IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 10, 2019, accepted June 2, 2019, date of publication June 6, 2019, date of current version June 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2921390

Exploration of Complementary Features for Speech Emotion Recognition Based on Kernel Extreme Learning Machine

LILI GUO[1], (Student Member, IEEE), LONGBIAO WANG[1], (Member, IEEE), JIANWU DANG[1,2], (Member, IEEE), ZHILEI LIU[1], (Member, IEEE), AND HAOTIAN GUAN[3]

[1]Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[2]Japan Advanced Institute of Science and Technology, Ishikawa 9231292, Japan
[3]Huiyan Technology (Tianjin) Co., Ltd., Tianjin 300384, China

Corresponding authors: Longbiao Wang (longbiao_wang@tju.edu.cn) and Jianwu Dang (jdang@jaist.ac.jp)

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1305200, in part by the National Natural Science Foundation of China under Grant 61771333, and in part by the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330.

ABSTRACT Previous studies of speech emotion recognition using either empirical features (e.g., F0, energy, and voice probability) or spectrogram-based statistical features. The empirical features can highlight the human knowledge of emotion recognition, while the statistical features enable a general representation, but they do not emphasize human knowledge sufficiently. However, the use of these two kinds of features together can complement some features that may be unconsciously used by humans in daily life but have not been realized yet. Based on this consideration, this paper proposes a dynamic fusion framework to utilize the potential advantages of the complementary spectrogram-based statistical features and the auditory-based empirical features. In addition, a kernel extreme learning machine (KELM) is adopted as the classifier to distinguish emotions. To validate the proposed framework, we conduct experiments on two public emotional databases, including Emo-DB and IEMOCAP databases. The experimental results demonstrate that the proposed fusion framework significantly outperforms the existing state-of-the-art methods. The results also show that the proposed method, by integrating the auditory-based features with spectrogram-based features, could achieve a notably improved performance over the conventional methods.

INDEX TERMS Speech emotion recognition, auditory-based features, spectrogram-based features, complementary features, kernel extreme learning machine.

I. INTRODUCTION

Human-computer interaction has become popular in various fields, especially for intelligent dialogue systems and voice assistants, such as Siri, Cortana, and Google Assistant. In these applications, intention understanding is one of the key parts of the whole dialog system. Previous research found that emotion can significantly help machines to understand user's intention [1], so accurately distinguishing a user's emotion can enable greater interactivity and improve user experiences. However, speech emotion recognition is still a challenging task. One of the difficulties is determining how to extract effective features [2]. Another challenge is that we cannot clearly ascertain which model is effective in distinguishing emotions [3]. In addition, humans do not express emotions in a unified way, so the features should have good robustness for different emotional expressions.

Researchers have proposed various methods for speech emotion recognition. Among them, the conventional methods use auditory-based features (e.g., Mel Frequency Cepstrum Coefficient (MFCC), F0, energy, voice probability, and zero-crossing rate) for this task. These auditory-based features are selected based on human auditory perception, so they have a certain physical meaning. People have focused on selecting different auditory-based features for a long time [4]. The most commonly used model is to first extract auditory-based features and then train a classifier to obtain the emotion labels [5]. There are some traditional methods for speech

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

emotion recognition, such as the Gaussian mixture model (GMM) [6], support vector machine (SVM) [7], hidden Markov model (HMM) [8] and bidirectional long short-term memory (BLSTM) [9]. Han *et al.* [10] proposed the DNN-ELM model, which utilized a deep neural network (DNN) to obtain the emotion state probability distribution. In addition, a simple classifier, the extreme learning machine (ELM), was then used to obtain the labels. Wang and Tashev [11] made improvements to the DNN-ELM model, in which the activation of the last hidden layer of the DNN replaced the probability distribution that is used to train the ELM. Lee and Tashev [12] proposed the recurrent neural network (RNN)-ELM model, which utilized the long contextual effect in emotional speech. These models have been regarded as the state-of-the-art models for many years in the field of speech emotion recognition. However, people's cognition of speech emotion recognition is limited [13]. It is difficult to extract abundant features using priori knowledge alone. Therefore, the auditory-based features are not sufficiently representative of emotional information.

With the development of deep learning (DL), there is a trend in the field of speech processing to use DL for automatically extracting features from speech signals [14], [15]. A CNN is adept at extracting local features from raw input data [16]. The CNN was initially applied to image field and was regarded as one of the representative models for image recognition systems [17]. In recent years, CNNs have been applied to speech processing and have achieved excellent results [18], [19]. A CNN-based speech emotion recognition model had been proposed in [20]. This paper utilized CNNs to extract deep acoustic features from spectrograms and then trained an SVM as classifier. Lim *et al.* [21] and Satt *et al.* [22] proposed the hybrid CNN-BLSTM model without using any traditional auditory-based features. Although using the CNN-BLSTM model on spectrograms derectly has obtained great achievements and has been regarded as the most commonly used method for this task over recent years, there are still many problems that exist in this model. First, the BLSTM model has a complicated structure and high complexity in training; therefore, it needs a large amount of training data [23]. For a task with insufficient data, this model tends to fall into overfitting. Furthermore, there is no sufficiently labeled public corpus of emotional speech at present [24], [25]. Second, the CNN-BLSTM model adopts a CNN to extract features automatically, and it uses the spectrogram-based statistical features alone. Although statistical features can give a general representation of emotion, they do not emphasize the human knowledge sufficiently. In addition, previous studies have indicated that some auditory-based empirical features (e.g., F0, energy, and voice probability) are very important to distinguish speech emotion [26].

In this work, we extend our previous work [27] and continue to explore complementary features for speech emotion recognition. To solve the first problem, this paper proposes the CNN-KELM model, which uses a CNN to extract deep features from spectrograms and then uses a kernel extreme learning machine (KELM) to distinguish emotions. The KELM is a learning algorithm for single-hidden layer feed-forward neural networks (SLFNs) [28], and it is a modified extreme learning machine (ELM) that was proposed by Huang *et al.* [29], [30]. ELM has been applied in various classification tasks due to the properties of high generalization capability and fast training [31], [32]. ELM as a classifier shows better performance than SVM for speech emotion recognition [10]. Moreover, ELM can perform well on small databases. To address the second problem, motivated by the powerful feature learning ability of some multimodal deep models [33]–[35], this paper proposes a whole dynamic fusion framework to utilize the potential advantages of the complementary spectrogram-based and auditory-based features, which is different from [27]. Paper [27] separated the feature extraction and feature fusion stages, which cannot guarantee the global optimal in tuning the parameters. In addition, decision-level fusion is also considered in this paper. In this way, the proposed framework can complement some parameters that may be unconsciously used by humans in daily life, but have not been realized yet. Furthermore, researches found that the raw auditory-based features are correlated, which results in a small inter-class distance [36], [37]. To avoid this problem, this paper extracts the discriminative bottleneck features from the raw auditory-based features using a deep neural network (DNN). To the best of our knowledge, it is leading edge work to explore the complementarity between spectrogram and auditory-based features. In addition, we adopt a KELM as the classifier to recognize emotion.

