IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 16, 2021, accepted March 25, 2021, date of publication March 30, 2021, date of current version April 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3069818

Autoencoder With Emotion Embedding for Speech Emotion Recognition

CHENGHAO ZHANG AND LEI XUE
School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China
Corresponding author: Lei Xue (16301098@163.com)

ABSTRACT An important part of the human-computer interaction process is speech emotion recognition (SER), which has been receiving more attention in recent years. However, although a wide diversity of methods has been proposed in SER, these approaches still cannot improve the performance. A key issue in the low performance of the SER system is how to effectively extract emotion-oriented features. In this paper, we propose a novel algorithm, an autoencoder with emotion embedding, to extract deep emotion features. Unlike many previous works, instance normalization, which is a common technique in the style transfer field, is introduced into our model rather than batch normalization. Furthermore, the emotion embedding path in our method can lead the autoencoder to efficiently learn a priori knowledge from the label. It can enable the model to distinguish which features are most related to human emotion. We concatenate the latent representation learned by the autoencoder and acoustic features obtained by the openSMILE toolkit. Finally, the concatenated feature vector is utilized for emotion classification. To improve the generalization of our method, a simple data augmentation approach is applied. Two publicly available and highly popular databases, IEMOCAP and EMODB, are chosen to evaluate our method. Experimental results demonstrate that the proposed model achieves significant performance improvement compared to other speech emotion recognition systems.

INDEX TERMS Speech emotion recognition, autoencoder, emotion embedding, instance normalization.

I. INTRODUCTION

In human speech interaction, people convey the underlying intent through paralinguistic characteristics such as emotions, intonations and styles. Therefore, speech emotion recognition (SER), the technique of recognizing emotions from speech, has gradually become a significant research interest. This technology has promising prospects and plays an important role in natural language understanding. For example, emotion recognition has been widely used in the process of human-computer interaction (HCI) and computer-dedicated human communication [1]. Recognizing these paralinguistic characteristics can help intelligent systems understand user intention and further improve the user experience. In this paper, an algorithm that analyzes the underlying emotions of speech with a deep learning algorithm is proposed.

Human emotions in speech are complex to model. The main reasons are as follows: 1) human emotions may be treated as noise and discarded in many current speech recognition methods due to their abstraction. 2) in general, human emotion in a long utterance can only be detected in some specific moments [2]. Early work on SER mainly focused on selecting speech acoustic features that can distinguish different emotions, such as statistical features and prosodic features. The most common approach is to extract a large number of statistical features from utterances and utilize basic machine learning algorithms (e.g., hidden Markov model (HMM) [3], Gaussian mixture model (GMM) [4], and support vector machine (SVM) [5]). Recently, with the increased interest in deep learning (DL) algorithms, the automatic extraction of useful features from speech signals by deep neural networks (DNNs), such as recurrent neural networks (RNNs) [6] and convolutional neural networks (CNNs) [7], has become a very popular technique. Prior researchers used DNNs have demonstrated that deep learning has the most promising results compared with traditional algorithms.

Encouraged by the recent success of autoencoder structures [19], [20] with deep unsupervised learning and the idea of word embedding [8], [9] in natural language processing (NLP), we propose a novel algorithm based on an autoencoder with emotion embedding for SER. The key

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

contributions of the proposed SER system are illustrated next.

1) A novel autoencoder structure, an autoencoder with emotion embedding, is proposed to improve both the learning and generalization capacities of deep networks. Emotion embedding in our method can lead the model to efficiently learn a priori emotion information from the label.
2) Encouraged by the fact that instance normalization (IN) can learn features that are invariant to appearance changes, it is introduced into our autoencoder.
3) In the emotion classification process, we concatenate the deep emotion feature extracted by the autoencoder and the IS10 feature set obtained by the openSMILE toolkit.
4) We tested our suggested SER model on benchmarks, which included the IEMOCAP and EMODB datasets. It achieved 71.2% and 95.6% recognition results. In the comparative analysis, our system showed outperformed recognition results.

The rest of this paper is organized as follows. Section II describes the related algorithm of the autoencoder. Section III presents our proposed novel algorithm in detail. Section IV shows the experimental details and databases. Section V demonstrates the experimental results.

II. RELATED WORK

Speech emotion recognition is considered a challenging task in the HCI domain. A large number of methodologies and corpora have been proposed in previous works [10]–[12]. The early stage of SER research used handcrafted speech features and low-level descriptors to train classic machine learning models. Recently, increasing attention has been drawn to the study of DNNs. However, there are two major issues observed in DL approaches: (1) a sufficient amount of labeled speech data. (2) extracting emotion related features from audio.

To address the scarcity of training data, multiple methodologies have been investigated. Generally, there are three approaches to address this obstacle. (1) Collecting and annotating new data. However, it is expensive and consumes much time to create a large enough dataset. (2) Data augmentation. This is the most common method that has been widely used in the DL field [13], [14]. (3) Transfer learning. This method is a popular research problem in DL that focuses on storing knowledge gained while training one model and applying it to another task. It has been successfully applied in various domains [15]–[18]. However, the mismatch between the datasets is the reason why the accuracy of the SER system has not been further improved. In this paper, a simple data augmentation method is applied to the proposed SER algorithm.

In recent years, to strengthen the feature extraction ability, many improvements to DL algorithms have been proposed. With the successful application of autoencoders in the DL field, they have also been introduced into SER tasks. An autoencoder is an unsupervised learning model used to reconstruct the input with minimum reconstruction error. The basic autoencoder has one input layer, one hidden layer and one output layer. The autoencoder first maps the input vector to the best latent representation through nonlinear mapping, and then this representation is mapped back to the output layer to reconstruct the input vector. If the number of hidden layers is greater than one, the network is considered to be deep. Many previous works directly utilized the latent representation learned by basic autoencoders for SER tasks. For example, in [21], a deep autoencoder based on a multilayer perceptron was proposed for SER. Pal and Baskar [22] proposed a deep dropout autoencoder based multilayer perceptron. Similar to [21] and [22], an autoencoder was also applied to extract the bottleneck features for dimensionality reduction in [23] and [24]. Finally, some machine learning algorithms, such as SVM and long short-term memory (LSTM), were applied for emotion classification.

