IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 9, 2020, accepted January 21, 2020, date of publication January 23, 2020, date of current version February 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969032

Learning Deep Binaural Representations With Deep Convolutional Neural Networks for Spontaneous Speech Emotion Recognition

SHIQING ZHANG[1], AIHUA CHEN[1], WENPING GUO[1], YUELI CUI[1], XIAOMING ZHAO[1], AND LIMEI LIU[2]

[1]Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, China
[2]Institute of Big Data and Internet Innovation, Hunan University of Commerce, Changsha 410205, 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 (NSFC) under Grant LZ20F020002 and Grant 61976149, in part by the Major Project for National Natural Science Foundation of China under Grant 71790615, in part by the Taizhou Science and Technology Project under Grant 1803gy08 and Grant 1802gy06, and in part by the Outstanding Youth Project of Taizhou University under Grant 2018JQ003 and Grant 2017PY026.

ABSTRACT Spontaneous speech emotion recognition is a new and challenging research topic. In this paper, we propose a new method of spontaneous speech emotion recognition on the basis of binaural representations and deep convolutional neural networks (CNNs). The proposed method initially employs multiple CNNs to learn deep segment-level binaural representations such as Left-Right and Mid-Side pairs from the extracted image-like Mel-spectrograms. These CNNs are fine-tuned on target emotional speech datasets from a pre-trained image CNN model. Then, a new feature pooling strategy, called block-based temporal feature pooling, is proposed to aggregate the learned segment-level features for producing fixed-length utterance-level features. Based on the utterance-level features, linear support vector machines (SVM) is adopted for emotion classification. Finally, a two-stage score-level fusion strategy is used to integrate the obtained results from Left-Right and Mid-Side pairs. Extensive experiments on two challenging spontaneous emotional speech datasets, including the AFEW5.0 and BAUM-1s databases, demonstrate the effectiveness of our proposed method.

INDEX TERMS Spontaneous speech emotion recognition, binaural representations, deep convolutional neural networks, temporal feature pooling.

I. INTRODUCTION

Speech signals are one of the most natural ways of human emotion expression. Speech emotion recognition (SER) has become an important and challenging task in the fields of signal processing, artificial intelligence, pattern recognition, *etc*, because of its potential applications to human-computer interaction [1].

Most prior works [2]–[4] in the past several decades focus on SER tasks with collected data in laboratory controlled environment, such as the popular acted EMO-DB [5] dataset developed to distinguish acted emotions. Although acted emotions are usually classified with good performance, they are easily exaggerated. Therefore, acted emotions fail to faithfully represent the characteristics of human emotion expression in real sceneries. In recent years, emotion recognition in the wild has drawn increasingly extensive attention, because such spontaneous emotions in the wild are more challenging and difficult to classify in comparison with conventional acted emotions.

Audio feature extraction is a key step in a fundamental SER system. It aims to extract effective feature representations characterizing human emotion expression. The early-used typical audio features [6]–[10] for SER are low-level descriptors (LLDs), such as prosody features like pitch and intensity, voice quality features like formants, and spectral features like Mel-frequency cepstral coefficients (MFCCs). In recent years, several extensive features with thousands of LLDs, including the INTERSPEECH 2010 [11], ComParE [12], AVEC-2013 [13], and GeMAPS [14] feature sets,

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

have been widely used for SER. However, these extracted LLDs are low-level hand-crafted features, so that they are not very effective to represent emotional characteristics of speech [4]–[6]. Therefore, it is needed to develop automatic feature learning methods for extracting high-level affective features that effectively characterize speakers' emotions.

To tackle the abovementioned problem, deep learning methods [15], [16] may provide an alternative solution. To date, two popular deep learning algorithms, *i.e.*, deep neural networks (DNNs) [15], and deep convolutional neural networks (CNNs) [17], have been employed to learn high-level feature representations for SER. When using DNNs for SER, hand-crafted acoustic features are often utilized. In [18], the authors used a DNN to learn high-level features from hand-crafted spectral features like MFCCs, followed by an extreme learning machine (ELM) for SER. In [19], MFCCs was used as inputs of DNNs, and then a hybrid of a cascaded Gaussian mixture model and deep neural network (GMM-DNN) was developed for SER. In [20], several perception features such as perceptual linear predictive cepstrum (PLPC), revised perceptual linear prediction coefficients (RPLPs) and inverted Mel-frequency cepstral coefficients (IMFCCs), were employed as inputs of DNNs for SER.

When using CNNs for SER, the raw waveforms or spectrograms are usually adopted as inputs of CNNs to learn high-level feature representations. In [21], the authors divided the raw waveform to fixed-length segments as inputs of a 2-layer CNN, and then conducted temporal 1D convolution, followed by a 1-layer long short-term memory (LSTM) [22] for modeling long-range dependencies. In [23], a 2-layer CNN combined with 2-layer LSTM was used for SER on the basis of the divided segments from the raw waveforms. In [24], the authors employed the spectrograms as inputs of a sparse auto-encoder, followed by a 1-layer CNN, to learn salient features for SER. Due to the limited emotional data, these works adopt shallow CNNs containing 1 or 2 convolutional layers to learn high-level features for SER.

Recently, a variety of deep CNNs like AlexNet [17], VGG [25], and ResNet [26], have been widely utilized to conduct various object detection and classification tasks. Moreover, these deep CNNs usually perform better than shallow CNNs. This is because deep CNNs adopt deep multi-level convolutional and pooling layers to capture mid-level feature representations from input image data. To date, exploring deep spectrum feature extraction with deep CNNs has become a new research trend in SER. In [27], [28], the authors leveraged attention-based bidirectional LSTM with fully convolutional networks to learn deep spectrum features for SER. In our recent work [29], we designed an image-like spectrogram as inputs of deep CNNs like AlexNet to learn high-level segment-level feature representations for SER. Such learned deep spectrum features benefit from the advantages of cross-media transfer learning, since they are developed by fine-tuning pre-trained deep CNNs on image classification tasks.

