IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 December 2022, accepted 12 January 2023, date of publication 16 January 2023, date of current version 19 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3237268

RESEARCH ARTICLE

Design of Efficient Speech Emotion Recognition Based on Multi Task Learning

LIU YUNXIANG AND ZHANG KEXIN
Department of Computer Science, Shanghai Institute of Technology, Shanghai 201418, China
Corresponding author: Zhang Kexin (1277767811@qq.com)

ABSTRACT Speech emotion recognition technology includes feature extraction and classifier construction. However, the recognition efficiency is reduced due to noise interference and gender differences. To solve this problem, this paper used two multi-task learning models based on adversarial multi-task learning(ASP-MTL). The first model took emotion recognition as the main task and noise recognition as the auxiliary task, and removed the noise part identified by the auxiliary task. After identifying the non-noise part, the second model was constructed. The second model took emotion recognition as the main task and gender classification as the auxiliary task. These two multi-task learning models can not only can use shared information to learn the relationship between different tasks, but also can identify specific tasks. This paper used Audio/Visual Emotion Challenge (AVEC) database and AFEW6.0 database,which were recorded in the field environment. Considering the problem of data imbalance between datasets, the data balance operation was carried out on the data sets in the process of data preprocessing. The paper shows an increase of around 10% in terms of accuracy and F1 score with the recent works using AVEC database and AFEW6.0 datasets, which proved that this paper has made a great progress in SER.

INDEX TERMS Speech emotion recognition, multi-task learning, noise reduction, eliminating gender differences, hidden layer sharing, data balance, specific task classification processing.

I. INTRODUCTION

Speech emotion recognition (SER) is widely used in education industry, service industry, assisted driving industry and criminal investigation industry [1], [2], [3]. At present, great progress has been made in feature extraction and emotion classification. Acoustic features, such as mel-scale Frequency Cepstral Coefficients (MFCC),short-time energy, fundamental frequency, short-time zero crossing rate, etc., are often extracted in previous work [4], [5]. When selecting a classifier, deep neural network(DNN), bi-directional long short-term memory(BLSTM),support vector machine (SVM) are usually proposed [6], [7]. Compared with machine learning models, deep learning models can extract complex features. Here are the achievements of the deep learning model. In [28], Mel spectrograms were extracted as speech features. Then the features were sent into CNN and TLEFuzzyNet model combined transfer learning by ensemble approach. In [29] they proposed MobileNet V2 with LSTM for the purpose of the precise classification of skin disease from the image. Their model can haddle gradient disappearing efficiently. In [30], by employing deep learning and swarm intelligence, they proposed a DSwarm-Net to conduct human action recognition. In [31], they proposed an ensemble model of 'CNN-net', 'CNN+LSTM','ConvLSTM',and 'Stacked LSTM-net. Then they made final prediction by tacking predictions from each of the four mentioned classification models.

However, Natural noise and gender differences have adverse effects on speech emotion recognition. This paper solved the key issues of the adverse effect of noise interface in natural environment and the existence of gender differences in SER. The two issues will reduce the accuracy of SER. For example, compared with men and women, women are more likely to expose their emotions than men. Men are more rational and women are more emotional. The database of natural environment are composed of many kinds of noise. In order to improve the robustness, the databases selected in this paper are AFEW6.0 database and 2016 audio visual

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

emotional challenge (AVEC) database, which are recorded in the wild natural environment and IEMOCAP database by adding different types of noise. Natural environment is filled with noise, so AFEW6.0 database and AVEC database are noisy speech data. For the problem of noise impact, previous researchers have also made improvements. In [8], the scheme of modulation spectrum feature pool is introduced. To reduce noise, there are two phrases. They extracted six modulation spectrum measures through modulation spectrum representation, and explored a new modulation spectrum feature pool scheme, which improved the ability of emotion recognition under noise conditions. In [9], histogram equalization was used to reduce the difference between feature vectors under clean and noisy conditions. They calculated the average histogram of pitch and MFCC, and then used it as a reference for equalization. In [10], speech activity detector (VAD) was used to discard prominent and noisy frames. The proposed VAD combined with nonnegative matrix decomposition technology played a key role in improving robustness. In [11], they used a weighted sparse representation model based on maximum likelihood estimation, which was an enhanced sparse representation classifier. And they compared it with k-neareat-neighbors (KNN), SVM and decision tree to prove that the robustness was improved under clean and noisy conditions. In view of the influence of gender differences, researchers usually use transfer learning. The basic idea of transfer learning is that human beings constantly acquire knowledge and learn new things. People can use previous experience to learn new things and adapt to new situations. In [12], considering that many works only consider the common information of the target domain and the source domain, however, the specific information of the two domains was always ignored. Therefore, they proposed double spatial transfer learning (DSTL) to make full use of these two kinds of information. DSTL can not only alleviate the differences between domains, but also make good use of specific information. In [13], they used a bi-hemispheric antagonistic neural network (BIDANN). In Bi-hemispheres, one function was to control source data and the other function was to control target data. The global discriminator attempts to reduce the possible domain differences between the source domain and the target domain in each hemisphere. An alignment related loss was proposed in [14]. The function of related alignment is to minimize domain offset. In [15], they proposed transfer subspace learning based on nonnegative matrix decomposition in order to find a shared feature subspace between source data and target data. Only in this way can the information of the source data be transmitted to the target data and the differences be combined.

However, transfer learning makes less use of the links between different tasks than multi-task learning. The method of noise removal in the above literature does not make use of the connection between noise and non-noise signals. The noise part may have information related to the non-noise part. If this information is removed, it will cause the loss of useful information. In order to make use of the relationship between different tasks, multi-task learning (MTL) has been widely used in speech emotion recognition. Multi-task learning associates the main task with several auxiliary tasks, which can improve the generalization of classification. In the neural network model of MTL, the bottom layer of the network is the shared hidden layer, the connection between learning tasks, and the top layer is the task specific layer to learn the unique attributes of each task [16], [17]. The difficulty of MTL is that it is difficult to distinguish which features are shared and which are private. In order to solve this problem, this paper learned from ASP-MTL in [18]. They used shared long short-term memory (LSTM) and private LSTM to extract common features and private features respectively and used confrontation training to make the shared features not exist in the private LSTM module, and uses orthogonal constraints to ensure that the private features do not exist in the common LSTM module. Therefore, our method to solve these the problems is to design two ASP-MTL models. There are two multi-task learning(ASP-MTL) models in our paper. The first ASP-MTL is used for noise cancellation. In the first ASP-MTL model, emotion recognition is primary task, noise recognition is auxiliary task. Noisy signal is the recognition results of the auxiliary task of ASP-MTL, so we remove these noisy signal and put the remaining signal into the second ASP-MTL model. In the second ASP-MTL model, the primary task is final emotion classification and the auxiliary task is gender classification. The ground truth used for noise cancellation is that making the use of the connection between noisy signal and non-noise signal and recognize noisy signal by the auxiliary task of the first ASP-MTL model and remove the noisy signal, then we input the remaining signal into the second ASP-MTL model.

