IEICE TRANSACTIONS

on Fundamentals of Electronics, Communications and Computer Sciences

DOI:10.1587/transfun.2024EAP1034

Publicized:2024/07/05

A PUBLICATION OF THE ENGINEERING SCIENCES SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

Speech Emotion Detection using Fusion on Multi-Source Low-Level Information Based Recurrent Branches

Jiaxin WU[†], **Bing LI**[†*a)], **Li ZHAO**[††], ***and*** **Xinzhou XU**[†††b)], ***Nonmembers***

SUMMARY The task of Speech Emotion Detection (SED) aims at judging positive class and negetive class when the speaker expresses emotions. The SED performances are heavily dependent on the diversity and prominence of emotional features extracted from the speech. However, most of the existing related research focuses on investigating the effects of single feature source and hand-crafted features. Thus, we propose a SED approach using multi-source low-level information based recurrent branches. The fusion multi-source low-level information obtain variety and discriminative representations from speech emotion signals. In addition, focal-loss function benifit for imbalance classes, resulting in reducing the proportion of well-classified samples and increasing the weights for difficult samples on SED tasks. Experiments on IEMOCAP corpus demonstrate the effectiveness of the proposed method. Compared with the baselines, MSIR achieve the significant performance improvements in terms of Unweighted Average Recall and F1-score.

key words: *Speech emotion detection, multi-source low-level information, recurrent branches, convolutional recurrent network*

1. Introduction

As the important medium for obtaining and disseminating information in human communication, speech signals not only contain linguistics information, but also contain rich paralinguistic information (e.g., emotion features) [1]. Typical paralinguistic topics such as *Speech Emotion Recognition* (SER) have been researched and applied extensively, which making *Human Computer Interaction* (HCI) [2] more intelligent and efficient. As another prominent topic in paralinguistic, *Speech Emotion Detection* (SED) learn implicit paralinguistic features from speech naturally and detect emotion states in speech signals by distinguishing the values of different parameters [3], [4], which providing valuable assistance in the field of stress-related occupations and psychological counseling.

The research in relation to the SED task relies on feature extraction module and classification module primarily. Most existing works focus on obtaining the accuracy of emotion detection through exploiting priori knowledges [5], [6], Mel spectrogram [7]–[9] and paralinguistic feature sets [10], [11] to identifying the paralinguistic content. The recent research of deep learning provides SER with *Deep Neural Networks* (DNNs) [12]–[14] to structure emotion representations from raw features. Then, the DNNs are applied as the feature extractor on Mel spectrogram to obtain the deep emotion representations [15]–[17], which are utilized to predict emotion classes.

Nevertheless, despite of these works in SER for the models above, the research in relation to the SED tasks still exist two issues to address. First, most existing DNN-based emotion detection tasks focus on utilizing the fixed types of features [18]–[20] as the underlying features which leads to the loss of high-level representations in speech emotion signals [21]. Second, most of the existing works achieved their detection performance through each sample's unweigted loss [22]–[24], without take into account the degradation of emotion detection performance due to the majority of negative class samples among the [25]–[27] and the lack of reasonable weights design between easily classified samples and difficult samples.

In order to address the first issue, we propose inclusion of the multi-source *Low-Level Information* (LLI) to obtain high-level representations and more prominent features in SED tasks. The LLI include three components: We utilise the *Convolutional Recurrent* (CR) branch to process the log-Mels features, while we employ the *Low-Level Descriptor* (LLD) and wav2vec branches to approach the low-level descripors and wav2vec 2.0 features, respectively. Afterwards, to address the second challenge, the *Focal-Loss* (FL) function is employed to emphasizes on the weights for difficult samples to reduce influence from easily classified training samples.

In this letter, we propose the *Multi-Source low-level information based recurrent Branches* (MSIR) approach for SED. The CR branch utilises *Convolutional Recurrent Neural Network* (CRNN) to approach the 2D-log-Mels features. The LLD branch employs *Recurrent Neural Network* (RNN) to process 3D-LLD features. The wav2vec branch makes use of RNN to address wav2vec 2.0 features. Further, we employ focal-loss to the concatenation of three branches. The major contribution of this paper can be summarized as follows:

• For the speech emotion detection tasks, we propose an MSIR approach using recurrent neural networks on low-level information for each utterance, while employing focal-loss to address the easily classified samples and difficult samples.

• Within the proposed fusion approach, we design the LLI containing 2D-log-Mels, 3D-LLD and wav2vec 2.0 fea-

[†]The author is with the School of Integrated Circuits, Southeast University, Nanjing, 210000, China.

[††]The author is with the School of Information Science and Engineering, Southeast University, Nanjing, 210000, China.

[†††]The author is with the School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, 210000, China.

[*]The author is with the School of Cyber Science and Engineering, Southeast University, Nanjing, 210000, China.

a) E-mail: bernie-seu@seu.edu.cn (Corresponding author)

b) E-mail: xinzhou.xu@njupt.edu.cn (Corresponding author)

tures as the input of recurrent branches with attention mechanism, which obtains more advanced depth features and more diverse emotional information for SED tasks.

• Within the proposed approach, we employ the focal-loss by control the weight parameters to balance the easily classified samples and difficult samples.

The rest of the paper is organized as follows. Section 2 reviews the related work, while Section 3 introduces the details of the proposed scheme in this paper. Then, we present the experiments and their corresponding analysis of results in Section 4.

2. Related Work

2.1 Emotion Detection in Speech

As previously works mentioned, spectrum features extracted from raw speech signals are extensively applied for SER. The prosody features and log-Mel spectrograms from audio samples [28], [29] are utilized as the input of DNN models to recognize emotions.

In view of the success of most SER tasks are applied in idealized scenarios, we consider a more detailed and targeted SED tasks. Lalitha et al. [30] analizes the performance of emotion detection on DNN, which adopts various perceptual features as the input to obtain important emotion information. Further, [31], [32] apply Mel spectrograms as the input of a *Gate Recurrent Unit* (GRU) based RNN, which is used to summarize emotion representations to detection depression emotions from audio.

On emotional spaces, in addition to conventional valence-arousal cases [33], Atmaja et al. [34] employ three-dimensional emotion model with valence, arousal, and dominance to characterize categorical emotion, where acoustic features extracted by CNN and LSTM, respectively. Additionally, the study in [35] utilizes four dimensions arousal-expectancy-power-valence to describe emotional states, and perform *Relevance Units Machine* (RUM) to predict emotions.