Moreover, to extract more robust features, a denoising autoencoder (DAE) and its deep structure, stacked denoising autoencoders, are introduced into the SER field. The major difference between DAE and traditional autoencoders is that DAE is trained to recover from corrupted inputs. Encouraged by the motivation behind this, Ghosh *et al.* [25] explored stacked DAEs for representation learning. Furthermore, Zhang *et al.* [26] proposed a memory-enhanced recurrent denoising autoencoder (rDA) that has shown that this method can significantly improve the performance.

In the aforementioned methods, the autoencoder is directly trained to learn a lower-dimensional distributed representation of the input data. However, one may note that the representation learned by basic autoencoder architecture also contains redundant information not related to human emotions. To address this problem, many researchers have explored many modified autoencoder networks. Xia and Liu [27] proposed a modified autoencoder method to project the input to two hidden spaces. One of them is meant to represent emotional information, whereas the other is used to capture redundant information. Deng *et al.* [28] proposed a shared hidden-layer autoencoder (SHLA) model for learning common feature representations shared across the training and test sets to reduce the discrepancy in them. In addition, Zong *et al.* [29] proposed a novel framework named multichannel autoencoder (MTC-AE) for emotion recognition. MTC-AE contains multiple local DNNs based on different low-level descriptors with different statistical functions that are partly concatenated together, by which the structure is enabled to consider both local and global features simultaneously. Wei *et al.* [30] proposed an algorithm based on an autoencoder, denoising autoencoder, and sparse autoencoder. The first layer of the structure uses a denoising autoencoder to learn a hidden feature with a larger dimension than the dimension of the input features, and the second layer employs a sparse autoencoder to learn sparse features.

Obviously, even if such methods can further improve the performance of SER, the high-level features extracted by reconstructing input mainly contain content information rather than emotion-oriented features. Moreover, the

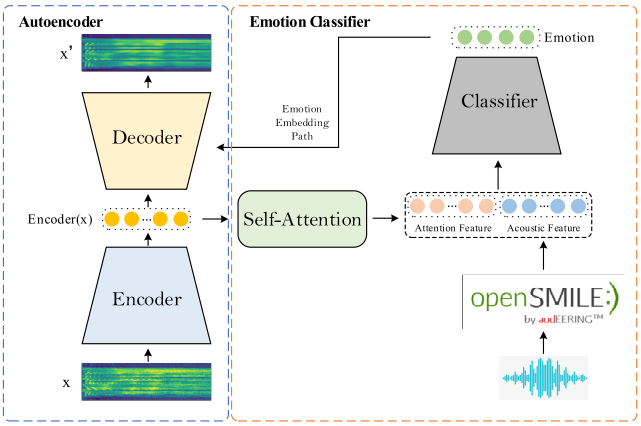

FIGURE 1. The framework of our proposed method.

FIGURE 2. Data augmentation.

TABLE 1. Encoder network details.

Conv1s	Conv1-1	513, 128, kernel_size=1
	Conv1-2	513, 128, kernel_size=2
	Conv1-3	513, 128, kernel_size=3
	Conv1-4	513, 128, kernel_size=4
	Conv1-5	513, 128, kernel_size=5
	Conv1-6	513, 128, kernel_size=6
	Conv1-7	513, 128, kernel_size=7
	Conv1-8	513, 128, kernel_size=8
Conv2		1537, 512, kernel_size=1
Conv3		512, 512, kernel_size=5
Conv4		512, 512, kernel_size=5, stride=2
Conv5		512, 512, kernel_size=5
Conv6		512, 512, kernel_size=5, stride=2
Conv7		512, 512, kernel_size=5
Conv8		512, 512, kernel_size=5, stride=2
FC1		512, 512
FC2		512, 512
FC3		512, 512
FC4		512, 512
FC5		768, 512
Dropout1-6		p=0.5
GRU		input_size=512, hidden_size=128 bidirectional

above-mentioned works do not consider the significance of a priori knowledge. To address this problem, our method does not rely on basic autoencoder architecture. In this work, we increase the modeling capacity by designing a new autoencoder architecture, emotion-embedded autoencoder. Emotion embedding layers in our method lead the model to efficiently learn a priori emotion information from the label, which allows the autoencoder to focus more on deep emotion features during the reconstruction process. Experimental results demonstrate that the proposed method can present performance improvement.

III. PROPOSED METHOD

In this section, we describe our proposed method. There are three parts: input speech feature, autoencoder with emotion embedding and emotion classification network. Fig. 1 depicts the model framework, which includes an autoencoder, an emotion embedding path, and an emotion classification net. Let us consider a dataset with N labeled samples $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ and M unlabeled samples $\{(x_{N+1}), (x_{N+2}), (x_{N+3}), \cdots, (x_{N+M})\}$, where x_i is denoted as the ith acoustic feature sequence of the speech sample and y_i is the emotion label corresponding to x_i. $y_i \in \{1, 2, 3, \cdots, K\}$, and K is the number of emotion categories.

A. INPUT SPEECH FEATURE

1) LOG MAGNITUDE SPECTROGRAM

A spectrogram is a useful expression for the analysis of speech and audio signals. Many applications are performed in the spectral domain using spectrograms rather than in the original time domain. Furthermore, the magnitude spectrograms of audio signals tend to be highly structured in terms of both spectral and temporal regularities [31]. Therefore, it is easier to deal with many problems by processing magnitude spectrograms than directly processing time-domain signals. In fact, magnitude spectrograms have been introduced into many audio processing fields including audio source separation and speech synthesis systems [32]–[36]. In this paper, the detailed spectral analysis was the same as in previous work [36].

