International Journal of Automation and Computing

16(4), August 2019, 437-448
DOI: 10.1007/s11633-019-1175-x

Semi-supervised Ladder Networks for Speech Emotion Recognition

Jian-Hua Tao[1,2,3] Jian Huang[1,2] Ya Li[1] Zheng Lian[1,2] Ming-Yue Niu[1,2]

[1] National Laboratory of Pattern Recognition, Beijing 100190, China

[2] School of Artificial Intelligence, University of Chinese Academy of Science (CAS), Beijing 100190, China

[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract: As a major component of speech signal processing, speech emotion recognition has become increasingly essential to understanding human communication. Benefitting from deep learning, many researchers have proposed various unsupervised models to extract effective emotional features and supervised models to train emotion recognition systems. In this paper, we utilize semi-supervised ladder networks for speech emotion recognition. The model is trained by minimizing the supervised loss and auxiliary unsupervised cost function. The addition of the unsupervised auxiliary task provides powerful discriminative representations of the input features, and is also regarded as the regularization of the emotional supervised task. We also compare the ladder network with other classical autoencoder structures. The experiments were conducted on the interactive emotional dyadic motion capture (IEMOCAP) database, and the results reveal that the proposed methods achieve superior performance with a small number of labelled data and achieves better performance than other methods.

Keywords: Speech emotion recognition, the ladder network, semi-supervised learning, autoencoder, regularization.

1 Introduction

As one of the main information mediums in human communication, speech contains not only basic language information, but also a wealth of emotional information. Emotion can help people understand real expressions and potential intentions. Speech emotion recognition (SER) has many applications in human-computer interactions, since it can help machines to understand emotional states like human beings do[1]. For example, speech emotion recognition can be utilized to monitor customers′ emotional state which reflects their service quality in call centers. The information can help promote service level and reduce the workload of manual evaluation[2].

Emotion is conventionally represented as several discrete human emotional moods such as happiness, sadness and anger over utterances[3]. In speech emotion recognition, the establishment of a speech emotional database is based on the reality that every speech utterance is assigned to a certain one of emotional categories. As a result, most researchers regard speech emotion recognition as a typical supervised learning task. Given the emotional database, the classification models are trained to predict exact emotional labels for each utterance. Thus, lots of conventional machine learning methods were applied successfully in speech emotion recognition. The models, hidden Markov models (HMMs) and Gaussian mixture models (GMMs) which emphasize the temporality of speech signal and had achieved great performance in speech recognition, were also applied in SER[4, 5]. Support vector machines (SVMs), which have the superiority of modeling small data sets, usually achieved better performance than other alterative models[6]. Inspired by the success of various tasks with deep learning[7, 8], numerous research efforts have been made to build an effective speech emotion recognition model with deep neural networks (DNN), leading to impressive achievement[9, 10].

However, speech emotion recognition still faces many challenges, such as the diversity of speakers, genders, languages and cultures which would influence the system performance. The difference of recording conditions is also bad for the stability of the system. While automatic systems have been shown to outperform naive human listeners on speech emotion classification[11], existing SER systems are not so mature compared with speech and image classification tasks. One of the serious problems is the shortage of emotional data that limits the robustness of the models.

Supervised classification methods estimate emotional class by learning the differences between different cat-

Research Article

Manuscript received October 12, 2018; accepted March 1, 2019; published online May 2, 2019

Recommended by Associate Editor Matjaz Games

egories. The guarantee of a large enough number of labelled speech emotional data is necessary for the exactitude of the separatrices. However, the acquisition of labelled data demands experts′ knowledge and is also highly time consuming. Even worse, there exists large ambiguity and subjectivity among the boundaries of the emotions since the expressions and perceptions of different people are different[12]. Thus, there is no definite standard for providing emotional labels. Due to these shortcomings, the quantity of speech emotion databases is limited, and cannot cover the diversity of different conditions[13].

Considering the scarcity of speech emotion data, it is beneficial to take full advantage of the information from unlabeled data. Unsupervised learning is one choice which extracts robust feature representations from the data automatically without depending on label information. This technique can depict the intrinsic structures of the data, and has stronger modeling and generalization ability for training better classification models[14]. Most of the existing unsupervised feature learning approaches have been explored to generate salient emotional feature representations for speech emotion recognition, such as autoencoders (AE)[15] and denoising autoencoders (DAE)[14]. The purpose of AE and DAE is to obtain intermediate feature representations which can rebuild the input data as much as possible. Other sophisticated methods, such as variational autoencoders (VAE)[16] and generative adversarial networks (GAN)[17], have achieved better performance in SER. They emphasize the modeling of the distribution of the data, explicit form such as normal distribution for VAE and inexplicit form for GAN, rather than the data itself.

The feature representations learning from unsupervised models are usually used as the inputs of supervised classification models to train speech emotion recognition systems. Nevertheless, such an approach has an underlying problem. The former unsupervised learning plays the role of the feature extractor, while the target of the model is to recover the input signals perfectly. It means all information would persist as much as possible. However, we only need to focus on emotionally relevant information. On the other hand, the later supervised learning only concentrates on the information that is good for classification prediction. The extra information which is maybe supplementary for SER would be dropped. Therefore, the feature representations learning from unsupervised learning may not necessarily support the supervised classification task. The objectives of two steps, unsupervised part and supervised part are not consistent because their trainings are parted.

To address this problem, deep semi-supervised learning is proposed to dispose of the difficulty[18–20]. Semi-supervised learning is the combination of unsupervised feature representation learning and supervised model training. The key is that these two parts are trained simultaneously so that the feature representations obtained from unsupervised learning can accord with the supervised model better. The typical structures, such as semi-supervised variational autoencoders[19] and ladder networks[18], have achieved competitive performance with less labelled training samples in other areas.

Benefiting from the unsupervised learning part, semi-supervised learning can introduce great feature representations with the aid of many unlabeled examples to improve the performance of supervised tasks. Due to the scarcity of speech emotional data and richness of speech data, it is appropriate to apply semi-supervised learning approaches to speech emotion recognition. Actually, the part of auxiliary unsupervised learning also plays the role of regularization in the semi-supervised learning model. The regularization is essential to develop speech emotion recognition systems that generalize across different conditions[21]. Conventional models obtained poor performance when the databases of training and testing are different[22, 23]. By training models that are optimized for primary and auxiliary tasks, the feature representations are more general, avoiding overfitting to a particular domain. It is appealing to create unsupervised auxiliary tasks to regularize the network.