The contributions of this paper are: (1) Although multitask learning approaches are already implemented and evaluated for SER task,seldom research on noise reduction use the multitask learning approaches. The commonly used approaches for noise cancellation are wavelet transform, extracting antinoise features, speech preprocessing method and spectral subtraction method. So the novelty of our work is use the multitask learning approaches for noise cancellation. Multi task learning is a new approach in noise cancellation. (2) Solving the problem that shared features are mistaken for private features and private features are mistaken for common features in multi-tasking learning. In our two ASP-MTL model, we used confrontation training and orthogonal constranint training to make sure there is no confusion between shared and private features. Traditional multitask learning methods seldom consider solving the problem that shared features are mistaken for private features and private features are mistaken for common features. (3) Compared with [32], our paper combined two multitask learning models together to solve the problem of noise interface and gender differences at the same time. Previous researchers seldom solved the two issues at the same time. After noise cancellation by the first

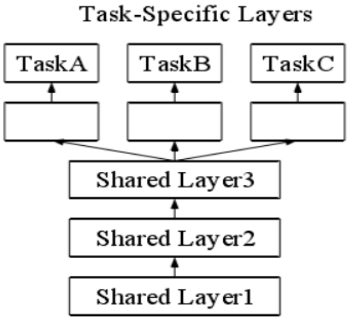

FIGURE 1. Multi-task learning model based on hard parameters.

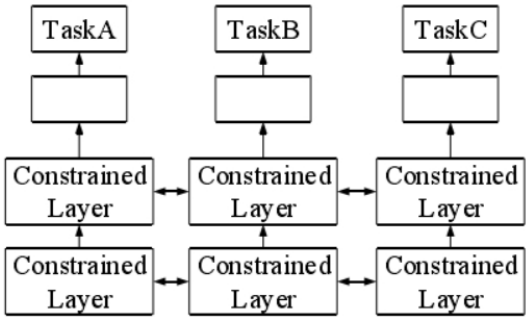

FIGURE 2. Multi-task learning model based on soft parameters.

ASP-MTL model, we use the second ASP-MTL model to make gender and emotion classification to reduce the impact of gender factors on the accuracy of speech emotion recognition. In the second ASP-MTL model, emotion classification was designed as main task, while gender classification was designed as auxiliary task. The practical implications of our research are considering the connection between non-noise signals and noise signals in the stage of noise reduction, making good use of the connection between gender classification and emotion classification,distinguishing shared features and private features in the two multi-task learning models.

The structure of this paper is as follows: the first part introduced the research progress of multitasking learning, the second part introduced the model of this paper, the third part introduces the experimental part, the fourth part discussed the experimental results, and the fifth part was the conclusion.

II. RELATED WORK

Multi-task learning is divided into hard parameter based multi-task learning and soft parameter based multi-task learning. The multi-task learning model diagram based on hard parameters and the multi task learning model diagram based on soft parameters are shown in Figure 1 and Figure 2. As for multi-task learning based on hard parameters, the parameters of the bottom shared layer are completely consistent, and the characteristics of the higher layer are completely independent. The bottom parameters of multi-task learning based on soft parameters do not have to be completely consistent, but the bottom parameters are optimized according to different tasks. Compared with the multi-task learning of hard parameters, the multi-task learning based on soft parameters is more flexible [19]. This section introduced the multi-task learning model used in previous research.

A. MULTI-TASK LEARNING BASED ON AUTOENCODER

The autoencoder is divided into encoding stage and decoding stage. The encoding stage is the operation from the input layer to the hidden layer, which compresses the input features from the high-dimensional space to the low-dimensional space. The formula of the encoding operation is:

$$h = f\left(Wx + b_1\right) \tag{1}$$

The decoding operation is the operation from the hidden layer to the output layer. The formula of the decoding operation is:

$$Y = g\left(W'h + b_2\right) \tag{2}$$

W represents the weight of encoder, b_1 represents the bias of encoder, W' represents the weight of decoder, b_2 represents the bias of decoder.,h represents the encoding process and Y represents the decoding process. The function of the decoding operation is to reconstruct the signal so that it is as close to the input signal as possible [20]. To solve the challenge that the shared feature space is always contaminated by domain-specific features, an adversarial multi-domain learning framework(MDANT) was proposed in [21]. MDANT takes advantages of both private and shared feature spaces of the domain representations, which encoded private features and shared features respectively. In [17], gender recognition and speaker recognition were selected as auxiliary tasks and emotion recognition was selected as primary task. They combined unsupervised learning with supervised learning. In the stage of unsupervised learning, they used autoencoder to generate a latent representation z by observing the statistical properties of a given prior distribution. After that, three supervised classifiers for gender recognition, speaker recognition and emotion recognition are connected with the autoencoder.

B. MULTI-TASK LEARNING BASED ON BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMER (BERT)

BERT model consists of input layer, transformer layer and output layer. Transformer is an improvement of sequence to sequence(seq2seq), which corporates the multi head attention mechanism. Seq2seq is composed of encoder and decoder. The encoder compresses the input text information into a fixed length vector, and the decoder outputs the vector as a vector sequence with the same length as the encoder. After calculating the multi-head attention mechanism, features are feed into the add & norm layer for weight addition and normalization. After passing through the feed forward layer, the features will be sent into the add & norm layer again. In this way, the text vectorization is completed through the BERT model. The pre-training of BERT model is divided into two tasks: masking word prediction and next sentence

judgment. In the masked word prediction task, some words will be masked with special symbols randomly, and then other words will be used to predict the masked words. The masked words account for 15%. 80% of these obscured words are replaced by special symbols and 10% are replaced by random words [22]. Because masking word prediction can not judge the relationship between sentences, it is necessary to use the next sentence judgment task to judge whether one sentence is the next sentence of another sentence.