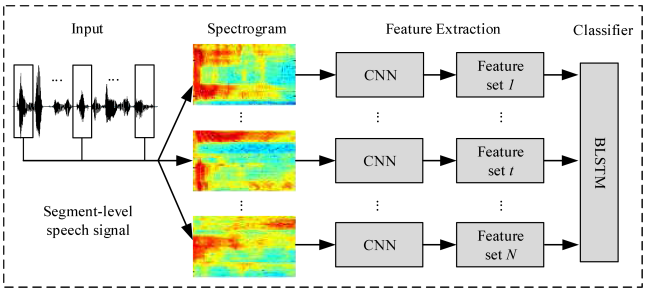


FIGURE 1. Structure of the CNN-BLSTM model.

The rest of this paper is organized as follows. Section II describes the background theory of the baseline CNN-BLSTM model. The proposed fusion framework is described in Section III. The experimental results and analysis are presented in Section IV. Finally, Section V gives the conclusions and prospects.

II. THEORETICAL BACKGROUND

Since extracting features manually has many problems such as being time-consuming and producing a limited number of feature categories, people begin to use CNNs to extract features automatically. In recent years, researchers have commonly used CNNs directly on spectrograms to extract deep

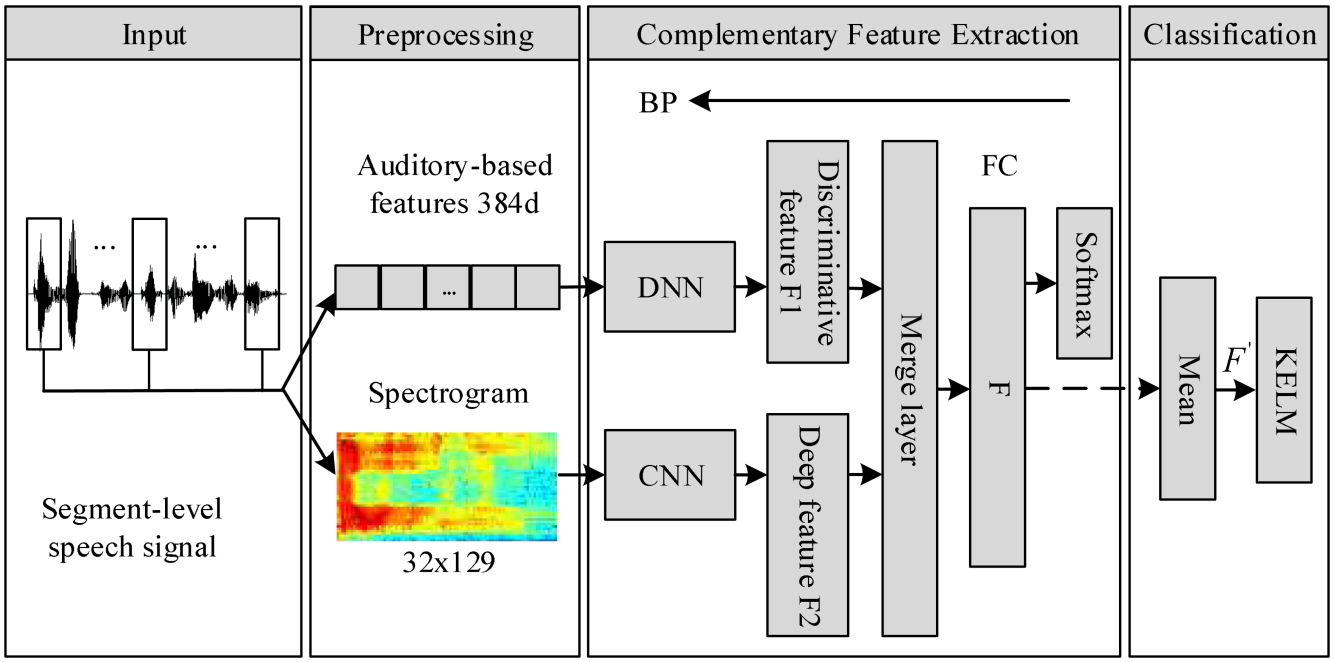


FIGURE 2. Proposed fusion framework to integrate the spectrograms and auditory-based features based on KELM.

acoustic features [38], [39], and then the BLSTM method was adopted to recognize emotions. The CNN-BLSTM model [22] has become the most commonly used method for speech emotion recognition at present. In this section, we will give a detailed introduction to the baseline CNN-BLSTM model.

Fig. 1 shows the structure of the CNN-BLSTM model. First, as emotional expression is dynamic, to utilize the dynamic information, the speech signal is divided into N fixed length segments. Then speech signals in the segment-level are transformed into spectrograms by using the Fourier transform since time-frequency analysis is widely used in the field of speech signal processing [40]. Next, CNNs are used to extract deep acoustic features from the spectrograms. The weights of each CNN feature map are shared, which could result in reducing the complexity of the network and the number of parameters. The activations of the full connection layer are the deep acoustic features that we would like to obtain. Finally, these deep acoustic features in the segment-level are fed into the BLSTM method to obtain the utterance-level labels. The main idea of BLSTM is to use the forward direction LSTM and backward direction LSTM to extract the contextual hidden information to form the final outputs [41]. This demonstrates that the BLSTM method can make good use of the contextual information, which is important in the speech processing field. Therefore, the BLSTM method is widely used in some sequence-based applications, including speech processing [42], [43].

Although the CNN-BLSTM model gets good results for many speech processing tasks, there are still many problems with this model. First, since the structure of the BLSTM method is complicated, it is easy to fall into overfitting when the database is small. Moreover, the CNN-BLSTM model utilizes the CNN to extract only acoustic features from the spectrograms, which does not emphasize the human knowledge sufficiently. However, some auditory-based empirical features (e.g., F0, energy, and voice probability) are key issues for distinguishing emotions [26].

III. THE FUSION FRAMEWORK BASED ON KELM

The proposed framework for speech emotion recognition is shown by the flowchart in Fig. 2, which consists of speech segmentation, data preprocessing, deep complementary feature extraction, and KELM classification. Different from the CNN-BLSTM model, this fusion framework considers the effects of auditory-based features. Furthermore, we adopt the CNN-KELM model in this framework; we use KELM to distinguish emotions because it has properties of high generalization capability and fast training. Moreover, the KELM can perform well on small databases. As there is no sufficiently labeled public corpus of emotional speech at present [24], [25], in order to get more training data, we divide speech into several segments. Moreover, we can use the dynamic information through speech segmentation. However, choosing a suitable segment size is a challenging problem for speech emotion recognition. Researchers have found that a segment speech signal that is greater than 250 ms includes sufficient emotional information [44], [45]. In this paper, we use a 265 ms window size and a slide window of 25 ms to transform an utterance into several segments, and all the segments in one utterance share the same label.

The detailed description of the proposed fusion framework is given as follows.

A. DATA PREPROCESSING

In this section, we would like to finish the extraction of auditory-based features and spectrograms. We use the openSMILE [46] tool to extract the auditory-based features with 384 dimensions proposed in [47]. The selected 16 low-level descriptors (LLDs) and their first-order derivatives are the basic features, and then 12 functionals are applied to these basic features. All of the LLDs and functionals set are shown in Table 1.

TABLE 1. Auditory-based feature set.

