

Table 8 SER performance comparison (%) (EMO-DB+white)

Method	SNR(dB)		
	-10	-5	0
Document[29]	52.3	57.6	61.9
Document[30]	54.2	59.8	**64.8**
Ours	**58.3**	**61.1**	63.5

Table 9 SER performance comparison (%) (EMO-DB+babble)

Method	SNR(dB)		
	-10	-5	0
Document[29]	56.7	61.5	65.4
Document[30]	58.2	63.4	**68.3**
Ours	**61.5**	**64.5**	66.7

Declarations

Ethical Approval Not applicable

Competing interests The authors declare that they have no competing interests

Authors' contributions Linhui Sun designed the core methodology of the study, carried out the implement, and drafted the manuscript. Shun WANG and Shuaitong Chen carried out the experiments and drafted the manuscript. Min ZHAO and Pingan Li conducted a formal analysis and validation. All authors read and approved the final manuscript.

Funding This work was supported by the National Natural Science Foundation of China (61901227, 62071242) and the China Scholarship Fund (202008320043).

Availability of data and materials Not applicable

References

[1] Fahad, M.S., Ranjan, A., Yadav, J., Deepak, A.: A survey of speech emotion recognition in natural environment. Digital Signal Processing 110, 102951 (2021)

[2] Mehmet, B.A., Kaya, O.: Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication 116, 56–76 (2020)

[3] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J., Schuller, B.W.: Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. IEEE Transactions on Affective Computing 13(2), 992–1004 (2022)

[4] Liu, K., Wang, D., Wu, D., Liu, Y., Feng, J.: Speech emotion recognition via multi-level attention network. IEEE Signal Processing Letters 29, 2278–2282 (2022)

[5] Sun, L., Zou, B., Fu, S., Chen, J., Wang, F.: Speech emotion recognition based on dnn-decision tree svm model. Speech Communication 115, 29–37 (2019)

[6] Zhou, Y., Liang, X., Gu, Y., Yin, Y., Yao, L.: Multi-classifier interactive learning for ambiguous speech emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 30, 695–705 (2022)

[7] Yi, L., Mak, M.-W.: Improving speech emotion recognition with adversarial data augmentation network. IEEE transactions on neural networks and learning systems 33(1), 172–184 (2020)

[8] Sajjad, M., Kwon, S., et al.: Clustering-based speech emotion recognition by incorporating learned features and deep bilstm. IEEE Access 8, 79861–79875 (2020)

[9] Fan, W., Xu, X., Cai, B., Xing, X.: Isnet: Individual standardization network for speech emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 30, 1803–1814 (2022)

[10] Abdelwahab, M., Busso, C.: Domain adversarial for acoustic emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26(12), 2423–2435 (2018)

[11] Perez-Toro, P.A., Vasquez-Correa, J.C., Bocklet, T., Noth, E., Orozco-Arroyave, J.R.: User state modeling based on the arousal-valence plane: Applications in customer satisfaction and health-care. IEEE Transactions on Affective Computing, 1 (2021)

[12] Zhu, H., Han, G., Shu, L., Zhao, H.: Arvanet: Deep recurrent architecture for ppg-based negative mental-state monitoring. IEEE Transactions on Computational Social Systems 8(1), 179–190 (2021)

[13] Dahmane, M., Alam, J., St-Charles, P.-L., Lalonde, M., Heffner, K., Foucher, S.: A multimodal non-intrusive stress monitoring from the pleasure-arousal emotional dimensions. IEEE Transactions on Affective Computing 13(2), 1044–1056 (2022)

[14] Barros, P., Barakova, E., Wermter, S.: Adapting the interplay between personalized and generalized affect recognition based on an unsupervised neural framework. IEEE Transactions on Affective Computing 13(3), 1349–1365 (2022)

[15] Mill ́an-Castillo, R.S., Martino, L., Morgado, E., Llorente, F.: An exhaustive variable selection study for linear models of soundscape emotions: Rankings and gibbs analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing 30, 2460–2474 (2022)

[16] Mansour, A., Lachiri, Z.: A comparative study in emotional speaker recognition in noisy environment. In: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), pp. 980–986. IEEE, Tunisia (2017)

[17] Mansour, A., Chenchah, F., Lachiri, Z.: Emotional speaker recognition in real life conditions using multiple descriptors and i-vector speaker modeling technique. Multimedia Tools and Applications 78, 6441–6458 (2019)

[18] Aharon, S., Shai, R., Ron, H.: Efficient emotion recognition from speech using deep learning on spectrograms. In: Proc. Interspeech 2017, pp. 1089–1093. Interspeech, Sweden (2017)

[19] Jing, S., Mao, X., Chen, L., Comes, M.C., Mencattini, A., Raguso, G., Ringeval, F., Schuller, B., Natale, C.D., Martinelli, E.: A closed-form solution to the graph total variation problem for continuous emotion profiling in noisy environment. Speech Communication 104, 66–72 (2018)

[20] Huang, Y., Ao, W., Zhang, G.: Novel sub-band spectral centroid weighted wavelet packet features with importance-weighted support vector machines for robust speech emotion recognition. Wireless Personal Communications 95, 2223–2238 (2017)

[21] Huang, Y., Tian, K., Wu, A., Zhang, G.: Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition. Journal of Ambient Intelligence and Humanized Computing 10(5), 1787–1798 (2019)

[22] Zhang, S., Chen, A., Guo, W., Cui, Y., Zhao, X., Liu, L.: Learning deep binaural representations with deep convolutional neural networks for spontaneous speech emotion recognition. IEEE Access 8, 23496–23505(2020)

[23] Atila, O., S ̧eng ̈ur, A.: Attention guided 3d cnn-lstm model for accurate speech based emotion recognition. Applied Acoustics 182, 108260 (2021)

[24] Xu, M., Zhang, F., Zhang, W.: Head fusion: Improving the accuracy and robustness of speech emotion recognition on the iemocap and ravdess dataset. IEEE Access 9, 74539–74549 (2021)

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

[26] Nam, Y., Lee, C.: Cascaded convolutional neural network architecture for speech emotion recognition in noisy conditions. Sensors 21(13), 4399 (2021)

[27] Sun, L., Wang, C., Liang, W., Li, P.: Monaural speech separation method based on deep learning feature fusion and joint constraints. Journal of Electronics Information Technology 44(9), 3266–3276 (2022)

[28] Zhao, J.J., Ma, R.L., Zhang, X.L.: Speech emotion recognition based on decision tree and improved svm mixed model. Beijing Ligong Daxue Xuebao/Transaction of Beijing Institute of Technology 37(4), 386–390 (2017)

[29] Xu, M., Zhang, F., Khan, S.U.: Improve accuracy of speech emotion recognition with attention head fusion. In: 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), pp. 1058–1064. IEEE, USA (2020)

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

Canada (2021)