Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks

Qirong Mao, *Member, IEEE*, Ming Dong, *Member, IEEE*, Zhengwei Huang, and Yongzhao Zhan

***Abstract*—As an essential way of human emotional behavior understanding, speech emotion recognition (SER) has attracted a great deal of attention in human-centered signal processing. Accuracy in SER heavily depends on finding good affect-related, discriminative features. In this paper, we propose to learn affect-salient features for SER using convolutional neural networks (CNN). The training of CNN involves two stages. In the first stage, unlabeled samples are used to learn local invariant features (LIF) using a variant of sparse auto-encoder (SAE) with reconstruction penalization. In the second step, LIF is used as the input to a feature extractor, salient discriminative feature analysis (SDFA), to learn affect-salient, discriminative features using a novel objective function that encourages feature saliency, orthogonality, and discrimination for SER. Our experimental results on benchmark datasets show that our approach leads to stable and robust recognition performance in complex scenes (e.g., with speaker and language variation, and environment distortion) and outperforms several well-established SER features.**

***Index Terms*—Affective-salient discriminative feature analysis, convolutional neural networks, feature learning, speech emotion recognition.**

I. Introduction

Speech is one of the most significant and natural means for human to communicate their emotions, cognitive states, and intentions to each other [44]. As an essential way of human emotional behavior understanding, in the past decades, speech emotion recognition (SER) has attracted a great deal of attention in human-centered computing. The increasing applications of SER, especially those in human computer interaction and affective computing [27], [45], make it a core component in the next generation of computer system, in which a natural human machine interface enables the automated provision of services that require a good appreciation of the emotional state of a user.

Although advances have been made recently in automatic SER in terms of speech emotion feature extraction and emotion recognition, robust and accurate SER is still a challenging problem due to complex factors such as the variations of speakers and contents, and environment distortion [44], [24]. In SER, one of the central research issues is how to extract discriminative, affect-salient features from speech signals [24]. In this direction, a number of speech emotion features have been proposed in the literature, and they can be roughly classified into four categories [20]: 1) acoustic features, 2) linguistic features (words and discourse), 3) context information (e.g., subject, gender, and turn-level features representing local and global aspects of the dialogue) [24], and 4) hybrid features that combine acoustic features with other information. However, it is unclear if these hand-tuned feature sets can sufficiently and efficiently characterize the emotional content of speech [24]. Moreover, their performance varies greatly in different scenarios. Finally, automatic extraction of some of these features can be difficult. For example, existing automatic speech recognition (ASR) systems cannot reliably recognize all the verbal content of emotional speech [1]. Extracting semantic discourse information is even more challenging, which, in many cases, has to be performed manually [44].

Thus, in SER, it is important to explore new strategies that can obtain the optimal feature set that is invariant to nuisance factors while maintaining discrimination with respect to the task of emotion recognition. In this work, we introduce feature learning in SER. Our idea here is inspired by the recent development of deep learning. Deep learning is part of a broader family of machine learning methods based on learning feature representations. It addresses the problem of what makes better representations and how to learn them. In many situations where labeled data is limited or not available, deep learning is shown to have the capability to generate good features, for example, for facial expression recognition [31] and ASR [43].

In the paper, we propose to learn affect-salient features for SER using convolutional neural networks (CNN). In CNN, simple features are learned in the lower layers, and affect-salient, discriminative features are obtained in the higher layers. More specifically, CNN has two learning phases. In the first stage, unlabeled samples are used to learn local invariant features (LIF) by a variant of sparse auto-encoder (SAE) with reconstruction penalization. In the second step,

Manuscript received March 13, 2014; revised June 27, 2014 and September 16, 2014; accepted September 22, 2014. Date of publication September 29, 2014; date of current version November 13, 2014. This work was supported in part by the National Nature Science Foundation of China under Grants 61272211 and 61170126, and by the Six Talent Peaks Foundation of Jiangsu Province under Grant DZXX-027. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gokhan Tur.

Q. Mao was with the Department of Computer Science, Wayne State University, Detroit, MI 48202 USA. She is now with the Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu Province 212013, China (e-mail: qr@mail.ujs.edu.cn).