Although these recent works [18]–[29] employed deep learning techniques such as DNNs or CNNs, and achieved good performance on SER tasks, they just concentrated on monaural audio signals and their related features. In particular, they usually made the record audios in stereo in real sceneries monaural by keeping its left channel while discarding its right channel prior to processing. As a result, these works [18]–[29] using monaural audio signals for SER ignore the advantages of binaural representations in stereo such as Left-Right or Mid-Side pairs. The Mid channel is calculated as $L + R$ and the Side channel is $L - R$. Fig. 1 shows the difference of extracted Mel-spectrograms of binaural representations from the same emotional speech sample. From Fig. 1, we can see that these binaural representations, *i.e.*, the Left-Right or Mid-Side pairs, contain much richer temporal-spatial information than monaural representations. In particular, Left Mel-spectrograms show different features in comparison with Right Mel-spectrograms. For Mid-Side pairs, they are similar due to the calculation with $L + R$ and $L - R$. Nevertheless, both of them are related to speech emotion expression, and may do good to identify speech emotions. This indicates that binaural representations, *i.e.*, the Left-Right or Mid-Side pairs, may provide complementary information for emotion identification. In addition, recent works [30]–[32] show that binaural representations performs better than monaural representations for speech segregation or speaker recognition. Therefore, learning deep binaural representations may present better performance than commonly-used monaural representations on SER tasks.

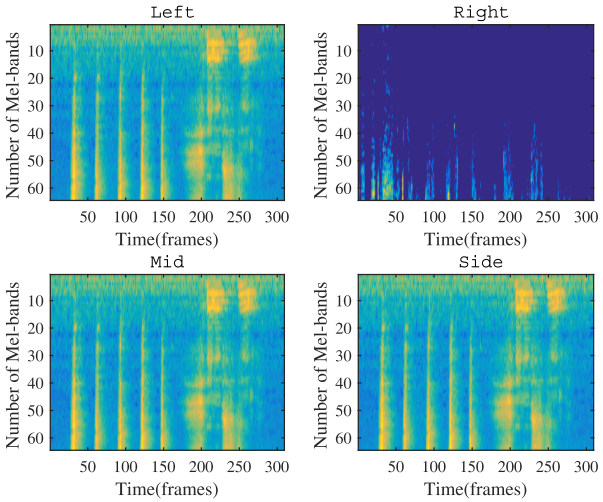

FIGURE 1. Different Mel-spectrograms of binaural representations.

Inspired by the abovementioned advantages of binaural representations, this paper proposes a new CNNs-based SER method with binaural representations. Initially, for the Left-Right or Mid-Side pairs, we use multiple deep CNNs to separately learn high-level segment-level features from the extracted image-like Mel-spectrograms. We fine-tune each CNN on target emotional speech datasets from a pre-trained image CNN model. Then, a new feature pooling strategy, called block-based temporal feature pooling, is proposed to produce fixed-length utterance-level features from the learned segment-level features, followed by linear support

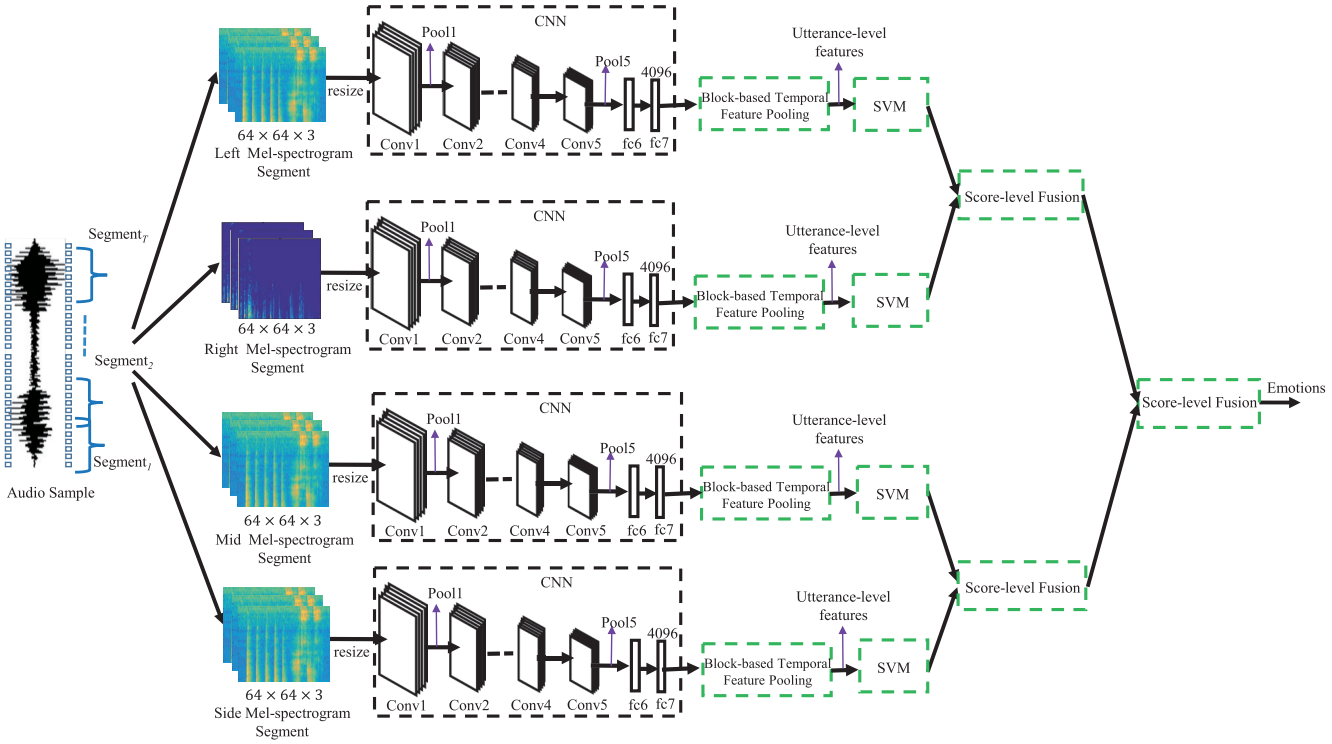

FIGURE 2. The flowchart of our proposed learning deep binaural representations with DCNNs for speech emotion recognition.

vector machines (SVM) for emotion classification. Finally, we implement two-stage score-level fusion to integrate the obtained results from the Left-Right and Mid-Side pairs. Two challenging spontaneous emotional speech datasets, including the AFEW5.0 [33] and BAUM-1s [34] databases, are employed to conduct SER experiments. Experiment results demonstrate the effectiveness of our proposed method on SER tasks.