Further, Mirheidari et al. [36] proposes emotion detection through recognition different degrees of Expressed Emotion, which exploits LLD extracted from audio segments. Moreover, Zou et al. [37] concatenated multiple levels acoustic features to predicte emotions, where spectrogram extracted by *Convolutional Neural Network* (CNN), MFCC extracted by *Bi-directional Long Short-Term Memory networks* (Bi-LSTM) and wav2vec 2.0 extracted by the transformer-based network, then fused by co-attention mechanism to achieve competitive performance. Although handcrafted features are very effective in distinguishing speech emotions, most of them are low-level features.

2.2 Deep Learning in Speech Emotion Analysis

DNNs are frequently used to learn hidden frequency and time domain representations in speech signals, also repeatedly employed in SER systems. Nevertheless, there is a lack of in-depth investigations of implementing aggregation approaches over different time steps in SER tasks [38]. To obtain time-dependent features, Luo et al. use RNN to learn long-time context from multiple frame-level LLDs [39], [40]. Meanwhile, attention mechanisms have been included to focus on the emotionally-relevant parts [41]–[43]. Furthermore, Liu et al. [44], [45] proposes *Long Short-Term Memory networks* (LSTM) with convolution filters on different scales, which is designed to extract emotion-relevant features from different domains in emotion classification task.

Similarly, DNNs have also been applied in the SED task, [46] using a CNN model for SED, achieving high average accuracy on three main emotions. Then, an RNN network [47] is used to emotion detection in dialogue, which provides better contextual information for the utterance and obtained better emotion detection results. After that, the Bi-LSTM [48] is applied to extract acoustic features in the classification model of neutral emotion detection.

Further, the combination of *Convolutional Neural Network and Temporal Convolution Network* (CNN-TCN) [49] is adopted as feature extraction module, which obtaines local spectral features for emotion detection. Currently, *Bidirectional Encoder Representations from Transformers* (BERT) [50] is employed to exploring contextual information to improve performance in emotion recognition, which indicating effectiveness and generalization capability of fine-tuning. Meanwhile, the low-rank adaptation [51] is utilized for parameter-efficient fine-tuning to reduce the training parameters of large language models, which achieves good performance in empathy detection and emotion classification tasks. In contrast, we employ a CNN-LSTM network with an attention mechanism for SED, containing a CNN network to capture information in the temporal dimension, and LSTM is used to capture temporal correlations between features. In addition, the attention mechanism assigns weights to emotional features with different strengths to obtain more distinctive features.

3. Proposed Methodology

In this section, we introduce our proposed MSIR for SED as shown in Fig. 1. First, the 2D-log-Mel sepctrograms (static features and first order delta) extracted from the raw speech signals are used as the input of the CR branch. Then, we generate the 3D-LLD and wav2vec 2.0 features from raw speech signals as the input to Bi-LSTM. Finally, through fusing the three branches, emotional states can be detected in speech.

3.1 The Convolutional Recurrent Branch

For the convolutional recurrent branch, the convolutional layers are adept in extracting locally invariant features from the input sequence, and the recurrent layer capture the temporal correlation between frames to obtain global features. Given a speech signal, we split the signal into frames with Hamming windows and *Fast Fourier Transform* (FFT) is performed for

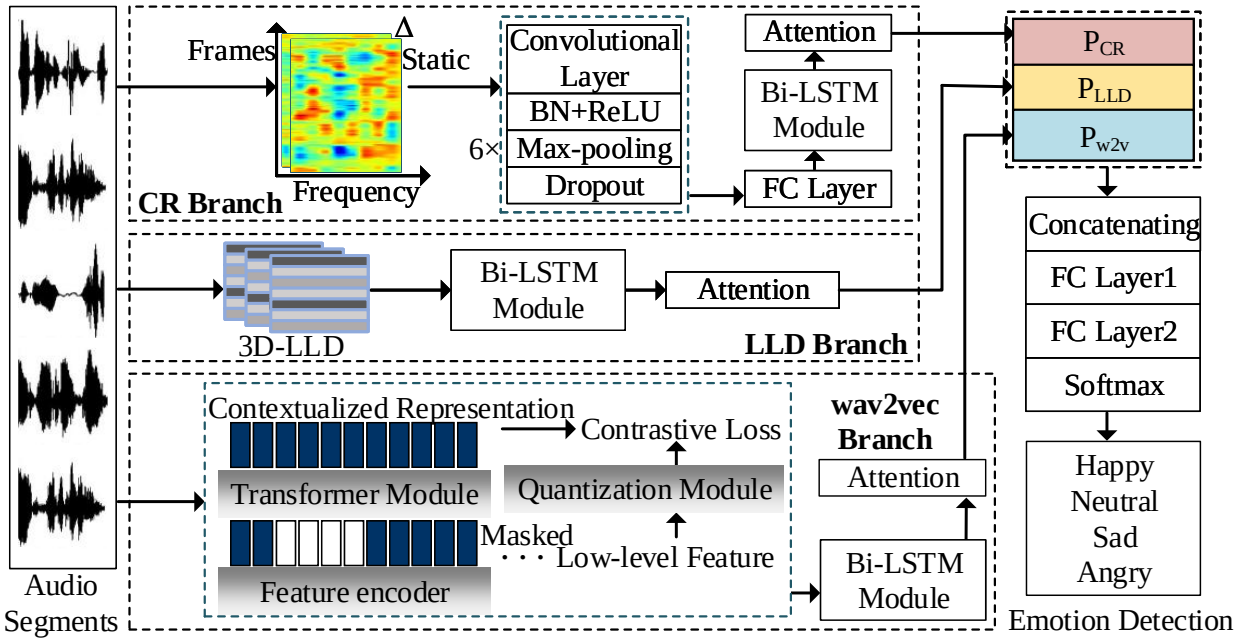

Fig. 1 The architecture of the proposed MSIR model. The 2D-log-Mels is first fed to Convolutional Recurrent Branch (CR Branch). The 3D-LLD and wav2vec 2.0 features are used as the inputs of Low-Level Descriptor Branch (LLD Branch) and wav2vec Branch, respectively. The multi-source low-level information with focal-loss are fed to the softmax classifier.

each frame to obtain frequency-domain information. Then, the frequency-domain information is weighted using Mel-filter bank n to obtain the energy y_n for each Mel frequency band. The energy of each Mel frequency band is logarithmic to obtain the log-Mel spectrogram l_n. In addition, the first order delta of the static log-Mels is calculated as l_n^d.