Classic semi-supervised learning structure is an autoencoder which introduces additional unsupervised learning. The autoencoder structure can be replaced by other structures like DAE and VAE. More layers can be stacked. A more advanced structure is a semi-supervised ladder networks[18, 24]. Similar to DAE, every layer of a ladder network is intended to reconstruct their corrupted inputs. Further, the ladder network adds the lateral connections between each layer of the encoder and decoder, which is different from DAE. Figuratively, this is also the meaning of the term “ladder”, and it indicates the deep multilayer structure of the ladder network. The attraction of hierarchical layer models is the ability of modeling latent variables to learn from low layers to high layers. Generally, low layers represent the specific information while high layers can generate abstract features which are invariant and relevant for classification tasks. This can model more complex nonlinear structures than conventions methods[25].

Most unsupervised methods aim to learn intermediate feature representations that may not support the underlying emotion classification task. This paper proposes to employ the unsupervised reconstruction of the inputs as an auxiliary task to regularize the network, while optimizing the performance of an emotion classification system. We efficiently achieve this goal with a semi-supervised ladder network. The addition of the unsupervised auxiliary task not only provides powerful discriminative representations of the input features, but is also regarded as the regularization of primary emotional supervised task. The core contributions of this paper can be summarized as follows:

1) In this paper, we utilize semi-supervised learning

with a ladder network for speech emotion recognition. We emphasize the importance of unsupervised reconstruction and skip connection modules. In addition, higher layers of the ladder network have a better ability to obtain discriminative features.

2) We show the benefit of semi-supervised ladder networks and that the promising results can be obtained with only a small number of labelled samples.

3) We compare the ladder network with DAE and VAE methods for emotion recognition from speech, showing superior performance of the ladder network. Besides, the convolutional neural network structure of the encoder and decoder has a better ability to encode emotional characteristics.

The remainder of the paper is organized as follows. Section 2 discusses the related work. In Section 3, we describe our proposed methods. We then present the dataset and acoustic features used for the experiments in Section 4. Section 5 presents the experimental results and analysis. Finally, Section 6 concludes this paper.

2 Related work

Traditional speech emotion recognition relies on well-established hand-crafted speech emotional features. The most popular acoustic features are frame-level low-level descriptors (LLD) such as mel-frequency cepstral coefficients (MFCC), followed by utterance-level information extraction with different functionals, such as mean and maximum etc.[26, 27]

With the great achievement of deep learning, a great number of approaches utilize DNN to extract effective emotional feature representations, then feed them as inputs to an emotional classifier. Kim et al.[28] captured high-order nonlinear relationships from multimodal data with four deep belief networks (DBN) architectures. Firstly, the audio features and video features were inputted to their individual layers, then their outputs were concatenated to generate final multimodal fusion emotional features in a later layer. Finally, the classifier SVM was used to evaluate their performance.

Various autoencoders have been widely applied in speech emotion recognition. Deng et al.[29] proposed shared hidden layer autoencoders for common feature transfers learning to cope with the mismatch between the corpora. Then, they extended to sparse autoencoders (SAE)[30]. In source domain, every emotional class trained its own individual SAE model. After that, all training data of the target domain was reconstructed with corresponding SAE of the same class to alleviate the difference. The new reconstructed data was regraded as training data to train the SVM model to predict test samples. Furthermore, they substituted SAE with DAE to obtain performance gain[6]. Xia and Liu[31] proposed a modified DAE to distinguish the emotional representations from non-emotional factors like speakers and genders. They designed two hidden layers separately to represent emotional and non-emotional representation in parallel. The non-emotional layer was trained firstly like normal autoencoder. Then, the emotional layer was trained with the non-emotional layer frozen. Finally, the emotional representations were the inputs of SVM for speech emotion classification. Next, they joined gender information to model emotional specific characteristics for further performance gain[32].

Ghosh et al.[33, 34] combined DAE and bidirectional long short-term memory (BLSTM) AE to get more robust emotional representations from the original wav spectrogram. They utilized a multilayer perceptron (MLP) to evaluate the performance of generated latent representations. Eskimez et al.[35] systematically investigated four kinds of unsupervised feature learning methods, DAE, VAE, adversarial autoencoder (AAE) and adversarial variational Bayes (AVB) for improving the performance of speech emotion recognition. They showed that the models which emphasized the distribution of speech emotional data, namely VAE, AAE and AVB, outperformed DAE.

Deng et al.[36] proposed semi-supervised autoencoders to improve the performance of SER. This was achieved by regarding the unlabeled data as an extra class, which explicitly aided the supervised learning by incorporating prior information from unlabeled samples. In this paper, our work builds upon the ladder network to further explore the influence of semi-supervised learning for speech emotion recognition.

Valpola[24] proposed the ladder network to reinforce autoencoder networks. The unsupervised tasks involve the reconstruction of hidden representations of a denoising autoencoder with lateral connections between the encoder and decoder layers. Rasmus et al.[18, 37] further extended this idea to support supervised learning. They included a batch normalization to reduce covariate shift. They also compared various denoising functions to be used by the decoder. The representations from the encoder are simultaneously used to solve the supervised learning problem. The ladder network conveniently solved unsupervised auxiliary tasks along with primary supervised tasks. Finally, Pezeshki et al.[38] explored different components of the ladder network, noting that lateral connections between encoder and decoder and the addition of noise at every layer of the network greatly contributed to their improved performance. The skip connections between the encoder and decoder ease the pressure of transporting information needed to reconstruct the representations to the top layers. Therefore, top layers can learn features that are useful for the supervised task, such as the emotional prediction.

Inspired by their work, we propose semi-supervised ladder networks for speech emotion recognition, showing their benefits for emotion prediction. This work is an extension of our previous work presented in [39], which fo-

cused on discrete emotion recognition. In similar work, Parthasarathy and Busso[40] utilized the ladder network to perform dimensional emotional recognition with multi-task learning. Notice that our work[39] was published first in AAAC Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia 2018) before Parthasarathy's work[40] in International Speech Communication Association (INTERSPEECH 2018).

3 Method

In this section, we will describe specific ladder network architecture with two hidden layers, as shown in Fig. 1. There are two encoders in the ladder network, that is, one is the noise encoder corrupted by noise which is similar to DAE, and the other is an original clean input signal with shared parameters. The ladder network combines a primary supervised task with an auxiliary unsupervised task. The auxiliary unsupervised task reconstructs the hidden representations of a clean encoder. The noise encoder is simultaneously used to train primary classification task.

