Speech Emotion Recognition by Combining a Unified First-Order Attention Network With Data Balance

GANG CHEN$^{1,2}$, SHIQING ZHANG$^1$, XIN TAO$^1$, AND XIAOMING ZHAO$^{1,2}$

$^1$Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, China
$^2$Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China

Corresponding author: Xiaoming Zhao (tzxyzxm@163.com)

This work was supported in part by the Zhejiang Provincial National Science Foundation of China and National Science Foundation of China under Grant LZ20F020002, Grant LQ21F020002, and Grant 61976149; in part by the Taizhou Science and Technology Project under Grant 1803gy08; and in part by the Outstanding Youth Project of Taizhou University under Grant 2018JQ003.

ABSTRACT In the domain of speech emotion recognition (SER), generally there is an unbalanced data distribution of emotional samples in existing emotional datasets. Moreover, different fragment areas in an utterance contribute diversely to SER. To address these two issues, this paper proposes a new SER method by combining a unified first-order attention network with data balance. The proposed method firstly utilizes the strategy of data balance to augment and balance the training data. Then, a pre-trained convolutional neural network (CNN) model (i.e., VGGish) is fine-tuned on target emotional datasets to learn segment-level speech features from the extracted Log Mel-spectrograms. Next, the unified first-order attention mechanism, including different feature-pooling strategies such as sum, min, max, mean, and standard deviation (std), is embedded into the output of a bi-directional long short-term memory (Bi-LSTM) network. This is used for learning high-level discriminative segment-level features, and simultaneously aggregating the learned segment-level features into fixed-length utterance-level features for SER. Finally, based on utterance-level features, the softmax layer in a Bi-LSTM network is adopted to conduct final emotion classification task. Extensive experiments are implemented on three public datasets such as BAUM-1s, AFEW5.0, and CHEAVD2.0, demonstrate the advantage of the proposed method.

INDEX TERMS Speech emotion recognition, first-order attention, data balance, convolutional neural network, bi-directional long short-term memory.

I. INTRODUCTION

Currently, automatic speech emotion recognition (SER) has drawn extensive attentions in the areas of speech signal processing, pattern recognition, affective computing and so on. This is attributed to the fact that automatic SER can be used in human-computer interaction [1], [2], smart home [3], smart healthcare [4], robots [5], real-time translation tools [6], etc. Speech signal is one of the most fundamental and effective methods of human communication. Due to the contained rich emotional information in speech signals, individuals can perceive emotional cues from speech signals to respond naturally. However, it is still difficult for automatic SER, since the emotional states of individuals is variable and complicated.

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang$^1$.

Feature extraction and emotion classification are two vital steps on SER tasks. Most prior works focus on learning hand-crafted acoustic features, which are fed into conventional classifiers for final emotion classification task. As far as feature extraction is concerned, these hand-crafted acoustic features are divided into three types: prosodic features [7], voice quality features [8], and spectral features [9], [10]. To date, spectral features have become one of the most well-known hand-crafted features for SER. In particular, Nwe et al. [11] adopted short time log frequency power coefficients (LFPC) as spectral features, and then employed hidden Markov models (HMM) as a classifier for SER. Neiberg et al. [12] proposed to use Mel-frequency cepstral coefficients (MFCCs), MFCC-low and pitch for SER. For emotion classification, Seltzer et al. [13] presented a Bayesian mask estimation method to implement SER. Morrison et al. [14] provided K-nearest neighbors (KNN)
to identify speech emotion categories. Schuller et al. [15] employed support vector machines (SVM) [16] based on radial basis functions to classify emotional acoustic features for SER. However, these traditional hand-crafted features cannot adequately discriminate human subjective emotions, since they are low-level. Therefore, automatic SER is still a challenging task, since it’s difficult to learn high-level speech features characterizing effectively speech emotions.

To address the above-mentioned issue, deep learning methods [17] have been widely used for SER in recent years. Intuitively, deep learning techniques perform high-level feature learning tasks with the aid of deep network structures. There is a great number of works [18]–[21] that concentrate on learning deep speech affective features for SER. The representative deep network models are convolutional neural networks (CNNs) [20], [22], deep belief networks (DBNs) [23], [24], recurrent neural networks (RNNs) [25], [26] as well as its variant called long short term memory network (LSTM) [27]. In addition, since speech signals are time-sequence data, and the saturation of emotion is different at different locations among the temporal sequences of speech [28]. In particular, silent fragments usually contain less speech emotional cues [18] compared with voiced fragments. Therefore, various speech features, derived from different speech fragments, differ in representation capabilities to identify emotions [28]. In other words, different fragment areas in an utterance contribute diversely to SER. In this case, it is helpful to promote SER performance by concentrating on learning important features with high discriminative power. Most deep learning models have powerful feature representation capabilities, but they cannot consider the difference of representation capability among various extracted speech features from different fragment areas.

To alleviate the above-mentioned problem, researchers have tried to integrate the human visual attention mechanism [29] into deep learning techniques for improving recognition performance further. Specially, the human brain’s attention to each area of an image is different, and it will thus focus on certain areas with obvious characteristics. Recently, the attention mechanism has been successfully applied for object detection [30], [31] and classification [32], nature language processing (NLP) [33], [34], machine translation [35], [36], etc. In order to extract more discriminative deep speech features for improving recognition performance, some attention-based deep models have been developed for SER in recent years. In particular, Yoon et al. [29] used two bi-directional long short-term memory (Bi-LSTM) networks with the multi-hop attention mechanism for SER. Ma et al. [37] adopted a multi-task attention-based deep neural networks (DNN) to efficiently learn speech emotional features. Chen et al. [38] used 3-D convolutional recurrent neural networks (3-D CRNN) with a attention model for SER. Huang et al. [39] employed deep convolutional recurrent neural networks (DCRNN) with the attention mechanism for SER. In essence, these attention-based works employ the typical sum-pooling strategy to construct the attention layer. However, different features are suited for different types of feature pooling methods for the attention layer. For instance, max-pooling is usually suitable for sparse features. Considering that different feature pooling methods have their own advantages, it is thus difficult to decide the optimal feature pooling strategy for different features in the attention layer.