M. Dong is with the Department of Computer Science, Wayne State University, Detroit, MI 48202 USA (e-mail: mdong@cs.wayne.edu).

Z. Huang and Y. Zhan are with the Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu Province 212013, China (e-mail: zhengwei.hg@gmail.com; yzzhan@mail.ujs.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2360798

the local invariant features are used as the input to a feature extractor, salient discriminative feature analysis (SDFA), to learn affect-salient, discriminative features. We propose a novel objective function in SDFA by encouraging feature saliency, orthogonality, and discrimination for SER. Our experimental results on several benchmark datasets show that our approach leads to stable and robust recognition performance in complex scenes (e.g., with speaker variation and noise), and outperforms several well-established SER features. The preliminary version of this work was first presented in a shortened form as a conference abstract [14]. The major contributions of this paper are:

1) To our best knowledge, this is the first paper introducing feature learning to SER, in which the optimal feature set can be effectively and automatically learned by CNN with a few labeled samples.
2) By introducing a novel objective function in SDFA, we can extract affect-salient features for SER by disentangling emotions from other factors such as speakers and noise. Specifically, the LIF from unsupervised learning are divided into two blocks, related to emotion and other remaining factors, respectively. The emotion-related features are discriminative and robust, leading to great performance improvement on SER.

The rest of the paper is organized as follows. We introduce the related work in Section II. Section III presents our CNN-based feature learning algorithm in detail. Section IV describes SER benchmark datasets and reports our experimental results. Conclusions and future directions are discussed in Section V.

II. Related Works

A typical SER system consists of two components: 1) a front-end processing unit that extracts the appropriate features from the speech data, and 2) a classifier that decides the emotion of the speech utterance. In the following, we first briefly review the classification strategies, and then focus on feature extraction methods, as they are more related to this work.

A. Classifiers

Various types of classifiers have been used for the task of SER, including hidden Markov model [26], Gaussian mixture model [38], support vector machine (SVM) [23], artificial neural networks [4], k-nearest neighbor [28] and many others [29]. Among these methods, SVM and HMM are widely used in almost all speech-related applications [25], [26], [9], [40]. However, experiments show that each classifier has its own advantages and limitations. In order to combine the merits of different classifiers, aggregating a group of classifiers has also been recently studied [21].

B. Feature Extraction

Two research issues must be considered in feature extraction for SER. The first one is regarding the Region of Interest (ROI) used for feature extraction. One possible approach is to divide the speech signal into many small intervals, i.e., frames, and construct a local feature vector for each frame. For example, prosodic speech features such as pitch and energy, can be extracted from each interval and are considered as local features [26], [30]. On the other hand, global features such as statistics, can be obtained from the whole speech utterance [22], [23], [16], which typically have a lower dimension than the local ones, leading to less computing time [13], [30].

The second research issue is to determine the most suitable type of features for SER. In general, speech features can be grouped into four categories: 1) acoustic features, 2) linguistic features, 3) context information, and 4) hybrid features that combining acoustic features with other information. Specifically, acoustic features can be further classified into four groups [24]: continuous features, qualitative features, spectral features, and Teager Energy Operator (TEO)-based features, and all of them have been intensively studied in SER. Koolagudi *et al.* [15] examined pitch-related features in Berlin emotion speech corpus. Four emotions from Chinese natural emotional speech corpus including anger, joy, sadness, and neutral are discriminated by combining prosody and voice quality features in [46]. In [47], the amplitude of emotional speech marginal spectrum is extracted. The discrimination ability of TEO for SER is studied in [37].

Linguistic content of the spoken utterance is also an important part of the conveyed emotion. In [34], a spotting algorithm that searches for emotional keywords or phrases in the utterances was employed. An alternative procedure for detecting emotions using lexical information is found in [18]. The context information generally includes subject, gender, and turn-level features representing local and global aspects of the dialogue. They have been investigated in [39] and [11] on audio affective recognition. More recently, there has been a focus on hybrid features, i.e., the integration of acoustic, linguistic, and context features [29], [42]

Although linguistic content of the spoken utterance can help the acoustic emotion features improve the accuracy of SER, current ASR systems still cannot reliably recognize all the verbal content of emotional speech. In addition, context information or transcripts are simply not available in many cases. Thus, the most popular feature representation for SER are acoustic features such as prosodic features (e.g., pitch-related feature, energy-related features and speech rate) and spectral features (e.g., Mel frequency cepstral coefficients (MFCC) and cepstral features). However, due to the tight coupling of speech emotion and other factors of variation such as speaker and other environment distortion, it is unclear if these hand-tuned acoustic features can sufficiently and efficiently characterize the emotional content of speech.