The key aspect of the ladder network is the lateral connections between the layers of the encoder and decoder. These lateral skip connections establish the relationships between each layer of the noisy encoder and its corresponding layer in the decoder. This operation enables the information to flow freely between the encoder and decoder. As a result, the feature representations from low layers to high layers would be from specific to abstract and emotional-relevant for speech emotion classification tasks. Formally, the ladder network is defined as follows:

$$\tilde{x}, \tilde{z}^{(1)}, \cdots, \tilde{z}^{(L)}, \tilde{y} = Encoder_{noisy}(x) \tag{1}$$

$$x, z^{(1)}, \cdots, z^{(L)}, y = Encoder_{clean}(x) \tag{2}$$

$$x, \hat{z}^{(1)}, \cdots, \hat{z}^{(L)}, y = Decoder\left(\tilde{z}^{(1)}, \cdots, \tilde{z}^{(L)}\right) \tag{3}$$

where the variables x, y, $\tilde{y}$ and y^* are the input, the noiseless output, the noisy output and the true target, respectively. The variables $z^{(l)}$, $\tilde{z}^{(l)}$ and $\hat{z}^{(l)}$ are the hidden representation, its noisy version, and its reconstructed version at layer l. In the following parts, we give a detailed description of the ladder network to introduce our proposed methods herein.

3.1 Encoder

The encoder of the ladder network is a fully connected MLP network. A Gaussian noise with variance σ^2 is added to each layer of the noisy encoder, as shown in Fig. 1. The representations from the final layer $\tilde{z}^{(l)}$ of the encoder are used for the supervised task. The decoder tries to reconstruct the latent representation $\hat{z}$ at every layer using a clean copy of the encoder z as target. In the

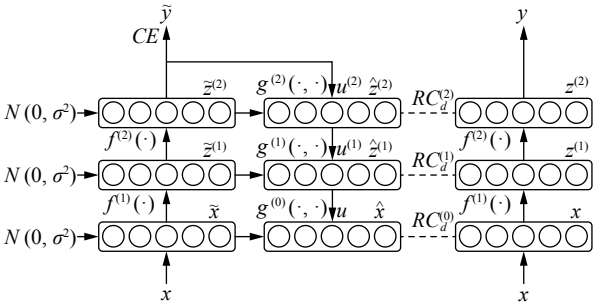


Fig. 1 The architecture using semi-supervised ladder networks[18] for speech emotion recognition. This figure illustrates the ladder network with two hidden layers. The noise feedforward path $\left(x \to \tilde{z}^{(1)} \to \tilde{z}^{(2)} \to \tilde{y}\right)$ shares the network parameters $f^{(l)}$ with the clean feedforward path $\left(x \to z^{(1)} \to z^{(2)} \to y\right)$. The decoder $\left(\tilde{z}^{(l)} \to \hat{z}^{(l)} \to \hat{x}\right)$ reconstruct the input vector of the encoder with denoising functions $g^{(l)}$. The lateral connections on every layer make $RC_d^{(l)}$ to minimize the difference between $\hat{z}^{(l)}$ and $z^{(l)}$. The output $\tilde{y}$ of the encoder and the true target y^* are utilized to calculate the supervised loss.

training phase, the supervised task is trained with the noisy encoder which further regularizes the supervised learning. Meanwhile, the clean encode is utilized to predict emotional class in the testing phase.

In the forward network, a single layer of the encoder includes three types of calculation. The inputs are first transformed with linear transformation, then batch normalization is applied with the mean and standard deviation of mini-batch, followed by a non-linear activation function. The detailed schematic diagram is illustrated in Fig. 2. Formally, the encoder is defined as follows:

$$\tilde{z}_{pre}^{(l)} = W^{(l)} \cdot \tilde{h}^{(l-1)} \tag{4}$$

$$\mu^{(l)} = mean\left(\tilde{z}_{pre}^{(l)}\right) \tag{5}$$

$$\sigma^{(l)} = std\left(\tilde{z}_{pre}^{(l)}\right) \tag{6}$$

$$\tilde{z}^{(l)} = \frac{\tilde{z}_{pre}^{(l)} - \mu^{(l)}}{\sigma^{(l)}} + \mathcal{N}\left(0, \sigma^2\right) \tag{7}$$

$$\tilde{h}^{(l)} = \Phi\left(\gamma^{(l)}\left(\tilde{z}^{(l)} + \beta^{(l)}\right)\right) \tag{8}$$

where $\tilde{h}^{(l-1)}$ is the post-activation at layer $l-1$ and $W^{(l)}$ is the weight matrix from layer $l-1$ to layer l. $\mu^{(l)}$ and $\sigma^{(l)}$ are the mean and standard deviation of mini-batch at layer $l-1$. The Gaussian noise with zero mean and σ^2 variance is added to post-normalization to get pre-activation $\tilde{z}^{(l)}$. The purpose of $\beta^{(l)}$ and $\gamma^{(l)}$ is to increase the diversity and robustness of the model. Finally, a non-linear activation function $\Phi(\cdot)$ is applied to obtain the output $\tilde{h}^{(l)}$. The difference between the noise encoder and

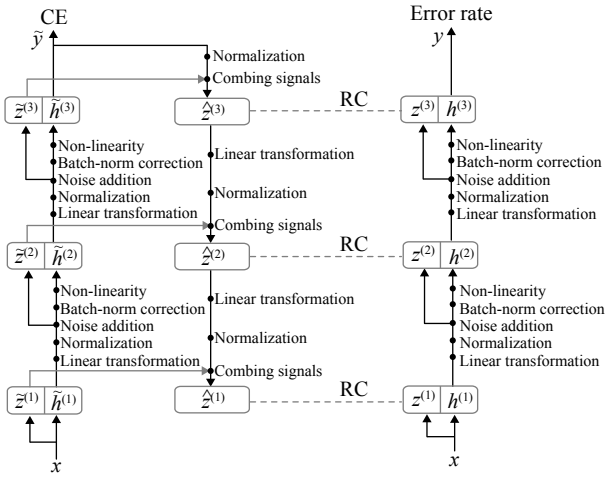


Fig. 2 The detailed calculation structure diagram of the ladder network[38]. Both sides of the figure are encoders, the left is noise encoder ((4) to (8)) and the right is clean encoder. At each layer of encoders, linear transformation and normalization are first applied to $h^{(l-1)}$ and $\tilde{h}^{(l-1)}$. The noise encoder injects noise addition to get $\tilde{z}^{(l)}$ while $z^{(l)}$ has no noise in clean encoder. Then, the batch normalization correction and nonlinearity activation is computed to get $h^{(l)}$ and $\tilde{h}^{(l)}$, respectively. At the decoder ((9) to (13)), the inputs of every layer are from two signals, that one is from above layer $\hat{z}^{(l+1)}$ and another is noise signal $\tilde{z}^{(l)}$ from corresponding layer in the encoder. The linear transformation and normalization are applied to $\hat{z}^{(l+1)}$ before combining singals. CE stands for supervised cross entropy cost and RC stands for unsupervised reconstruction cost. The total objective function is a weighted sum of the supervised loss and unsupervised loss.

clean encoder is the second item of (7). It is a noisy encoder with noise $\mathcal{N}\left(0,\sigma^2\right)$ just as mentioned above, and it is a clean encoder without noise $\mathcal{N}\left(0,\sigma^2\right)$, $\tilde{h}$ and $\tilde{z}$ are replaced with h and z, respectively.