Additionally, in SER domain, there are also two challenging issues which should be addressed. First, generally there is an unbalanced data distribution of emotional samples for each emotion in existing limited emotional datasets. However, most speech data augmentation methods [40]–[43] are employed to increase the quantity of speech data. They do not present a possible solution to the unbalanced data distribution of emotional samples. For instance, in an emotional dataset, there are usually much more samples about two typical emotions, i.e., joy and sadness, in comparison with other emotions, because joy and sadness are more conventional than other emotions. Intuitively, the phenomenon of unbalanced data distribution will make the trained models prone to recognize such common emotional categories with a large number of samples. This will greatly reduce the generalization of the trained models. To alleviate this problem, this paper proposes to employ the strategy of “data balance” to enlarge the number of data samples and simultaneously balance the data distribution of emotional samples. Therefore, the used data balance method can not only augment the training data, but also make the number of emotional samples for each emotion as close as possible. As a result, this will increase the generalization and robustness of the trained models. The details will be illustrated in the Section II-A.

Inspired by the above-mentioned advantages of the attention mechanism and data balance, this paper proposes a new method of SER that combines data balance with a unified first-order attention network, including sum, min, max, mean, standard deviation (std). First, speech signals are firstly converted into the Log Mel-spectrogram, and then divided into fixed-length Mel-spectrogram segments. Second, the divided Mel-spectrogram segments is augmented and balanced by using the strategy of data balance. Third, the augmented Mel-spectrogram segments are fed into the pre-trained VGGish [44] network model, which is fine-tuned to learn high-level segment-level features. Therefore, to produce utterance-level features, we aggregate all segment-level features in an audio utterance by using a Bi-LSTM model embedded with the unified first-order attention mechanism. As a result, the obtained utterance-level features are fed into a softmax layer in the Bi-LSTM model to implement speech emotion classification. Experiment results on three public emotional datasets such as BAUM-1s, AFEW5.0 and CHEAVD2.0, show that our proposed method obtains promising performance on SER tasks.

II. RELATED WORK
The main component in a SER system is feature extraction. Existing speech affective features are divided into two categories: hand-crafted features and deep features.
In this section, we hence review related works about feature extraction.

A. FEATURE EXTRACTION

1) HAND-CRAFTED FEATURE EXTRACTION

Most previous works in the last two decades focus on employing hand-crafted features to identify speech emotions. For SER, it is popular to derive acoustic features to capture speech affective cues, since acoustic features carry plentiful emotional information [45]. The well-known acoustic features are low-level descriptors (LLDs). LLDs are sampled from short-term speech segments, in which the window size and overlapping size are 25 ms and 10 ms [46], respectively. The typical feature sets, such as the eGeMAPS [45] and ComParE [47] features, include MFCCs, harmonic-to-noise ratio (HNR), pitch and jitter, shimmer, loudness, and so on. LLDs are usually extracted by professional software tools such as openSMILE [48] and openEAR [46].

Kamaruddin et al. [49] extracted MFCCs for SER. Yang and Lugger [50] used new harmony features to characterize pitch intervals for SER. Bitouk et al. [51] proposed a set of class-level spectral features that contain the mean and std of MFCCs over stressed vowel, unstressed vowel and consonant regions for SER. Shen et al. [52] adopted energy and pitch related features to perform speech emotion classification. However, these hand-crafted features are low-level, since they can not sufficiently recognize emotions well owing to its poor robustness and low discriminative ability for SER.

2) DEEP FEATURE EXTRACTION

Owing to the above-mentioned limitations of hand-crafted features, many recent studies in SER domain have employed deep learning techniques to learn high-level speech features, which contain high-level emotional semantic information. The representative deep models on SER tasks are CNNs [53]–[55], DBNs [23], and RNNs [18], [54], [56], [57]. It is known that CNNs were originally used for image recognition. Recently, CNNs are also employed for deep feature extraction from speech signals, due to its powerful feature representation capabilities. Huang et al. [58] employed CNNs to learn affect-salient discriminative features for SER. However, directly training deep CNNs with large network parameters is prone to result in overfitting, since most existing speech emotion datasets are relatively small. To relieve this issue, Zhang et al. [55] proposed to adopt a deep image model pre-trained on a large ImageNet dataset [59], and then implemented a fine-tuning strategy to extract deep speech affective features for SER.

Since speech signals belong to time-sequence data, researchers begin to apply RNNs or LSTMs to model temporal dynamic sequence cues for SER. Xie et al. [18] extracted frame-level speech features and then generated deep speech affective features learned by attention-based LSTM. Fayek et al. [57] employed a LSTM-RNN model trained in a way of Sequence-to-One for SER, and indicated that the trained network models on long sequences obtained better performance. Mirsamadi et al. [25] learned short-time frame-level acoustic features with deep recurrent neural networks, and used the feature pooling strategy based on the local attention to focus on learning the emotionally relevant regions of speech signals.

However, these above-mentioned works do not consider the effect of the unbalanced data distribution in an emotional dataset. To address this problem, this paper proposes to combine a unified first-order attention network with data balance for SER. The proposed method is verified on three public speech emotional datasets, including BAUM-1s, AFEW5.0 and CHEAVD2.0. Experimental results demonstrate the proposed method can present comparable performance to state-of-the-arts.

III. OUR METHOD

Fig. 1 provides the flowchart of the proposed SER method by combining a unified first-order attention network with data balance for speech emotion classification. The proposed method contains five steps: (1) speech preprocessing, (2) data balance, (3) segment-level feature extraction, (4) utterance-level feature extraction, (5) speech emotion classification. The detail of each step will be described in the following.