The major contributions of this paper can be summarized as follows:

(1) Considering the rich temporal-spatial information of binaural representations, we propose a new SER method based on CNNs and binaural representations. To the best of our knowledge, we are the first to consider and employ binaural representations for SER.

(2) To form fixed-length utterance-level features, we propose a new feature pooling method, *i.e.,* block-based temporal feature pooling. It takes the temporal clues of an utterance into account in the process of aggregating the learned segment-level features. Extensive experiments on two spontaneous datasets demonstrate that our method outperforms state-of-the-arts.

The remainder of this paper is organized as follows. Section 2 describes the details of our proposed method. Experiment results and analysis are presented in Section 3. Section 4 provides the conclusions and future work.

II. OUR PROPOSED METHOD

Fig. 2 provides the flowchart of our proposed learning deep binaural representations with CNNs for speech emotion recognition. Our method contains four steps: (1) generation of audio CNN inputs, (2) learning segment-level features with deep CNNs, (3) block-based temporal feature pooling, (4) two-stage score-level fusion. In the followings, we describe the abovementioned four steps of our method in details.

A. GENERATION OF AUDIO CNN INPUTS

To employ the pre-trained image CNNs (AlexNet) [17] for cross-media transfer learning, from the original 1D audio sample we create three channels of Mel-spectrogram segments similar to the color RGB image. The created Mel-spectrogram segments are then transformed into suitable inputs of AlexNet with a fixed input size of $227 \times 227 \times 3$.

Following in [29], we adopt 64 Mel-filter banks spanning from 20 Hz to 8000 Hz to produce the whole log Mel-spectrogram by using a Hamming window size of 25 ms with an overlap of 10 ms. Then, we use a context window of 64 frames to divide the whole log Mel-spectrogram into fixed-length segments with a size of 64×64. Finally, on the basis of the static segment of 64×64, we compute its first order and second order regression coefficients along the time axis, thereby producing the delta and delta-delta coefficients of the static segment. Consequently, similar to the color RGB image, three channels (static, delta, and delta-delta) of Mel-spectrogram segments with a size of $64 \times 64 \times 3$ can be created as inputs of audio CNNs. Subsequently, we resize $64 \times 64 \times 3$ with bilinear interpolation into the fixed size of $227 \times 227 \times 3$ as inputs of AlexNet. In this

way, we can generate individually the Left, Right, Mid, and Side Mel-spectrogram segments for learning deep binaural representations, as depicted in Fig. 2.

B. LEARNING SEGMENT-LEVEL FEATURES WITH DEEP CNNs

As described in Fig. 2, the used deep CNNs are the same as the original AlexNet [17]. It includes five convolution layers (Conv1, Conv2, Conv3, Conv4, and Conv5), three max-pooling layers (Pool1, Pool2, and Pool5), and three fully connected (FC) layers (fc6, fc7, and fc8). The fc6 and fc7 have 4096 neurons, whereas fc8 denotes a label vector corresponding to data categories. It is noted that fc8 in AlexNet equals to 1000 image categories on the ImageNet data. We do not employ other CNN models with deeper structure than AlexNet, because there is no any improvement with them due to the limited emotional datasets.

To deeply learn segment-level features, we adopt a fine-tuning strategy for cross-media transfer learning. In computer vision [35], [36] have proved that it is feasible to fine-tune the pre-trained CNNs on target data to relieve the problem of data insufficiency. To achieve segment-level feature learning with CNNs, we use target emotional speech data to fine-tune the AlexNet [17] model pre-trained on the large-scale ImageNet data. Specially, we initialize the used CNN network by means of copying the network parameters from pre-trained AlexNet. Next, the fc8 layer in AlexNet is replaced with a new class label vector corresponding to speech emotion categories used in our experiments. Finally, we retrain the CNN models by using the standard back propagation strategy. The following minimizing problem is solved to update the CNN network parameters:

$$\min_{W,\vartheta} \sum_{i=1}^{N} H\left(\text{softmax}\left(W \cdot \Upsilon\left(a_i;\vartheta\right)\right), y_i\right), \tag{1}$$

where W represents the weight values of the softmax layer for the network parameters ϑ. $\gamma(v_i;\vartheta)$ denotes the 4096-D output of the fc7 layer for input data a_i, and y_i represents the class label vector of the i-th segment. H denotes the softmax log-loss function:

$$H(\vartheta, y) = -\sum_{j=1}^{C} y_j \log\left(y_j\right), \tag{2}$$

where C denotes the total number of speech emotion categories. After fine-tuning the CNN models, the 4096-D outputs of their fc7 layers are the learned deep segment-level feature representations in audio Mel-spectrogram segments.

It is pointed out that we split the whole Mel-spectrogram from an audio sample into a certain number of Mel-spectrogram segments to conduct segment-level feature learning with CNNs. In this situation, we set the emotion category of each Mel-spectrogram segment to be the utterance-level emotion category.

C. BLOCK-BASED TEMPORAL FEATURE POOLING

Once the CNN training is finished, the 4096-D outputs of their fc7 layers represent the learned segment-level features. Because the duration time for utterances is unfixed, the number of divided segments in utterances varies. Therefore, it is needed to aggregate segment-level features in an utterance into utterance-level features with fixed dimensionality. This process is also called feature pooling.

So far, there are two popular feature pooling methods, *i.e.*, average-pooling and max-pooling, which aim to calculate the average and maximum values for local feature maps to produce global feature representations, respectively. For instance, in computer vision they are used to aggregate frame-level features on videos into video-level features [37]. Likewise, they can be also employed to perform average and maximum operation for all divided segment-level features in an utterance to form fixed-length utterance-level features. However, owing to the useful temporal clues of an utterance for emotion expression, these two feature pooling methods discard the temporal information in an utterance for speech emotion recognition. To make use of the temporal clues of an utterance, we propose a simple yet effective feature pooling strategy, *i.e.*, block-based temporal feature pooling. Here, we take max-pooling as an example, Fig. 3 shows the flowchart of the proposed feature pooling method.