3.2 Decoder

The structure of the decoder is similar to the encoder. The goal of the decoder is to denoise the noisy latent representations. Instead of using a nonlinear activation function, the denoising function $g(\cdot,\cdot)$ combines top-down information from the decoder and the lateral connection from the corresponding encoder layer. With lateral connections, the ladder network performs similarly to hierarchical latent variable models. Lower layers are mostly responsible for reconstructing the input vector and higher layers can learn more abstract, discriminative features for speech emotion recognition.

Similarly, batch normalization is also employed at each layer of the decoder. In the back network of the decoder, the inputs of every layer are from two signals, that one is from above layer $\hat{z}^{(l+1)}$ and another is the noise signal $\tilde{z}^{(l)}$ from the corresponding layer in the encoder. The detailed schematic diagram is illustrated in Fig. 2. Formally, the decoder is defined by the following equations:

$$u_{pre}^{(l+1)} = V^{(l)} \cdot \hat{z}^{(l+1)} \tag{9}$$

$$\mu^{(l+1)} = mean\left(u_{pre}^{(l+1)}\right) \tag{10}$$

$$\sigma^{(l+1)} = std\left(u_{pre}^{(l+1)}\right) \tag{11}$$

$$u^{(l+1)} = \frac{u_{pre}^{(l+1)} - \mu^{(l+1)}}{\sigma^{(l+1)}} \tag{12}$$

$$\hat{z}^{(l)} = g\left(\tilde{z}^{(l)}, u^{(l+1)}\right) \tag{13}$$

where $V^{(l)}$ is a weight matrix from layer l+1 to layer l.

The function $g(\cdot,\cdot)$ is also called the combinator function as it combines the vertical $u^{(l+1)}$ and the lateral $\tilde{z}^{(l)}$. We use the function proposed by Pezeshki et al.[38], modeled by an MLP with inputs $[u, \tilde{z}, u \odot z]$, where u is the batch normalized projection of the layer above and $\odot$ represents the Hadamard product.

3.3 Objective function

The objective function of the ladder network consists of two parts which correspond to the supervised part and unsupervised part respectively. The goal of the unsupervised part is to reconstruct the input signals, whose impact is to obtain effective intermediate hidden representations automatically that accord with speech emotion classification better. Besides, the unsupervised objective can regularize the supervised speech emotional recognition task. The unsupervised objective and supervised objective are optimized simultaneously, which makes the system integrated into a whole model to train the ladder network and avoids the discordance of the optimization objective of two parts.

The supervised loss is cross entropy cost calculated between the noisy output $\tilde{y}$ from the top of noise encoder and the true target y^*. The unsupervised loss is reconstruction loss between every layer of clean encoder and its corresponding layer of the decoder with lateral connections. The total objective function is a weighted sum of the supervised loss and unsupervised loss:

$$C = CE + \lambda_l \sum_{l=1}^{L} RC_d^{(l)} \tag{14}$$

where λ_l is a hyper-parameter weight for the unsupervised loss and CE is the supervised loss:

$$CE = -\sum_{n=1}^{N} \log P\left(\tilde{y}(n) = y^*(n) \,|x(n)\right) \tag{15}$$

and $RC_d^{(l)}$ is the reconstruction loss at layer l:

$$RC_d^{(l)} = ReconsConst\left(z^{(l)}, \hat{z}^{(l)}\right) = \left\| \frac{\tilde{z}^{(l)} - u^{(l)}}{\sigma^{(l)}} - z^{(l)} \right\|^2 \tag{16}$$

where $u^{(l)}$ and $\sigma^{(l)}$ are mean and standard deviation of the samples in the encoder. The output of the decoder $\hat{z}^{(l)}$ is normalized to release the effect of unwanted noise introduced by the limited batch size of batch normalization.

3.4 Variational autoencoder (VAE)

In this paper, we utilize VAE[16], another version of AE, as a comparison. Unlike DAE which aims to reconstruct the data, VAE emphasizes the modeling of the explicit distribution form of the data to generate more intrinsic feature representations, as shown in Fig. 3. Formally, VAE is defined as follows:

$$(z_\mu, z_\sigma) \sim f_\theta(z|x) \tag{17}$$

$$z = z_\mu + z_\sigma \odot N(0, I) \tag{18}$$

where x is the input, f_θ means the encoder. z_μ, z_σ are the mean and standard deviation of normal distribution learning from the encoder network. $N(0, I)$ is the Gaussian distribution with zero mean and unit standard deviation.

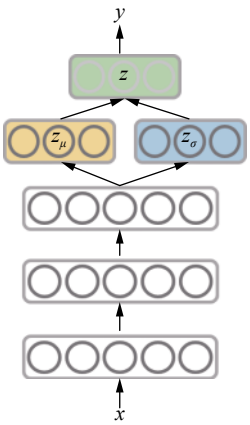


Fig. 3 Variational autoencoder models the data using an explicit distribution represented by z_μ and z_σ

During the training of VAE, besides the reconstruction loss, Kullback-Leibler (KL) divergence loss is also used.

$$Loss = KL(f_\theta(z|x)||m(z)) \tag{19}$$

where $m(z) = N(z; 0, I)$ is the prior multivariate Gaussian distribution.

4 Experiments

4.1 Database

In this paper, we conduct the experiments on the interactive emotional dyadic motion capture (IEMOCAP)[41]. The database has multimodal data of 12 hours duration from audio, visual and textual data. We focus on emotion recognition from speech data. It was recorded by ten actors; five males and five females. The recording condition was based on the form of dyadic interaction acting in two different scenarios: scripted play and spontaneous dialog. After completing the conversations, the recordings would be cut into the sentence levels. Valid and valuable sentences would be selected and annotated as emotional labels and neutral label with at least three annotators. The database has nine emotion classes in total, namely angry, excited, happy, sad, neutral, frustrated, fearful, surprised, and disgust. For this study, we use four categories to evaluate the system performance including "angry", "happy", "sad" and "neutral" which are researched frequently and own most samples. Like other researchers do[9, 10], we regard the "excited" class as "happy" class. Only the sentences satisfying the condition that at least two annotations are agreed would be selected. In total we collect 5 531 utterances. The basis for partitioning the training set and test set is leave-one-speaker-out. The class distribution is: 20.0% "angry", 19.6% "sad", 29.6% "happy", and 30.8% "neutral".