A. SPEECH PREPROCESSING

Since the audio files in different datasets have different sampling frequencies, all audio files are firstly resampled to the mono audios of 16 kHz. This is used to standardize the sampling frequencies of all audio files. Therefore, following in Hershey et al. [44], a 25ms Hann window [60] and a 10ms frame shift is used to perform a short-time Fourier transform [61] for producing corresponding Mel-spectrogram. Next, the Log Mel-spectrogram is calculated by:

\[
\text{Mel}_{\log} = \log (\text{Mel} + 0.01)
\]

where the Log Mel-spectrogram is denoted by \(\text{Mel}_{\log}\), and \(\text{Mel}\) is the Mel-spectrogram. \(\log (\cdot)\) is the logarithmic operation, and the bias of 0.01 is used to avoid taking the logarithm of 0. Eventually, each sample consists of 96 frames. There are no overlapping frames, and the duration of each frame is 10ms. Subsequently, mapping the spectrogram into the 64 Mel-filter banks is performed to obtain the Mel-spectrogram with 64 frequency bands. Therefore, each utterance could be transformed into a feature dimension of \(n \times 96 \times 64\), where \(n\) is calculated by:

\[
n = T / 0.96
\]

where \(T\) is the duration of audio file.

B. DATA BALANCE

The number of emotional samples for each emotion in an emotional dataset is usually different. Specially, there is even a big difference in quantity for certain emotions. Therefore, the strategy of data balance is employed for balancing the
sample number of each emotion. After data balance, each emotion category has almost the same sample number.

A speech emotional dataset should be divided into the training and testing sets for experiments. The training set is employed to train a deep model, whereas the testing set is used to evaluate the performance of deep models. However, the size and quality of the used datasets significantly influences the performance of deep models. Therefore, researchers start to increase the quantity of training data or corrupt the original speech signals with noise [42], in order to avoid overfitting and improve the robustness of deep models [62].

All speech Mel-spectrogram segments can be divided into two categories. One is the duration of speech segments shorter than 0.96s, the other is longer than 0.96s. The former needs to replicate new frames with zero values, so that the obtained log Mel-spectrogram segment contains 96 frames. In particular, the detail of data balance for BAUM-1s [55] is described in Fig. 2. In addition, Fig. 3 and 4 separately show the details of data balance on AFEW5.0 [63] and CHEA VD2.0 [64].

As shown in Fig. 2, the white arrow represents a standard speech segment with a length of 0.96s. The length of color bar means the whole duration of audio file. Different operations are denoted by different colors. For instance, for these two emotions, joy and sadness, down-sampling is performed to reduce the number of samples, since they have the largest number of samples. The specific operation is to extract Mel-spectrogram segments with 96 frames at the beginning and end of the audio file, respectively. For these emotions with fewer samples, the resampling method is used to enlarge the number of speech segments. Specially, for fear and surprise, we resample Log Mel-spectrogram segments in a length of 24 frames per sampling interval. Eventually, the entire dataset is close to a “balance state”. Similarly, the data balancing operations on the AFEW5.0 and CHEA VD2.0 datasets are illustrated in Fig. 3 and 4, respectively.

C. SEGMENT-LEVEL FEATURE EXTRACTION

Inspired by the success of VGG [65] networks in the ImageNet challenge, Hershey et al. [44] presented the VGGish network trained by the Google’s AudioSet. AudioSet is a large data set with 2.1 million audio files and the sum of all audio files durations is 5.8 thousand hours. All audio files can be divided into 527 classes. They are presented in the form of audio files of music, voice, musical instruments, etc.
The idea of transfer learning [66], [67] is to fine-tune the pre-trained models on target datasets. This is used to initialize the network model parameters, thereby accelerating network training and alleviating the pressure of data insufficiency. In this paper, the pre-trained VGGish network model is employed for fine-tuning on target emotional speech datasets, thereby learning high-level segment-level features from the extracted Mel-spectrogram. The VGGish network model consists of six convolutional layers (Conv1, Conv2, Conv3, Conv4, Conv5 and Conv6) and four max-pooling layers (Pool1, Pool2, Pool3, and Pool4), with ReLU activation functions, followed by four fully connected layers (fc1, fc2, fc3 and fc4). The number of last fully connected layer’s neurons (fc4) is equal to emotion categories. Note that the third fully connected layer (fc3) has 128 neurons, thereby producing a 128-dimension feature vector for each extracted Mel-spectrogram segment after fine-tuning the pre-trained VGGish model. Since the learned deep features by VGGish contain high-level emotional information, these deep features are highly relevant to human emotion recognition in speech signals.

Given the $i$-th Mel-spectrogram segment $s_i$, the fine-tuning process of the VGGish network is equivalent to minimize the following problem:

$$
\min_{W, \theta} \sum_{i=1}^{N} L \left( \text{softmax} \left( W^S \cdot \phi \left( s_i; \theta \right) \right), y_i \right)
$$

(3)

where $N$ denotes the number of Mel-spectrogram segments. $W^S$ is the weight values of softmax layer, and $\phi$ is the output of VGGish network. $\theta$ is the parameter of VGGish network. $y_i$ is the true label of $i$-th input. The loss function $L$ is expressed as:

$$
L \left( \psi, y \right) = - \sum_{i=1}^{N} y_i \log (y'_i)
$$

(4)

where $y'_i$ is the $i$-th output value of the softmax layer for the network $\psi$.