(△)LLDs (16×2)	MFCC (12): Mel Frequency Cepstrum Coefficient, RMSenergy (1): root-mean-square signal frame energy, F0 (1): fundamental frequency, ZCR (1): zero-crossing rate, voicing probability (1)
Functionals (12)	max, min, mean, range, standard deviation, kurtosis, skewness, offset, slope, MSE, absolute position of min/max

A time-frequency analysis is commonly used in speech signal processing [40], so we transform the speech signal into a spectrogram for training the CNN. First, pre-emphasis is applied to improve the high frequency and better maintain the speech information rather than eliminating the noise completely. Then, the framing and windowing operations are adopted. In this paper, we use a Hamming window with a size of 16 ms and a 50% overlap. Finally, short time Fourier transform (STFT) with default values for 256 points is adopted to obtain the spectrogram. As the speech signal is divided into many fixed-length segments of 265 ms, the size of the spectrogram is 32×129.

B. COMPLEMENTARY FEATURE EXTRACTION

The proposed complementary feature extraction method consists of an auditory-based features channel and a spectrogram channel followed by a merge layer, a fully connected layer, and a softmax layer.

Since the raw auditory-based features are correlated, which will reduce the inter-class distance, we should not use these features directly but rather extract the discriminative features from the raw auditory-based features using a DNN. Therefore, the auditory-based features with 384 dimensions are fed into DNN to extract the discriminative features, $F1$. It is well known that the deep belief network (DBN) is able to model natural signals [37], and so a DBN consisting of superimposed restricted Boltzmann machines (RBMs) is used for pre-training DNN [48]. Meanwhile, CNNs are adopted to extact deep acoustic features from the spectrograms. The structure of the CNN contains the convolutional layer, pooling layer, flatten layer, and fully connected layer, and the outputs of the fully connected layer are the deep acoustic features, $F2$.

Then, the discriminative features, $F1$, and the deep acoustic features, $F2$, are spliced into a large vector, V, by a merge layer. The representation is as follows.

$$V = [F1, F2]. \tag{1}$$

The last two layers are the fully connected layer and the softmax layer. The whole network is trained by using the error back propagation technique. By adjusting the parameters, the auditory-based features and spectrogram can constrain each other to extract more robust complementary features. When the model converges to an ideal state, the outputs of the fully connected layer are the desired complementary features, F.

In this work, the utterance is divided into N segments; thus, all the features that were extracted are segment-level. To get utterance-level features, we perform the mean-operation as follows.

$$F_i' = \frac{1}{N}\sum_{t=1}^{N} F_i^t, \tag{2}$$

where F_i' is the feature set of the i-th utterance, N is the number of segments in utterance i, and F_i^t is the feature set of the t-th segment of the i-th utterance. Finally, the fusion feature set, F', is fed into the KELM for speech emotion recognition.

In fact, we have used max-pooling, min-pooling and mean pooling to obtain the utterance-level features during our experiments, while the mean-pooling gave the best result. Therefore, in this paper, we only give the results of mean-pooling, which are similar to those of [11] and [35].

C. THE KELM-BASED CLASSIFIER

ELM is a learning algorithm for single-hidden layer neural networks, which was proposed by Schuller *et al.* [49] and Zhu *et al.* [50]. ELM has been used for many classification tasks and achieves better results than some of the traditional classifiers such as SVM [10]. Suppose a network contains n input layer nodes, l hidden layer nodes and m output layer nodes, and there are N random samples, $(x_i, y_i) \in R_n \times R_m$, $(i = 1, \ldots, N)$. The training process of the ELM method contains three steps:

The first step is to set the number of hidden layer nodes. Furthermore, the bias values and weights, w, that are used for connecting the input layer and hidden layer are randomly initialized.

The next step is to compute the output matrix, H, of the hidden layer as follows.

$$H = \begin{bmatrix} g(w_1, b_1, x_1) & \ldots & g(w_l, b_l, x_1) \\ g(w_1, b_1, x_2) & \ldots & g(w_l, b_l, x_2) \\ \vdots & \ddots & \vdots \\ g(w_1, b_1, x_N) & \ldots & g(w_l, b_l, x_N) \end{bmatrix}_{N \times L}, \tag{3}$$

where $g(.)$ is the activation function, and the commonly used activation function is the sigmoid function.

Finally, the weights, β, that are between the hidden layer and output layer are computed using the least squares method, as shown in (4):

$$\beta = H^T H, \tag{4}$$

where H^T is the generalized inverse matrix of H.

We can see that the whole training process of ELM only contains a pseudo-inverse calculation without parameter adjustment [51], [52]. The training process finishes in a single iteration, which is faster than the training time for conventional back propagation (BP)-based algorithms such as BLSTM method. KELM is a modification of the original ELM, which defines the kernel function of the inner product for the hidden layer outputs, H^T and H, and does not need to give the number of hidden layer nodes. Previous studies have shown that KELM is better than ELM [28], so in our method, the KELM is used as the classifier to distinguish emotions.

IV. EXPERIMENTS

A. EXPERIMENTAL DATABASES

The evaluation and comparison of different methods are challenging due to the lack of a sufficiently labeled public corpus of emotional speech. In this paper, we use two publicly available databases of emotional speech, the Berlin Emotional Database (Emo-DB) [53] and Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [54]. The following sections provide detailed descriptions of both databases.

1) EMO-DB DATABASE

The Emo-DB database contains the emotional utterances produced by 10 German actors (five females/five males); they read one of 10 pre-selected sentences typical of daily conversation using different emotional expressions. This database contains 535 utterances in German with seven emotions: anger, boredom, fear, disgust, happiness, sadness, and neutral. All utterances are sampled at 16 kHz and are approximately 2-3 seconds long.

TABLE 2. Emotion distribution of the Emo-DB database.

Emotion	Number of utterances	Percentage
Fear	69	12.90%
Disgust	46	8.60%
Happiness	71	13.27%
Boredom	81	15.14%
Neutral	79	14.77%
Sadness	62	11.59%
Anger	127	23.74%

Table 2 shows the emotion distribution of this database. As it is a small database, similar to [21], we adopt random 10-fold cross-validation to conduct the experiments in this paper. In addition, we also conduct speaker-independent experiments, which are usually adopted in most real applications [55]. The sentences from 8 speakers are used for training, and the sentences from remaining 2 speakers are used for testing.

2) IEMOCAP DATABASE

The IEMOCAP database is one of the most commonly used corpora for speech emotion recognition; it contains scripted and improvised dialogs. This database contains approximately 12 hours of audiovisual data including video, speech, motion capture of faces and text transcriptions performed by 10 skilled actors. All utterances are sampled at 16 kHz and are approximately 3-15 seconds long. Each utterance from either of the actors in the interaction has been evaluated categorically over the set of: angry, happy, sad, neutral, frustrated, excited, fearful, surprised, and disgusted by three different human annotators. Since three annotators may give different labels for an utterance, we use the utterances with at least two agreed-upon emotion labels for our experiments. In addition, in this paper, we only select the utterances with labels for four emotions: neutral, anger, sadness, and happiness, which are often used in previous studies [56].

TABLE 3. Emotion distribution of the IEMOCAP database.