The input of convolutional layer is the 2D-log-Mels feature $\mathbf{M}$ ($L \times W \times K$), where L is the frame number, and $W = 40$ is the channels number of Mels while $K = 2$ represent the static features and first order delta of Mels. The output feature map of the convolutional layer is $\mathbf{C}$ ($L \times W_c \times K$), where W_c represents the feature dimension after convolutional operation and the activation function is Leaky-ReLU. Then, the max-pooling layer is used to reduce dimensionality and control overfitting, the output feature maps $\mathbf{P}$ is expressed as

$$\mathbf{P} = \mathrm{POOL}\left(\sigma_{\text{Leaky-ReLU}}\left(\mathrm{CONV}_{m\times n}(\mathbf{M})\right)\right) \tag{1}$$

where $\sigma_{\text{Leaky-ReLU}}(\cdot)$ represents the activation function Leaky-ReLU, $m \times n$ represents the size of the convolution kernel is 5×3. $\mathrm{POOL}(\cdot)$ represents the max-pooling operation.

Before passing the output feature of CNN modules to the long short-term memory network, the *Fully-Connected* (FC) layer is added for each low-level unit to reduce feature dimension with no loss in accuracy. The LSTM module updates the value of the cell through operations between gate functions, which effectively storing and acquiring contextual information. In this work, the Bi-LSTM is adopted to obtain the present and future information in an utterance. Additionally, in order to obtain discriminative utterance-level representations, we employ an attention layer to focus on emotion relevant parts for SED. Accordingly, the output of convolutional recurrent branch can be presented as

$$\mathbf{P}_{\text{CR}} = \mathbf{D}_{\text{BiLSTM}}(\mathbf{P}) \odot \left(\omega_{\mathbf{P}} \cdot \sigma_{\tanh}(\mathbf{D}_{\text{BiLSTM}}(\mathbf{P}))\right), \tag{2}$$

with

$$\mathbf{D}_{\text{BiLSTM}}(\mathbf{P}) = \left[\mathbf{D}^{(\text{F})}_{\text{LSTM}}(\mathbf{P})^{\text{T}}, \mathbf{D}^{(\text{B})}_{\text{LSTM}}(\mathbf{P})^{\text{T}}\right]^{\text{T}}, \tag{3}$$

where $\mathbf{P}$ is the output of CNN moduel, $\mathbf{D}_{\text{BiLSTM}}(\mathbf{P})$ consists of $\mathbf{D}^{(\text{F})}_{\text{LSTM}}(\mathbf{P})$ and $\mathbf{D}^{(\text{B})}_{\text{LSTM}}(\mathbf{P})$ representing the forward and backward output of Bi-LSTM respectively, and $\omega_{\mathbf{P}}$ represents weight vector learned from $\mathbf{D}_{\text{BiLSTM}}(\mathbf{P})$. Then, $\sigma_{\tanh}(\cdot)$ represents the tanh activation function [31]. Accorrdingly, the $\mathbf{P}_{\text{CR}}$ is the final output of convolutional recurrent branch with attention.

3.2 The Low-Level Descriptor and wav2vec Branches

We obtain 65 low-level descriptors and their delta descriptors from the input speech signal. The LLDs contain the categories of loudness, energy, *Zero-Crossing Rate* (ZCR), the 1–26 *RelAtive Spectral TrAnsform* (RASTA) auditory band, *Mel Frequency Cepstrum Coefficient* (MFCC) 1–14 without 0-th coefficient, spectral features, and F0-related features. We extract the 130-dimensional LLDs using the openSMILE toolkit [52].

The pre-trained wav2vec branch through self-supervised learning employ the wav2vec 2.0 to obtain meaningful speech representations from raw signals [37]. Fig. 1

shows the wav2vec branch composed of three blocks. The feature encoder contains several convolutional blocks, which encodes the raw audio $\mathbf{G}$ into latent speech representations $\mathbf{L_1}, \mathbf{L_2}, ..., \mathbf{L_j}$, where j is the time step. $\mathbf{L_j}$ is normalized to zero mean and unit variance. Specifically, CNN blocks are composed of temporal convolution layer, layer normalization and gaussian error linear unit activation function.

Then, $\mathbf{L_j}$ are fed to the transformer-based contextualized encoder module which gain contextualized representations $\mathbf{R_1}, \mathbf{R_2}, ..., \mathbf{R_j}$ by aggregating multiple time steps. Finally, we take the latent speech representations $\mathbf{L_j}$ from the CNN blocks as input to the quantization module and obtain embedding $\mathbf{Q_j}$ after discretization by product quantization. The contrastive loss is employed to optimize the context representations deriving from contextualized encoder module and the discretization embeddings. We add the L2 regularization and diversity loss to increase the use of the quantized codebook representations [53]. Finally, contrastive loss is optimized at each step to obtain the trained contextualized representation $\mathbf{R_j}$ as part of SED acoustic features. Note that we freeze the wav2vec 2.0 model (with built-in contrastive loss) and regard it as the feature extractor.

Finally, the LLD features and wav2vec 2.0 features are utilized as input to two Bi-LSTM networks with attention layer, respectively. The output of low-level descriptor branch and wav2vec branch are present as

$$\mathbf{P}_{\mathrm{LLD}} = \omega_{\mathrm{LLD}} \cdot \sigma_{\tanh} \left(\mathbf{D}_{\mathrm{BiLSTM}}(\mathbf{LLD}) \right), \tag{4}$$

$$\mathbf{P}_{\mathrm{w2v}} = \omega_{\mathrm{R_j}} \cdot \sigma_{\tanh} \left(\mathbf{D}_{\mathrm{BiLSTM}}(\mathbf{R_j}) \right), \tag{5}$$

where $\mathbf{D}_{\mathrm{BiLSTM}}(\mathbf{P}) = \mathbf{D}^{(\mathrm{F})}_{\mathrm{LSTM}}(\mathbf{P}) + \mathbf{D}^{(\mathrm{B})}_{\mathrm{LSTM}}(\mathbf{P})$ represents the forward and backward output of Bi-LSTM, respectively. Specifcally, ω_{LLD} and $\omega_{\mathrm{R_j}}$ represent the weights for $\mathbf{D}_{\mathrm{BiLSTM}}(\mathbf{LLD})$ and $\mathbf{D}_{\mathrm{BiLSTM}}(\mathbf{R_j})$, respectively. In addition, $\sigma_{\tanh}(\cdot)$ represents the tanh activation function [31]. Accordingly, the $\mathbf{P}_{\mathrm{LLD}}$ and $\mathbf{P}_{\mathrm{w2v}}$ are the final output of LLD and wav2vec branches, respectively.