As depicted in Fig. 3, the proposed feature pooling method contains three steps:

(1) For the obtained CNN-learned segment-level features X, we calculate its first-order regression coefficients along the time axis, *i.e.*, *delta_X*. This aims to exhibit the temporal dynamic information among adjacent segments in an utterance.

(2) X and *delta_X* are equally divided into successive non-overlapping sub-blocks, respectively, and then a feature pooling operation like max-pooling is implemented on each divided sub-block. Here, two sub-blocks associated with max-pooling are taken as an example.

(3) The achieved sub-blocks based features are concatenated and produce global utterance-level features $f^m(X)$. This can be expressed as

$$f^m(X) = \left(f_l^m(X), f_l^m(delta\_X)\right), \tag{3}$$

where $f_l^m(X)$ and $f_l^m(delta\_X)$ are max-pooling operations performed on l successive non-overlapping sub-blocks from X and $(delta\_X)$, respectively.

Both X and $(delta\_X)$ are equally split into l successive sub-blocks along the time axis at multiple levels with $l = 1, 2, 3, \ldots, L$. This can be expressed as

$$X = (X_1, X_2, \cdots, X_L), \tag{4}$$

$$delta\_X = (delta\_X_1, delta\_X_2, \cdots, delta\_X_L), \tag{5}$$

Given a sub-block $X_l = (x_1, x_2, \cdots, x_n) \in R^{d\times n}$ with n segments associated with d dimensionality, max-pooling on

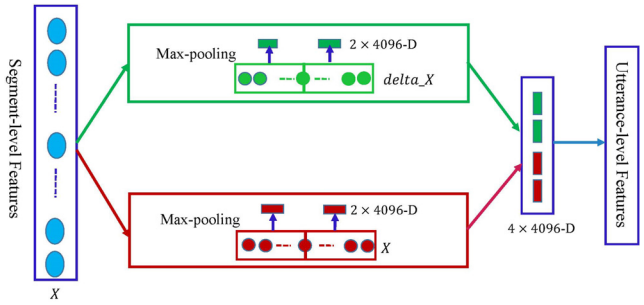

FIGURE 3. The flowchart of block-based temporal feature pooling. Here, two sub-blocks associated with max-pooling are taken as an example.

this sub-block is calculated as

$$f_l^m(X_l) = \sum_{j=1}^{n} \max \left| x_j \right|, \tag{6}$$

For a sub-block $delta\_X_l = \left(x_1', x_2', \cdots, x_n'\right) \in R^{d \times n}$, we also compute max-pooling by

$$f_l^m(delta\_X_l) = \sum_{j=1}^{n} \max \left| x_j' \right|, \tag{7}$$

Our block-based temporal feature pooling strategy considers the useful temporal clues for emotion expression in two aspects. First, the first-order regression of segment-level features along the time axis, *i.e.*, *delta_X*, is used to reveal the temporal dynamic emotional information among adjacent segments in an utterance. Second, the divided non-overlapping sub-blocks from *X* and *delta_X* along the time axis also contain certain temporal emotional information to some extent.

When we obtain fixed-length utterance-level features with the proposed block-based temporal feature pooling strategy, the linear SVM is employed to conduct the final emotion classification. In detail, we adopt the obtained utterance-level features from training samples to train the linear SVM, and then evaluate the trained SVM model on each testing sample with utterance-level features, producing final emotion classification results. Similarily, in Fig. 3 average-pooling can be used instead of max-pooling.

It is noted that for the final emotion classification we do not employ other typical emotion classifiers such as Random Forest, K-Nearest Neighbor, *etc.*, because SVM usually performs better than them, as demonstrated in [38].

D. TWO-STAGE SCORE-LEVEL FUSION

Considering the complementarity between binaural representations, such as Left-Right Mel-spectrogram, and MS Mel-spectrogram, we adopt a two-stage score-level fusion strategy to implement final emotion classification. In particular, we initially separately combine the emotion recognition results from Left-Right, and Mid-Side pipelines at score-level. Then, we implement score-level fusion further on the basis of the first-stage obtained results. This can be expressed as

$$score_1^{fusion} = \alpha * score^{left} + (1-\alpha) * score^{right}, \tag{8}$$

$$score_2^{fusion} = \beta * score^{mid} + (1-\beta) * score^{side}, \tag{9}$$

$$score_{final}^{fusion} = \lambda * score_1^{fusion} + (1-\lambda) * score_2^{fusion}, \tag{10}$$

where α, β and λ are the weight values corresponding to the emotion classification score values. For simplicity, α, β and λ are decided by an exhaustive search in a range of [0,1] with an interval of 0.1 in this work. The optimal weight values correspond to the best performance of score-level fusion.

III. EXPERIMENT STUDIES

To evaluate the performance of the proposed method on spontaneous SER tasks, we conduct experiments on two challenging spontaneous emotional speech datasets, *i.e.*, the AFEW5.0 [33] and BAUM-1s [34] databases. It is noted that in this work we concentrates on spontaneous SER rather than the conventional acted SER, because spontaneous emotions in the wild are more challenging and difficult in comparison with acted emotions. Hence, we just employ the abovementioned spontaneous emotional speech datasets for experiments.

A. DATASETS

AFEW5.0: AFEW5.0 [33] was developed for audiovisual emotion recognition in the wild challenge in 2015. It is a spontaneous audiovisual video dataset, and comprises of seven emotional categories,*i.e.*, anger, joy, sadness, disgust, surprise, fear, and neutral. Three annotators are invited to annotate these emotions. This dataset includes three parts: the Train (723 samples), Val (383 samples), and Test (539 samples) sets. This work adopts the Train and Val sets for experiments, because the Test set is only open for the participants in emotion recognition competitions.