Emotion	Number of utterances	Percentage
Neutral	1708	30.88%
Anger	1103	19.94%
Sadness	1084	19.60%
Happiness	1636	29.58%

Table 3 lists the emotion distribution of IEMOCAP database. This database has five sessions and includes scripted and improvised utterances. In addition, there are two speakers for each session, and there is no speaker overlapping between the different sessions. Therefore, we utilize this setup to conduct the speaker-independent 5-fold cross validation. In each fold, the data from four sessions is used for training the model, and the data from the remaining session is used for testing.

TABLE 4. Important parameters of CNN.

Layer	value
Convolution 1	$32@5 \times 5$
Max-pooling 1	2×2
Convolution 2	$64@5 \times 5$
Max-pooling 2	2×2
Dense layer	1024
Dropout	0.5

B. EXPERIMENTAL SETUP

To get more training data, the utterances are divided into several segments, and all segments in the same utterance share the same label. However, it is an open problem to choose the length of a segment. Researchers have shown that a segment longer than 250 ms contains enough emotional information [44], [45]. Similar to [10], in this work, the length of a segment is set to be 265 ms. We also attempted longer segments such as 3 s [22] and 655 ms [39], and did not achieve better results. This outcome means that segment of 265 ms is more suitable for our method. Moreover, we conducted many trials with different numbers of hidden layer nodes,

layers, etc., to select the optimal structure for all the comparison methods. The parameters that are used in CNN in this work are shown in Table 4. The CNN contains two convolutional layers, two max-pooling layers, a flatten layer, a fully connected layer, and a dropout layer. There are two pairs of alternate convolutional layer with a size of 5×5 and max-pooling layer with a size of 2×2; the number of filters for these two convolutional layers are 32 and 64, respectively. After the last max-pooling layer, all the feature maps are changed to one dimensional vector by the flatten layer. Then follows a fully connected layer with 1024 hidden units, which contain the deep acoustic features. In addition, to avoid overfitting, a dropout layer with a factor of 0.5 is used before the output layer.

As the database and the selected utterances are different in most studies, we cannot compare them under different conditions. To make the experimental results more convincing, our experimental setup is consistent with that of [10]. Furthermore, for other comparison methods, we attempted many times to choose the optimal parameters under the same conditions. All the experimental methods are listed as follows.

- **CNN-BLSTM**: This is the baseline model of this paper. The structure of the CNN, as shown in Table 4, is utilized to extract deep acoustic features with 1024 dimensions from the segment-level spectrograms. Then, these segment-level features are fed into the BLSTM method to recognize the emotion with utterance-level label. After many trials on Emo-DB and IEMOCAP databases, the selected optimal structure of BLSTM method contains two hidden layers, and each layer has 200 nodes.
- **CNN-ELM**: It is a novel method for speech emotion recognition that was introduced in our previous work [27]. This model is used to verify the effect of the ELM classifier by comparing it with the CNN-BLSTM model. This model uses CNN to extract deep acoustic features from the spectrograms. Then, the ELM is used as a classifier to recognize emotion. For the ELM structure, the number of hidden layer nodes is set to 2100 for the Emo-DB database; meanwhile the number of hidden layer nodes is set to 100 for the IEMOCAP database.
- **CNN-KELM**: This is the adopted method for speech emotion recognition in this work. We adopt KELM as the classifier to distinguish emotions. For Emo-DB database, the KELM parameters, including the regularization coefficient and kernel parameter, are all set to 100. For IEMOCAP database, the regularization coefficient and kernel parameter of the KELM are set to 10000 and 10, respectively.
- **DNN-ELM**: This is a commonly used model that uses only auditory-based features, which used to be compared with spectrogram-based methods. All the auditory-based features with 384 dimensions are fed into the DNN to extract discriminative features. The structure of the DNN contains four hidden layers, and each layer has 512 nodes. Then, mean-pooling is performed to obtain the utterance-level features. Finally, these discriminative features are fed into the ELM to distinguish emotions. The number of hidden layer nodes for the ELM is set to 2100 and 100 for Emo-DB and IEMOCAP databases, respectively.
- **DNN-KELM**: This method is used to verify KELM by comparing it with the DNN-ELM model. This model uses DNN to extract features from the auditory-based features and then adopts the as the classifier. The structure of the DNN contains three hidden layers, and each layer has 512 nodes. For Emo-DB database, the regularization coefficient and kernel parameter of the KELM are set at 10 and 1, respectively. For IEMOCAP database, all the parameters of the KELM are set at 1.
- **Decision-Level Fusion**: To compare with the proposed feature-level fusion framework, we also consider the decision-level fusion method. First, we use the KELM to calculate the classification scores for the different types of features including the auditory-based bottleneck features, $Score1$, and the spectrogram-based deep acoustic features, $Score2$. Then, the scores are fused to form a decision rule for the classification [57]. A weighted summation is adopted in terms of the obtained class score values as follows.

$$S = max\left\{a \cdot Score1 + (1-a) \cdot Score2)\right\}, \qquad (5)$$

where $a(0 < a < 1)$ defines the weight between two kinds of features. After repeated experiments, the weights for Emo-DB and IEMOCAP databases are 0.4 and 0.6, respectively.
- **The Proposed Fusion Framework Based on KELM**: This framework uses the CNN structure, as shown in Table 4, to extract the deep acoustic features from the spectrograms and uses the DNN to extract the auditory-based discriminative features. The structure of the DNN contains three hidden layers, each having 512 nodes, followed by a merge layer, a fully connected layer with 1024 nodes, and a softmax layer. The outputs of the fully connected layer are the complementary features. Finally, these complementary features are fed into the KELM to distinguish emotions. For Emo-DB database, the regularization coefficient and kernel parameter of the KELM are set to 10000 and 10, respectively, and the regularization coefficient and kernel parameter are set to 100 and 1000 for IEMOCAP database.

C. ANALYSIS OF THE DIFFERENT FEATURES

To analyze the effects of different features, Fig. 3 shows their visualization maps. All types of features are listed as following:

- **Raw Auditory-Based Features**: These features with 384 dimensions consist of the LLDs and their statistical features, which are shown in Table 1.
- **Discriminative Features**: The discriminative features with 512 dimensions are extracted by the DNN from the raw auditory-based features.