4.2 Acoustic features

The inputs of the networks are speech acoustic features which are traditional hand-crafted emotional features for speech emotion recognition. We refer to the baseline features of the INTERSPEECH 2009 Emotion Challenge[42]. As shown in Table 1, it contains 16 acoustic low-level descriptors (LLDs) including zero-crossing-rate (ZCR), root mean square (RMS) frame energy, pitch frequency (normalized to 500 Hz), harmonics-to-noise ratio (HNR) by autocorrelation function, and mel-frequency cepstral coefficient (MFCC) 1–12. Their first order delta regression coefficients are utilized to double the LLDs resulting in 32 LLDs. 12 functionals – mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and ranges as well as two linear regression coefficients with their mean square error (MSE) are applied to 32 LLDs to calculate 384 dimensional features. The extraction of the LLDs and the computation of the functionals are done using the openSMILE toolkit[43].

4.3 Experimental setup and evaluation metrics

In the experiments, three hidden layers with the size of 500-300-100 from low layers to high layers are employed. The size of input layer is 384 corresponding to speech acoustic feature dimension and the final prediction layer is 4 corresponding to emotional classes. For the hyper-parameter weight of the unsupervised loss $\lambda^{(l)}$ in (14), we optimize them with search grid {0.1, 0.2, 0.5,

Table 1 Features sets including 32 low-level descriptors (LLDs) and 12 functions

LLD (16×2)	Functions (12)
(△) ZCR	Mean
(△) RMS Energy	Standard deviation
(△) F0	Kurtosis, skewness
(△) HNR	Extremes: value, real, position, range
(△) MFCC 1–12	Linear regression: offset, slope, MSE

1, 2, 5}. Since every layer has individual $\lambda^{(l)}$, the global search would consume much time. Thus, this parameter is optimized layer by layer. The ADAM optimization algorithm[44] is utilized. The batch size is set as 32. The initial learning rate is 0.02 for 50 iterations followed by 25 iterations with a learning rate decaying linearly to 0. In the following part, we use SVM to evaluate the performance of the feature representations learning from the networks. To determine the parameters of SVM, we use a grid search in the range of [1.0, 100.0] and [0.000 1, 0.1] for C and g, respectively. Each experiment is repeated five times to account for instability. The evaluation measure is the unweighted accuracy (UAR).

5 Results

For semi-supervised learning, the true benefit of the ladder network is that only a few labelled samples are available for the primary supervised task. Thus, we conduct four semi-supervised emotion recognition tasks with 300, 600, 1 200 and 2 400 labelled samples. Labelled samples are chosen randomly from the training set but the number of training samples in each class is balanced. The left training samples are utilized to pretrain the network without label information. To evaluate the performance of the semi-supervised ladder networks, we utilize three other methods as a comparison. SVM is utilized to evaluate the performance of acoustic features as baseline results. The DAE method has similar network structure to the ladder network shown in Fig. 1, however it has no skip connections and reconstruction loss of high layers. Further, the VAE method is achieved by replacing the encoder and decoder of the DAE structure with the part described in Fig. 3. The network settings of DAE and VAE are the same as the ladder network.

Table 2 lists the average accuracies with standard deviation over five trials for four models. In addition, the best accuracy of five trials is presented in the table as well. The following analyses are mostly based on the average accuracies. With the increase of training samples, the performance is gradually improved and the best performance is achieved with all training samples for all situations, showing the increasing number of training samples is beneficial to the emotional classification task. Furthermore, our proposed methods can achieve superior performance with a small number of labelled data. The VAE method using only 600 training samples, achieves better performance 53.7% than SVM 52.4% using all training samples, while the ladder network only needs 300 training samples to reach 53.6%. This suggests semi-supervised learning has a positive influence on performance improvement for speech emotion recognition. It is worth noticing that the performance is improved faster for three network methods when fewer training samples are available, specifically from 300 to 600 and from 600 to 1 200. This suggests the auxiliary unsupervised task is essential to performance improvement. Overall, three network methods achieve better performance than SVM baseline results in all situations. Thus, the representations learning from deep autoencoder structures achieves better performance than conventional models when using similar hand-crafted acoustic features. As can be seen from Table 2, VAE yields better accuracy than the DAE method, which shows that VAE would model intrinsic structure of speech emotional data to generate better feature representations. However, there is instability for VAE, since their standard deviations are greater than DAE. We can also observe that the ladder network achieves best performance among all methods. Noticing the difference between DAE and the ladder network is the existence of lateral connections. The results verify that lateral connections between encoder and decoder greatly contribute to the improved performance. The ladder network yields better performance than VAE with smaller standard deviation, showing its superiority to speech emotion recognition.

The results of Table 2 are based on the structure whose encoder and decoder are composed of MLP. Next, we replace the MLP layer with convolutional neural net-

Table 2 Average accuracies (in percent) with standard deviation over five trials on the testing set with 300, 600, 1 200, 2 400 and all labelled samples using MLP

Systems	Number of labelled samples				
	300	600	1 200	2 400	All
SVM	46.3±1.3 (47.2)	47.3±1.4 (48.0)	49.5±0.9 (49.9)	51.1±1.1 (52.0)	52.4±0.8 (53.8)
DAE	50.1±1.6 (51.1)	50.9±2.1 (52.1)	53.9±1.5 (54.3)	54.9±1.2 (55.7)	55.8±1.4 (56.4)
VAE	51.9±2.0 (53.0)	53.7±2.0 (54.8)	55.4±1.6 (56.1)	56.7±1.5 (57.3)	57.4±1.7 (58.2)
Ladder network	53.6±1.8 (54.7)	55.6±1.7 (56.5)	57.3±2.1 (58.4)	58.0±1.0 (58.5)	58.6±1.1 (59.1)

works (CNN) layer for three network models to improve the performance. The inputs are frame-level features which replace utterance-level features when using MLP. Specifically, the encoder is replaced with three 2D convolutional layers and the decoder is replaced with three 2D deconvolutional layers. The filter number and kernel size are shown in Table 3. The number of parameters is increased from 2.8 M with MLP to 5.8 M with CNN for three networks. Therefore, the training time is also increased from less than one hour to about two hours for three networks. The training time of the ladder network is larger than DAE because of the addition of lateral connections and reconstruction loss. The training time of VAE is sometimes larger than the ladder network or smaller than the ladder network due to its instability.