BAUM-1s: BAUM-1s [34] is a recently-developed spontaneous audio-visual emotional dataset. It comprises of not only six basic emotional categories, *i.e.*, anger, joy, sadness, disgust, fear, surprise, but also other mental states, such as unsure, thinking, concentrating, and bothered. Following in [34], this work focus on identifying six basic emotions, thereby producing 521 video samples in total from 31 Turkish subjects.

B. SETTINGS

For CNN training, the mini-batch size of input data is 30. The learning rate is 0.001 and the maximum of epoch number is 300. An NVIDIA GTX TITAN X GPU with 12GB memory is used to accelerate the CNN's training. We employ the MatConvNet [39] software to perform deep CNNs. We conduct SER experiments by using a subject-independent cross-validation strategy, which is most utilized in real sceneries. Specially, on the AFEW5.0 database, the Train and Val sets divided by the data developers are used for experiments. On the BAUM-1s database with 31 Turkish subjects, we adopt a leave-one-subject-group-out (LOSGO)

TABLE 1. Recognition accuracy (%) of different monaural representations with block-based temporal Max-pooling.

MONAURAL REPRESENTATIONS	AFEW5.0	BAUM-1s
Left	32.64	43.37
Right	32.90	33.77
Mid	32.89	40.16
Side	25.32	38.72

strategy with five subject groups for experiments. And the average recognition accuracies in five test-runs are obtained to evaluate the performance of all used methods. For block-based temporal feature pooling, $l = 2$ is used for its good performance and computation efficiency.

We divide an audio sample into a certain number of audio segments as inputs of CNNs, thereby augmenting the amount of training data to some extent. In particular, on the AFEW5.0 database we produce 5,141 segments from the Train set (723 samples), and 2,484 segments from the Val set (383 samples), respectively. On the BAUM-1s database, we produce 6,386 segments from 521 video samples.

C. RESULTS AND ANALYSIS

1) EFFECTS OF DIFFERENT MONAURAL REPRESENTATIONS

We initially evaluate the performance of different monaural representations. Table 1 gives the recognition results of four monaural representations such as Left-Right or Mid-Side pairs with block-based temporal max-pooling.

As shown in Table 1, we can see that: (1) the obtained performance with Left-Right pairs on the AFEW5.0 dataset is very close, but very different on the BAUM-1s dataset. This is because the difference of used datasets. In particular, on the AFEW5.0 dataset our method separately presents an accuracy of 32.64% for Left Mel-spectrogram and 32.90% for Right Mel-spectrogram. On the BAUM-1s dataset, we obtain an accuracy of 43.37% for Left Mel-spectrogram, and 33.77% for Right Mel-spectrogram, respectively, (2) for Mid-Side pairs, Mid Mel-spectrogram outperforms Side Mel-spectrogram. On the AFEW5.0 dataset Mid Mel-spectrogram is significantly higher than Side Mel-spectrogram with an accuracy of 7.57%. On the BAUM-1s dataset Mid Mel-spectrogram exceeds Side Mel-spectrogram with an accuracy of 1.44%.

To verify whether the complementarity between Left-Right or Mid-Side pairs exists or not, Fig. 4 shows the confusion matrices of recognition results of different monaural representations on the AFEW5.0 database. From Fig. 4, we can see that for Left-Right pairs, Right Mel-spectrogram obtains an accuracy of 15.38% for disgust emotion, whereas Left Mel-spectrogram gives an accuracy of 0%. This indicates that Left-Right pairs are complementary when identifying disgust to some extent.

Similarily, for Mid-Side pairs, the obtained results are also complementary when classifying disgust and surprise. For instance, Mid Mel-spectrogram achieves an accuracy of 6.90% for disgust, and 37.04% for surprise, respectively. By contrast, Side Mel-spectrogram provides an accuracy of 35.29% for disgust and 10.34% for surprise, respectively. The results in Fig. 4 demonstrate that fusing different monaural representations may improve the recognition performance for certain emotions further due to the existing complementarity between Left-Right or Mid-Side pairs.

2) EFFECTS OF DIFFERENT FEATURE POOLING

To evaluate the effectiveness of block-based temporal feature pooling, Tables 2-5 separately present the performance of four feature pooling methods, including block-based temporal max-pooling, block-based temporal average-pooling, global max-pooling, and global average-pooling. Note that global max-pooling or average-pooling demonstrates the global max or average operation is implemented on the whole divided segments.

From Tables 2-5, we can observe that block-based temporal feature pooling methods perform better than global feature pooling methods on the AFEW5.0 and BAUM-1s datasets, when using Left, Right, Mid and Side Mel-spectrograms. Specially, block-based temporal max-pooling and average-pooling are much better than global max-pooling and average-pooling for Mid and Side Mel-spectrograms. This indicates that block-based temporal pooling methods, which consider the useful temporal clues, are capable of producing better feature representations when aggregating segment-level features in an utterance into utterance-level features. Besides, max-pooling outperforms average-pooling at most cases, as max-pooling may be more suitable for the learned segment-level sparse features.

3) EFFECTS OF TWO-STAGE SCORE-LEVEL FUSION

Table 6 lists recognition results of two-stage score-level fusion, associated with the corresponding weight values α, β and λ in Eqs. (8)-(10). We initially fuse the results obtained with block-based temporal max-pooling from Left-Right pairs, and Mid-Side pairs, respectively. Then, we perform the second integration at score-level on the basis of the obtained results of the initial fusion. The results in Table 6 show that fusing Left, Right, Mid, and Side Mel-spectrograms outperforms integrating Left-Right pairs or Mid-Side pairs. In particular, on the AFEW5.0 dataset, fusing Left, Right, Mid, and Side Mel-spectrograms presents an accuracy of 36.29%, thereby making an improvement of 1.78% over Left-Right pairs, and 1.3% over Mid-Side pairs. On the BAUM-1s dataset, fusing Left, Right, Mid, and Side Mel-spectrograms achieves an accuracy of 44.31%, which is higher than Left-Right pairs by over 0.61%, and Mid-Side pairs by over 1.59%, respectively. In addition, compared with monaural representations as shown in Table 2-5, binaural presentations perform better on SER tasks because they contain more temporal-spatial information. This demonstrates the advantages of fusing binaural presentations on SER tasks.