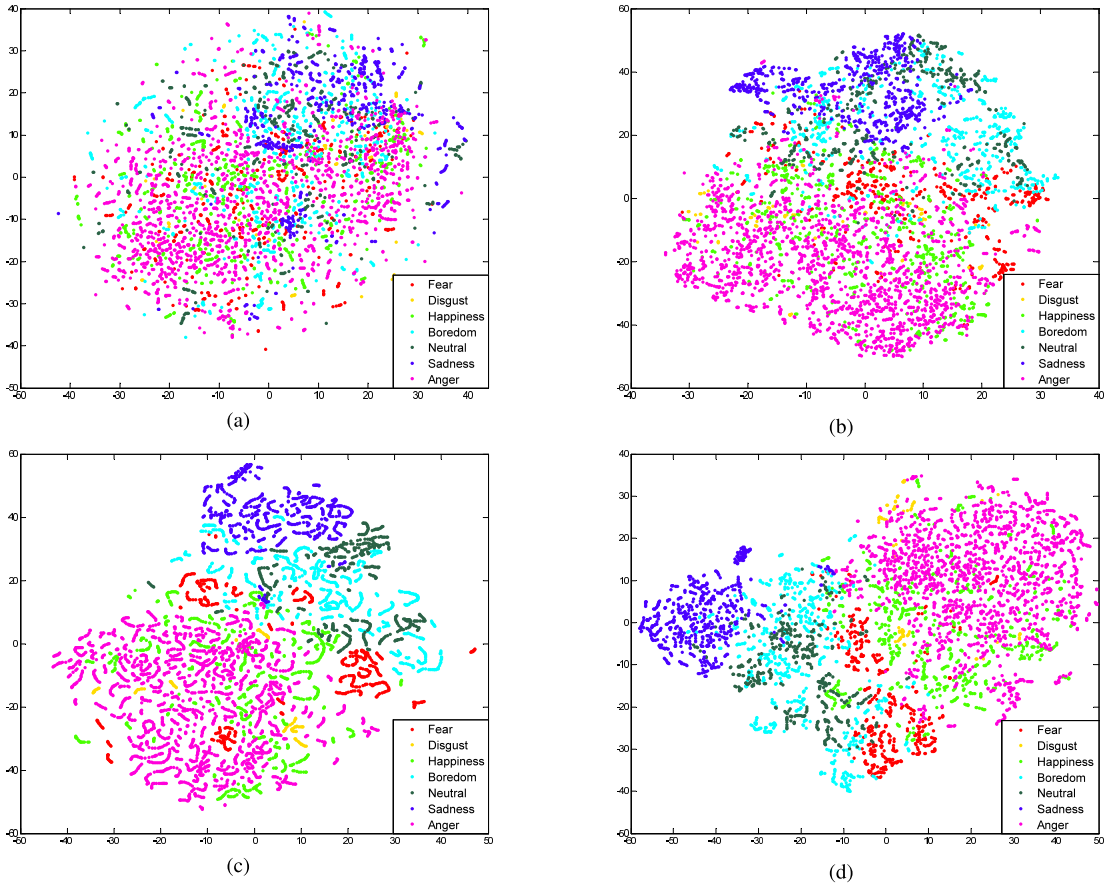


FIGURE 3. Visual distributions of the different features. (a) Raw auditory-based features. (b) Discriminative features. (c) Spectrogram-based features. (d) Complementary features.

- **Spectrogram-Based Features**: These deep acoustic features with 1024 dimensions are extracted from the spectrograms using the CNN.
- **Complementary Features**: The complementary features with 1024 dimensions are extracted by the fusion framework from the auditory-based features and spectrograms.

We use data from the Emo-DB database for the feature analysis. First, we need to reduce the features to two dimensions. There are many techniques for dimensionality reduction; among them, the t-distributed stochastic neighbor embedding (t-SNE) [58] is a commonly used method for dimensionality reduction. In particular, the t-SNE technique has been successfully applied for visualization. In this paper, we use an optimization of t-SNE called fast t-SNE [59] for dimensionality reduction. Then, we illustrate the distributions of the seven emotions using different colors in a two-dimensional plane.

From Fig. 3(a), we can see that the data distributions of the different classes are significantly overlapped, and it is difficult to distinguish the different emotions. Furthermore, the boundaries of the different emotions in Fig. 3(b) are clearer than those in Fig. 3(a). In particular, there is clear boundary between anger and sadness. This results indicates that the discriminative features are more effective than the raw auditory-based features for speech emotion recognition, so it is necessary to extract the discriminative features using the DNN. Fig. 3(c) shows that most of the classes are clustered together, especially for anger, sadness, and boredom classes. However, the distributions of fear and happiness are rather scattered. In addition, there is a large inter-class distance in each emotion. Fig. 3(d) performs the best performance among those types of features. First, there are clear contours for sadness, anger, boredom, and fear. Furthermore, the inter-class distances in Fig. 3(d) are less than those of other features. To summarize, the complementary features show strong discriminative ability for emotions.

D. RESULTS AND DISCUSSION

In this work, we use two common evaluation criteria [22] to validate the overall effect of the proposed fusion framework, as following:

- Weighted accuracy (WA) - this is the classification accuracy for the whole test set.
- Unweighted accuracy (UA) - the classification accuracy for each emotion is first calculated and then averaged.

The evaluation results for Emo-DB and IEMOCAP databases are illustrated in Table 5 and Table 6, respectively, and some conclusions can be drawn as following:

TABLE 5. Accuracy of Emo-DB database.

Method	Random 10-fold		Speaker-indep	
	UA(%)	WA(%)	UA(%)	WA(%)
DNN-ELM	83.53	84.49	79.19	80.56
DNN-KELM	84.09	84.67	80.68	81.31
CNN-BLSTM	86.66	87.66	80.96	81.31
CNN-ELM	90.83	91.21	81.93	82.24
CNN-KELM	91.07	91.96	84.40	85.05
Decision-level fusion	91.59	92.33	86.69	86.92
Feature-level fusion	**92.45**	**92.90**	**87.49**	**87.85**

TABLE 6. Accuracy of IEMOCAP database.

Method	UA(%)	WA(%)
DNN-ELM	55.07	54.19
DNN-KELM	55.28	54.53
CNN-BLSTM	51.44	50.41
CNN-ELM	54.49	53.12
CNN-KELM	55.41	53.84
Decision-level fusion	56.76	56.01
Feature-level fusion	**57.99**	**56.55**

1) As shown in Table 5, the results of the random 10-fold cross validation are better than those of the speaker-independent experiments because normalizing features on a per-speaker basis can significantly improve the performance [60]. Furthermore, we can observe that they show a consistent trend in all the methods. This outcome means that the proposed method is still effective under the speaker-independent condition. Therefore, in the following experiments, we only report the results of the random 10-fold cross validation for Emo-DB database.
2) The spectrogram-based methods (i.e., CNN-BLSTM, CNN-ELM, and CNN-KELM) all outperform the perceptual feature-based methods (i.e., DNN-ELM, and DNN-KELM) for Emo-DB database, but for IEMOCAP database, the spectrogram-based methods are not better than perceptual feature-based methods. We believe the reason is that the utterances of the IEMOCAP database contain more noise and silent segments [54], so CNN cannot extract effective emotion-relevant features from spectrograms with noise and silence. In addition, there are differences in size, annotations, speech quality, speaker, etc., for these two corpora.
3) Both the CNN-ELM and CNN-KELM models perform better than the CNN-BLSTM model on Emo-DB and IEMOCAP databases. For Emo-DB database, the CNN-KELM model outperforms the CNN-BLSTM model in terms of UA and WA by an absolute 4.41% (from 86.66% to 91.07%) and 4.3% (from 87.66% to 91.96%), respectively. For IEMOCAP database, the CNN-KELM achieves an absolute 3.97% (from 51.44% to 55.41%) and 3.43% (from 50.41% to 53.84%) improvements over the CNN-BLSTM model in terms of UA and WA. The results prove that the proposed CNN-KELM model is effective for emotion recognition and KELM/ELM models are excellent classifiers, at least in this work. We think there are two reasons contribute to these results. First of all, as a relatively abundant number of features have been extracted by the CNN, the emotional utterances can be classified by a simple static classifier. Furthermore, KELM/ELM can perform well on small databases. Meanwhile, we can see that using the KELM as a classifier is better than the ELM on both spectrogram-based and perceptual feature-based methods, so the KELM is used as the classifier in our proposed fusion framework.
4) Although the decision-level fusion framework obtain better performances than other comparison methods, it still performs worse than the proposed feature-level fusion framework on Emo-DB and IEMOCAP databases. We think the reason is that decision-level fusion cannot capture the mutual correlation among different types of features because auditory-based features and spectrogram-based features are independent in this framework.
5) For Emo-DB database, the results of the proposed feature-level fusion framework are better than those of the other methods in terms of UA and WA. For example, the proposed fusion framework outperforms the state-of-the-art model, the CNN-BLSTM, by an absolute 5.79% (from 86.66% to 92.45%) and 5.24% (from 87.66% to 92.90%) in terms of UA and WA, respectively. We think it is because the utterances of the Emo-DB database are clean and the labels are uncontroversial. In addition, the CNN can extract relatively rich and effective features from clean spectrograms. Finally, the fusion framework can extract more robust complementary features based on the spectrogram-based features and auditory-based features, which can highlight the weights of the emotion-relevant features.
6) For IEMOCAP database, the proposed feature fusion model based on KELM obtains the best results in terms of UA and WA. Compared with the CNN-BLSTM model, the fusion framework achieves an absolute 6.55% (from 51.44% to 57.99%) and 6.14% (from 50.41% to 56.55%) improvement in terms of UA and WA, respectively. In addition, the proposed model also performs better than the perceptual features-based methods. For example, the fusion framework outperforms DNN-KELM by an absolute 2.71% (from 55.28% to 57.99%) and 2.02% (from 54.53% to 56.55%) in terms of UA and WA, respectively. From the above results, we can see that the fusion framework is effective on IEMOCAP database, which indicates that the spectrogram-based features and auditory-based discriminative features are complementary.