Table 3 Architecture of the encoder and decoder. Conv2D is a 2D convolution layer, and Conv2D-d is a 2D deconvolution layer

	Layers	Filter number	Kernel size	Strides
Encoder	Conv2D	32	9×9	2×2
	Conv2D	64	7×7	2×2
	Conv2D	128	5×5	2×2
Decoder	Conv2D-d	128	5×5	2×2
	Conv2D-d	64	7×7	2×2
	Conv2D-d	32	9×9	2×2

The corresponding experimental results are shown in Table 4. By comparing Table 2 with Table 4, we can observe that the models with MLP have better performance when fewer training samples (300 and 600) are available, while the models with CNN have better performance when more training samples (1 200 and more) are available. In addition, the standard deviation is relatively decreased. This suggests CNN has better and robust ability to encode emotional characteristics when training data is enough. Similarly, the ladder network achieves better performance than VAE, which achieves better performance than DAE. The results also verify the significance of unsupervised learning when few training samples are available.

The autoencoder structures have superior ability to extract effective feature representations compared with other network models. We extract the features from the highest hidden layer of trained models and feed them to the SVM classifier to assess their quality. Fig. 4 reveals the classifier performance of three network methods on the testing set with 300, 600, 1 200 and 2 400 labelled samples using MLP and CNN as feature extractor. We also explore the influence of supervised learning. In Fig. 4, the symbol "1" represents training without supervised learning while "2" represents training with supervised learning. At every setting, the ladder network achieves best performance followed by VAE, and DAE is worst. Comparing Fig. 4 (a) with Fig. 4 (b), the performance of CNN structure achieves better performance than MLP structure. For example, when using 2 400 training samples with supervised learning, the ladder network using CNN structure achieves 60.3%, better than the MLP structure′s result of 59.8%. The results show the models with supervised learning yield better performance than the models without supervised learning. The "Ladder2" using CNN achieves better performance of 60.3% than "Ladder1" with 59.4%, which shows the supervised information is beneficial to guide better feature representations. The results of Fig. 4 are better than the results of Tables 2 and 4, verifying the ability of autoencoder structures to generate more discriminating feature representations for speech emotion recognition.

After comparing the performance of the features from the highest layer, we turn our attention from low layers to high layers with three network structures. Similarly, SVM is utilized to evaluate the quality of the features. This part is based on CNN structure using all training samples and the experimental results are shown in Fig. 5. The performances of the first layer "384" are the experimental results of acoustic features which are similar to Table 2 and final layer "4" are accuracies of supervised learning which are similar to Table 4. The results show that the accuracies of last layer "4" are worse than last hidden layer "100" which is the same as the experimental results of Fig. 4. The accuracies are improved with the increase of the layers for three network methods. We can observe that DAE achieves better performance than VAE and the ladder network in first hidden layer "500" while VAE and the ladder network outperform DAE in the following layers "300" and "100". Therefore, high layers have the advantage of generating more salient emotional representations. Further, the ladder network achieves better performance than VAE in all hidden layers, which verifies the effectiveness of our proposed methods.

Table 4 Average accuracies (in percent) with standard deviation over five trials on the testing set with 300, 600, 1 200, 2 400 and all labelled samples using CNN

Systems	Number of labelled samples				
	300	600	1 200	2 400	All
DAE	48.4±2.0 (49.5)	50.4±1.8 (51.4)	54.3±1.1 (55.3)	55.4±1.1 (56.1)	56.3±1.0 (57.0)
VAE	47.3±2.1 (48.5)	51.2±1.7 (52.1)	55.8±1.3 (56.3)	57.0±1.4 (57.7)	58.0±1.3 (58.5)
Ladder network	49.1±1.3 (49.9)	53.5±0.9 (53.9)	57.7±1.2 (58.2)	58.4±1.1 (58.8)	59.4±0.8 (59.7)

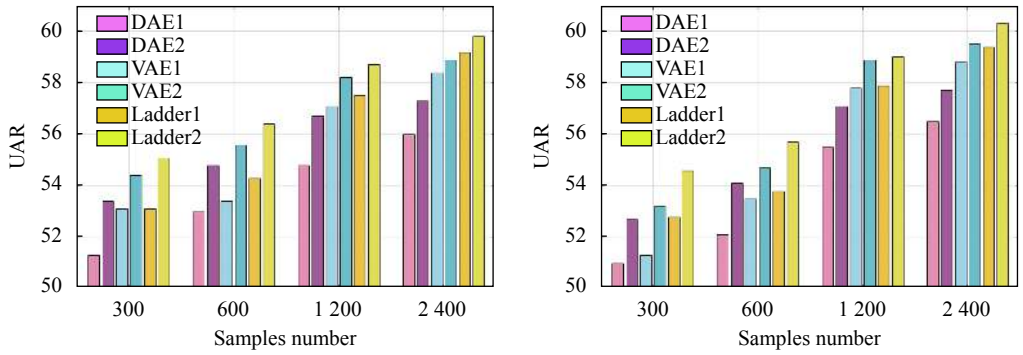


Fig. 4 Classification performance on the testing set with 300, 600, 1200 and 2400 labelled samples using MLP and CNN as feature extractor. SVM is used as classifier. The symbol "1" represents training without supervised learning while "2" represents training with supervised learning.

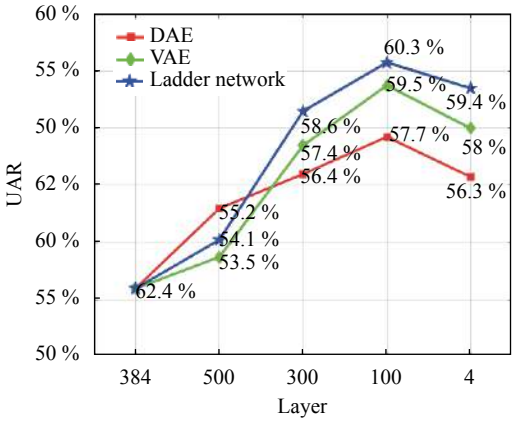


Fig. 5 Classification performance from low layers to high layers for DAE, VAE and the ladder network

Finally, we demonstrate the performance on different speech emotional categories. Table 5 shows the results of different models corresponding to the best results using all training samples in Table 4. The performance of DAE, VAE and the ladder network are 57.0%, 58.5% and 59.7% respectively. Compared with DAE, the results show that the VAE method yields better performance on "angry", "happy" and "sad". The ladder network achieves its best performance on "angry", "happy" and "neutral", while the performance of "sad" is decreased slightly. Thus, the enhanced network structure is beneficial to the performance improvement of "angry" and "happy".

We also compare the proposed method with other methods in the literature. We also compare the proposed method with other methods in the literature, as shown in Table 6. Our proposed method achieves better performance of 59.7% than the 56.1% of Michael's work in [9], which uses an attentive convolutional neural network to recognize emotions. Fayek et al.[10] introduce a frame-based formulation to model intra-utterance dynamics with end-to-end deep learning, achieving better performance of 60.9%. The feature representations from the top layer with SVM achieve 60.3% in Fig. 5, which is a comparable result to [10].