Recently, it has been shown that unsupervised feature leaning is very helpful for ASR [43] and image understanding [2], [10]. For SER, Stuhlsatz *et al.* [36] used generatively pre-trained artificial neural networks (ANNs) to learn discriminative features of low dimension and found improvement in both weighted and unweighted recall on multiple emotion corpora. Schmidt and Kim [33] used deep belief networks (DBNs) to learn high-level features directly from magnitude spectra and achieved good performances on emotional music recognition compared to other feature extraction schemes. More recently, Le *et al.* [17] investigated dynamic frame-level modeling with hybrid DBN-HMM classifiers on the FAU Aibo spontaneous emotion corpus [35] and achieved state-of-the-art results on the 5-class problem. In [6], a common emotion-specific mapping rule is learnt from a

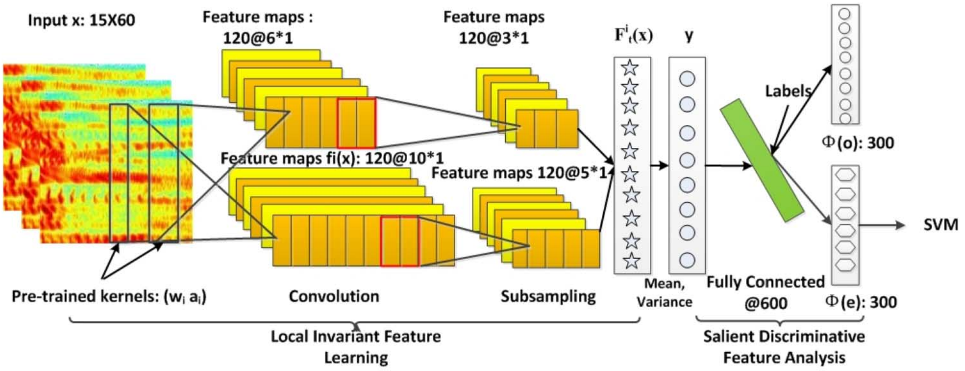

Fig. 1. System pipeline. Left: Input spectrogram at two different resolutions. The next stage is the local invariant feature learning containing the output of one long feature vector. The salient discriminative feature learning produces the last stage of affect-salient features $\phi^{(e)}$ and nuisance features $\phi^{(o)}$, and the former is then fed to a linear SVM for SER.

small set of labeled data in a target domain. Then, newly reconstructed data are obtained by applying this rule on the emotion-specific data in a different domain. Wollmer *et al.* [41] presents a method to systematically investigate the number of past and future utterance-level observations that are considered to generate an emotion prediction for a given utterance, and to examine to what extent this temporal bidirectional context contributes to the overall performance.

While these previous works studied the problem of feature learning for SER using various techniques, their focus is mainly on learning discriminative features from the input. How to learn affect-salient features by disentangling factors of variation, a very important aspect in SER, is not yet addressed. In this paper, we introduce a feature learning framework to SER using CNN, in which we directly deal with the issue of entangle factors of variation through a novel object function that integrates feature saliency, discrimination, orthogonality, and reconstruction.

III. Learning Salient Features for SER

In this section, we present our feature learning algorithm using CNN in SER. The architecture of CNN is shown in Fig. 1, which has an input layer, one convolutional layer, one fully connected layer, and a SVM classifier. We use the spectrogram of the speech signal as the input of CNN. The main idea of feature learning is to learn high-level representations from the low-level raw features, and the spectrogram is well-suited for this task. As a low-level feature, spectrogram is widely used in speech recognition and audio-based speaker and gender recognition. For example, in [19], the spectrogram is used together with convolutional deep belief networks for various audio classification tasks, e.g., speaker identification, gender classification, and phone classification. In [43], spectrogram features are used to conduct speech recognition and achieved solid performance.