To evaluate the effects for each emotion, we give the F1 for all the methods. The F1 score is the most commonly used evaluation criterion for testing accuracy because it has a balance between recall (R) and precision (P). Equation (6) gives the expression for F1. Table 7 and Table 8 give the F1 results

TABLE 7. F1 (%) of each emotion for Emo-DB database.

Method	Fear	Disgust	Happiness	Boredom	Neutral	Sadness	Anger	Average
DNN-ELM	79.07	91.76	72.73	83.64	84.61	90.08	88.24	84.30
DNN-KELM	79.41	90.91	73.44	84.07	86.08	90.90	87.08	84.56
CNN-BLSTM	91.05	91.96	76.03	85.53	87.34	90.90	89.61	87.49
CNN-ELM	**92.65**	**96.63**	83.46	91.14	88.75	96.12	91.51	91.47
CNN-KELM	91.04	90.70	87.22	93.75	90.68	95.38	93.23	91.72
Decision-level fusion	91.73	91.95	**88.06**	93.75	90.68	95.38	**93.58**	92.16
Feature-level fusion	91.85	95.56	84.62	**95.71**	**92.99**	**97.64**	92.54	**92.99**

TABLE 8. F1 (%) of each emotion for IEMOCAP database.

Method	Neutral	Anger	Sadness	Happiness	Average
DNN-ELM	51.21	59.48	58.03	51.16	54.97
DNN-KELM	51.96	59.61	58.03	51.53	55.28
CNN-BLSTM	44.63	60.96	50.43	48.45	51.12
CNN-ELM	47.96	64.87	57.29	47.71	54.46
CNN-KELM	46.91	66.40	57.22	49.91	55.11
Decision-level fusion	**52.40**	62.68	58.92	**53.57**	56.89
Feature-level fusion	50.90	**66.64**	**60.84**	52.53	**57.73**

for Emo-DB and IEMOCAP databases, respectively.

$$F1 = \frac{2 \cdot P \cdot R}{P + R}. \tag{6}$$

From Table 7, we can see that for Emo-DB database, the proposed feature-level fusion framework performs best in the boredom, neutral and sadness classes. For utterances with labels of fear, disgust, happiness and anger, the F1 scores of the fusion framework are not the best, but the results are still significantly better than those of the baseline CNN-BLSTM model. CNN-ELM performs best in the fear and disgust classes, indicating that using ELM as the classifier is useful for speech emotion recognition. Additional, the decision-level fusion method obtains the best results in the happiness and anger classes. Furthermore, the fusion framework outperforms the state-of-the-art method, the CNN-BLSTM model, by an absolute 5.5% (from 87.49% to 92.99%) in terms of the average F1.

From Table 8, we can see that for IEMOCAP database, the feature-level fusion framework achieves the best results in most of the emotion classes (i.e., anger, sadness, and happiness) and especially in the sadness class where it outperforms the CNN-BLSTM model by an absolute 10.41% (from 50.43% to 60.84%). In addition, for neutral and happiness classes, the decision-level fusion method performs best. In terms of the average F1, our methods (''CNN-ELM'', ''CNN-KELM'', ''Decision-level fusion'', ''Feature-level fusion'') each obtain a better performance than the CNN-BLSTM model, and the proposed feature-level fusion framework significantly outperforms the CNN-BLSTM and DNN-KELM models by an absolute 6.61% (from 51.12% to 57.73%) and 2.45% (from 55.28% to 57.73%), respectively. The results indicate that the decision-level fusion framework and feature-level fusion framework are effective in IEMOCAP database. In addition, we can observe that the extracted complementary features are more useful for distinguishing emotions than the decision fusion strategy.

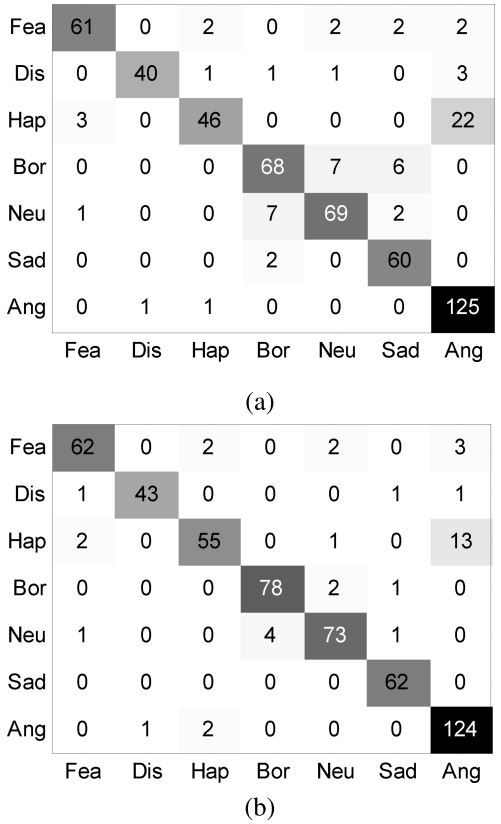


FIGURE 4. Confusion matrices for Emo-DB database. (a) CNN-BLSTM. (b) The fusion framework.

Finally, to analyze the relation between each emotion class, we give the confusion matrices. Fig. 4 and Fig. 5 give the confusion matrices for Emo-DB and IEMOCAP databases, respectively. The abscissa is the detected label and the ordinate is the actual label.