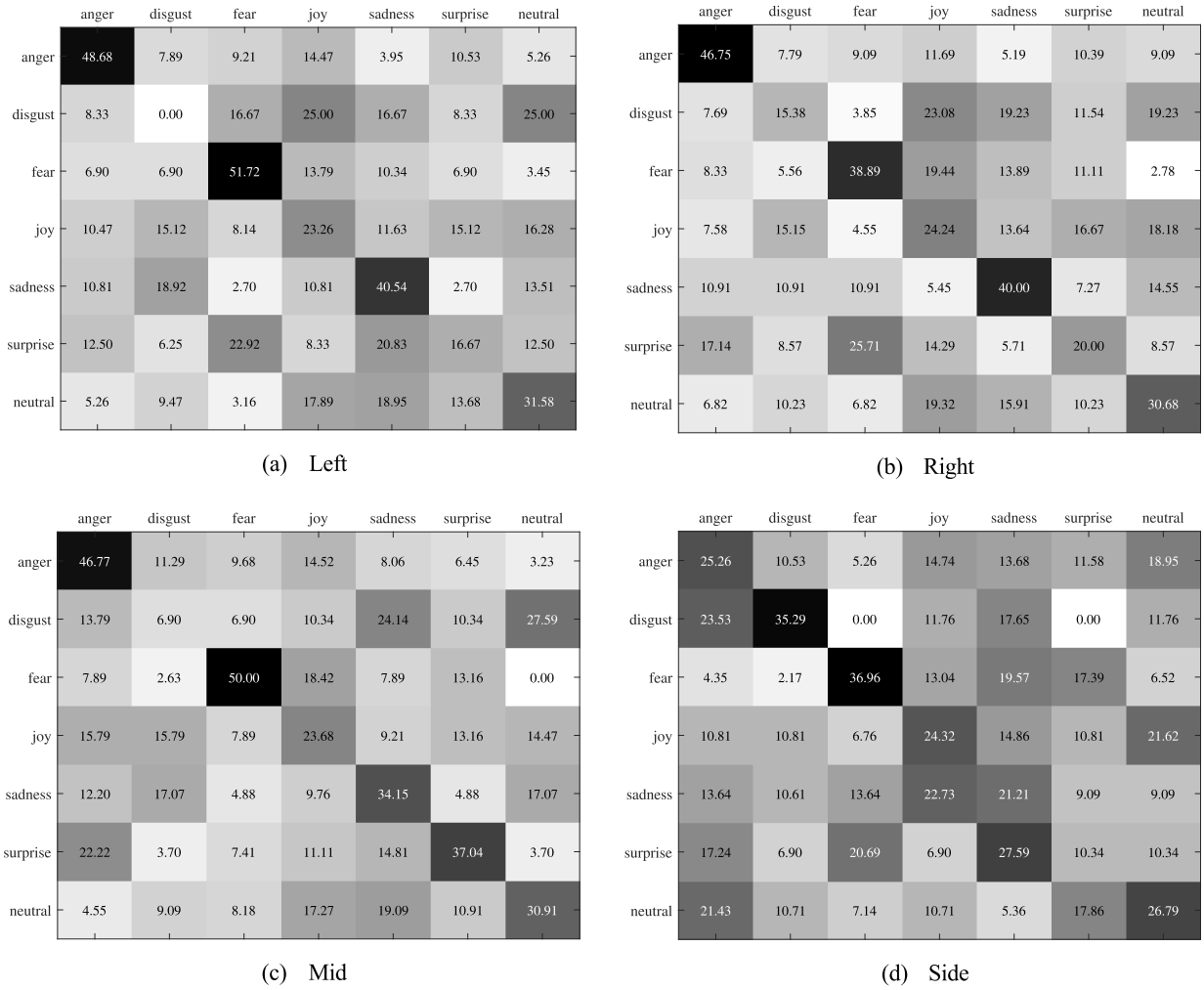

FIGURE 4. Comparisons of confusion matrices obtained by different monaural representations on the AFEW5.0 database.

TABLE 2. Precognition accuracy (%) of different feature pooling methods with left mel-spectrogram.

Pooling method	AFEW5.0	BAUM-1s
Global max-pooling	29.57	42.31
Global average-pooling	29.24	42.26
Block-based temporal max-pooling	32.64	43.10
Block-based temporal average-pooling	31.33	43.04

TABLE 3. Precognition accuracy (%) of different feature pooling methods with Right mel-spectrogram.

Pooling method	AFEW5.0	BAUM-1s
Global max-pooling	30.88	32.04
Global average-pooling	30.35	32.82
Block-based temporal max-pooling	32.90	33.77
Block-based temporal average-pooling	31.59	35.69

In addition, we also conduct emotion classification experiments at feature-level fusion. Specially, we concatenate utterance-level features of binaural presentations, followed by the linear SVM for emotion classification. Table 7 shows the results of feature-level fusion. From Table 6 and 7, we can see that our two-stage score-level fusion method performs better than feature-level fusion method. This indicates the effectiveness of two-stage score-level fusion in our work.

To further present the recognition performance per emotion, Fig. 5 and 6 individually provide the confusion matrices of recognition results when our method performs best on two used datasets. The results in Fig. 5 indicate that on the AFEW5.0 dataset "anger", "neutral" and "fear" are identified with an accuracy of 59.38%, 49.21% and 45.65%, respectively, whereas other three emotions are identified with an accuracy of less than 40%. The results in Fig. 6 show that on the BAUM-1s dataset "sadness" and "joy" are classified with an accuracy of 70.90%, and 55.49%, respectively. The remaining three emotions are distinguished with an accuracy of less than 30%.

TABLE 4. Precognition accuracy (%) of different feature pooling methods with mid mel-spectrogram.

Pooling method	AFEW5.0	BAUM-1s
Global max-pooling	30.85	38.18
Global average-pooling	29.33	37.73
Block-based temporal max-pooling	32.89	41.52
Block-based temporal average-pooling	32.37	41.24

TABLE 5. Precognition accuracy (%) of different feature pooling methods with side mel-spectrogram.