3.3 Multi-Source Concatenated Features with Focal-Loss

Then, we fuse the three branches as the multi-source low-level information. The utterance-level representations $\mathbf{S}$ as

$$\mathbf{S} = [\mathbf{P}_{\mathrm{CR}}{}^{\mathrm{T}}, \mathbf{P}_{\mathrm{LLD}}{}^{\mathrm{T}}, \mathbf{P}_{\mathrm{w2v}}{}^{\mathrm{T}}]^{\mathrm{T}} \tag{6}$$

The concatenations of LLI is the input of fully-connected layers. Similar to the previous works, the softmax activation function is applied to compute the emotion predictive probabilities.

Focal-loss is commonly used to solve category imbalance and classification difficulties in object detection tasks, which also exist in SED tasks. Therefore, this paper uses the *Focal-Loss* (FL) function to minimize the divergence between predicted labels and the ground truth.

First, in order to solve the problem of imbalance between positive and negative samples, the weighting factor $\zeta \in [0, 1]$ is increased as shown in Eq.(7). When the number of positive samples is much larger than the negative samples, then control $\zeta \in [0, 0.5]$ to increase the weight of negative samples and decrease the weight of positive samples. When $\zeta = 0.5$, it is the standard cross-entropy function.

Although the weighting factor balances the positive and negative samples, it does not address the balance problem of simple and hard samples. Therefore, the modulating factor $(1-p)^{\theta}$ is added to the function, where p represents the network's estimated probability for positive cases and $1-p$ is negative cases, the θ is focus factor. By adjusting θ to reduce the loss contribution of simple samples and increase the weight of hard samples. The FL is generated as shown in

$$\mathcal{L}_{FL}(p) = -\zeta (1-p)^{\theta} \log(p), \tag{7}$$

where ζ is weighting factor, $\theta \geq 0$ is focus factor, and $(1-p)^{\theta}$ is the modulating factor. When simple positive samples are correctly classified, p tends to 1, $(1-p)^{\theta}$ tends to 0, and the contribution to the total loss is very small.

4. Experiments

4.1 Experimental Preparation

1) Date and Features: To evaluate the performance of our proposed method, we test it on the *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) [54] database collected by the University of Southern California. The database contains approximately 12 hours of data in total including 10 professional actors. Each dialogue is performed by two actors of different genders, contains a total of 5 dialogues, and each dyadic dialogue is segmented into utterances. The average duration of each utterance is 4.5 seconds. The utterances are labeled with six emotion labels (happy, sad, neutral, angry, excited, and frustrated). In this paper, happy (595) and excited (1 041) is combined into happy, so we predict the four most representative emotions among them: neutral (1 708), sad (1 084), angry (1 103), and happy (1 636), with a total of 5 531 utterances as in related works on IEMOCAP [32], [37], [43], [44].

In the experiments, the log-Mels features are extracted by python-speech-features 0.6 toolkit. The openSMILE feature extract toolkit [52] (2.4.1 version) is utilized to extract LLDs (static features, first order delta (Δ) and second order delta (Δ^2)). Note that the transformers 4.26.1 toolkit is employed to extract wav2vec 2.0 features, based on the pre-trained wav2vec2-base-960h model. Moreover, this SED system is implemented in PyTorch 0.4.0 version, accelerated by CUDA 9.0. Specifically, the dimensionality of the input features for each branch is shown in Table 2.

2) Evaluation Setups: In the experiment, we employ z-score to normalize the experimental data. In order to reduce the influence of over-fitting, the *Session-Independent* (SI) strategy is adopted in our experiment, setting the first four sessions as the training set and the last session as the test set. The *Unweighted Average Recall* (UAR) is given by the average of four classes recall values as the metric to evaluate

the performance. It is consistent with most other works on IEMOCAP. In addition, we use F1-score to evaluate the classification accuracy of our proposed model.

3) Experimental Parameters: For the CR branch, the first convolutional layer contains 128 kernels accompanied by batch normalization (momentum of 0.99, the weight decay of 0.001), and each of the other convolutional layers includes 256 kernels. Note that, We only use a max-pooling layer after the first convolutional layer, with the size 2 and the stride of 2, employing zero padding.

For the wav2vec branch, the feature encoder module contains 7 blocks and each block with 512 channels. The convolutional with strides $(5, 2, 2, 2, 2, 2, 2)$ and kernel widths $(10, 3, 3, 3, 3, 2, 2)$. The raw audio encodes into a sequence of embeddings with a stride of 20ms and a receptive field of 25ms. The contextualized encoder module uses 12 transformer blocks with 8 attention heads each, model dimension 768. The proposed MSIR is implemented using Python platform and TensorFlow framework. We use the *Adaptive moment estimation* (Adam) optimizer in our experiments and the range of initial learning rate is $\{5\times10^{-6}, 10^{-6}, 5\times10^{-5}, 10^{-5}, 5\times10^{-4}, 10^{-4}, 5\times10^{-3}, 10^{-3}, 5\times10^{-2}, 10^{-2}\}$. The architecture is trained with the batch size of 64. The model parameters are optimized by minimizing the loss function within 100 epochs.

4.2 Experimental Results and Analysis

4.2.1 Comparison between Approches

First, we examine the UAR and F1-score performances using different low-level information and their concatenations for different recurrent branches with different loss functions. The low-level information consist of 2D-log-Mels, 3D-log-Mels, 1D-LLD, 2D-LLD, 3D-LLD, and wav2vec 2.0, while we consider different RNN branches with different loss. The feature sets of the *Computational Paralinguistics Challenge* (ComParE) [55] and the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) are the baselines. Table 1 lists the UAR and F1-score of different approaches (including baselines and the proposed MSIR) on SED. The results suggest that RNN branches outperform baselines and obtain absolute improvements of UAR and F1-score on SED. This indicates that low-level information with recurrent branches can retain and obtain effective emotional information. Furthermore, the results demonstrate that the UAR of 2D-log-Mels on *Attention-based Convolutional Recurrent Neural Network* (ACRNN) is better than [18]. Thus, we employ this feature in our following experiments. We utilize LLDs (including 1D-LLD, 2D-LLD, 3D-LLD) as the input of RNN (the module refer to as LLF-LSTM in [36]) and concatenate the obtained features into ACRNN (2D), respectively. Compared with LLF-LSTM (1D-LLD) and LLF-LSTM (2D-LLD), the concatenation of ACRNN (2D) and LLF-LSTM (3D-LLD) obtains relative improvement of 1.5% and 0.9% on UAR, 2.7% and 1.5 % on F1-score, respectively.