2) IS10 FEATURE SET

We utilize the openSMILE [37] toolkit to extract statistical features that were used in the INTERSPEECH 2010 Paralinguistic Challenge [38]. The open-source media interpretation by large feature-space extraction (openSMILE) toolkit is a modular tool for signal processing and machine learning applications. It can flexibly extract the features of signals and is mainly used for audio signal feature extraction. Therefore, 1582-demensional features are generated by extracting 38 kinds of LLDs and applying 21 statistical functions. Details about these features can be found in [38].

B. AUTOENCODER WITH EMOTION EMBEDDING

In this part, we interpret the complete scheme of the autoencoder with emotion embedding in detail. Fig. 3-4 and Table 1-2 depict the detailed autoencoder architecture. In this letter, due to the 2D representation of the spectrogram, our proposed autoencoder is mainly based on a CNN, and it contains two parts: encoder and decoder. Each part mainly consists of four building blocks: convolution parts, instance normalization, dropout layers and gated recurrent

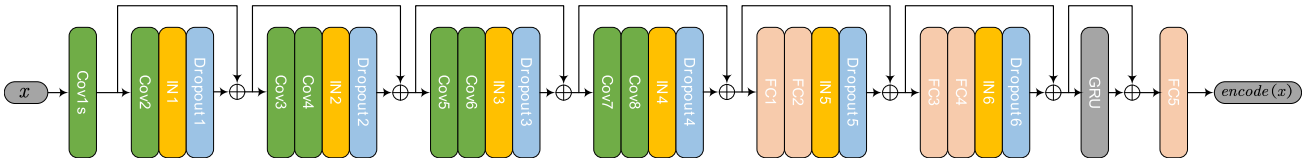

FIGURE 3. Encoder network architecture.

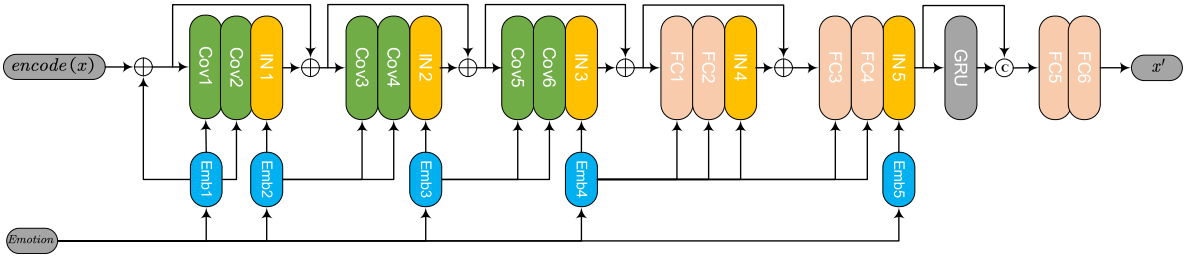

FIGURE 4. Decoder network architecture. Emotion means emotion label. Emb1-5 are the emotion embedding layers, and they are trained to map an emotion to a 512-dimensional latent representation.

TABLE 2. Decoder network details.

Conv1	512, 1024, kernel_size=3
Conv2	512, 512, kernel_size=3
Conv3	512, 1024, kernel_size=3
Conv4	512, 512, kernel_size=3
Conv5	512, 1024, kernel_size=3
Conv6	512, 512, kernel_size=3
FC1	512, 512
FC2	512, 512
FC3	512, 512
FC4	512, 512
FC5	1536, 512
FC6	512, 513
Emb1-5	(number of classes, 512)
GRU	input_size=512, hidden_size=256, bidirectional

unit (GRU). CNNs are one of the most popular deep learning models that have demonstrated great success in various research fields. In SER, CNNs have also been widely used to learn salient features, also directly for classification. Generally, the basic components of a CNN are convolution layers, pooling layers, batch normalization (BN) and activation layers. In our method, pooling layers are discarded since we do not expect to lose any high-level information that may be related to emotions. In addition, we replace BN with instance normalization (IN) [39] in our network. BN [40] is one of the most common components in many CNNs, and it normalizes the features by the mean and variance computed within a batch. It enables a larger learning rate and faster convergence by reducing the internal covariate shift during the training process. Unlike BN, the key difference between BN and IN is that the latter applies normalization to an individual sample instead of a whole batch of samples. Generally, IN is mainly used in the style transfer field, for instance, image style transfer [41]. Some methods [42], [43] employ IN to help remove image contrast. Moreover, many existing works disclose that IN learns features that are invariant to appearance changes, such as colors, styles, and virtuality, while BN is essential for preserving content related information [44]. To this end, we introduce IN into our autoencoder network with the purpose of leading the model to attract more attention to features related to emotion while maintaining discrimination of the learned features. In addition, considering that emotions in speech are context-dependent, the ability to model contextual information makes RNNs suitable for SER. Therefore, GRU [45], which is a special case of RNN, is utilized in our network. Finally, residual connection [46] are utilized in the network to address vanishing/exploding gradients.