Following the hierarchy of CNN, the features learned at each layer become increasingly invariant to nuisance factors while maintaining affect-salience with respect to the goal of SER. Specifically, the training can be broken down into the following three steps:

1) *Local Invariant Feature Learning:* We use a sparse auto-encoder to learn local invariant features from emotional speech signal at multiple scales in an unsupervised fashion. First, we use a sparse auto-encoder to learn kernels with different scales. Then, the entire emotional spectrogram fragment is convolved with the learned kernels to form a series of feature maps. These feature maps are then subsampled through mean-pooling and stacked into one feature vector as the final output of the convolutional layer.
2) *Salient Discriminative Feature Analysis:* The local invariant features obtained through the auto-encoder are used as the input to the fully connected layer, in which they are divided into two blocks. While all features are trained to cooperate to reconstruct the input, the affect-salient feature block is also trained to predict the emotion classes based on the labeled samples. Our objective in segregating the features is to disentangle the affect-salient features that learn to encode useful information about speech emotion from nuisance features (that are complementary but not affect-salient). To this end, we propose a novel objective function in SDFA that locally encourages the affect-salient features and non-discriminative features to encode distinct directions of variation in the input space.
3) *SVM Training:* Finally, the affect-salient features are used as the input to train a linear SVM based on the labeled training data.

More specifically, the input layer of CNN in our system has 900 neurons to receive normalized (in range of [0,1]) spectrogram fragments of size 15 × 60. Generally speaking, normalizing the input will help the network converge quicker. The convolutional layer consists of three consecutive operations: convolution with kernels, non-linear activation function, and pooling. The convolutional layer contains 120 kernels with two fixed sizes of 6 × 60 and 10 × 60, respectively. The kernel is pretrained by unsupervised auto-encoder feature learning patchwise. The size of the feature maps in the convolutional layer is 10 × 1 and 6 × 1, respectively. The feature is given as follows:

$$h(x) = sigmoid(Wx + \alpha) \tag{1}$$

where W is the weight matrix (kernel), $x \in D_x$ is a spectrogram fragment (D_x is the set of the input), and α is the bias.

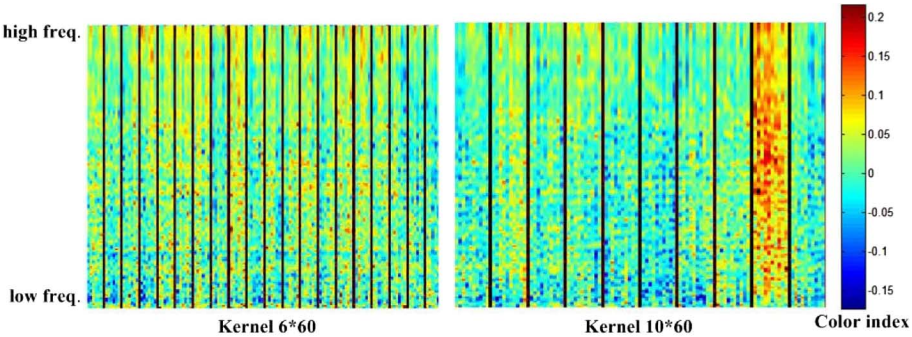

Fig. 2. Convolutional kernels learned by sparse auto-encoder. The horizontal axis represents time, and the vertical axis denotes frequency. Each kernel corresponds to a rectangular area in the image, and the color of the kernel is determined by the element value based on the color index. Left: 6×60. Right: 10×60.

In feature pooling, one of the frequently used functions is down-sampling. We perform down-sampling using a mean operation with a window size 2×1

$$B = \sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\|h(x)\|_2^2} \tag{2}$$

where $m \times n$ is the size of the pooling window. The output of the convolutional layer is employed as the input of the fully-connected layer to disentangle factors of variation. The affect-salient feature block of the final feature vector is passed to a SVM classifier to determine the emotion class of the speech utterance. The output layer has 600 fully connected neurons, each neuron corresponding to one feature.