Meanwhile, in order to obtain more plentiful emotional representations, we obtain the wav2vec 2.0 feature, while concatenate with ACRNN (2D) and LLF-LSTM (3D-LLD) features (This architecture is the *Multi-Source low-level information based recurrent Branches without Focal-Loss* (MSIR (w/o FL)). The results of Table 1 expresses that, the UAR and F1-score of (MSIR (w/o FL)) are 80.9% and 71.1% which obtain an absolute improvement for baselines. Besides, compared with the concatenation of ACRNN (2D) and LLF-LSTM (3D-LLD) features, we also concatenate wav2vec 2.0 (which as the input of w2v-EN is similar to [56] without dense layer) and LLF-LSTM (3D-LLD), which gain improvement of 0.9% UAR and 2.2% F1-score, indicating that contextual emotional representation contains more emotional information and effective for improving SED tasks. In particular, compared with the concatenation of two branches, multi-source low-level information have better performance for the SED task.

Table 1 The Unweighted Average Recall (%) and F1-score (%) comparison between the proposed MSIR and other different RNN architectures with multi-source low-level information on SED, where the baselines are obtained by the Support Vector Machines.

Approaches \ Metrics	UAR (%)	F1-score (%)
ComParE [54]	64.8	60.1
eGeMAPS [55]	62.6	65.3
ACRNN (2D) [18]	75.0	60.7
ACRNN (3D) [18]	74.8	60.7
ACRNN (2D) +LLF-LSTM (1D-LLD) [39]	77.4	62.9
ACRNN (2D) +LLF-LSTM (2D-LLD) [39]	78.0	64.1
ACRNN (2D) +LLF-LSTM (3D-LLD) [39]	78.9	65.6
MLAI-Co.Att [37]	79.1	69.1
w2v-EN [56]	74.3	58.9
LLF-LSTM (3D-LLD) [39]	79.4	66.3
LLF-LSTM (3D-LLD)+w2v-EN [56]	79.8	67.8
MSIR (w/o FL)	80.9	71.1
MSIR	**82.1**	**71.2**

Table 2 Dimensionality of input features or representations for each branch.

Branches	Descriptors	# Features
Convolutional Recurrent Branch	log-Mels	40
	log-Mels (Δ)	40
Low-Level Descriptor Branch	LLD	65
	LLD (Δ)	65
	LLD (Δ^2)	65
wav2vec Branch	wav2vec 2.0	768

4.2.2 Ablation Studies

Further, in order to process the imbalance between positive samples and negative samples as well as simple samples and hard samples, we employ focal-loss function for SED. Table 1 demonstrates that we obtain 1.2% UAR and 0.1% F1-score further improvement compared with MSIR (w/o FL), which indicate that the focal-loss is effectiveness for imbalance

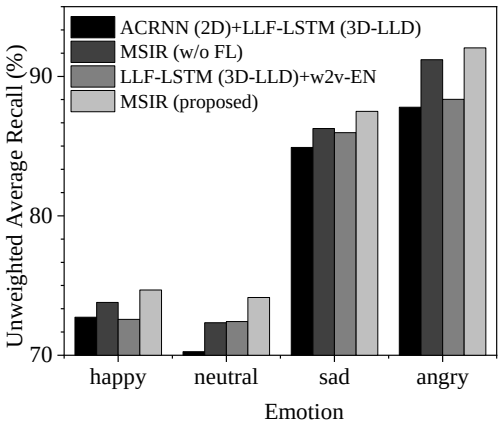

Fig. 2 The column graphs of UAR for the proposed approach MSIR and MSIR (w/o FL).

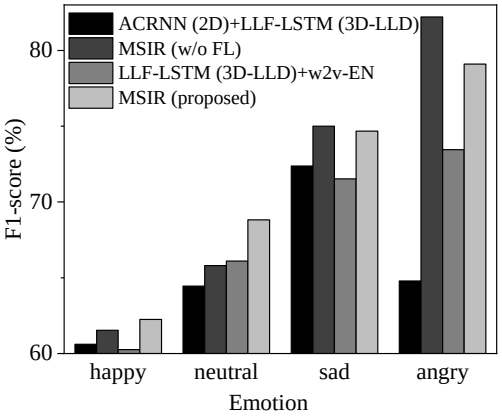

Fig. 3 The column graphs of F1-score for the proposed approach MSIR and MSIR (w/o FL).

classes. Specifically, we present the column graph results of UAR and F1-score for the proposed approach MSIR and MSIR (w/o FL) in Fig. 2 and Fig. 3.

We make a comparison between our proposed system MSIR and other recurrent branches for "happy", "neutral", "sad", "angry" emotions on Sensitivity (True Positive Rate noted as 'TPR'; %) and Specificity (True Negative Rate noted as 'TNR'; %). The results of Table 3 and Table 4 imply that MSIR achieved the highest Specificity of happy, neutral, sad, and angry, which are 87.3%, 86.0%, 96.6%, and 98.4%, respectively. Besides, the Sensitivity of neutral, sad, and angry with the MSIR (w/o FL) architecture are 63.1%, 67.7%, and 78.8%, which higher than only using two branches. This verifies that multi-source fusion features can achieve better performance.

For the purpose of making class-wise comparison, the confusion matrices of four emotions on the proposed system MSIR in Fig. 4a, 4b, 4c, and 4d, respectively. The results

Table 3 The Sensitivity (True Positive Rate noted as 'TPR'; %) of "happy", "neutral", "sad", "angry" emotions obtained by four methods respectively.