In the encoder process, the encoder is trained to map an input sequence x to a latent representation $encoder(x)$. In the reconstruction process, the decoder network is equipped with the emotion embedding path, leading the model to efficiently learn emotion information from the label, as shown in Fig. 4. The fact that the latent representation from conventional autoencoder learned by reconstructing input mainly contains content information is the reason why the SER accuracy in many works has not been further improved. The main attraction of emotion embedding for SER is that it allows the network to distinguish which deep features are related to emotion. In this paper, the decoder is trained to generate x' which is a reconstruction of x from $encoder(x)$ given the emotion label y, as shown in (1).

$$x' = decoder(encoder\,(x)\,, y) \tag{1}$$

The mean absolute error (MAE) is used as the reconstruction loss since it generates a sharper output than the mean square error [47], as shown in (2).

$$L_{AE}\,(\theta_{Enc}, \theta_{Dec}) = \sum\nolimits_{(x,y)\epsilon D} \left\| x' - x \right\|_1 \tag{2}$$

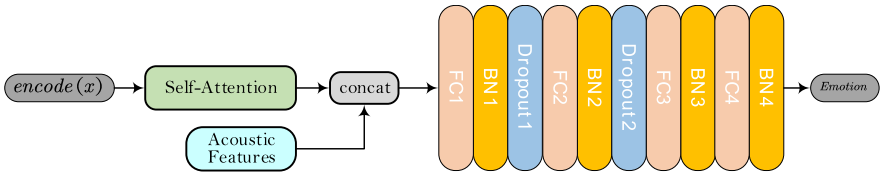

FIGURE 5. Emotion classifier architecture.

TABLE 3. Emotion classifier network details.

Self-attention	num_heads	1 (Best Performance)
Classifier	FC1	2478, 1024
	FC2	1024, 512
	FC3	512, 128
	FC4	128, number of classes
	Dropout1-2	p=0.5

where θ_{Enc} and θ_{Dec} are the parameters of the encoder and decoder respectively.

C. EMOTION CLASSIFICATION WITH FEATURE FUSION

In the classification process, the classification network takes the encoder's output and learns the links between it and the emotion label, as shown in Fig. 5 and Table 3. The best representation from the encoder is fed into the self-attention [48] layer first, and the details of the attention layer are the same as in previous work [48]. With the attention mechanism, the network can focus more on the emotion-oriented feature in the best representation obtained by the autoencoder.

Moreover, while the progressive downsampling of CNNs provides strong capability in local context modeling and emotion-related pattern detection, Li *et al.* [49] believed that the temporal structure of speech that is highly related to emotions will gradually be lost in the downsampling process [50]. To overcome this problem, we concatenate the deep attention emotion feature extracted from the attention layer and acoustic features obtained by the openSMILE toolkit. These features contain global information of speech. Finally, the concatenated feature vector is fed into the fully connected network for emotion classification.

The emotion classification network takes the concatenated feature vector as input and outputs the predicted emotion class. The classifier is trained to minimize the negative log-probability, as shown in (3).

$$L_{EC}(\theta_{Att}, \theta_{Cla}) = \sum\nolimits_{(x,y)\epsilon D} -logP_{EC}(y|encoder(x)) \quad (3)$$

where θ_{Att} and θ_{Cla} are the parameters of the attention layer and classification network, respectively.

During the training process, the object function of our network is a joint function decided by both reconstruction error and the negative log-probability:

$$L_{Total}\left(\theta_{Total}\right) = L_{AE}\left(\theta_{Enc}, \theta_{Dec}\right) + \lambda L_{EC}\left(\theta_{Att}, \theta_{Cla}\right) \quad (4)$$

where λ is a constant controlling the weighting between the encoder path and the classify path.

IV. EXPERIMENT

A. DATA AUGMENTATION

Currently, there are two common problems with datasets in the SER field: the typical inherent mismatch between the dataset and the difficulty in creating corpora. This mismatch means different emotion annotation schemes in different datasets, and high data collection often comes with high annotation costs. Therefore, one of the serious obstacles to the applications of SER systems in real-life settings is the lack of a sufficient amount of labeled speech data. Inspired by data augmentation, in this paper, a simple data augmentation method is presented to make use of data effectively. Data augmentation has been proposed as a method to generate additional training data for computer vision. Artificial data have also been augmented for many previous works in automatic speech recognition (ASR) [51], [52]. In [53], Navdeep Jaitly *et al.* proposed a method named vocal tract length normalization for data augmentation. The approach of superimposing clean audio with a noisy audio signal was adapted in [54]. In LVSCR tasks [55], they applied speed perturbation in their work. In addition, the use of an acoustic room simulator [56] and generative adversarial networks (GANs) [57] have also been proposed for data augmentation. However, the aforementioned approaches all operated on the raw audio itself rather than the spectrogram. In [58], D. S. Park *et al.* proposed a simple and computationally cheap method for data augmentation, which directly acted on the log mel spectrogram and did not require any additional data. Three deformations of the spectrogram were chosen in their work: time warping, frequency masking and time masking. More generally, many works have demonstrated that data augmentation techniques have achieved state-of-the-art performance in ASR. In this paper, to avoid losing local context information of the speech signal, we randomly sampled 128 frames of log magnitude spectrogram with overlap. This means that 128 consecutive time steps $[t, t + 128)$ are termed a training sample, where t is chosen from a uniform distribution $[0, T - 128)$, and T is the length of the log magnitude spectrogram, as shown in Fig. 2.

B. DATASET

To investigate the performance of the proposed method, two publicly available and highly popular databases,

namely the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [59] and Berlin Emotional Speech Database (EMODB) [60] are chosen as source sets. We first briefly introduce the database.

1) IEMOCAP

IEMOCAP was collected by SAIL lab at USC, USA, and it consists of 5 sessions. It has 10 professional actors (5 male and 5 female) acting in two different scenarios: scripted play and spontaneous dialog. This corpus has approximately 12 hours of audiovisual data, including video, speech, motion capture of face, and text transcription. Each interaction is segmented into sentences that are labeled by at least 3 annotators. In this paper, we used four emotion categories: angry, happy, sad and neutral. Note that, like many previous works, happy and excited in the original annotation were merged into one class: happy. Only the audio signals were used in the experiments.