D. UTTERANCE-LEVEL FEATURE EXTRACTION

Since the segment-level features learned by the VGGish network only contain short-time emotional information. However, an audio utterance is composed of several segment-level features along the time dimension. The contribution of each segment-level feature in an utterance is different for the final emotional recognition task, since the discriminative power of each segment-level feature is distinguishing [18]. Therefore, the significance of segment-level features for SER can be represented by the weight coefficients computed by the attention mechanism [25]. In this paper, we integrate the Bi-LSTM network with the unified first-order attention to model the long-time cues of affective speech dynamic information along the time dimension. The used Bi-LSTM network consists of one hidden layer with 512 neurons. Each direction contains 256 cells and the dimension of output is 512.

Given a speech sequence $N(x_1, x_2, \ldots, x_N)$, a Bi-LSTM computes the forward hidden sequence $\overrightarrow{h}$, the backward hidden sequence $\overleftarrow{h}$, the output sequence $y$ by iterating the forward layer from $t = 1$ to $N$, and the backward layer from...
t = N to 1. Then, updating the output layer is implemented. The forward process is depicted in Eq. (5). The backward process is described in Eq. (6). And Eq. (7) denotes the process of output.

$$\overrightarrow{h}_t = H \left( W_{h} \overrightarrow{x}_t + W_{f} \overrightarrow{h}_{t-1} + b_{h} \right)$$ (5)

$$\overrightarrow{h}_t = H \left( W_{h} \overrightarrow{x}_t + W_{f} \overrightarrow{h}_{t-1} + b_{h} \right)$$ (6)

$$\overrightarrow{y}_t = W_{o} \overrightarrow{h}_t + W_{o} \overrightarrow{h}_t + b_{o}$$ (7)

where $W_{h}$, $W_{f}$ and $W_{o}$ are respectively the weight parameter of input gate, forget gate and output gate. H denotes the hidden layer function, and $b$ is a bias vector.

Following in Gers et al. [68], H is implemented by the following five composite functions:

$$i_t = \text{sigmoid} \left( W_{i} [x_t, h_{t-1}, c_{t-1}] + b_{i} \right)$$ (8)

$$f_t = \text{sigmoid} \left( W_{f} [x_t, h_{t-1}, c_{t-1}] + b_{f} \right)$$ (9)

$$o_t = \text{sigmoid} \left( W_{o} [x_t, h_{t-1}, c_{t-1}] + b_{o} \right)$$ (10)

$$c_t = f_t c_{t-1} + i_t \tanh \left( W_{c} [x_t, h_{t-1}] + b_{c} \right)$$ (11)

$$h_t = o_t \tanh \left( c_t \right)$$ (12)

where the sigmoid symbol is the logistic sigmoid function. $i_t$, $f_t$, $o_t$, and $c_t$ are separately the input gate, forget gate, output gate and cell activation. $W_{i}$, $W_{f}$, $W_{o}$ and $W_{c}$ are the weight matrices of input gate, forget gate, output gate and cell state update, respectively. $\tanh (\cdot)$ is the hyperbolic tangent function. The $b$ term is a bias vector, and $h_t$ is the $t$-th time step output of the corresponding hidden layer.

An attention layer aims to learn highly discriminative features from the extracted segment-level features, since all segment-level features do not contribute equally to the final emotion classification task [38]. In this paper, we use a unified first-order attention model to score the contribution of segment-level features for SER. As shown in Fig. 5, the used unified first-order attention mechanism contains five different feature pooling strategies: sum, mean, standard deviation (std), min and max. They are defined in Eq. (15). Note that the sum-pooling based attention mechanism is commonly used in most previous works [10], [18], but it does not always perform better in contrast to other feature-pooling strategies. This is verified in the following experiments.

In Fig. 5, the output of a Bi-LSTM is $h_t = \left[ \overrightarrow{h}_t, \overrightarrow{h}_t \right]$ at time step $t$. First, the normalized contribution weight $\alpha_t$ is calculated by a softmax function in Eq. (13) [38]. Then, as shown in Eq. (14), the utterance-level feature $E$ is computed by the function $\Phi (\cdot)$, which is considered as the mathematical operation of a first-order pooling. The detail of $\Phi (\cdot)$ is described in Eq. (15).

$$\alpha_t = \frac{\exp (W_a h_t)}{\sum_{t=1}^{N} \exp (W_a h_t)}$$ (13)

$$E = \Phi (\alpha_t, h_t)$$ (14)

where $\exp (\cdot)$ is an exponential function. $W_a$ is an attention weight vector.

E. SPEECH EMOTION CLASSIFICATION

After extracting utterance-level features with a dimension of 512, we feed these utterance-level features into a fully connected layer with 512 neurons in a Bi-LSTM network to obtain higher-level affective representations. Then, the softmax classifier is used to map utterance-level features into C different space for emotion classification, where C denotes the number of emotion categories.

IV. EXPERIMENTS

To verify the performance of the proposed approach, we employ three public video emotional datasets including BAUM-1s, AFEW5.0 and CHEAVD2.0 for experiments.

A. EXPERIMENT SETTINGS

Since the data distribution of the used three datasets involved in the experiment is different, the same model may have different parameter settings. Table 1 presents the parameters settings and the corresponding computational complexity of VGGish and Bi-LSTM when obtaining the highest accuracy on each dataset. Note that these parameters are
empirically selected for its best performance. All deep network models are built by the Pytorch toolbox [69], and trained on the NVIDIA QUADRO M6000 with 24GB memory. The subject-independent Leave-One-Speakers-Group-Out (LOSGO) cross-validation strategy [70] is employed to conduct experiments. In detail, as done in Zhalehpour et al. [71], the LOSGO strategy with five groups is employed for experiments on the BAUM-1s dataset. On the AFEW5.0 and CHEAVD2.0 datasets, the original training set is used to train deep network models, and the original validation set is adopted to evaluate the performance of the proposed method.