Pooling method	AFEW5.0	BAUM-1s
Global max-pooling	22.71	38.72
Global average-pooling	21.35	36.96
Block-based temporal max-pooling	25.06	41.09
Block-based temporal average-pooling	24.54	39.86

TABLE 6. Precognition accuracy (%) of two-stage score-level fusion.

Score-level fusion	AFEW5.0	BAUM-1s
Left, Right	34.51 (α=0.4)	43.70 (α=0.6)
Mid, Side	34.99 (β=0.7)	42.72 (β=0.6)
Left, Right, Mid, Side	36.29 (λ=0.6)	44.31 (λ=0.5)

TABLE 7. Precognition accuracy (%) of feature-level fusion.

Feature-level fusion	AFEW5.0	BAUM-1s
Left, Right	33.95	43.56
Mid, Side	34.18	41.45
Left, Right, Mid, Side	35.47	44.02

TABLE 8. Performance (%) comparisons of state-of-the-arts.

Datasets	Refs.	Methods	Accuracy
AFEW5.0	[11]	INTERSPEECH 2010 set + SVM	31.85
	[40]	INTERSPEECH 2010 set + RNN	33.20
	[41]	INTERSPEECH 2010 set + PLSR	33.96
	[42]	INTERSPEECH 2010 set + PCA + SVM	35.51
	Ours	Binaural representations + CNN	**36.29**
BAUM-1s	[29]	CNN + SVM	42.46
	[34]	MFCC, PLP + SVM	29.41
	[43]	CNN + SVM	42.38
	Ours	Binaural representations + CNN	**44.31**

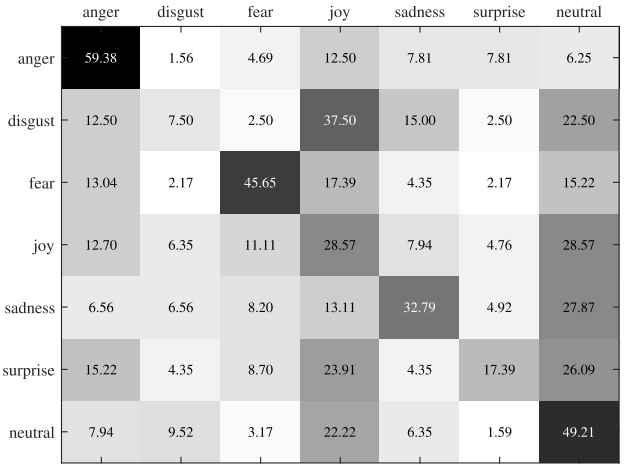

FIGURE 5. Confusion matrix of recognition results with two-stage score-level fusion on the AFEW5.0 database.

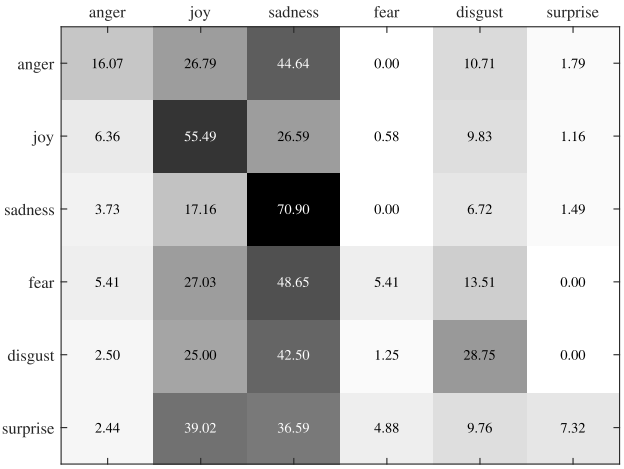

FIGURE 6. Confusion matrix of recognition results with two-stage score-level fusion on the BAUM-1s database.

4) COMPARISONS OF STATES-OF-THE-ARTS

To evaluate the effectiveness of our proposed method, we directly compare our reported results with previous works on used two emotional datasets. The experimental settings in these comparing works are similar to ours. Specially, the subject-independent cross-validation strategy is used. Table 8 presents the comparisons of state-of-the-art methods. From Table 8, we can see that our proposed method performs better than state-of-the-art methods on used two datasets. This exhibits the superiority of our proposed method over previous works. In particular, on the AFEW5.0 dataset [11], [40]–[42] adopted typical hand-crafted INTERSPEECH 2010 set (1582 LLDs), which were extracted from monaural representations such as the left channel of audio signals. They used common emotion classifiers, such as SVM, recurrent neural network (RNN), partial least squares regression (PLSR), and principal component analysis (PCA) associated with SVM. In comparison with [11], [40]–[42], our learned deep binaural representations with CNNs gives much higher performance. This demonstrates the advantages of deep learned features over hand-crafted features. On the BAUM-1s dataset, our method based on learned binaural representations with CNNs is also

superior to other CNN-learned methods [29], [43] which employ monaural representations. This indicates the effectiveness of learning binaural representations compared with monaural representations.

IV. CONCLUSION AND FUTURE WORK

Motivated by the fact that binaural representations in stereo contain much more information than conventional monaural representations, this paper proposes a new spontaneous SER method by learning deep binaural representations with CNNs.

Our method consists of three key steps: (1) multiple CNNs are separately employed to learn deep segment-level features from the extracted image-like Mel-spectrograms, (2) a block-based temporal feature pooling strategy is proposed to aggregate the learned segment-level features to fixed-length utterance-level features, (3) a two-stage score-level fusion strategy is adopted to combine the obtained results with different binaural representations. Experiment results on two challenging spontaneous emotional datasets, *i.e.*, AFEW5.0 and BAUM-1s, show that our proposed method outperforms state-of-the-art methods. In future, we will investigate more advanced deep models to learn deep discriminative features for SER. It is also interesting to extend our work to develop a real-time SER system. Additionally, our proposed method is not an end-to-end learning since we finish each step individually. It is thus interesting to develop an automatic end-to-end learning system to improve performance further. Besides, this work performs a simple score-level fusion strategy to integrate different binaural representations. We will explore other advanced fusion methods such as graph-based fusion graph-based fusion with metric learning [44].