In [23], the text is vectorized by BERT, and the attention mechanism of Bi-GRU and sentence level is used to extract the shared feature part. Specific task is an aspect level based attention mechanism, which is an independent part of each subtask. In reference [24], the sharing part is the fine-tuning Bert model and the shared BLSTM. The function of shared BLSTM is to extract shared features. Private task layer uses private BLSTM to extract private features. If it is the main task, the private features and shared features are spliced together.

C. MULTI-TASK LEARNING BASED ON CONVOLUTIONAL NEURAL NETWORK (CNN) AND LSTM

CNN is good at extracting global features and LSTM has the potential of extracting local features. Combining the two can make full use of their advantages. In [19], they proposed shared encoder and private encoder. Shared encoders were used to extract sentence features that are shared with each task, while private encoders are used to extract features that are more relevant to tasks. The encoder was composed of CNN and LSTM. CNN was used to extract global features, LSTM was used to extract local features, and then they were spliced to realize the coding of shared features and private features. In order to prevent private features from entering shared features, they use generative countermeasure network (GAN). GAN's generator constantly generates features to ''cheat'' the discriminator. The discriminator discriminates whether the features are irrelevant to the task. The generator and the discriminator confront each other.

III. OUR PROPOSED MODEL

The structure diagram of ASP-MTL model used in this paper is shown in Figure 3. Feature space is divided into shared LSTM and private LSTM, which are used to extract shared features and private features respectively. ASP-MTL model is divided into feature extraction layer, confrontation and orthogonal constraint layer and specific task layer. The functions of each layer of ASP-MTL model will be introduced in detail in sections II-A, II-B and II-C. Both model 1 and model 2 are based on ASP-MTL. In model 1, the auxiliary task indentifies the noise category and discards the signal of noise category, so as to achieve the effect of eliminating noise through multi-task learning. Then the non-noise signal is input into model 2. The auxiliary task of model 2 is gender classification and the primary task of model to is emotion classification by learning emotion classification and gender classification at the same time, which reduces the impact

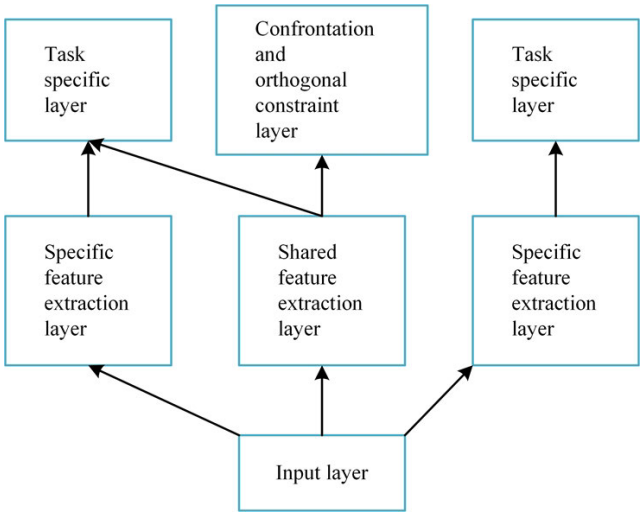

FIGURE 3. Model structure of ASP-MTL.

of gender factors on the recognition accuracy. We train two multi-tasking learning models by minimizing the average error.

A. FEATURE EXTRACTION LAYER

The feature extraction layer is composed of shared LSTM module and private LSTM module. The feature extraction layer is stacked by multiple layers of LSTM. The low-level LSTM learns grammatical features and the high-level LSTM learns semantic features. For each LSTM module, the network parameters will change during the training process. In order to prevent variable transformation caused by change, layer standardization is required. The specific calculation process of batch standardization are as followings:

$$\mu_i = \frac{1}{m}\sum_{j=1}^{m} x_{ij} \tag{3}$$

$$\sigma_i^2 = \frac{1}{m}\sum_{j=1}^{m} \left(x_{ij} - \mu_i\right)^2 \tag{4}$$

$$\hat{x}_{i,j} = \frac{x_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \varepsilon}} \tag{5}$$

$$LN\ (x) = \gamma \hat{x}_{i,j} + \beta \tag{6}$$

x_{ij} represents the input features,m represents the number of features, μ_i represents average value of features, σ_i^2 represents variance of features, $\hat{x}_{i,j}$ represents features after standardization, ε represents very number, $LN\ (x)$ represents activate inverse transformation operation formula, γ represents scale operation, β represents shift operation. The pseudocode of batch standardization is shown in the following.

The shared and private hidden features of the L-th feature extraction layer at time t are as following:

$$s_{l,t}^k = LN\left(LSTM\left(s_{l-1,t}^k, s_{l,t-1}^k\right)\right) \tag{7}$$

$$h_{l,t}^k = LN\left(LSTM\left(h_{l-1,t}^k, h_{l.t-1}^k\right)\right) \tag{8}$$

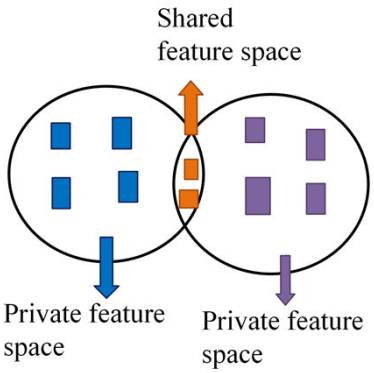

FIGURE 4. Feature space of ASP-MTL.

L represents the number of layers of the feature layer, $s_{l,t}^k$ represents the shared feature at time t of the feature extraction layer of layer L, and $h_{l,t}^k$ represents the private feature at time t of the feature extraction layer of layer L. The output of each feature extraction layer is used as the input of the next layer. The hidden states of shared and private multilayer LSTMs are weighted using trainable parameters α_s, α_h respectively. The expression of the fused shared feature and private feature are:

$$S_t^k = \alpha_s \left[s_{1,t}^k, \ldots, s_{L,t}^k \right]^T \tag{9}$$

$$h_t^k = \alpha_h \left[h_{1,t}^k, \ldots, h_{L,t}^k \right]^T \tag{10}$$

S_t^k represents the fused shared features, h_t^k represents the fused private features.