### V. RESULTS AND ANALYSIS

The strategy of data balance is adopted to balance the data distribution on training data. TABLE 2, TABLE 3, and TABLE 4 separately present the number of data distribution before and after data balance on the used three datasets. Before data balance, the number of speech Mel-spectrogram segments is very different for each emotion. After data balance, not only the number of speech Mel-spectrogram segments is balanced, but also the number of speech segments for each emotion is balanced. The sample distribution of the Train set is anger (118), fear (77), joy (145), disgust (72), neutral (131), sadness (107), and surprise (73). The sample distribution of the Val set is anger (64), fear (46), joy (63), disgust (40), neutral (63), sadness (61), and surprise (46). However, the Test set can only be obtained by participants of emotion recognition competition. Therefore, this work conducts experiments on the Train and Val sets on the AFEW5.0 dataset.

CHEAVD2.0: CHEAVD2.0 [64] was developed for audio-visual emotion recognition in the multimodal emotion recognition challenge in 2017. This dataset includes three parts: the Train, Val and Test sets, and each part contains eight basic emotional categories, i.e., neutral, anger, joy, sadness, worry, anxiety, surprise and disgust. The sample distribution of the Train set is neutral (1400), anger (884), joy (828), sadness (462), worry (567), anxiety (457), surprise (175), and disgust (144). Similarly, the sample distribution of the Val set is neutral (200), anger (128), joy (119), sadness (462), worry (81), anxiety (66), surprise (25), and disgust (21).

#### TABLE 1. The parameter settings for training VGGish and Bi-LSTM on each dataset.

| Model       | VGGish | Bi-LSTM |
|-------------|--------|---------|
| Datasets    | BAUM-1s| AFEW5.0 | CHEAVD2.0 |
| Training time (s) | 1045  | 2232 | 10868 |
| GFOPs       | 0.87   | 0.03   | 0.01   | 0.04   |
| Model size (million parameters) | 72.14 | 0.7936 | 0.7941 | 0.7946 |
| Learning rate | 0.0001 | 0.0005 |
| Batch size  | 64     | 32     |
| Epoch       | 300    | 5000   |

#### TABLE 2. Data distribution for each emotion before and after data balance on the BAUM-1s dataset.

| Data Balance | Anger | Disgust | Fear | Joy | Sadness | Surprise |
|--------------|-------|---------|------|-----|---------|----------|
| Before       | 144   | 268     | 93   | 491 | 487     | 83       |
| After        | 314   | 522     | 326  | 405 | 345     | 309      |

#### TABLE 3. Data distribution for each emotion before and after data balance on the AFEW5.0 dataset.

| Data Balance | Anger | Disgust | Fear | Joy | Neutral | Sadness | Surprise |
|--------------|-------|---------|------|-----|---------|---------|----------|
| Before       | 246   | 180     | 141  | 276 | 286     | 248     | 146      |
| After        | 246   | 320     | 241  | 276 | 286     | 248     | 254      |

#### TABLE 4. Data distribution for each emotion before and after data balance on the CHEAVD2.0 dataset.

| Data Balance | Anger | Anxiety | Disgust | Joy | Neutral | Sadness | Surprise | Worry |
|--------------|-------|---------|---------|-----|---------|---------|----------|-------|
| Before       | 3096  | 1798    | 470     | 2597| 523     | 2433    | 468      | 2345  |
| After        | 3096  | 3352    | 1663    | 2597| 2777    | 2433    | 1580     | 2345  |
TABLE 5. Recognition accuracy (%) of different first-order attention methods before and after data balance on the BAUM-1s, AFEW5.0 and CHEAVD2.0 datasets.

| Dataset  | Data balance          | Max     | Mean   | Min     | Std     | Sum     | No-attention |
|----------|-----------------------|---------|--------|---------|---------|---------|--------------|
| BAUM-1s  | Accuracy(before)      | 45.81   | 46.95  | 45.82   | 46.14   | 44.09   | 43.14        |
|          | Accuracy(after)       | 44.19   | 48.79  | 44.93   | 46.38   | 47.27   | 45.78        |
| AFEW5.0  | Accuracy(before)      | 35.25   | 36.29  | 35.77   | 36.81   | 35.25   | 34.20        |
|          | Accuracy(after)       | 35.77   | 36.81  | 37.08   | 37.60   | 36.55   | 36.03        |
| CHEAVD2.0| Accuracy(before)      | 42.72   | 42.29  | 42.15   | 41.73   | 41.73   | 39.89        |
|          | Accuracy(after)       | 43.85   | 42.57  | 42.72   | 42.29   | 43.42   | 42.29        |

To verify the effectiveness of the proposed method integrating a unified first-order attention with data balance, we implement ablation experiments to show whether the used data balance and first-order attention methods can improve the performance of the proposed method or not. Table 5 presents the recognition results on the BAUM-1s, AFEW5.0 and CHEAVD2.0 datasets.

From the results of Table 5, we can see that: (1) The highest accuracies obtained by the proposed method are separately 48.79%, 37.60% and 43.85% on the BAUM-1s, AFEW5.0 and CHEAVD2.0 datasets. (2) As far as data balance is concerned, the recognition accuracies achieved by data balance (i.e., after data balance) are better than the obtained accuracies without data balance (i.e., before data balance). The reason may be that the balanced data distribution of each emotion helps to improve the robustness of trained deep models. This indicates that the used strategy of data balance is effective to promote the performance of trained deep models. (3) The obtained performance with a unified first-order attention is usually higher than the given results without any attention. For example, after data balance, the recognition performance with the mean attention is 48.79% on the BAUM-1s dataset, making an improvement of 3.01% over the obtained results without any attention. The reason may be that the adopted attention mechanism helps trained deep models focus on learning more discriminative features for SER. This shows that integrating the unified first-order attention mechanism with deep models is useful to improve the performance of deep models on SER tasks. (4) Among all the used first-order attention mechanisms, three attention pooling strategies including mean, std and max, separately perform best on the BAUM-1s, AFEW5.0 and CHEAVD2.0 datasets. This demonstrates that different first-order attention methods may present different performance on the same dataset. Each first-order attention method does not always perform best on the used three datasets. The reason may be that different features may need different optimal feature pooling strategies. This results in the performance diversity of different first-order attention methods on SER tasks.