2) EMODB

This dataset was collected by the Institute of Communication Science at the Technical University of Berlin. It was spoken by 10 actors and comprises 535 utterances divided into seven emotion classes, namely, anger, fear, happiness, sadness, disgust, boredom, and neutral. We use all data in this study.

C. EXPERIMENT SETUP

Since there were 10 speakers in IEMOCAP and each session consisted of 2 speakers, leave-one-speaker-out cross-validation was applied in our experiments so that there was no speaker overlap between the training and test data. Moreover, 10-fold cross-validation strategies were also used to evaluate the proposed method. For EMODB, we performed all evaluations using 5-fold and 10-fold cross-validation, to stay in the same manner as most approaches. For performance comparison, we utilize unweighted accuracy (UA) and weighted accuracy (WA) [61], which have been used in several previous emotion challenges. Weighted accuracy is the accuracy over all testing utterances in the dataset, and unweighted accuracy is the average accuracy over each emotion category. They are quite good measurements in this case since the class distribution is imbalanced. In addition, to further measure the proposed method, the metrics of precision, recall and F1-score are also computed.

We used sampled log magnitude spectrogram as the inputs, and trained the network using Adam optimizer with $lr = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The model was trained for 40 epochs on the dataset. All the experiments were performed using an Nvidia GTX 1080Ti with 11 GB memory.

V. RESULTS AND ANALYSIS

A. IMPACTS OF INSTANCE NORMALIZATION AND EMOTION EMBEDDING

In this part, the IS10 feature set is combined with the traditional machine learning algorithm SVM (IS10+SVM) to serve as a comparison baseline. Moreover, to verify that our modified autoencoder ($AE_{+EE,IN}$) can efficiently improve SER performance, contrast experiments are performed on three different models (AE, AE_{+EE}, AE_{+IN}). In addition, to better verify the effectiveness of emotion embedding, two models (AE_{-OS}, $AE_{-OS,+EE,IN}$) without the usage of the external openSMILE toolkit are performed.

TABLE 4. The impacts of emotion embedding and instance normalization.

Method	emotion embedding	instance normalization	UA	
			10-fold	LOSO
IS10+SVM	--	--	--	58.6
AE_{-OS}	×	×	66.17	59.27
$AE_{-OS,+EE,IN}$	✓	✓	68.10	62.11
AE	×	×	69.98	63.9
AE_{+EE}	✓	×	71	64.9
AE_{+IN}	×	✓	70.55	66.34
$AE_{+EE,IN}$	✓	✓	**71.2**	**66.7**

TABLE 5. Performance (%) measure for each emotion on IEMOCAP (10-fold LOSO).

Emotion	Precision	Recall	F1
anger	72.87	70.90	71.64
happiness	66.12	59.79	62.23
neutral	63.37	56.41	59.14
sadness	58.95	77.01	66.10

TABLE 6. Performance (%) measure for each emotion on IEMOCAP (10-fold).

Emotion	Precision	Recall	F1
anger	75.69	74.65	74.94
happiness	69.75	65.68	67.53
neutral	68.57	66.30	67.20
sadness	69.10	78.20	73.22

TABLE 7. Performance (%) measure for each emotion on EMODB (5-fold).

Emotion	Precision	Recall	F1
anger	90.44	92.57	91.43
boredom	97.00	90.84	93.41
disgust	96.00	93.81	94.42
fear	94.07	95.37	94.64
happiness	88.41	77.77	82.64
neutral	90.80	96.92	93.35
sadness	92.29	97.01	94.21

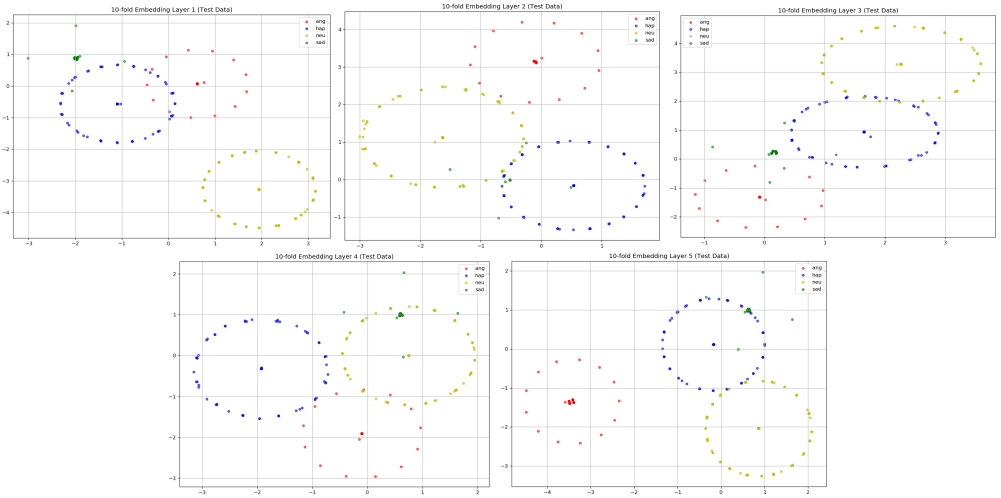

FIGURE 6. T-SNE visualization of emotion embedding on IEMOCAP (10-fold).

TABLE 8. Performance (%) measure for each emotion on EMODB (10-fold).