CNN extracts local invariant features from the input by using SAE and disentangles affect-salient factors from others by using SDFA. In local invariant feature learning, to preserve the neighborhood relations, a neuron in the current layer is only connected to the neurons in a small neighborhood of the previous layer. In early layers, neurons can extract fine-grained features such as upward-slanting or intensity pattern in the spectrogram. In the later layers, higher level features (global features) are learned based more on their relations to the other features and less on their exact locations. In the pooling step, the saliency of the output features to emotion is increased while their sensitivity to speaker variation and environment distortion is reduced. In SDFA, the affect-salient features are learned by encouraging the weight saliency in the local invariant features.

A. Local Invariant Feature Learning

Unsupervised feature learning has shown impressive results in many applications, e.g., image classification. An auto-encoder is commonly used to learn a compressed representation for a set of data. The auto-encoder has two parts: encoder and decoder [32]. The encoder learns a function h to map an input $x \in D_x$ to a feature vector $h(x) \in D_{x'}$, and the decoder reconstructs the input by minimizing the reconstruction error. Usually, the learning is done by training each layer individually and using the current layer codes to feed the next layer.

In our system, the encoder function h maps the input spectrogram fragment x to a latent representation $x' \in D_{x'}$. The encoder function can be written in the form

$$x' = h(x) = s(Wx + \alpha) \tag{3}$$

where s is a nonlinear activation function, typically a logistic sigmoid $sigmoid(z) = 1/(1 + e^{-z})$ [6], [5], W is the weight matrix, and α is the bias vector.

The decoder function g reconstructs x based on the feature vector x' and is written as

$$g(x') = g(h(x)) = s'(W^T x' + \beta) \tag{4}$$

where s' is the decoder's activation function, typically either the identity (yielding linear reconstruction) or the sigmoid. In our work, s' is selected as a logistic sigmoid ($s' = s$). W^T is the weight matrix shared with the encoder and β is the bias vector. Auto-encoder trains the network by adjusting the parameters $\theta = (W, \alpha, \beta)$ on the collected training set to minimize the total reconstruction error

$$min_\theta \sum_{x \in D_x} L(x, g(h(x))) \tag{5}$$

where L is the reconstruction error that is computed using squared error $L(x, x') = \|x - x'\|^2$.

In order to encourage units to maintain a low average activation, an additional penalty term is added into (5)

$$min_\theta \sum_{x \in D_x} L(x, g(h(x))) + \lambda_1 \sum_{k=1}^{s_1} KL(\rho\|\hat{\rho}) \tag{6}$$

where $KL(\rho\|\hat{\rho}_j) = \rho log\frac{\rho}{\hat{\rho}_j} + (1 - \rho)log\frac{1-\rho}{1-\hat{\rho}_j}$ is the sparse penalty term, $\hat{\rho}_j = \frac{1}{\rho}\sum_{t=1}^{p} h_j(x_t)$ is the average activation of hidden unit j (averaged over the training set), S_1 denotes the number of active units, and P denotes the number of training samples. Non-negative parameter ρ is the sparsity level, and λ_1 controls the weight of the sparsity penalty term. They are set at 0.05 and 1 respectively in our experiments.

Kernel Learning: To capture the structure of the input spectrogram at different scales, we consider kernels with multiple sizes. Instead of generating kernel randomly, we pre-train kernels by sparse auto-encoder. Sparse auto-encoder is trained with patches extracted randomly at different locations, the size of which matches that of the convolutional kernels being learned. Assuming that we have q different kernel sizes denoted as l_i-by-h_i, $(i = 1, 2, 3, \ldots, q)$, we can get the kernel (W^i, α^i) after we pre-train independently a sparse auto-encoder for each kernel size using (6).

In our system, we employ two kernel sizes: 6×60 and 10×60. Fig. 2 shows examples of the learned kernels, in which the horizontal axis represents time, and the vertical axis denotes frequency. The kernels are log scaled to the full range of the color map before the visualization. Each kernel corresponds to a rectangular area in the image, and the color of the patch is determined by the element value based on the color index. From Fig. 2, we can clearly observe the distribution of learned kernels. In general, kernels with a higher frequency have a larger value (indicated by a warmer color), and thus they play a more important role in feature reconstruction.