B. CONFRONTATION AND ORTHOGONAL CONSTRAINT LAYER

As shown in Figure 4, ASP-MTL divides space into private space and shared space. In order to avoid redundant sharing features in private space, shared space has redundant private features, confrontation training was used to make sure that the shared space contains only shared features. Orthogonal constraint training was used to make the private space contain only private features. The generated countermeasure network is composed of generator and discriminator. The generator tries its best to generate samples that cannot be discriminated by the discriminator, and the discriminator tries its best to distinguish the true and false samples. The discriminator uses shared features to judge the source of the task:

$$D(s^k, \theta_D) = soft\max(Us^k + b) \tag{11}$$

U represents learnable parameters, b represents bias.

U is the parameter learned by the task discriminator, and b is the offset. By introducing the countermeasure loss function, the shared feature learning discriminator can not judge the shared feature of the source, and the task discriminator can judge the source as much as possible. The formula of countermeasure loss function is:

$$L_{Adv} = \min\left(\max\left(\sum_{k=1}^{K} \sum_{i=1}^{N_k} d_i^k \log\left[D(E)\left(X^k \right) \right] \right) \right) \tag{12}$$

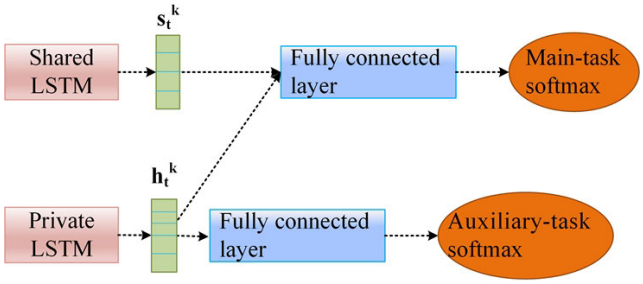

FIGURE 5. Structure diagram of specific task layer.

L_{Adv} represents countermeasure loss, d_i^k represents label of task category.

In order to avoid shared features appearing in private features, orthogonal constraint was applied to private features and shared features. The calculation formula of orthogonal constraint loss is

$$L_{Diff} = \sum_{K=1}^{K} \left\| S^{k^T} H^k \right\|_F^2 \tag{13}$$

$||\ ||_F^2$ represents Frobenius norm, S^k represents shared feature matrix, H^k represents private feature matrix.

C. SPECIFIC TASK LAYER

The structure of specific task layer is shown in Figure 5. There are two softmax classifiers, and there is a full connection layer in front of each softmax classifier. In the main task, the shared features and private features are spliced and sent to the full connection layer to generate the feature vector of the main task. Then the generated feature vector is sent to the softmax classifier. The process formula of main task classification are as followings:

$$h^k = \left[s_t^h, h_t^k \right] \tag{14}$$

$$y^{main} = soft\max\left(Wh^k + b \right) \tag{15}$$

h^k represents the input features of main task, s_t^h represents shared features, h_t^k represents private features, y^{main} represents classification results of main task, W and b represents the weight and bias respectively.

In the auxiliary task, the private features of the auxiliary task are sent to the full connection layer to generate the feature vector of the auxiliary task. Then the feature vector of the auxiliary task is sent to the softmax classifier of the auxiliary task. The classification formula of auxiliary tasks is

$$y^{auxiliary} = soft\max\left(Wh^k + b \right) \tag{16}$$

$y^{auxiliary}$ represents classification results of auxiliary task, h_t^k represents private features of auxiliary task,W and b represents the weight and bias respectively.

D. THE TRAINING PROCESS OF OUR PROPOSED MODEL

In this paper, cross entropy is used as the loss function of auxiliary task prediction and main task prediction. The

calculation formula of loss function of main task is

$$L_{main} = -\left[y^{main} \log \hat{y}^{main} + \left(1 - y^{main}\right) \log \left(1 - \hat{y}^{main}\right)\right] \tag{17}$$

y^{main} represents the real value of the main task and $\hat{y}^{main}$ represents the predicted value of the main task. The loss function calculation formula of auxiliary task is

$$L_{auxiliary} = -\left[y^{auxiliary} \log \hat{y}^{auxiliary} + \left(1 - y^{auxiliary}\right) \log \left(1 - \hat{y}^{auxiliary}\right)\right] \tag{18}$$

$y^{auxiliary}$ represents the real value of the auxiliary task and $\hat{y}^{auxiliary}$ represents the predicted value of the auxiliary task.The loss of every ASP-MTL model is

$$L = \alpha L_{main} + \beta L_{auxiliary} + \gamma L_{Adv} + \mu L_{Diff} \tag{19}$$

$\alpha, \beta, \gamma, \mu$ represents the weight of each loss part.

IV. EXPERIMENT

A. DATASETS

In order to obtain the effect of robustness, in my experiment, these databases are collected in the natural environment affected by noise. We chose the most recent 2016 audio visual emotional challenge (AVEC) database and AFEW6.0 database. Both of the two databases are recorded in the natural environment. In natural environment, there exists noise. AFEW6. 0 database is recorded in a natural environment, which solves the disadvantage that the voice recorded in the experimental environment is not close to nature. This database is closer to our real life. It is a realistic and uncontrolled state. It is an official database raised in the emotion recognition in the wild competition. It contains the facial expressions and voice of the photographed contestants, and contains emotions such as fear, surprise, disgust, happiness, neutrality and sadness, Angry, calm. The AVEC database was recorded by 16 men and 16 women in a natural environment. The emotions included fear, surprise, disgust, happiness, neutrality, sadness and anger. To evaluate the signal strength after noise cancellation, IEMOCAP database was used to make a comparison of our proposed method and other method in noise cancellation. The IEMOCAP database was recorded by the University of Southern California. The emotion database contains four kinds of emotions: happiness, anger, sadness and neutrality. The emotion database contains four kinds of emotions: happiness, anger, sadness and neutrality. We added four types of noise, such as white, pink, babble and factory noise at different SNR level on IEMOCAP database and made compared our SNR_{out} with other methods.