Fig. 4 shows the confusion matrices of CNN-BLSTM and the proposed feature-level fusion framework for Emo-DB database. We can see that much confusion is concentrated between the happiness and anger classes; as seen in Fig. 4(a), there are approximately 30% happiness utterances detected as anger. Although our model makes a great improvement on happiness recognition, there are still approximately 18% happiness utterances detected as anger in Fig. 4(b). We assume the reason is that both happiness and anger have the high energy and arousal [61]. However, there are few anger

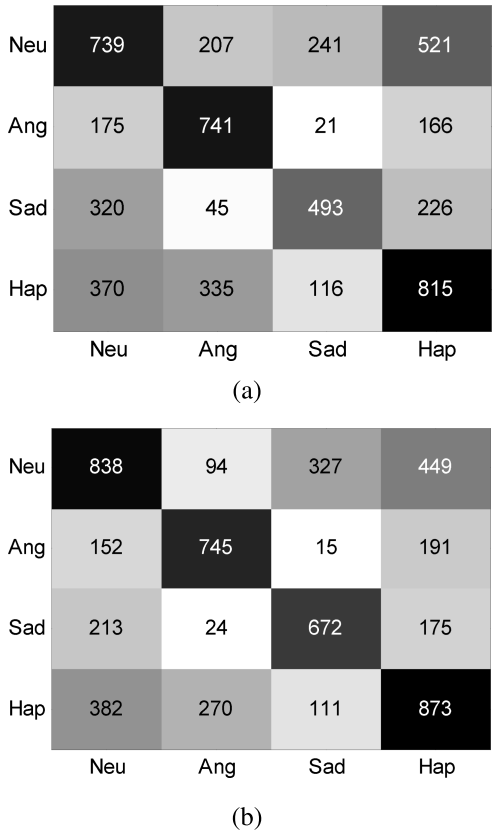


FIGURE 5. Confusion matrices for IEMOCAP database. (a) CNN-BLSTM. (b) The fusion framework.

utterances detected as happiness. This is mainly because the anger utterances have the highest proportion in this database. In addition, we found that the confusion between the boredom and neutral classes is mutual. Fig. 4(a) shows that there are approximately 8.6% boredom utterances detected as neutral and 8.8% neutral utterances detected as boredom. We assume this result is because both the boredom and neutral emotions have the peaceful mood and low arousal [61]; it is difficult to distinguish them, while our method can significantly weaken the confusion between them.

Fig. 5 shows the confusion matrices for IEMOCAP database. Much confusion is concentrated between the happiness and neutral/anger classes. Different from Fig. 4, the confusion between happiness and anger is mutual. Furthermore, the proposed method achieves a higher accuracy than the CNN-BLSTM model in all classes. However, the improvement in the anger emotion is limited. We assume this result is because there is a low percentage of anger utterances in this database.

To summarize, the proposed fusion framework is effective for speech emotion recognition, which indicates that the spectrogram-based features and auditory-based features are complementary to some extent. However, the results for IEMOCAP database are obviously worse than those for Emo-DB database. There are three reasons for this. First, the speech quality is different. The IEMCOAP database contains much noise and silent segments [54]. In addition, the IEMOCAP database contains scripted and improvised utterances, and as the script text exhibits strong correlation with the labeled emotions, it may give rise to lingual content learning, which has a side effect on speech emotion recognition [62]. Finally, there are three different human annotators, which may give rise to different labels. Therefore, some labels are controversial.

V. CONCLUSIONS AND PROSPECTS

In this paper, we focused on improving speech emotion recognition by using complementary features. To utilize the potential advantages of two types of features (i.e., the spectrogram-based statistical features and auditory-based empirical features), we proposed a dynamic fusion framework to extract the complementary features based on spectrograms and the auditory-based features. In addition, the CNN-KELM model was adopted in this work, which utilized the KELM as the classifier to distinguish emotions since the KELM can perform well on small databases. After obtaining the utterance-level feature by using mean-operation, the complementary features were fed into the KELM. In this paper, to build a better fusion framework, we also considered the decision-level fusion strategy. Experiments were conducted on Emo-DB and IEMOCAP databases. The experimental results showed that the CNN-KELM model was effective for speech emotion recognition. Furthermore, the proposed feature-level fusion framework outperformed the decision-level fusion model. This is because a framework using feature-level fusion can capture the mutual correlation among the different types of features. The results also demonstrated that the fusion of these two kinds of features (i.e., spectrogram-based features and auditory-based features) performed better than using either one alone. This results means that these two kinds of features are complementary to some extent. Furthermore, the proposed fusion framework can also be used for other similar tasks such as language recognition, speaker recognition, dialog act detection, and spoken language recognition.

To further improve speech emotion recognition, some aspects of this model should be improved. First, we will have a stricter requirement in selecting the auditory-based features. In addition, as emotional expressions are dynamic, it is important to capture the key emotional segments in the stage of feature extraction.

REFERENCES

[1] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 99–117, 2012.

[2] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. INTERSPEECH*, 2010, pp. 2794–2797.

[3] 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, 2011.

[4] F. Eyben, M. Wöllmer, and B. Schuller, "OpenEAR—Introducing the munich open-source emotion and affect recognition toolkit," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. Workshops*, 2009, pp. 1–6.