Emotion	Precision	Recall	F1
anger	93.96	96.40	95.13
boredom	96.78	95.79	96.10
disgust	96.00	95.00	94.59
fear	95.66	98.33	96.73
happiness	94.48	87.86	90.57
neutral	98.75	93.64	95.58
sadness	97.22	100.00	98.50

TABLE 9. Confusion matrix on IEMOCAP (10-fold LOSO).

	anger	happiness	neutral	sadness
anger	70.90	14.99	11.41	2.70
happiness	10.28	59.79	16.79	13.14
neutral	8.46	17.38	56.41	17.75
sadness	1.60	6.33	12.29	79.78

TABLE 10. Confusion matrix on IEMOCAP (10-fold).

	anger	happiness	neutral	sadness
anger	74.65	13.75	9.13	2.46
happiness	8.78	65.68	15.27	10.26
neutral	6.54	14.58	66.30	12.58
sadness	2.27	6.90	12.63	78.20

Table 4 shows the performance of different classifiers on the IEMOCAP speech database. For 10-fold cross validation, the UA obtained with $AE_{-OS,+EE,IN}$ is improved by 1.84% compared with AE_{-OS}. Furthermore, the UA obtained with the proposed method is improved by 1.22%, 0.2% and 0.65% compared with AE, AE_{+EE} and AE_{+IN}, respectively. For 10-fold leave-one-speaker-out cross validation, the UA obtained with $AE_{-OS,+EE,IN}$ is improved by 2.93% compared with AE_{-OS}. Meanwhile, the UA obtained with the proposed method is improved by 8.1%, 1.24%, 1.8% and 0.36% compared with IS10+SVM, AE, AE_{+EE} and AE_{+IN}, respectively. In summary, the performance of SER is further improved by introducing emotion embedding and instance normalization.

B. VISUALIZING THE EMOTION EMBEDDING USING T-DISTRIBUTED NEIGHBOR EMBEDDING (T-SNE)

T-SNE is an algorithm developed for visualizing multidimensional data based on the idea of dimensionality reduction. We use t-SNE plots to visualize the emotion embedding of our modified autoencoder model. In our method, there are five emotion embedding layers in the decoder network, as shown in Fig. 4. Each test sample is now a multidimensional data point (512 dimensions). T-SNE was then used to reduce the dimensions to only two for a 2D plot, as shown in Fig. 6 and Fig. 7. From Fig. 6 and Fig. 7, we can clearly see the separation between ''ang'', ''hap'' and ''neu''. Such a result is expected since there are obviously different characteristics between them. However, we can also see that it is not clearly separated between ''sad'' and the other three emotions. One possible explanation is that the low-energy state of the emotion ''sad'' does not have salient characteristics compared with the other emotions. In summary, the experimental results demonstrate that our proposed autoencoder naturally learns useful emotion representations from the label; in turn, the learning process discovers the intrinsic attributes necessary to solve emotion recognition.

C. THE PERFORMANCE OF THE PROPOSED METHOD

Tables 5-8 show the metrics of precision, recall and F1-score of our model on two datasets. From Tables 5-8, it can be

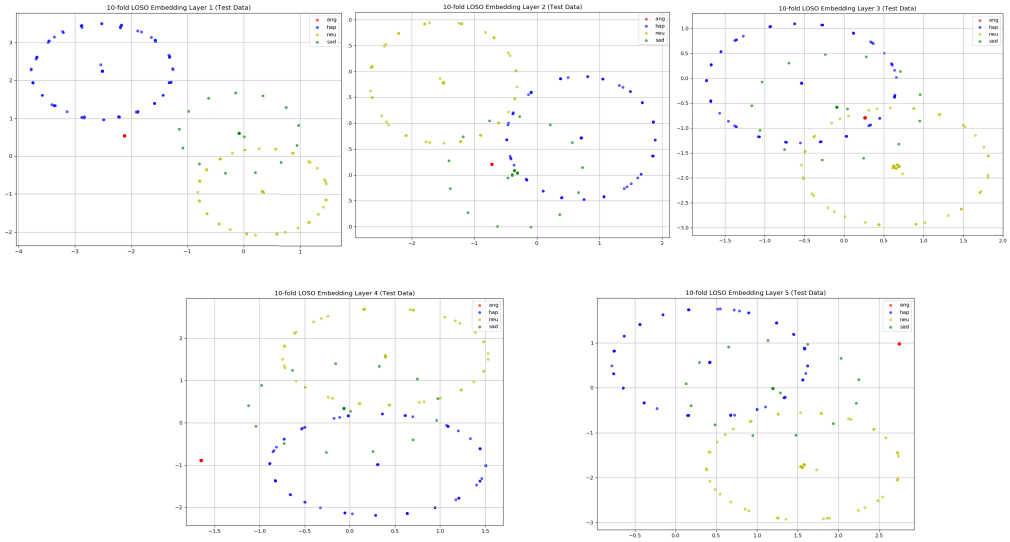

FIGURE 7. T-SNE visualization of emotion embedding on IEMOCAP (10-fold LOSO).

TABLE 11. Confusion matrix on IEMOCAP (5-fold).

	anger	boredom	disgust	fear	happiness	neutral	sadness
anger	92.57	0.00	0.83	0.77	5.82	0.00	0.00
boredom	0.00	90.84	1.43	0.00	0.00	6.40	1.33
disgust	0.00	0.00	93.81	3.33	0.00	2.86	0.00
fear	0.00	0.00	0.00	95.37	0.00	1.43	3.21
happiness	17.64	0.00	0.00	3.48	77.77	1.11	0.00
neutral	0.00	3.08	0.00	0.00	0.00	96.92	0.00
sadness	0.00	1.18	0.00	0.00	0.00	1.82	97.01

TABLE 12. Confusion matrix on IEMOCAP (10-fold).

	anger	boredom	disgust	fear	happiness	neutral	sadness
anger	96.40	0.00	1.11	0.00	2.49	0.00	0.00
boredom	0.00	95.79	1.43	0.00	0.00	1.67	1.11
disgust	0.00	1.67	95.00	3.33	0.00	0.00	0.00
fear	0.00	0.00	0.00	98.33	0.00	0.00	1.67
happiness	10.14	0.00	0.00	2.00	87.86	0.00	0.00
neutral	0.00	3.25	0.00	1.11	2.00	93.64	0.00
sadness	0.00	0.00	0.00	0.00	0.00	0.00	100.00

seen that ''anger'' and ''sadness'' are easier to distinguish than other emotions. In the EMODB dataset, ''boredom'', ''disgust'' and ''fear'' can also be recognized well. However, ''happiness'' and ''neutral'' are the most difficult to identify.