B. PREPROCESSING OF SPEECH SIGNAL

In order to achieve efficient speech emotion recognition, Not only do we need to remove the signal interfered by noise, the signal without emotional meaning also need to be removed. The collected datasets have the problem of data imbalance. To solve the problem of data imbalance, the method of data balance proposed in [25] was used. For example, for these two emotions, happiness and sadness, down sampling is performed to reduce the number of samples because they have the largest number of samples. The specific operation is to extract the beginning and end of Mel spectrum segment audio file in 96 frames. For these few emotion samples, resampling method is used to expand the number of speech segments. In particular, out of fear and surprise, we resampled the logarithmic Mel spectrum with a segment of 24 frames in each sampling interval. Finally, the whole data set is close to the ''balance state''. The preprocessing operations of speech signal include pre-emphasis, framing and windowing, and endpoint detection. Pre emphasis is realized by digital filter to enhance the high-frequency part of speech, make the signal smooth and retain the original information. Because the speech signal is in a changing process. In order to analyze the short-term speech energy, the speech signal is divided into frames in the range of 10ms and 30ms, and the framing operation is realized through Hamming window. In order to prevent the loss of frame continuity, the overlapping framing method is used. In order to prevent the signal frame of non speech part from increasing the redundancy of speech emotion recognition, endpoint detection technology is used to detect zero crossing rate and short-term energy, and distinguish speech signal from non speech signal. Then remove the non voice signal.

C. FEATURE EXTRACTION

If the features are too large, it is easy to lead to dimension disaster. So it is necessary to select the features. The selection of features should be conductive to the problem to be solved. For the removal of noise and non speech signals, the prosodic features zero crossing rate, short-term energy and spectral features with strong noise resistance, such as MFCC and spectrogram, need to be extracted. The frequency of speech emotion signal spectrum is converted into Mel scale, and then the Mel spectrum is obtained. The Mel spectrum is calculated as following:

$$\mathrm{f_{mel}} = 1127\mathrm{lg}(1 + \mathrm{f}/700). \tag{20}$$

The advantage of MFCC is that it can effectively reduce the noise interference of high frequency, so it can well describe speech emotion. In order to reduce the impact of gender factors on the efficiency of speech emotion recognition, a potential mechanism to improve the recognition accuracy of different gender and language groups is spectral features. The pitch change in the audio signal can distinguish gender. We called Python's librosa package to extract features. To sum up, the speech features extracted with this package include MFCC, spectrogram, pitch, zero crossing rate and short-term energy.

D. EVALUATING CRITERION

The criterions selected in this paper to evaluate emotion classification and the signal strength after noise cancellation include accuracy, recall F1 value, SNR_{out},and unweighted average recall(UAR). Accuracy (P) refers to the proportion

of the classifier's prediction to the positive and the prediction samples to all the prediction samples. Recall rate (R) refers to the proportion of samples with positive prediction and correct prediction by the classifier to all samples with true positive prediction. F1 is a harmonic average based on accuracy and recall. The formula of SNR_{out} can be calculated as formula (23). We set the pure voice signal as s(k), additive noise as n(k),f(k) is the noisy signal. f(k) becomes $\hat{s}(k)$ after noise cancellation and $\hat{s}(k)$ is estimation value of pure voice signal s(k).

$$F1 = \frac{2*P*R}{P+R} \tag{21}$$

$$UAR = \frac{\sum_{i=1}^{N} R}{N} \tag{22}$$

$$SNR_{out} = 10\lg \frac{\sum_{k=1}^{L} s^2(k)}{\sum_{k=1}^{L} \left[\hat{s}(k) - s(k)\right]^2} \tag{23}$$

In experiment 1,we compared the methods mentioned in this paper with the previous methods in [8], [9], [10], and [11] on AVEC database and AFEW6.0 database to verify the advantages of our method in removing noise compared with other literatures. The methods used in [8], [9], [10], and [11] were described in the introduction. In the second experiment, the methods mentioned in this paper are compared with references [12], [13], [14], [15] on AVEC database and AFEW6.0 database in the field of eliminating the differences between different databases. The methods used in [12], [13], [14], and [15] were as described in the introduction. This paper repeated their methods and compares them with their own methods. Experiment 3 is the comparison result between the multi-task learning method used in this paper and other multi-task learning methods. The comparison methods included the multi task learning of DNN model sharing hidden layer in [17], [21], and [26]. In the experiment 4, we compared our SNR_{out} value with other noise reduction work on noisy IEMOCAP database.In experiment 5,we carried ablation experiment to demonstrate the importance of noise reduction and gender differences classification on IEMOCAP database.

E. EXPERIMENTAL ENVIRONMENT CONFIGURATION AND PARAMETER SETTING

The running environment of the experiment is Windows system. The programming language is python, the framework of deep learning are tensorflow and keras. And the hardware environment of the experiment is PC. In all of these methods, dropout is set to 0.5, and the learning rate is set to 0.001. We use Matlab simulation tool to add different types of noise under -5dB,-10dB,5dB,10dB signal –noise ratio respectively to the IEMOCAP database.

The multitasking model of this paper: Batch_ size was set to 256, the hidden layer size was set to 32.

The method of [25]: The hidden layer is set as two layers with 256 units in each layer. Relu activation function is selected.

The method of [17]: The encoder part was composed of three convolution layers. Each convolution layer is followed by a pool layer. The RELUactivation function was selected. The structure of the decoder part was the same as that of the encoder part. The three classifiers were composed of convolution layer, maximum pooling layer and softmax classifier.

The method of [26]: They used wavelet soft threshold de-noising method to filter noisy speech signals, and perform wavelet threshold processing on all wavelet coefficients. The speech signal after wavelet threshold de-noising was used as the input signal of variable step size LMS algorithm for the second filtering, and the noise reduction accuracy was appropriately reduced under the premise of ensuring the convergence speed.