Approaches \ Emotions	Happy	Neutral	Sad	Angry
ACRNN (2D) +LLF-LSTM (3D-LLD) [39]	65.2	56.7	64.0	51.1
MSIR (w/o FL)	60.2	63.1	67.7	78.8
LLF-LSTM (3D-LLD) +w2v-EN [56]	63.6	62.2	59.6	65.7
MSIR	57.5	59.3	64.1	70.7

Table 4 The Specificity (True Negative Rate noted as 'TNR'; %) of "happy", "neutral", "sad", "angry" emotions obtained by four methods respectively.

Approaches \ Emotions	Happy	Neutral	Sad	Angry
ACRNN (2D) +LLF-LSTM (3D-LLD) [39]	84.7	81.3	94.7	98.0
MSIR (w/o FL)	86.1	80.2	95.1	97.8
LLF-LSTM (3D-LLD) +w2v-EN [56]	81.5	80.8	96.4	97.4
MSIR	87.3	86.0	96.6	98.4

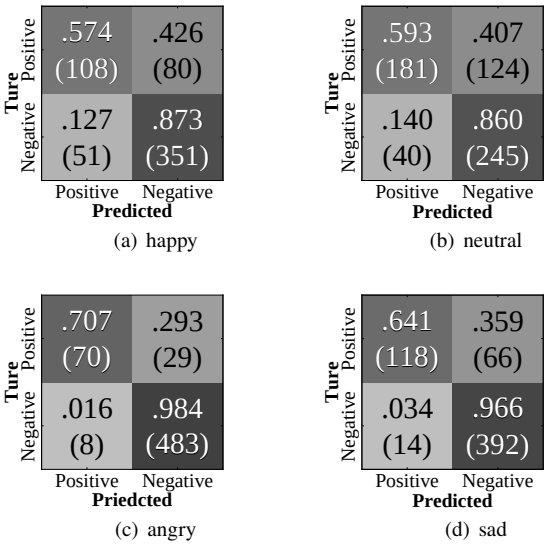

Fig. 4 Confusion matrices (including recalls and the numbers) of the proposed MSIR on four emotons.

express that the proposed approach obtain the recalls of four emotions, where "happy" is 57.4% (for 'positive') and 87.3% (for 'negative'), "neutral" is 59.3% (for 'positive') and 86.0% (for 'negative'), "sad" is 64.1% (for 'positive') and 96.6% (for 'negative'), and "angry" is 70.7% (for 'positive') and 98.4% (for 'negative'). This verifies the performance of the proposed approach on both of the classes.

Finally, we consider comparing the parameter ζ of FL

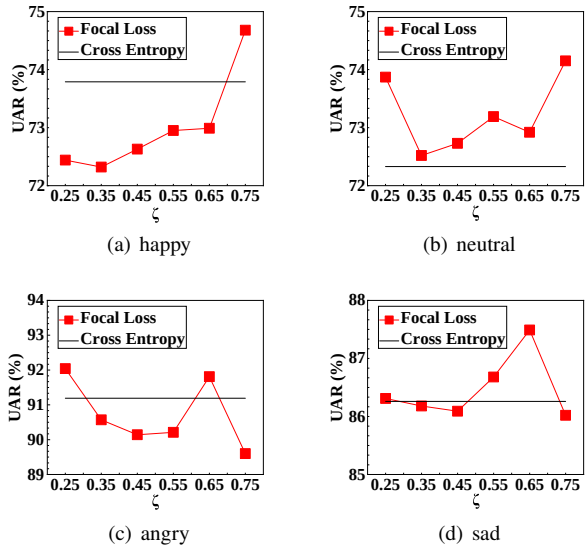

Fig. 5 The line graph of the focal-loss with different value of weighting factor (where focusing parameter θ=2).

in four emotions (where we set $\theta = 2$ as the same as most of the previous works), and the performance are all better than *Cross Entropy* (CE) (i.e., we examine the best UAR results for the proposed MSIR approaches compared with MSIR (w/o FL)), as shown in Fig. 5a, 5b, 5c, and 5d, respectively. The UAR of "happy" with the ζ=0.75 is 0.9% better than CE. The UAR of "sad" with the ζ=0.25, 0.55, 0.65 are better than CE, where UAR is 1.2% better than CE when ζ=0.65. The UAR of "angry" with the alpha=0.25, 0.65 are better than CE, where UAR is 0.9% better than CE when ζ=0.25. In particular, the UAR of "neutral" for all parameters of ζ is better than CE. When ζ=0.75, the UAR is 1.8% better than CE which obtain the largest performance improvement. By adding focusing parameter θ, the contribution of simple samples in the loss is reduced and expands the range of samples accepting low loss. By adjusting the value of weighting factor ζ, which balances the importance of positive and negative samples and improves the performance of the model.

5. Conclusion

This paper presented a novel approach for detecting speech emotions using recurrent branches through fusing on low-level information. We first extracted low-level information from audio segments to generate emotion representations. These low-level information were input to different recurrent branches, outputting the concatenation of the features. Next, we utilized the focal-loss to dispose the imbalance of classes in SED tasks. The experimental results on IEMOCAP dataset show the superior performance of the proposed approach, compared with existing research and baselines. Future work may focus on investigating more effective low-level infomation for the different branches, it is also expected the other pre-trained models. Additionally, we would like to explore transfer learning for solving cross-domain speech emotion detection problem [57]–[59].

Acknowledgments

This work is supported by the State Key Program of National Natural Science Foundation of China (U2003207), National Natural Science Foundation of China (NSFC) (62174150), and China Postdoctoral Science Foundation (2022M711693).