Additionally, the metric of precision, recall and F1-score are also computed on these three datasets, so as to further measure the performance of the proposed method. These results are given in Tables 6, 7, and 8, respectively.

From Tables 6-8, it can be seen that “anger”, “joy” and “sadness” are easier to be distinguished than other emotions on these three datasets. However, “surprise” is the most difficult to be identified, since it is prone to be confused with other emotions. Besides, on the CHEAVD2.0 dataset, “disgust” can’t be recognized well. This may be attributed to the fact that there are less samples for this emotion.
In order to further observe the performance of the proposed method for each emotion, we also present the classification confusion matrix, as shown in Fig. 6-8. These results correspond to the BAUM-1s, AFEW5.0 and CHEAVD2.0 datasets, respectively. From these results in Fig. 6-8, the recognition performance of certain emotions on different datasets is generally similar. However, there are also some special cases. For instance, the recognition accuracies of “anger” and “joy” on different datasets vary greatly. This is because the data distribution of each dataset is different.

To evaluate the superiority of the proposed method, Table 9 presents the performance comparisons of different methods on the BAUM-1s, AFEW5.0 and CHEAVD2.0 datasets. These comparing works employ the same settings as ours. In particular, subject-independent test-runs are used to report the results.

From the results in Table 9, we can see that the proposed method outperforms hand-crafted features and some deep features on these three datasets. This shows the advantage of our proposed method over other methods. In particular, the best performance of our proposed method is 48.79% for BAUM-1s, 37.60% for AFEW5.0, 43.85% for CHEAVD2.0, respectively. Compared with hand-crafted features in [64], [71], [73], it is obvious that the performance of our method is much better. This indicates the learned deep features are superior to hand-crafted features. In addition, in comparison with other CNN-based methods [55], [74], [75], our method also gives better performance. Note that [75] employed principal component analysis (PCA) and linear discriminant analysis (LDA) to reduce the feature dimensionality of the extracted CNN features. Likewise, [78] also adopted PCA to compress the extracted audio features. This shows the effectiveness of the used attention mechanism and data balance in our method. Besides, our method obtains a little lower performance than [76] on the CHEAVD2.0 dataset.

| Table 9. Recognition performance (%) comparisons with state-of-the-arts on three used datasets. |
|--------------------------------------------------------------------------------------------------|
| **Datasets** | **Ref.** | **Methods** | **Accuracy** |
|-------------------------------------------------|---------|-------------|--------------|
| BAUM-1s | Zalepshis et al., 2014 [71] | MFCC, RASTA-PLP | 79.41 |
| | Zhang et al., 2017 [55] | CNN + SVM | 42.26 |
| | Ours | CNN + Bi-LSTM + Attention | 48.79 |
| AFEW5.0 | Kaya et al., 2017 [77] | INTERSPEECH 2010 + ELM | 36.59 |
| | Cai et al., 2019 [73] | INTERSPEECH 2010 + PCA + SVM | 35.51 |
| | Zhang et al., 2020 [78] | CNN + SVM | 36.29 |
| | Ours | CNN + Bi-LSTM + Attention | 37.60 |
| CHEAVD2.0 | Li et al., 2018 [64] | eGEMAPS | 39.90 |
| | Xi et al., 2019 [76] | Residual Adapter Model | 43.96 |
| | Ours | CNN + Bi-LSTM + Attention | 43.85 |

Note that Xi et al. [76] proposed a residual adapter model to recognize speech emotions on the CHEAVD2.0 dataset, and gives an accuracy of 43.96%, whereas our method presented an accuracy of 43.85%. Nevertheless, the used speaker adaption method in [76] is complicated due to its computation complexity. In addition, due to the used spontaneous emotional datasets, the reported performance on these three
datasets is relatively still low. This is because spontaneous emotional datasets are much more difficult to identify than acted emotional datasets, as shown in [20].

VI. CONCLUSION AND FUTURE WORK

In this study, a new SER method, which integrates a unified first-order attention network with data balance, is proposed. The strategy of data balance aims to keep the number of data samples for each category as close as possible, thereby improving the generalization ability of trained deep models. The unified first-order attention mechanism is embedded into the output of Bi-LSTM so as to effectively capture high-level segment-level discriminative features for speech emotion classification. The process of the proposed method is summarized as follows. Initially, speech preprocessing related to data balance is performed on the Mel-spectrogram in the datasets. Next, we use the pre-trained VGGish network model to perform segment-level feature learning from the extracted Mel-spectrogram segments. Then, we leverage a unified attention-based feature pooling strategy to produce utterance-level features. Finally, the utterance-level features are fed into the softmax layer in the Bi-LSTM network for emotion identification. Experimental results on three challenging emotional datasets, i.e. BAUM-1s, AFEW5.0 and CHEAVD2.0, indicate that the proposed method obtains competitive performance to state-of-the-art methods. This demonstrates the advantages of the proposed method. In comparison with other methods, our method has two characteristics. On one hand, before training deep models it employs a data balance strategy to alleviate the issue of an unbalanced data distribution of emotional samples. On the other hand, it designs a unified first-order attention network for concentrating on discriminative feature learning so as to consider the difference of representation capability on different speech fragment areas.

In future, we will investigate the performance of more advanced deep network models on utterance-level feature learning for SER. It is interesting that the proposed method can be extended to develop a real-time SER system, when considering the imbalanced characteristics of speech emotion data. In addition, it is also interesting to employ feature selection methods to show that which features are selected for its importance from all the learned features. This can be used to improve performance further.