Algorithm 1 Input: Network N With Trainable Parameters θ

Output: Batch-normalized network for interface $N_{BN}^{\inf}$

1. $N_{BN}^{\inf} \leftarrow N$ // Training BN network
2. For k=1…K do
3. Add transformation $y^{(k)} = BN_{\gamma(k),\beta(k)}(x^{(k)})$ to N_{BN}^{tr}
4. Modify each layer in N_{BN}^{tr} with input $(x^{(k)})$ to make $y^{(k)}$ instead
5. End for
6. Train N_{BN}^{tr} to optimize the parameters $\theta \cup (\gamma(k), \beta(k))_{k=1}^{K}$
7. $N_{BN}^{\inf} \leftarrow N_{BN}^{tr}$ // Interface BN network with frozen parameters
8. forK=1…K do
9. Process multiple training mini-batches β,each of size m,and average over them:

$E[x] \leftarrow E_{\beta}[\mu_{\beta}]$

$Var[x] \leftarrow \frac{m}{m-1} E_{\beta}[\sigma_{\beta}^2]$

10. In N_{BN}^{inf},replace the transform $y = BN_{\gamma,\beta(x)}$ with $y = \frac{\gamma}{\sqrt{var[x]+\varepsilon}} + (\beta - \frac{\gamma E[x]}{\sqrt{var[x]+\varepsilon}})$

The method of [27]: Compared with traditional spectral subtraction, his method has two additional parameters: α And $\beta_{\circ}$ is a power correction factor.After enhancement, it can effectively improve the signal-to-noise ratio, but also increase the distortion of voice signal. β is the noise correction factor. The value of α is 2,the value of β is 0.5.

It can effectively improve the signal-to-noise ratio, but also increase the distortion of voice signal.

F. EXPERIMENTAL RESULTS

The comparison results of experiment 1 on AVEC database are shown in Table 1 and comparison results of experiment 1 on AFEW6.0 database are shown in Table 2. The

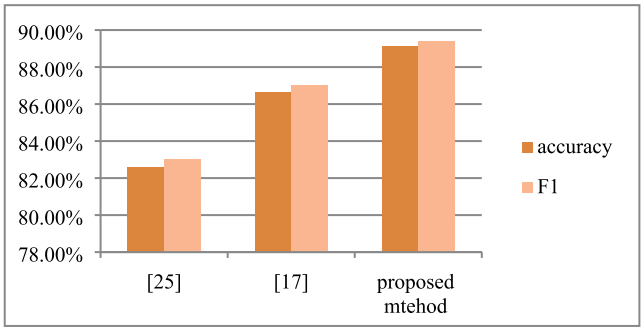

FIGURE 6. Comparison results of multi-task learning model on AVEC database.

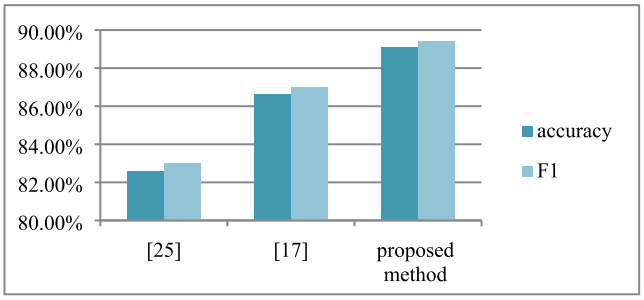

FIGURE 7. Comparison results of multi-task learning model on AFEW6.0 database.

comparison results of experiment 2 on AVEC database are shown in Table 3 and comparison results of experiment 2 on AFEW6.0 database are shown in Table 4. Table 5 shows the comparison results of adding white noise under different signal-noise ratio on IEMOCAP. Table 6 shows the comparison results of adding pink noise under different signal-noise ratio on IEMOCAP. Table 7 shows the comparison results of adding babble noise under different signal-noise ratio on IEMOCAP. Table 8 shows the comparison results of adding factory noise under different signal-noise ratio on IEMOCAP. The comparison results of experiment 3 on AVEC database and AFEW6.0 database are shown in Figure 6 and figure 7 respectively. Experiment 1, experiment 2 and Experiment 3 divided 80% of the data into training set and 20% of the data into test set. Experiment 3 was conducted in AVEC database and AFEW6.0.In experiment 4,we compared our SNR_{out} with literature [26] and [27] on IEMOCAP database.As shown in Table 1, the accuracy rates of literature [8], literature [9], literature [10], literature [11] and the methods mentioned in this paper were 80.02%, 81%, 77.86%, 71.67% and 89.13% respectively. The F1 values were 0.804, 0.816, 0.78, 0.718 and 0.894 respectively. As shown in Table 2, the accuracy rates of literature [8], literature [9], literature [10], literature [11] and the methods mentioned in this paper were 78.4%, 81.6%, 80.9%, 81.98% and 90.13% respectively. As shown in Table 3, the accuracy rates of literature [12], literature [13], literature [14], literature [15] and the methods mentioned in this paper were 73.2%, 79.12%, 70.81%, 70.01% and 89.13% respectively. The F1 values were 0.73, 0.8, 0.71, 0.72 and 0.894 respectively. The accuracy rates of the methods mentioned in literature [12], literature [13], literature [14], literature [15] and this paper were 75.2%, 77.5%, 79.8%, 80.04% and 90.13% respectively, and the F1 values are 0.776, 0.791, 0.71, 0.815 and 0.92 respectively The experimental results in tables 3 and 4 showed that the method used in this paper has the highest accuracy and F1 value compared with other methods to eliminate gender differences. Figure 6 and figure 7 showed that the accuracy and F1 value of the proposed multi-task learning method have been improved. According to table 5,we can see that our proposed method performed the best among these methods, followed by [27]. Reference [26] and [28] performed poorly compared with other comparative reference. The reasons for these results were discussed in the discussion section. As is shown in table 5 to table 8,our proposed method has the highest SNR_{out} value among the comparison methods under white noise, pink noise, babble noise and factory noise. It demonstrated that our proposed noise cancellation method is superior than comparison methods.

TABLE 1. Comparison results of Experiment 1 on AVEC database.

METHOD	ACCURACY	F1
[8]	80.02%	0.804
[9]	81%	0.816
[10]	77.86%	0.78
[11]	71.67%	0.718
PROPOSED METHOD	89.13%	0.894

TABLE 2. Comparison results of Experiment 1 on AFEW6.0 database.

METHOD	ACCURACY	F1
[8]	78.4%	0.78
[9]	81.6%	0.81
[10]	80.9%	0.809
[11]	81.98%	0.819
PROPOSED METHOD	90.13%	0.92

TABLE 3. Comparison results of Experiment 2 on AVEC database.

METHOD	ACCURACY	F1
[12]	73.2%	0.73
[13]	79.12%	0.8
[14]	70.81%	0.71
[15]	70.01%	0.72
PROPOSED METHOD	89.13%	0.894

TABLE 4. Comparison results of Experiment 2 on AFEW6.0 database.