To further observe the performance of the proposed method for each emotion, we also present the classification confusion matrix, as shown in Tables 9-12. These results correspond to the IEMOCAP (10-fold, 10-fold LOSO) and EMODB (5-fold, 10-fold) datasets, respectively. Tables 9-12 clearly indicate that recognizing ''happiness'' is the most difficult task since ''happiness'' is easily confused with other emotions. In addition, the recognition accuracy of ''neutral'' in different datasets varies greatly. This is because the data distribution of each dataset is different.

To evaluate the superiority of the proposed method, Tables 13 and 14 present the performance comparisons of

TABLE 13. Performance comparisons with other methods (IEMOCAP).

Method	Validation Setting	UA
[62]	10-fol d LOSO	60.9
[63]	10-fold LOSO	57.4
[64]	10-fold	61.3
[64]	10-fold LOSO	62.4
[29]	10-fold LOSO	64.8
[65]	10-fold LOSO	64.2
[66]	10-fold LOSO	62.8
[67]	10-fold LOSO	59.54
[68]	10-fold	68.8
Proposed Method	10-fold LOSO	**66.7**
Proposed Method	10-fold	**71.2**

TABLE 14. Performance comparisons with other methods (EMODB).

Method	Validation Setting	WA
[69]	10-fold	74.6
[18]	10-fold	72.4
[70]	10-fold	83.42
[71]	10-fold	78.48
[72]	5-fold	76.06
Proposed Method	5-fold	**92.2**
Proposed Method	10-fold	**95.6**

different methods on the IEMOCAP and EMODB datasets. From Tables 13 and 14, the best performance of our proposed method is 71.2% for IEMOCAP and 95.6% for EMODB. Additionally, we find that the performances of all models on the IEMOCAP dataset are relatively low. This is because the IEMOCAP dataset is collected in two different scenarios, scripted play and spontaneous dialog, and spontaneous dialog is much more difficult to identify than acted dialog. Finally, compared with other deep learning methods, it clearly shows that the proposed method achieves significant performance improvement.

VI. CONCLUSION

In this paper, we proposed a novel algorithm that combines both autoencoder and emotion embedding. The emotion embedding path focuses on learning strong emotional information from labels. This allows the latent representation from the autoencoder to learn which deep features are related to emotion. In the emotion classification process, the IS10 feature set was fused with the deep emotion feature from the autoencoder. Experimental results with two publicly available corpora show that the proposed algorithm further enhances the classification accuracy.

In future work, considering the powerful capabilities of BERT [73] in natural language processing tasks, we will consider introducing it into SER tasks to help the model extract deep attention features. In addition, the use of text information can be a measure to further improve the accuracy of SER.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] B. T. Atmaja and M. Akagi, "Speech emotion recognition based on speech segment using LSTM with attention model," in *Proc. IEEE Int. Conf. Signals Syst. (ICSigSys)*, Jul. 2019, pp. 40–44.

[3] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, Nov. 2003.

[4] M. M. H. El Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using Gaussian mixture vector autoregressive models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2007, pp. IV-957–IV-960.

[5] A. K. Samantaray, K. Mahapatra, B. Kabi, and A. Routray, "A novel approach of speech emotion recognition with prosody, quality and derived features using SVM classifier for a class of north-eastern languages," in *Proc. IEEE 2nd Int. Conf. Recent Trends Inf. Syst. (ReTIS)*, Jul. 2015, pp. 372–377.

[6] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. 16th Annu. Conf. Int. speech Commun. Assoc.*, 2015, pp. 6–10.

[7] A. Rawat and P. K. Mishra, "Emotion recognition through speech using neural network," *Int. J*, vol. 5, pp. 422–428, May 2015.

[8] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Feb. 2003.

[9] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*. [Online]. Available: http://arxiv.org/abs/1802.05365

[10] 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.

[11] 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.

[12] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. W. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," in *Proc. Interspeech*, Sep. 2019, pp. 1691–1695.

[13] F. Bao, M. Neumann, and N. T. Vu, "CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition," in *Proc. Interspeech*, Sep. 2019, pp. 2828–2832.

[14] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "CNN+LSTM architecture for speech emotion recognition with data augmentation," in *Proc. Workshop Speech, Music Mind*, Sep. 2018, pp. 21–25.

[15] M. Ezzeldin A. ElShaer, S. Wisdom, and T. Mishra, "Transfer learning from sound representations for anger detection in speech," 2019, *arXiv:1902.02120*. [Online]. Available: http://arxiv.org/abs/1902.02120

[16] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 265–275, Apr. 2019.

[17] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Cross corpus speech emotion classification—An effective transfer learning technique," 2018, *arXiv:1801.06353*. [Online]. Available: http://arxiv.org/abs/1801.06353

[18] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," in *Proc. Interspeech*, Sep. 2018, pp. 257–261.

[19] W. Wang, Y. Huang, Y. Wang, and L. Wang, "Generalized autoencoder: A neural network framework for dimensionality reduction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 490–497.

[20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ. San Diego La Jolla Inst. Cogn. Sci., San Diego, CA, USA, Tech. Rep. ICS-8506, 1985.

[21] N. E. Cibau and M. L. Enrique Albornoz Hugo Rufiner, "Speech emotion recognition using a deep autoencoder," *Anales de La XV Reunion de Procesamiento de la Informacion y Control*, vol. 16, pp. 934–939, May 2013.