References

[1] S. Li, X. Xing, W. Fan, B. Cai, P. Fordson, and X. Xu, "Spatiotemporal and frequential cascaded attention networks for speech emotion recognition," Neurocomputing, vol.448, pp.238–248, 2021.
[2] R.S. Sudhakar and M.C. Anil, "Analysis of speech features for emotion detection: a review," Proc. International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, pp.661–664, IEEE, 2015.
[3] A. Koduru, H.B. Valiveti, and A.K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," International Journal of Speech Technology, vol.23, no.1, pp.45–55, 2020.
[4] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), Stockholm, Sweden, pp.1089–1093, ISCA, 2017.
[5] K. Hartmann, I. Siegert, D. Philippou-Hübner, and A. Wendemuth, "Emotion detection in HCI: From speech features to emotion space," IFAC Symposium on Analysis, Design, and Evaluation of Human-Machine Systems, vol.46, no.15, pp.288–295, 2013.
[6] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii State, USA, pp.2852–2861, IEEE, 2017.
[7] S. Bedoya-Jaramillo, E. Belalcazar-Bolaños, T. Villa-Cañas, J. Orozco-Arroyave, J. Arias-Londoño, and J. Vargas-Bonilla, "Automatic emotion detection in speech using mel frequency cesptral coefficients," Proc. Symposium of Image, Signal Processing, and Artificial Vision (STSIVA), Medellin, Antioquia, Colombia, pp.62–65, IEEE, 2012.
[8] S. Lalitha, D. Geyasruti, R. Narayanan, and M. Shravani, "Emotion detection using MFCC and cepstrum features," Procedia Computer Science, vol.70, pp.29–35, 2015.
[9] I. Shahin, O.A. Alomari, A.B. Nassif, I. Afyouni, I.A. Hashem, and A. Elnagar, "An efficient feature selection method for arabic and english speech emotion recognition using Grey Wolf Optimizer," Applied Acoustics, vol.205, p.109279, 2023.
[10] M. Sajjad, S. Kwon, *et al.*, "Clustering-based speech emotion Recognition by incorporating learned features and Deep BiLSTM," IEEE Access, vol.8, pp.79861–79875, 2020.
[11] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyderabad, India, pp.3683–3687, ISCA, 2018.
[12] S.P. Mishra, P. Warule, and S. Deb, "Variational mode decomposition based acoustic and entropy features for speech emotion recognition," Applied Acoustics, vol.212, p.109578, 2023.
[13] N. Scheidwasser-Clow, M. Kegler, P. Beckmann, and M. Cernak, "SERAB: A Multi-Lingual Benchmark for Speech Emotion Recognition," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual and Singapore, pp.7697–7701, IEEE, 2022.
[14] A.S. Tehrani, N. Faridani, and R. Toosi, "Unsupervised representations improve supervised learning in speech emotion recognition,"

ArXiv Preprint ArXiv:2309.12714, 2023.

[15] M. Baruah and B. Banerjee, "Speech emotion recognition via generation using an attention-based variational recurrent neural network," Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), Brno, Czechia, pp.4710–4714, ISCA, 2022.

[16] G.A. Prabhakar, B. Basel, A. Dutta, and C.V.R. Rao, "Multichannel CNN-BLSTM architecture for speech emotion recognition system by fusion of magnitude and phase spectral features using DCCA for consumer applications," IEEE Transactions on Consumer Electronics, vol.69, no.2, pp.226–235, 2023.

[17] S. Sarker, K. Akter, and N. Mamun, "A text independent speech emotion recognition based on convolutional neural network," Proc. International Conference on Electrical, Computer and Communication Engineering (ECCE), Swansea, UK, pp.1–4, IEEE, 2023.

[18] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," IEEE Signal Processing Letters, vol.25, no.10, pp.1440–1444, 2018.

[19] D.M. Schuller and B.W. Schuller, "A review on five recent and near-future developments in computational processing of emotion in the human voice," Emotion Review, vol.13, no.1, pp.44–50, 2021.

[20] C. Marechal, D. Mikolajewski, K. Tyburek, P. Prokopowicz, L. Bougueroua, C. Ancourt, and K. Wegrzyn-Wolska, "Survey on AI-based multimodal methods for emotion detection," High-performance Modelling and Simulation for Big Data Applications, vol.11400, pp.307–324, 2019.

[21] A. Triantafyllopoulos, S. Liu, and B.W. Schuller, "Deep speaker conditioning for speech emotion recognition," Proc. International Conference on Multimedia and Expo (ICME), Shenzhen, China, pp.1–6, IEEE, 2021.

[22] H. Zhou, K. Liu, and P. Shenzhen, "Speech emotion recognition with discriminative feature learning," Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), Shanghai, China, pp.4094–4097, ISCA, 2020.

[23] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Graz, Austria, pp.7405–7409, ISCA, 2019.

[24] P. Kumar, S. Jain, B. Raman, P.P. Roy, and M. Iwamura, "End-to-end Triplet loss based emotion embedding system for speech emotion recognition," Proc. International Conference on Pattern Recognition (ICPR), Virtual Event / Milano, Italy, pp.8766–8773, Springer, 2021.

[25] T.Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," Proc. International Conference on Computer Vision (ICCV), Venice, Italy, pp.2980–2988, IEEE, 2017.

[26] J. Cai, Z. Meng, A.S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," Proc. International Conference on Automatic Face & Gesture Recognition (FG), Xi'an, China, pp.302–309, IEEE, 2018.

[27] X.Y. Jing, X. Zhang, X. Zhu, F. Wu, X. You, Y. Gao, S. Shan, and J.Y. Yang, "Multiset feature learning for highly imbalanced data classification," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.43, no.1, pp.139–156, 2019.

[28] Y. Chang, Z. Ren, T.T. Nguyen, K. Qian, and B.W. Schuller, "Knowledge transfer for on-device speech emotion recognition with neural structured learning," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, p.no pagination, IEEE, 2023.

[29] P. Pérez-Toro, D. Rodríguez-Salas, T. Arias-Vergara, S. Bayerl, P. Klumpp, K. Riedhammer, M. Schuster, E. Nöth, A. Maier, and J. Orozco-Arroyave, "Transferring quantified emotion knowledge for the detection of depression in alzheimer's disease using forestnets," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, p.no pagination, IEEE, 2023.

[30] S. Lalitha, S. Tripathi, and D. Gupta, "Enhanced speech emotion detection using deep neural networks," International Journal of Speech Technology, vol.22, no.3, pp.497–510, 2019.

[31] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: an emotional Audio-Textual corpus and a Gru/Bilstm-based model," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual and Singapore, pp.6247–6251, IEEE, 2022.

[32] W. Wu, M. Wu, and K. Yu, "Climate and weather: Inspecting depression detection via emotion recognition," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual and Singapore, pp.6262–6266, IEEE, 2022.

[33] Y. Feng and L. Devillers, "End-to-end continuous speech emotion recognition in real-life customer service call center conversations," Proc. International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp.1–8, IEEE, 2023.

[34] B.T. Atmaja and M. Akagi, "Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning," APSIPA Transactions on Signal and Information Processing, vol.9, p.e17, 2020.