METHOD	ACCURACY	F1
[12]	75.2%	0.776
[13]	77.5%	0.791
[14]	79.8%	0.71
[15]	80.04%	0.815
PROPSOSED METHOD	90.13%	0.92

TABLE 5. The comparison results of adding white noise under different signal-noise ratio on IEMOCAP.

METHOD	$SNR_{IN} = 10$	$SNR_{IN} = 5$	$SNR_{IN} = -10$	$SNR_{IN} = -5$
[26]	15.87	12.92	3.74	7.28
[27]	13.04	10.89	3.09	6.04
PROPOSED METHOD	18.01	13.84	4.56	8.76

TABLE 6. The comparison results of adding pink noise under different signal-noise ratio on IEMOCAP.

METHOD	$SNR_{IN} = 10$	$SNR_{IN} = 5$	$SNR_{IN} = -10$	$SNR_{IN} = -5$
[26]	14.57	11.37	2.54	5.01
[27]	12.11	9.89	3.06	5.19
PROPOSED METHOD	16.45	12.41	4.32	6.76

TABLE 7. The comparison results of adding babble noise under different signal-noise ratio on IEMOCAP.

METHOD	$SNR_{IN} = 10$	$SNR_{IN} = 5$	$SNR_{IN} = -10$	$SNR_{IN} = -5$
[26]	14.49	11.54	2.36	5.13
[27]	12.18	9.91	3.13	5.24
PROPOSED METHOD	16.51	13.09	4.39	7.02

TABLE 8. The comparison results of adding factory noise under different signal-noise ratio on IEMOCAP.

METHOD	$SNR_{IN} = 10$	$SNR_{IN} = 5$	$SNR_{IN} = -10$	$SNR_{IN} = -5$
[26]	15.76	11.97	4.09	6.98
[27]	12.99	9.02	3.08	4.49
PROPOSED METHOD	18.84	13.26	5.02	8.76

TABLE 9. Ablation experiment results on IEMOCAP database.

METHOD	ACCURACY
A	81.98%
B	87.12%
PROPOSED METHOD	90.13%

We then carried out ablation experiment to demonstrate the importance of noise reduction and gender differences classification. Method A means our method with removing model 1.Method B means our method with removing model 2.The comparision results are shown in table 9.The experiment results showed that noise reduction is more important than gender classification, because the interface of noise will impair the clarity of speech signal. The results also demonstrated that both noise reduction and gender classification are vital for SER, ignoring either of them will lead to poor performance. Only by solving the problem of noise interface and gender differences at the same time can increase the accuracy.

V. DISCUSSION

This paper used ASP-MTL multi-task learning model. We used shared LSTM and private LSTM to extract common features and private features respectively. And confrontation training was used to make the shared features do not exist in the private LSTM module, and orthogonal constraints was proposed to ensure that the private features do not exist in the common LSTM module. The accuracy rate and F1 score of our proposed method have been improved compared with other similar literature methods. It is proved that the effect of multi task learning to eliminate the influence of noise and gender differences is better than the method without multi task learning. If we consider the multi task learning with constraints on common features and private features, the classification accuracy of multi task learning with constraints on common features and private features is higher than that of traditional multi task learning. The reasons for these results are described in the following analysis:

(1) According to table 1 and table 2 and from table 5 to table 8, it is proved that the proposed noise reduction method performed better than other noise reduction methods without multi task learning. The reason is that other methods only consider eliminating noise, while do not consider the shared information of noise signal and non noise signal. As a result, they lost the correlation between noise signal and non-noise signal, which will cause the loss of correlation information. The reason why [27] performed worse than [26] is that compared with wavelet transform, The disadvantage of spectral subtraction is that the subtraction process needs to be very careful to avoid speech distortion.

(2) According to the results of tables 3 and 4, the reason why the accuracy and F1 of our proposed method performed better than other gender differences reduction method is that other methods to eliminate gender differences only focus on how to reduce the differences, but do not make use of the links between the differences. Moreover, although the method of transfer learning can apply the knowledge learned from the source data set to the target data set and reduce the difference between the source data set and the target data set, it can not share information by the tasks.

(3) According to the results of figure 6 and figure 7, the accuracy and F1 value of literature [25] and literature [17] are lower than other multitasking learning methods. Because in [25], they learned the shared features through the underlying shared hidden layer, and the top layer performed multi task classification. In [17], they encoded and decoded speech signals in the unsupervised learning stage, and constructed three multi task learning classifiers in the supervised learning stage. However, they do not restrict shared features and private features, so it is difficult to distinguish which features are shared features and which are private features. Sometimes, shared features are learned as private features, and private features are learned as shared features, which will result in low recognition rate.

VI. CONCLUSION

SER is an vital part in artificial intelligence, which can make machine understand human emotions. SER can be widely applied in education industry, service industry, assisted driving industry and criminal investigation industry.

Gender differences and noise interface are two major problems will have adverse effect on the accuracy of SER. This paper solved the two problems at the same time. This paper used two multi task learning models of ASP-MTL to reduce noise interference and gender differences. Multitasking learning can take advantage of the relevance between different tasks. The first ASP-MTL model considered emotion recognition as the main task and noise recognition as the auxiliary task. Then we removed the noise signal identified by model 1. Then the non-noise part were put into model 2. The main task of model 2 was emotion classification, and the auxiliary task was gender classification. ASP-MTL model restricted shared features and private features, so that shared features only exist in shared feature space and private features only exist in private feature space. Compared with other methods of removing noise, eliminating gender differences and multitasking learning, the accuracy and F1 value of the proposed method have been improved.

The limitation of our present work are as followings(1)The problem of limited data is not solved.(2)Compared with emotion types in real life, the SER database used in the exper-

iment contains too few emotion types.(3) It is not enough to use acoustic modal only, which will cause confusion of different emotions. Most cross-corpus SER only use acoustic modal, and the research on multi-modal cross-corpus SER is rare. In the future, These issues need further in-depth study. According to the first issue, semi-supervised learning model is needed in the further study. According to the second issue, we need to design emotion databases that can simulate more human emotions. Last but not least, it is an urgent need to consider other modal in the multimodal SER. We need to fuse acoustic features with other modals such as facial expression, blood pressure or heart rate.

REFERENCES

[1] Z. S. Abdallah, M. M. Gaber, B. Srinivasan, and S. Krishnaswamy, "StreamAR: Incremental and active learning with evolving sensory data for activity recognition," in *Proc. IEEE 24th Int. Conf. Tools Artif. Intell.*, Nov. 2012, pp. 1163–1170.