REFERENCES

[1] R. A. Calvo and S. D’Mello, “Affect detection: An interdisciplinary review of models, methods, and their applications,” IEEE Trans. Affect. Comput., vol. 1, no. 1, pp. 18–37, Jan. 2010.
[2] 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, Feb. 2015.
[3] B. L. Stojkoska and K. V. Trivodaliev, “A review of Internet of Things for Smart home: Challenges and solutions,” J. Cleaner Prod., vol. 140, pp. 1454–1464, Jan. 2017.
[4] S. Tuli, N. Basumaty, S. S. Gill, M. Kahani, R. C. Arya, G. S. Wander, and R. Buyya, “HealthFog: An ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments,” Future Gener. Comput. Syst., vol. 104, pp. 187–200, Mar. 2020.
[5] P.-C. Yang, K. Sasaki, K. Suzuki, K. Kase, S. Sugano, and T. Ogata, “Repeatable folding task by humanoid robot worker using deep learning,” IEEE Robot. Autom. Lett., vol. 2, no. 2, pp. 397–403, Apr. 2017.
[6] H. H. O. Nasereddin and A. A. R. Omari, “Classification techniques for automatic speech recognition (ASR) algorithms used with real time speech translation,” in Proc. Comput. Conf., London, U.K., Jul. 2017, pp. 200–207.
[7] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, “Speech emotion recognition using both spectral and prosodic features,” in Proc. Int. Conf. Inf. Eng. Comput. Sci., Dec. 2009, pp. 1–4.
[8] I. Idris and M. S. H. Salam, “Voice quality features for speech emotion recognition,” J. Inf. Assurance Secur., vol. 10, no. 4, pp. 183–191, 2015.
[9] H. Hu, M.-X. Xu, and W. Wu, “GMM supervisor based SVM with spectral features for speech emotion recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Honolulu, HI, USA, Apr. 2007, pp. IV-413–IV-416.
[10] S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” Speech Commun., vol. 53, no. 5, pp. 768–785, May 2011.
[11] T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.
[12] D. Neiberg, K. Elenius, and K. Kask, “Emotion classification in spontaneous speech using GMMs,” in Proc. 9th Int. Conf. Spoken Lang. Process., Pittsburgh, PA, USA, 2006, pp. 1–4.
[13] M. L. Seltzer, B. Raj, and R. M. Stern, “A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition,” Speech Commun., vol. 43, no. 1, pp. 279–303, Sep. 2004.
[14] D. Morrison, R. Wang, and L. C. De Silva, “Ensemble methods for spoken emotion recognition in call-centres,” Speech Commun., vol. 49, no. 2, pp. 98–112, Feb. 2007.
[15] B. Schuller, G. Rigoll, and M. Lang, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Montreal, QC, Canada, May 2004, pp. 1–577.
[16] J. K. Suykens and J. Vandewalle, “Least squares support vector machine classifiers,” Neural Process. Lett., vol. 9, no. 3, pp. 293–300, Jan. 1999.
[17] G. E. Hinton, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
[18] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, “Speech emotion classification using attention-based LSTM,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 27, no. 11, pp. 1675–1685, Nov. 2019.
[19] J. Lee and I. Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,” in Proc. 16th Annu. Conf. Int. Speech Commun. Assoc., Dresden, Germany, 2015, pp. 1–4.
[20] S. Zhang, S. Zhang, T. Huang, and W. Gao, “Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching,” IEEE Trans. Multimedia, vol. 20, no. 6, pp. 1576–1590, Jun. 2018.
[21] J. Zhao, X. Mao, and L. Chen, “Speech emotion recognition using deep 1D & 2D CNN LSTM networks,” Biomed. Signal Process. Control, vol. 47, pp. 312–323, Jan. 2019.
[22] Y. LeCun and Y. Bengio, “Convolutional networks for images, speech, and time series,” in The Handbook of Brain Theory and Neural Networks, vol. 3361, no. 10, pp. 9, 1995.
[23] Y. Huang, K. Tian, A. Wu, and G. Zhang, “Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition,” J. Ambient Intell. Humanized Comput., vol. 10, no. 5, pp. 1787–1798, May 2019.
[24] A.-R. Mohamed, G. E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 1, pp. 14–22, Jan. 2012.
[25] S. Masalmadi, E. Bazdast, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), New Orleans, LA, USA, Mar. 2017, pp. 2227–2231.
[26] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[27] M. Sundermeyer, R. Schlüter, and H. Ney, “LSTM neural networks for language modeling,” in Proc. 13th Annu. Conf. Int. Speech Commun. Assoc., Portland, OR, USA, 2012, pp. 194–197.
VOLUME 8, 2020

H. Xu and K. Saenko, “Ask, attend and answer: Exploring question-based speech emotion mechanisms,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Brighton, U.K., May 2019, pp. 2822–2826.

X. Wang, Z. Cai, D. Gao, and N. Vasconcelos, “Towards universal object detection by domain attention,” in Proc. IEEE/CFV Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 7299–7308.

W. Sun, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, “Salient object detection with pyramid attention and salient edges,” in Proc. IEEE/CFV Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 1448–1457.

A. Ando, S. Kobashikawa, H. Kamiyama, R. Masumura, Y. Iijima, and Y. Aono, “Soft-target training with ambiguous emotional utterances for DNN-based speech emotion classification,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Calgary, AB, Canada, Apr. 2018, pp. 4964–4968.

D. Yu, J. Fu, T. Mei, and Y. Rui, “Multi-level attention networks for visual question answering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 4709–4717.

H. Xu and K. Saenko, “Ask, attend and answer: Exploring question-guided spatial attention for visual question answering,” in Proc. Eur. Conf. Comput. Vis., Amsterdam, The Netherlands, 2016, pp. 451–466.