[5] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. ICASSP*, May 2011, pp. 5688–5691.
[6] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs," in *Proc. INTERSPEECH*, 2006, pp. 809–812.
[7] H. Hu, M. Xu, and W. Wu, "GMM supervector based SVM with spectral features for speech emotion recognition," in *Proc. ICASSP*, Apr. 2007, pp. 413–416.
[8] S. Ntalampiras and N. Fakotakis, "Modeling the temporal evolution of acoustic parameters for speech emotion recognition," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 116–125, Oct. 2012.
[9] G. Keren and B. Schuller, "Convolutional RNN: An enhanced model for extracting features from sequential data," in *Proc. Int. Joint Conf. Neural Netw.*, 2016, pp. 3412–3419.
[10] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. INTERSPEECH*, Jul. 2014, pp. 223–227.
[11] Z. Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *Proc. ICASSP*, 2017, pp. 5150–5154.
[12] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. INTERSPEECH*, 2015, pp. 1537–1540.
[13] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Proc. INTERSPEECH*, 2018, pp. 152–156.
[14] 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. ICASSP*, 2016, pp. 5200–5204.
[15] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *Proc. ICASSP*, Mar. 2017, pp. 5115–5119.
[16] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. CVPR*, Jun. 2015, pp. 2625–2634.
[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[18] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proc. ICMI*, 2015, pp. 467–474.
[19] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. ICASSP*, Apr. 2015, pp. 4580–4584.
[20] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proc. ACM MM*, 2014, pp. 801–804.
[21] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Proc. Signal Inf. Process. Assoc. Summit Conf.*, Dec. 2017, pp. 1–4.
[22] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. INTERSPEECH*, 2017, pp. 1089–1093.
[23] S. Han, J. Kang, H. Mao, Y. Hu, X. Li, Y. Li, and H. Yang, "Ese: Efficient speech recognition engine with sparse LSTM on FPGA," in *Proc. ACM/SIGDA Int. Symp. Field-Program. Gate Arrays*, 2017, pp. 75–84.
[24] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1062–1087, Nov./Dec. 2011.
[25] W. J. Han, H. F. Li, H. B. Ruan, and L. Ma, "Review on speech emotion recognition," *J. Softw.*, vol. 25, no. 1, pp. 37–50, 2014.
[26] Z. Liu, J. Xu, M. Wu, W. Cao, L. Chen, X. Ding, M. Hao, and Q. Xie, "Review of emotional feature extraction and dimension reduction method for speech emotion recognition," *Chin. J. Comput.*, vol. 40, no. 123, pp. 1–23, 2017.
[27] L. Guo, L. Wang, J. Dang, L. Zhang, and H. Guan, "A feature fusion method based on extreme learning machine for speech emotion recognition," in *Proc. ICASSP*, 2018, pp. 2666–2670.
[28] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.
[29] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2005, pp. 985–990.
[30] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, 2006.
[31] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015.
[32] J. Cao, K. Zhang, M. Luo, C. Yin, and X. Lai, "Extreme learning machine and adaptive sparse representation for image classification," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 91–102, 2016.
[33] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. ICASSP*, May 2013, pp. 3687–3691.
[34] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.
[35] 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.
[36] F. Grezl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *Proc. ICASSP*, Mar. 2008, pp. 4729–4732.
[37] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Proc. INTERSPEECH*, Florence, Italy, Aug. 2011, pp. 237–240.
[38] 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.
[39] 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, Oct. 2017.
[40] L. B. Almeida, "The fractional Fourier transform and time-frequency representations," *IEEE Trans. Signal Process.*, vol. 42, no. 11, pp. 3084–3091, Nov. 1994.
[41] T. Thireou and M. Reczko, "Bidirectional long short-term memory networks for predicting the subcellular localization of eukaryotic proteins," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 4, no. 3, pp. 441–446, Jul. 2007.
[42] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. INTERSPEECH*, 2014, pp. 338–342.
[43] K. Greff, R. K. Srivastava, J. Koutnìk, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.
[44] Y. Kim and E. M. Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," in *Proc. ICASSP*, May 2013, pp. 3677–3681.
[45] E. M. Provost, "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," in *Proc. ICASSP*, May 2013, pp. 3682–3686.
[46] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. ACM MM*, 2010, pp. 1459–1462.
[47] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. INTERSPEECH*, 2009, pp. 312–315.
[48] G. E. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, and T. Sainath, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Oct. 2012.
[49] G. Feng, G.-B. Huang, Q. Lin, and R. Gay, "Error minimized extreme learning machine with growth of hidden nodes and incremental learning," *IEEE Trans. Neural Netw.*, vol. 20, no. 8, pp. 1352–1357, Aug. 2009.
[50] Q.-Y. Zhu, A. K. Qin, P. N. Suganthan, and G.-B. Huang, "Evolutionary extreme learning machine," *Pattern Recognit.*, vol. 38, no. 10, pp. 1759–1763, Oct. 2005.
[51] G.-B. Huang and C.-K. Siew, "Extreme learning machine with randomly assigned RBF kernels," *Int. J. Inf. Technol.*, vol. 11, no. 1, pp. 16–24, 2005.

[52] G.-B. Huang, N.-Y. Liang, H.-J. Rong, P. Saratchandran, and N. Sundararajan, "On-line sequential extreme learning machine," in *Proc. IASTED Int. Conf. Comput. Intell.*, Calgary, AB, Canada, Jul. 2005, pp. 232–237.
[53] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, Lisbon, Portugal, Sep. 2005, pp. 1517–1520.
[54] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
[55] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, 2009, pp. 552–557.
[56] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, "Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition," in *Proc. INTERSPEECH*, Lisbon, Portugal, Sep. 2018, pp. 272–276.
[57] B. Schuller, R. Müller, B. Hörnler, A. Höthker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *Proc. ICMI*, Aichi, Japan, 2007, pp. 30–37.
[58] L. van der Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2431–2456, Nov. 2008.
[59] L. van der Maaten, "Fast optimization for t-SNE," in *Proc. Neural Inf. Process. Syst.*, Lisbon, Portugal, Sep. 2010, pp. 1–5.
[60] C. Busso, A. Metallinou, and S. S. Narayanan, "Iterative feature normalization for emotional speech detection," in *Proc. ICASSP*, May 2011, pp. 5692–5695.
[61] E. Schubert, "Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space," *Austral. J. Psychol.*, vol. 51, no. 3, pp. 154–165, 1999.
[62] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *Proc. INTERSPEECH*, Lisbon, Portugal, Sep. 2018, pp. 3087–3091.

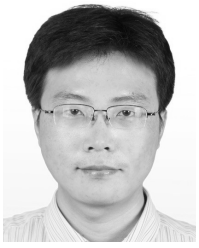
LILI GUO received the B.S. degree from the School of Computer Science and Technology, Linyi University, Linyi, China, in 2013, and the M.S. degree from the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China, in 2016. She is currently pursuing the Ph.D. degree with the College of Intelligence and Computing, Tianjin University, Tianjin, China. Her research interests are in the fields of speech emotion recognition, deep learning, and acoustic signal processing.

LONGBIAO WANG received the B.E. degree from Fuzhou University, Fuzhou, China, in 2000, and the M.E. and Dr. Eng. degrees from the Toyohashi University of Technology, Toyohashi, Japan, in 2005 and 2008, respectively.

He was an Assistant Professor with the Faculty of Engineering, Shizuoka University, Japan, from 2008 to 2012. He was an Associate Professor with the Nagaoka University of Technology, Japan, from 2012 to 2016. Since 2016, he has been a Professor with Tianjin University, Tianjin, China. He has also been a Visiting Professor with the Japan Advanced Institute of Science and Technology, since 2017. He has published more than 100 journals and international conference papers. His research interests include robust speech recognition, speaker recognition, and acoustic signal processing.

JIANWU DANG received the B.E. and M.E. degrees from Tsinghua University, China, in 1982 and 1984, respectively, and the Ph.D. degree from Shizuoka University, Japan, in 1992.

He was with Tianjin University, Tianjin, China, as a Lecturer, from 1984 to 1988. From 1992 to 2001, he was with ATR Human Information Processing Laboratories, Japan. Since 2001, he has been on the faculty of the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), as a Professor. He joined the Institute of Communication Parlee (ICP), Center of National Research Scientific, France, as a Research Scientist (first class), from 2002 to 2003. Since 2009, he has also been with Tianjin University. He has published more than 300 journal and international conference papers. His research interests are in all the fields of speech production, speech synthesis, speech recognition, and spoken language understanding. He constructed MRI-based physiological models for speech and swallowing and endeavors to apply these models on clinics.



ZHILEI LIU received the Ph.D. degree in computer science from the University of Science and Technology of China, in 2014. In 2013, he visited the Intelligent Systems Laboratory, Rensselaer Polytechnic Institute, USA. In 2017, he joined Nanyang Technological University, Singapore, as a Research Fellow. He is currently an Assistant Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. His research interests include multimedia computing, affective computing, machine learning, and pattern recognition.

HAOTIAN GUAN received the B.S. degree from the School of Information Science and Engineering, Shenyang University of Technology, Shenyang, China, in 2015, and the M.S. degree from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2018. He is currently with Huiyan Technology (Tianjin) Co., Ltd. as an Engineer. His research interests are in the fields of speech emotion recognition and speech synthesis.