[35] F. Wang, H. Sahli, J. Gao, D. Jiang, and W. Verhelst, "Relevance units machine based dimensional and continuous speech emotion prediction," Multimedia Tools and Applications, vol.74, pp.9983–10000, 2015.

[36] B. Mirheidari, A. Bittar, N. Cummins, J. Downs, H.L. Fisher, and H. Christensen, "Automatic detection of expressed emotion from five-minute speech samples: Challenges and opportunities," Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), Incheon, Korea, pp.2458–2462, ISCA, 2022.

[37] H. Zou, Y. Si, C. Chen, D. Rajan, and E.S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual and Singapore, pp.7367–7371, IEEE, 2022.

[38] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN," Speech Communication, vol.120, pp.11–19, 2020.

[39] M. Luo, H. Phan, and J. Reiss, "Cross-modal fusion techniques for utterance-level emotion recognition from text and speech," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, p.no pagination, IEEE, 2023.

[40] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based LSTM," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.27, no.11, pp.1675–1685, 2019.

[41] L. Tarantino, P.N. Garner, A. Lazaridis, *et al.*, "Self-attention for speech emotion recognition," Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, Austria, pp.2578–2582, ISCA, 2019.

[42] Z. Zhao, H. Wang, H. Wang, and B. Schuller, "Hierarchical network with decoupled knowledge distillation for speech emotion recognition," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, p.no pagination, IEEE, 2023.

[43] S. Kakouros, T. Stafylakis, L. Mošner, and L. Burget, "Speech-based emotion recognition with self-supervised models using attentive channel-wise correlations and label smoothing," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, p.no pagination, IEEE, 2023.

[44] K. Liu, D. Wang, D. Wu, and J. Feng, "Speech emotion recognition via two-stream pooling attention with discriminative channel weighting," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, p.no pagina-

tion, IEEE, 2023.
[45] M. Rayhan Ahmed, S. Islam, A. Muzahidul Islam, and S. Shatabda, "An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition," Expert Systems with Applications, vol.218, p.119633, 2023.
[46] D. Bertero and P. Fung, "A first look into a Convolutional Neural Network for speech emotion detection," Proc. IEEE international conference on acoustics, speech and signal processing (ICASSP), New Orleans, LA, USA, pp.5115–5119, IEEE, 2017.
[47] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," Proc. AAAI Conference on Artificial Intelligence, Hawaii, USA, pp.6818–6825, AAAI Press, 2019.
[48] J. Santoso, T. Yamada, K. Ishizuka, T. Hashimoto, and S. Makino, "Performance improvement of speech emotion recognition by neutral speech detection using autoencoder and intermediate representation," Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), Incheon, Korea, pp.4700–4704, ISCA, 2022.
[49] W. Li, J. Xue, R. Tan, C. Wang, Z. Deng, S. Li, G. Guo, and D. Cao, "Global-local-feature-fused driver speech emotion detection for intelligent cockpit in automated driving," IEEE Transactions on Intelligent Vehicles, vol.8, no.4, pp.2684–2697, 2023.
[50] X. Qin, Z. Wu, T. Zhang, Y. Li, J. Luan, B. Wang, L. Wang, and J. Cui, "BERT-ERC: Fine-tuning BERT is enough for emotion recognition in conversation," Proc. AAAI Conference on Artificial Intelligence, Washington, DC, USA, pp.13492–13500, 2023.
[51] Y. Wang, J. Wang, and X. Zhang, "YNU-HPCC at WASSA-2023 Shared Task 1: Large-scale language model with LoRA fine-tuning for empathy detection and emotion classification," Proc. Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis (WASSA), Toronto, Canada, pp.526–530, Association for Computational Linguistics, 2023.
[52] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," Proc. ACM International Conference on Multimedia, Barcelona, Spain, pp.835–838, ACM, 2013.
[53] Y. Wang, A. Boumadane, and A. Heba, "A Fine-tuned Wav2vec 2.0/HuBERT Benchmark for speech emotion recognition, speaker verification and spoken language understanding," ArXiv Preprint ArXiv:2111.02735, 2021.
[54] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol.42, no.4, pp.335–359, 2008.
[55] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), Lyon, France, pp.148–152, ISCA, 2013.
[56] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," 2021 IEEE Spoken Language Technology Workshop (SLT), pp.373–380, 2021.
[57] S. Li, P. Song, and W. Zheng, "Multi-source discriminant subspace alignment for cross-domain speech emotion recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, p.no pagination, 2023.
[58] W. Zhang, P. Song, D. Chen, C. Sheng, and W. Zhang, "Cross-corpus speech emotion recognition based on joint transfer subspace learning and regression," IEEE Transactions on Cognitive and Developmental Systems, vol.14, no.2, pp.588–598, 2021.
[59] W. Zhang and P. Song, "Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.28, pp.307–318, 2019.

Jiaxin Wu received the M.S. degree in Electronic and Communication Engineering from Northwest Normal University in 2018. Since 2018, she has been studying for her doctor degree in Electronics Science and Technology (Integrated Circuit Design) from Southeast University. Her research interests include speech signal processing, affective computing, and the area of information security.

Bing Li received the B.S. degree in Electronics Science and Technology from Southeast University in 1991, and Ph.D from Southeast University in 2004. Professor and tutor of doctoral students at the School of Microelectronics, the School of Cyber Science and Engineering, Southeast University. Director of the joint research center for advanced cloud system of Southeast University. His main research is the efficient and secure integrated circuits and system, include data compression, data encryption, Physical Unclonable Functions, the blockchain, Internet of Things, and the area of information security.

Li Zhao received the B.S. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1982, the M.S. degree from Soochow University, Suzhou, China, in 1988, and the Ph.D. degree from the Kyoto Institute of Technology, Kyoto, Japan, in 1998. He is currently a Professor with the School of Information Science and Engineering, Southeast University, Nanjing. His research interests include spoken signal processing and affective computing.

Xinzhou Xu received the B.S. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2009, and the M.S. and Ph.D. degrees from Southeast University, Nanjing, in 2012 and 2017, respectively. He is currently a Associate Professor with the School of Internet of Things, Nanjing University of Posts and Telecommunications. Previously, he was with the Machine Intelligence and Signal Processing Group, MMK, Technical University of Munich (TUM), Munich, Germany, from 2014 to 2016, and the Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany, from 2015 to 2016. His research interests include audio signal processing, pattern recognition, machine learning, and affective computing.