M.-T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” 2015, arXiv:1508.04025. [Online]. Available: https://arxiv.org/abs/1508.04025.

K. Chen, R. Wang, M. Utiiyama, E. Sumita, and T. Zhao, “Syntactically directed attention for neural machine translation,” in Proc. 32nd AAAI Conf. Artif. Intell., New Orleans, LA, USA, 2018.

F. Ma, W. Gu, W. Zhang, S. Ni, S.-L. Huang, and L. Zhang, “Speech emotion recognition via attention-based DNN from multi-task learning,” in Proc. 16th ACM Conf. Embedded Networked Sensor Syst., Shenzhen, China, Nov. 2018, pp. 363–364.

M. Chen, X. He, D. Yang, and H. Zhang, “3-D convolutional recurrent neural networks with attention mechanism for speech emotion recognition,” IEEE Signal Process. Lett., vol. 25, no. 10, pp. 1440–1444, Oct. 2018.

C.-W. Huang and S. S. Narayanan, “Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Hong Kong, Jul. 2017, pp. 583–588.

A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales, “Data augmentation for low resource languages,” in Proc. 15th Annu. Conf. Int. Speech Commn. Assoc. (INTERSPEECH), Singapore, 2014, pp. 810–814.

M. J. F. Gales, A. Ragni, H. AllDarmaki, and C. Gautier, “Support vector machines for noise robust ASR,” in Proc. IEEE Workshop Acoust. Speech Recognit. Understand., Merano, Italy, Dec. 2009, pp. 205–210.

A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, “Deep speech: Scaling up end-to-end speech recognition,” 2014, arXiv:1412.5567. [Online]. Available: http://arxiv.org/abs/1412.5567.

R. Takahashi, T. Matsubara, and K. Uehara, “Data augmentation using random image cropping and patching for deep CNNs,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 9, pp. 2917–2931, Sep. 2020.

S. Hershys, S. Chaudhuri, S. Dey, and K. Jung, “Speech emotion recognition using multi-hop attention mechanism,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Brighton, U.K., May 2019, pp. 1–5.

F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Lautka, S. S. Narayanan, and K. P. Truong, “The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” IEEE Trans. Affect. Comput., vol. 7, no. 2, pp. 190–202, Apr. 2016.

F. Eyben, M. Schuller, and B. Schuller, “OpenEAR-introducing the Munich open-source emotion and affect recognition toolkit,” in Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops, Amsterdam, The Netherlands, 2009, pp. 1–6.

B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, and E. Marchi, “The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism,” in Proc. 14th Annu. Conf. Int. Speech Commn. Assoc. (INTERSPEECH), Lyon, France, 2013, pp. 1–5.
[71] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, “BAUM-1: A spontaneous audio-visual face database of affective and mental states,” IEEE Trans. Affect. Comput., vol. 8, no. 3, pp. 300–313, Jul. 2017.

[72] A. Dhall, O. R. Murthy, R. Goecke, J. Joshi, and T. Gedeon, “Video and image based emotion recognition challenges in the wild: Emotiw 2015,” in Proc. ACM Int. Conf. Multimodal Interact., 2015, pp. 423–426.

[73] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O’Reilly, S. Han, P. Liu, M. Chen, and Y. Tong, “Feature-level and model-level audiovisual fusion for emotion recognition in the wild,” in Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR), San Jose, CA, USA, Mar. 2019, pp. 443–448.

[74] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (AVEF): A deep efficient weighted approach," Inf. Fusion, vol. 46, pp. 184–192, Mar. 2019.

[75] J. Y. R. Cornejo and H. Pedrini, “Audio-visual emotion recognition using a hybrid deep convolutional neural network based on census transform,” in Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC), Oct. 2019, pp. 3396–3402.

[76] Y. Xi, P. Li, Y. Song, Y. Jiang, and L. Dai, “Speaker to emotion: Domain adaptation for speech emotion recognition with residual adapters,” in Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC), Lanzhou, China, Nov. 2019, pp. 513–518.

[77] H. Kaya, F. Gürpınar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," Image Vis. Comput., vol. 65, pp. 66–75, Sep. 2017.

[78] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (AVEF): A deep efficient weighted approach," Inf. Fusion, vol. 46, pp. 184–192, Mar. 2019.

[79] J. Y. R. Cornejo and H. Pedrini, “Audio-visual emotion recognition using a hybrid deep convolutional neural network based on census transform,” in Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC), Oct. 2019, pp. 3396–3402.

SHIQING ZHANG received the Ph.D. degree from the School of Communication and Information Engineering, University of Electronic Science and Technology of China, in 2012. He was a Post-Doctoral Researcher with the School of Electronic Engineering and Computer Science, Peking University, Beijing, China, from 2015 to 2017. He is currently a Professor with the Institute of Intelligent Information Processing, Taizhou University, China. His research interests include affective computing and pattern recognition. He has published over 40 articles in journals, such as IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He is an Associate Editor of IEEE ACCESS.

XIN TAO received the bachelor’s degree in computer science and technology and the Ph.D. degree in technology of computer application from Peking University, China, in 2008 and 2014, respectively. He is currently a Lecturer with the Institute of Intelligent Information Processing, Taizhou University, China. His research interests include affective computing and digital document analysis.

GANG CHEN received the bachelor’s degree from the School of Electronics and Information Engineering, Hunan Institute of Technology of China, in 2018. He is currently pursuing the master’s degree with the School of Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou, China. His research interests include affective computing and pattern recognition.

XIAOMING ZHAO received the B.S. degree in mathematics from Zhejiang Normal University in 1990 and the M.S. degree in software engineering from Beihang University in 2006. He is currently a Professor with the Institute of Intelligent Information Processing, Taizhou University, China. His research interests include audio and image processing, machine learning, and pattern recognition.

* * *