[22] A. Pal and S. Baskar, "Speech emotion recognition using deep dropout autoencoders," in *Proc. IEEE Int. Conf. Eng. Technol. (ICETECH)*, Mar. 2015, pp. 1–6.

[23] K.-Y. Huang, C.-H. Wu, T.-H. Yang, M.-H. Su, and J.-H. Chou, "Speech emotion recognition using autoencoder bottleneck features and LSTM," in *Proc. Int. Conf. Orange Technol. (ICOT)*, Dec. 2016, pp. 1–4.

[24] W. Fei, X. Ye, Z. Sun, Y. Huang, X. Zhang, and S. Shang, "Research on speech emotion recognition based on deep auto-encoder," in *Proc. IEEE Int. Conf. Cyber Technol. Autom., Control, Intell. Syst. (CYBER)*, Jun. 2016, pp. 308–312.

[25] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in *Proc. Interspeech*, Sep. 2016, pp. 3603–3607.

[26] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks," in *Proc. Interspeech*, Sep. 2016, pp. 3593–3597.

[27] R. Xia and Y. Liu, "Using denoising autoencoder for emotion recognition," in *Proc. Interspeech*, Sep. 2013, pp. 2886–2889.

[28] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4818–4822.

[29] Z. Zong, H. Li, and Q. Wang, "Multi-channel auto-encoder for speech emotion recognition," 2018, *arXiv:1810.10662*. [Online]. Available: http://arxiv.org/abs/1810.10662

[30] P. Wei and Y. Zhao, "A novel speech emotion recognition algorithm based on wavelet kernel sparse classifier in stacked deep auto-encoder model," *Pers. Ubiquitous Comput.*, vol. 23, nos. 3–4, pp. 521–529, Jul. 2019.

[31] K. Oyamada, H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, and H. Ando, "Generative adversarial network-based approach to signal reconstruction from magnitude spectrogram," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2514–2518.

[32] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 66–75, May 2014.

[33] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 125–144, Mar. 2015.

[34] H. Kameoka, "Non-negative matrix factorization and its variants for audio signal processing," in *Applied Matrix and Tensor Variate Data Analysis*. Tokyo, Japan: Springer, 2016, pp. 23–50.

[35] S. Takaki, H. Kameoka, and J. Yamagishi, "Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis," in *Proc. Interspeech*, Aug. 2017, pp. 1128–1132.

[36] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," 2017, *arXiv:1703.10135*. [Online]. Available: http://arxiv.org/abs/1703.10135

[37] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 1459–1462.

[38] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 1–4.

[39] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022

[40] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[41] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6924–6932.

[42] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," 2016, *arXiv:1610.07629*. [Online]. Available: http://arxiv.org/abs/1610.07629

[43] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

[44] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via IBN-Net," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 464–479.

[45] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*. [Online]. Available: http://arxiv.org/abs/1406.1078

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

[47] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[49] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6675–6679.

[50] S. J. Mozziconacci and D. J. Hermes, "Expression of emotion and attitude through temporal speech variations," in *Proc. 6th Int. Conf. Spoken Lang. Process.*, vol. 2, 2000, pp. 373–378.

[51] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 309–314.

[52] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales, "Data augmentation for low resource languages," in *Proc. Interspeech*, Sep. 2014, pp. 810–814.

[53] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop Deep Learn. Audio, Speech Lang.*, vol. 117, Jun. 2013, pp. 1–5.

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

[55] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, Sep. 2015, pp. 1–4.

[56] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. N. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google home," in *Proc. Interspeech*, Aug. 2017, pp. 1–5.

[57] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing, "Recurrent topic-transition GAN for visual paragraph generation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3362–3371.

[58] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," 2019, *arXiv:1904.08779*. [Online]. Available: http://arxiv.org/abs/1904.08779

[59] 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, p. 335, Dec. 2008.

[60] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 1–4.

[61] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 312–315.

[62] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of SVM trees for multimodal emotion recognition," in *Proc. Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2012, pp. 1–4.

[63] M. Shah, C. Chakrabarti, and A. Spanias, "A multi-modal approach to emotion recognition using undirected topic models," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Jun. 2014, pp. 754–757.

[64] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 439–448.

[65] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," in *Proc. Interspeech*, Sep. 2018, pp. 3683–3687.

[66] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," 2017, *arXiv:1712.08708*. [Online]. Available: http://arxiv.org/abs/1712.08708

[67] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7390–7394.

[68] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," 2019, *arXiv:1907.06078*. [Online]. Available: http://arxiv.org/abs/1907.06078

[69] M. Sidorov, S. Ultes, and A. Schmitt, "Emotions are a personal thing: Towards speaker-adaptive emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4803–4807.

[70] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using machine learning," in *Social Media and Machine Learning*. London, U.K.: IntechOpen, 2019.

[71] G. Assuncao and P. Menezes, "Intermediary fuzzification in speech emotion recognition," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2020, pp. 1–6.

[72] X. Zeng, L. Dong, G. Chen, and Q. Dong, "Multi-feature fusion speech emotion recognition based on SVM," in *Proc. IEEE 10th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2020, pp. 77–80.

[73] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

CHENGHAO ZHANG was born in Wuxi, Jiangsu, China, in 1992. He is currently pursuing the M.E. degree in communication and information systems with Shanghai University. His research interests include signal processing, pattern recognition, and deep learning.

LEI XUE was born in Beijing, China, in 1963. He received the B.E. degree from Henan University, China, in 1985, the M.E. degree in computer science from Fudan University, China, in 1999, and the Ph.D. degree in pattern recognition and intelligent control from the Huazhong University of Science and Technology, China, in 2004. His research interests include signal processing and intelligent control.