Table 5 Accuracy (in percent) for each class using three different methods

Systems	Class			
	Ang	Hap	Neu	Sad
DAE	65.9	52.9	53.9	63.3
VAE	69.4	55.4	52.9	65.3
Ladder network	70.6	56.3	56.8	62.9

Table 6 Performance comparison between our method with other methods

Model	Accuracy
Attentive CNN[9]	56.1
Frame-based SER[10]	60.9
Ladder network	59.7

6 Conclusions

In this paper, we apply semi-supervised learning to speech emotion recognition to explore the effect of the ladder network. The unsupervised reconstruction of the inputs is an auxiliary task to regularize the network, which can generate more powerful representations for speech emotion recognition system. We conduct the experiments on the IEMOCAP database and the results demonstrate that the proposed methods achieve superior performance with a small number of labelled data. We also compare the ladder network with two classic network structures DAE and VAE, showing the ladder network outperforms them significantly. The results suggest lateral connections between encoder and decoder greatly contribute to the improved performance. The skip connections between the encoder and decoder ease the pressure of transporting information needed to reconstruct the representations to the top layers. Thus, a higher layer has the ability to generate discriminative features for speech emotion recognition. Meanwhile, the supervised

learning task is beneficial to generate more effective feature representations. Besides, CNN has better and robust ability to encode emotional characteristics compared to the MLP structure. Finally, our proposed methods are beneficial to the performance improvement of "angry" and "happy". In the future, we will try to utilize more available unlabeled speech data to improve the performance of SER. Deeper semi-supervised learning and other network structures like recurrent neural networks (RNNs) will be explored.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Nos. 61425017 and 61773379), the National Key Research & Development Plan of China (No. 2017YFB1002804).

References

[1] J. H. Tao, T. N. Tan. Affective computing: A review. In *Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction*, Springer, Beijing, China, pp. 981–995, 2005. DOI: 10.1007/11573548_125.

[2] H. Bořil, A. Sangwan, T. Hasan, J. H. Hansen. Automatic excitement-level detection for sports highlights generation. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, ISCA, Makuhari, Japan, pp. 2202–2205, 2010.

[3] H. Gunes, B. Schuller. Categorical and dimensional affect analysis in continuous input: current trends and future directions. *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013. DOI: 10.1016/j.imavis.2012.06.016.

[4] T. L. Nwe, S. W. Foo, L. C. De Silva. Speech emotion recognition using hidden Markov models. *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003. DOI: 10.1016/S0167-6393(03)00099-2.

[5] M. M. H. El Ayadi, M. S. Kamel, F. Karray. Speech emotion recognition using Gaussian mixture vector autoregressive models. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, USA, pp. 957–960, 2007. DOI: 10.1109/ICASSP.2007.367230.

[6] J. Deng, Z. X. Zhang, F. Eyben, B. Schuller. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014. DOI: 10.1109/LSP.2014.2324759.

[7] B. Zhao, J. S. Feng, X. Wu, S. C. Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, vol. 14, no. 2, pp. 119–135, 2017. DOI: 10.1007/s11633-017-1053-3.

[8] Z. J. Yao, J. Bi, Y. X. Chen. Applying deep learning to individual and community health monitoring data: a survey. *International Journal of Automation and Computing*, vol. 15, no. 6, pp. 643–655, 2018. DOI: 10.1007/s11633-018-1136-9.

[9] M. Neumann, N. T. Vu. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, ISCA, Stockholm, Sweden, pp. 1263–1267, 2017.

[10] H. M. Fayek, M. Lech, L. Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, vol. 92, pp. 60–68, 2017. DOI: 10.1016/j.neunet.2017.02.013.

[11] S. E. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Y. Duan, W. Heinzelman. Emotion classification: How does an automated system compare to Naive human coders? In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, pp. 2274–2278, 2016. DOI: 10.1109/ICASSP.2016.7472082.

[12] B. Jou, S. Bhattacharya, S. F. Chang. Predicting viewer perceived emotions in animated GIFs. In *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, USA, pp. 213–216, 2014. DOI: 10.1145/2647868.2656408.

[13] M. El Ayadi, M. S. Kamel, F. Karray. Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011. DOI: 10.1016/j.patcog.2010.09.020.

[14] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, vol. 11, no. 12, pp. 3371–3408, 2010.

[15] G. E. Hinton, R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, vol. 313, no. 5786, pp. 504–507, 2006. DOI: 10.1126/science.1127647.

[16] D. P. Kingma, M. Welling. Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*, ICLR, Ithaca, USA, 2013.

[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, MIT Press, Montreal, Canada, pp. 2672–2680, 2014.

[18] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, T. Raiko. Semi-supervised learning with ladder networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, MIT Press, Montreal, Canada, pp. 3546–3554, 2015.

[19] J. Weston, F. Ratle, H. Mobahi, R. Collobert. Deep learning via semi-supervised embedding. *Neural Networks: Tricks of the Trade*, 2nd ed., G. Montavon, G. B. Orr, K. R. Müller, Eds., Berlin Heidelberg, Germany: Springer, pp. 639–655, 2012. DOI: 10.1007/978-3-642-35289-8_34.

[20] D. P. Kingma, D. J. Rezende, S. Mohamed, M. Welling. Semi-supervised learning with deep generative models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, MIT Press, Montreal, Canada, pp. 3581–3589, 2014.

[21] C. Busso, M. Bulut, S. Narayanan. Toward effective automatic recognition systems of emotion in speech. *Social Emotions in Nature and Artifact: Emotions in Human and Human Computer Interaction*, J. Gratch and S. Marsella, Eds., New York, USA: Oxford University Press, pp. 110–127, 2014.

[22] S. Parthasarathy, C. Busso. Jointly predicting arousal, valence and dominance with multi-task learning. In *Pro-*

ceedings of the 18th Annual Conference of the International Speech Communication Association, ISCA, Stockholm, Sweden, pp. 1103–1107, 2017.

[23] M. Shami, W. Verhelst. Automatic classification of expressiveness in speech: a multi-corpus study. *Speaker Classification II: Selected Projects*, C. Müller, Ed., Berlin Heidelberg, Germany: Springer-Verlag, vol. 4441, pp. 43–56, 2007. DOI: 10.1007/978-3-540-74122-0_5.

[24] H. Valpola. From neural PCA to deep unsupervised learning. *Advances in Independent Component Analysis and Learning Machines*, E. Bingham, S. Kaski, J. Laaksonen, J. Lampinen, Eds., Amsterdam, Netherlands: Academic Press, pp. 143–171, 2015. DOI: 10.1016/B978-0-12-802806-3.00008-7.