REFERENCES

[1] R. Li, Z. Wu, J. Jia, J. Li, W. Chen, and H. Meng, "Inferring user emotive state changes in realistic human-computer conversational dialogs," in *Proc. ACM Multimedia Conf. Multimedia (MM)*, Seoul, Republic Korea, 2018, pp. 136–144.

[2] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[3] 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.

[4] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.

[5] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 1–4.

[6] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition," *IEEE Trans. Affective Comput.*, vol. 10, no. 2, pp. 265–275, Apr. 2019, doi: 10.1109/taffc.2017.2705696.

[7] S. Demircan and H. Kahramanli, "Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech," *Neural Comput. Appl.*, vol. 29, no. 8, pp. 59–66, Apr. 2018.

[8] X. Zhao and S. Zhang, "Spoken emotion recognition via locality-constrained kernel sparse representation," *Neural Comput. Appl.*, vol. 26, no. 3, pp. 735–744, Apr. 2015.

[9] D. Gharavian, M. Sheikhan, A. Nazerieh, and S. Garoucy, "Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network," *Neural Comput. Appl.*, vol. 21, no. 8, pp. 2115–2126, Nov. 2012.

[10] Z. Zixing, E. Coutinho, D. Jun, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 115–126, 2015.

[11] M. Kayaoglu and C. Eroglu Erdem, "Affect recognition using key frame selection based on minimum sparse reconstruction," in *Proc. ACM Int. Conf. Multimodal Interact.*, Seattle, WA, USA, 2015, pp. 519–524.

[12] 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. INTERSPEECH*, Lyon, France, 2013, pp. 148–152.

[13] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, Barcelona, Spain, 2013, pp. 3–10.

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

[15] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[18] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. Interspeech*, 2014, pp. 223–227.

[19] I. Shahin, A. B. Nassif, and S. Hamsa, "Emotion recognition using hybrid Gaussian mixture model and deep neural network," *IEEE Access*, vol. 7, pp. 26777–26787, 2019.

[20] S. Lalitha, S. Tripathi, and D. Gupta, "Enhanced speech emotion detection using deep neural networks," *Int. J. Speech Technol.*, vol. 22, no. 3, pp. 497–510, Sep. 2019.

[21] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5200–5204.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 5089–5093.

[24] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, San Diego, CA, USA, 2015, pp. 1–14.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[27] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang, and C. Li, "Deep spectrum feature representations for speech emotion recognition," in *Proc. Joint Workshop 4th Workshop Affective Social Multimedia Comput., 1st Multi-Modal Affective Comput. Large-Scale Multimedia Data*, Seoul, Republic Korea, 2018, pp. 27–33.

[28] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97515–97525, 2019.

[29] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, Oct. 2018.

[30] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 5, pp. 1075–1084, May 2017.

[31] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 2112–2121, Dec. 2014.
[32] R. Venkatesan and A. Balaji Ganesh, "Binaural classification-based speech segregation and robust speaker recognition system," *Circuits Syst Signal Process*, vol. 37, no. 8, pp. 3383–3411, Aug. 2018.
[33] A. Dhall, O. Ramana 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.*, Seattle, WA, USA, 2015, pp. 423–426.
[34] 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. Affective Comput.*, vol. 8, no. 3, pp. 300–313, Jul. 2017.
[35] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun, "Object detection networks on convolutional feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1476–1481, Jul. 2017.
[36] V. Campos, B. Jou, and X. Giró-i-Nieto, "From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction," *Image Vis. Comput.*, vol. 65, pp. 15–22, Sep. 2017.
[37] B. Banerjee and V. Murino, "Efficient pooling of image based CNN features for action recognition in videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 2637–2641.
[38] S. Casale, A. Russo, G. Scebba, and S. Serrano, "Speech emotion classification using machine learning algorithms," in *Proc. IEEE Int. Conf. Semantic Comput.*, Santa Clara, CA, USA, Aug. 2008, pp. 158–165.
[39] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 689–692.
[40] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, 2015, pp. 467–474.
[41] J. Wu, Z. Lin, and H. Zha, "Multiple models fusion for emotion recognition in the wild," in *Proc. ACM Int. Conf. Multimodal Interact.*, Seattle, WA, USA, 2015, pp. 475–481.
[42] 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.*, San Jose, CA, USA, Mar. 2019, pp. 443–448.
[43] 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.
[44] M. Angelou, V. Solachidis, N. Vretos, and P. Daras, "Graph-based multimodal fusion with metric learning for multimodal classification," *Pattern Recognit.*, vol. 95, pp. 296–307, Nov. 2019.

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. From 2015 to 2017, he held a postdoctoral position at the School of Electronic Engineering and Computer Science, Peking University, Beijing, China. He currently works as 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 the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He is an Associate Editor of IEEE ACCESS.

AIHUA CHEN received the Ph.D. degree in mechanical and electronic engineering from the Changchun Institute of Optics, Precision Mechanics and Physics, Chinese Academy of Sciences, in 2009. She currently works as a Lecturer with the Institute of Intelligent Information Processing, Taizhou University, China. Her research interests include image processing and pattern recognition.

WENPING GUO received the B.S. degree in mechatronic engineering and the M.S. degree in computer science from Southwest Jiaotong University, China, in 1998 and 2005, respectively. He is currently an Associate Professor with the Department of Mathematics and Information Engineering, Taizhou University, China. His research interests include machine learning and multimedia content analysis.

YUELI CUI received the B.S. degree from the Zhejiang University City College, Hangzhou, in 2006, and the M.S. degree from Hebei University, Baoding, in 2009, both in electronics and communication engineering. He currently works as a Lecturer with the Department of Physics and Electronics Engineering, Taizhou University, China. His research interests include image processing 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.

LIMEI LIU received the Ph.D. degree from the School of Information Science and Engineering, Central South University, in 2011. She held a postdoctoral position at the Business School of Hunan University, Changsha, China, from 2013 to 2017. She is currently a Professor with the Institute of Big Data and Internet Innovation, Hunan University of Commerce, China. Her research interests include machine learning and pattern recognition.