[2] I. Alnujaim, H. Alali, F. Khan, and Y. Kim, "Hand gesture recognition using input impedance variation of two antennas with transfer learning," *IEEE Sensors J.*, vol. 18, no. 10, pp. 4129–4135, May 2018.

[3] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga, "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey," in *Proc. 23th Int. Conf. Architecture Comput. Syst.*, Feb. 2010, pp. 1–10.

[4] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digital Signal Process.*, vol. 22, pp. 1154–1160, Dec. 2012.

[5] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[6] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *Artif. Intell. Rev.*, vol. 43, no. 2, pp. 155–177, 2015.

[7] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, Nov. 2006.

[8] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," *Int. J. Speech Technol.*, vol. 23, no. 1, pp. 45–55, Mar. 2020.

[9] Ł. Juszkiewicz, "Improving noise robustness of speech emotion recognition system," in *Intelligent Distributed Computing VII*. Cham, Switzerland: Springer, 2014, pp. 223–232.

[10] M. Pandharipande, R. Chakraborty, A. Panda, and S. K. Kopparapu, "Robust front-end processing for emotion recognition in noisy speech," in *Proc. 11th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Nov. 2018, pp. 324–328.

[11] X. Zhao, S. Zhang, and B. Lei, "Robust emotion recognition in noisy speech via sparse representation," *Neural Comput. Appl.*, vol. 24, nos. 7–8, pp. 1539–1553, Jun. 2014.

[12] Y. Chen, Z. Xiao, X. Zhang, and Z. Tao, "DSTL: Solution to limitation of small corpus in speech emotion recognition," *J. Artif. Intell. Res.*, vol. 66, pp. 381–410, Oct. 2019.

[13] Y. Li, W. Zheng, Y. Zong, Z. Cui, T. Zhang, and X. Zhou, "A bi-hemisphere domain adversarial neural network model for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 494–504, Apr. 2021.

[14] E. N. N. Ocquaye, Q. Mao, H. Song, G. Xu, and Y. Xue, "Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition," *IEEE Access*, vol. 7, pp. 93847–93857, 2019.

[15] H. Luo and J. Han, "Nonnegative matrix factorization based transfer subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2047–2060, 2020.

[16] J. Wen, "Research on classification algorithm for multi-task learning," *Guangdong Univ. Technol.*, to be published.

[17] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 992–1004, Apr. 2022.

[18] W. Zhao, H. Gao, S. Chen, and N. Wang, "Generative multi-task learning for text classification," *IEEE Access*, vol. 8, pp. 86380–86387, 2020.

[19] N. Jin, J. Wu, X. Ma, K. Yan, and Y. Mo, "Multi-task learning model based on multi-scale CNN and LSTM for sentiment classification," *IEEE Access*, vol. 8, pp. 77060–77072, 2020.

[20] L. H. Meng, "Theoretical research and application of autoencoder," *China Univ. Mining Technol.*, to be published.

[21] N. N. Wang, "Research on text representation model based on BERT improvement," *Southwest Univ.*, to be published.

[22] Y. Ma, "Research on aspect-level sentiment analysis algorithm based on multi-task learning," *Wuhan Univ. Technol.*, to be published.

[23] K. L. Wan, "Research on online offensive speech detection and recognition based on multi-task learning," *Sichuan Univ.*, to be published.

[24] G. Chen, S. Zhang, X. Tao, and X. Zhao, "Speech emotion recognition by combining a unified first-order attention network with data balance," *IEEE Access*, vol. 8, pp. 215851–215862, 2020.

[25] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proc. Interspeech*, 2017, pp. 1103–1107.

[26] L. Qingqiang, "A varible step size LSM speech denoising algorithm based on wavelet threshod," *J. Jinlin Univ.*, vol. 60, no. 4, pp. 945–949, 2022.

[27] L. Junhui, "Comparison and analysis of two improved spectral noise reduction processing algorithms," *J. Shandong Agricult. Univ.*, vol. 50, no. 5, pp. 849–851, 2019.

[28] K. K. Sahoo, I. Dutta, M. F. Ijaz, M. Wozniak, and P. K. Singh, "TLEFuzzyNet: Fuzzy rank-based ensemble of transfer learning models for emotion recognition from human speeches," *IEEE Access*, vol. 9, pp. 166518–166530, 2021.

[29] P. N. Srinivasu, J. G. SivaSai, M. F. Ijaz, A. K. Bhoi, W. Kim, and J. J. Kang, "Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM," *Sensors*, vol. 21, no. 8, p. 2852, 2021.

[30] H. Basak, R. Kundu, P. K. Singh, M. F. Ijaz, M. Woźniak, and R. Sarkar, "A union of deep learning and swarm-based optimization for 3D human action recognition," *Sci. Rep.*, vol. 12, no. 1, pp. 1–17, Mar. 2022.

[31] D. Bhattacharya, D. Sharma, W. Kim, M. F. Ijaz, and P. K. Singh, "EnsemHAR: An ensemble deep learning model for smartphone sensor-based human activity recognition for measurement of elderly health monitoring," *Biosensors*, vol. 12, no. 6, p. 393, Jun. 2022.

[32] C. Alves, B. Carlotto, and B. Dias, A. Garcia, and B. Gianesi, "Transfer learning and data augmentation techniques applied to speech emotion recognition in SE&R 2022," Tech. Rep., 2022.

LIU YUNXIANG is currently pursuing the Ph.D. degree. He is also a Professor, a Master Supervisor, and the Leader of key disciplines of the Institute. He is also a Senior Member of China Computer Society, a Reviewer of *Journal of Computer Software*, and an Editorial Board of *Journal of Computer Measurement and Control*. He is mainly engaged in the research work in the fields of artificial intelligence, computer software and theory, information fusion, and intelligent information processing. He has achieved a series of important results in the theory and application of fuzzy set, the theory and application of rough set, intelligent decision support system, data fusion system testing technology, and the research and development of intelligent instruments.

ZHANG KEXIN received the bachelor's degree from the Dalian Institute of Science and Technology, in 2021. She is currently pursuing the master's degree with the Shanghai Institute of Technology. Her research interests include natural language processing and speech emotion recognition.