[25] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009. DOI: 10.1561/2200000006.

[26] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, K. P. Truong. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016. DOI: 10.1109/TAFFC.2015.2457417.

[27] J. Huang, Y. Li, J. H. Tao. Effect of dimensional emotion in discrete speech emotion classification. In *Proceedings of the 3nd International Workshop on Affective Social Multimedia Computing*, ASMMC, Stockholm, Sweden, 2017.

[28] Y. Kim, H. Lee, E. M. Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, pp. 3687–3691, 2013. DOI: 10.1109/ICASSP.2013.6638346.

[29] J. Deng, R. Xia, Z. X. Zhang, Y. Liu, B. Schuller. Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, pp. 4818–4822, 2014. DOI: 10.1109/ICASSP.2014.6854517.

[30] J. Deng, Z. X. Zhang, E. Marchi, B. Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *Proceedings of Humaine Association Conference on Affective Computing and Intelligent Interaction*, IEEE, Geneva, Switzerland, pp. 511–516, 2013. DOI: 10.1109/ACII.2013.90.

[31] R. Xia, Y. Liu. Using denoising autoencoder for emotion recognition. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, ISCA, Lyon, France, pp. 2886–2889, 2013.

[32] R. Xia, J. Deng, B. Schuller, Y. Liu. Modeling gender information for emotion recognition using denoising autoencoder. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, pp. 990–994, 2014. DOI: 10.1109/ICASSP.2014.6853745.

[33] S. Ghosh, E. Laksana, L. P. Morency, S. Scherer. Learning representations of affect from speech. In *Proceedings of International Conference on Learning Representations*, ICLR, San Juan, Puerto Rico, 2016.

[34] S. Ghosh, E. Laksana, L. P. Morency, S. Scherer. Representation learning for speech emotion recognition. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, ISCA, San Francisco, USA, pp. 3603–3607, 2016.

[35] S. E. Eskimez, Z. Y. Duan, W. Heinzelman. Unsupervised learning approach to feature analysis for automatic speech emotion recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, 2018. DOI: 10.1109/ICASSP.2018.8462685.

[36] J. Deng, X. Z. Xu, Z. X. Zhang, S. Frühholz, B. Schuller. Semisupervised autoencoders for speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 31–43, 2018. DOI: 10.1109/TASLP.2017.2759338.

[37] A. Rasmus, H. Valpola, T. Raiko. *Lateral Connections in Denoising Autoencoders Support Supervised Learning*, [Online], Available: https://arxiv.org/abs/1504.08215, April, 2015.

[38] M. Pezeshki, L. X. Fan, P. Brakel, A. Courville, Y. Bengio. Deconstructing the ladder network architecture. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, ACM, New York, USA, pp. 2368–2376, 2016.

[39] J. Huang, Y. Li, J. H. Tao, Z. Lian, M. Y. Niu, J. Y. Yi. Speech emotion recognition using semi-supervised learning with ladder networks. In *Proceedings of the 1st Asian Conference on Affective Computing and Intelligent Interaction*, IEEE, Beijing, China, 2018. DOI: 10.1109/ACII-Asia.2018.8470363.

[40] S. Parthasarathy, C. Busso. *Ladder Networks for Emotion Recognition: Using Unsupervised Auxiliary Tasks to Improve Predictions of Emotional Attributes*, [Online], Available: https://www.isca-speech.org/archive/Interspeech_2018/abstracts/1391.html, 2018.

[41] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008. DOI: 10.1007/s10579-008-9076-6.

[42] B. Schuller, S. Steidl, A. Batliner. The Interspeech 2009 emotion challenge. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association*, ISCA, Brighton, UK, pp. 312–315, 2009.

[43] F. Eyben, M. Wöllmer, B. Schuller. Opensmile: The Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, ACM, Florence, Italy, pp. 1459–1462, 2010. DOI: 10.1145/1873951.1874246.

[44] D. P. Kingma, J. L. Ba. Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representations*, ICLR, Ithaca, USA, 2015.

Jian-Hua Tao received the Ph. D. degree in computer science from Tsinghua University, China in 2001. He is winner of the National Science Fund for Distinguished Young Scholars and the deputy director in National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China. He has directed many national projects, including "863", National Natural Science Foundation of China. He has published more than eighty papers on journals and proceedings including *IEEE Transactions on ASLP*, and *ICASSP, INTERSPEECH*. He also serves as the steering committee member for *IEEE Transactions on Affective Computing* and the 470 chair or program committee member for major con-

ferences, including *International Conference on Pattern Recognition (ICPR), INTERSPEECH*, etc.

His research interests include speech synthesis, affective computing and pattern recognition.

E-mail: jhtao@nlpr.ia.ac.cn (Corresponding author)
ORCID ID: 0000-0002-9437-7188

Jian Huang received the B. Eng. degree in automation from Wuhan University, China in 2015. He is a Ph. D. degree candidate in pattern recognition and intelligent system at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, China. He had published the papers in *INTERSPEECH* and *ICASSP*.

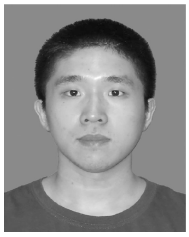
His research interests include affective computing, deep learning and multimodal emotion recognition.

E-mail: jian.huang@nlpr.ia.ac.cn

Ya Li received the B. Eng. degree in automation from University of Science and Technology of China (USTC), China in 2007, and the Ph. D. degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China in 2012. She is currently an associate professor with CASIA, China. She has published more than 50 papers in the related journals and conferences, such as *Speech Communication, ICASSP, INTERSPEECH* and *Affective Computing and Intelligent Interaction (ACII)*. She has won the Second Prize of Beijing Science and Technology Award in 2014. She has also won the Best Student Paper in *INTERSPEECH 2016*.

Her research interests include affective computing and human-computer interaction.

E-mail: yli@nlpr.ia.ac.cn



Zheng Lian received the B. Eng. degree in telecommunication from Beijing University of Posts and Telecommunications, China in 2016. He is a Ph. D. degree candidate in pattern recognition and intelligent system at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China.

His research interests include affective computing, deep learning and multimodal emotion recognition.

E-mail: lianzheng2016@ia.ac.cn

Ming-Yue Niu received the M. Sc. degree in information and computing science from Department of Applied Mathematics, Northwestern Polytechnical University (NWPU), China in 2017. Currently, he is a Ph. D. degree candidate in pattern recognition and intelligent system at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China.

His research interests include affective computing and human-computer interaction.

E-mail: niumingyue2017@ia.ac.cn