Feature Mapping: After we get the kernel (W^i, α^i), we compute the corresponding feature maps on the whole spectrogram by applying (W^i, α^i) to each l_i-by-h_i patch of the input spectrogram

$$f^i(x) = s(conv(W^i, x) + \alpha^i) \tag{7}$$

where $conv()$ denotes the convolution operation. Then, we perform down-sampling using mean operation with window size $m \times n$ and get the feature map: $F^i_{t,j} = mean_{k \in win_j}(f^i_k(x))$, where win_j denotes the jth window. All the pooled features for patch t and kernel i are stacked into one feature vector with length N^i_w

$$F^i_t(x) = [F^i_{(t,1)}(x), \ldots, F^i_{(t,N^i_w)}(x)] \tag{8}$$

where N^i_w is the number of windows associated with kernel i. Since each utterance may contain a different number of patches (segments), we calculate the mean and variance of $F^i_t(x)$ over all the patches of a given utterance. The result is a feature vector $F^i(x)$, which has a fixed length $2 * N^i_w$ for all the utterances

$$F^i(x) = [mean(F^i_t(x))|_t, var(F^i_t(x))|_t]. \tag{9}$$

Finally, we concatenate $F^i(x)$ for all the kernels. That is, the final feature vector y generated by the convolutional layer is

$$y = [F^1(x), \ldots, F^q(x)] \tag{10}$$

and it is used as the input of SDFA to disentangle emotion-salient factors from others. Note that in this stage, CNN is trained with unlabeled data, abundant in real-world speech applications.

B. Salient Discriminative Feature Analysis

Unsupervised learning provides a network for speech signal reconstruction. In SER, the network needs to be trained to learn good features to identify speech emotion classes. SDFA is more task-related than the initial unsupervised training as some training samples are labeled according to the emotion classes. SDFA promotes the disentangling of emotion discriminative factors of variation in the data from other prominent factors that may well dominate the discriminative factors, and separates the factors of the speech that are discriminative with respect to the SER task from factors that characterize speakers and environment distortion.

While the data is encoded into a single feature vector y after unsupervised feature learning, an input is mapped into two distinct blocks of features: one $(\phi^{(e)}(y))$ that encodes affect-salient factors of its input, and one $(\phi^{(o)}(y))$ that encodes all other factors. Both feature blocks are trained to cooperate to reconstruct their common input y with a reconstruction loss function

$$\begin{aligned}\hat{y} &= g([\phi^{(e)}(y), \phi^{(o)}(y)]) \\ &= s'(U^T[\phi^{(e)}(y), \phi^{(o)}(y)] + \delta)\end{aligned} \tag{11}$$

where U^T is the weight matrix in the fully connected network, and δ_i is an offset to capture the mean value of y.

Given (x, z), the labeled training set with input spectrogram fragment x and emotion label z, the $\phi^{(e)}(y)$ block is also trained to predict the emotion label $z(y)$ when the label is available. The class prediction is given by the logistic regression of the discriminative block $\phi^{(e)}(y)$, which is learned by the sigmoid function over an affine transformation of the $\phi^{(e)}(y)$ block

$$\hat{z}_i = s(A_i \phi^{(e)}(y) + \beta_i) \tag{12}$$

where the weight matrix A_i maps the $\phi^{(e)}(y)$ block to prediction for class i, and β_i is the class specific bias. The corresponding discriminant component of the overall loss function is

$$L_{DISC}(z, \hat{z}) = -\sum_{i=1}^{C} z_i log(\hat{z}_i) + (1 - z_i) log(1 - \hat{z}_i) \tag{13}$$

where C is the number of emotion classes, and z and $\hat{z}$ are the ground truth and the logistic regression output, respectively.

To encourage $\phi^{(e)}(y)$ and $\phi^{(o)}(y)$ to present different directions of variation in the input y, we ask each sensitivity vector $(\partial \phi^{(e)}_i(y))/\partial y$ of the ith discriminant feature $\phi^{(e)}_i$ to prefer being orthogonal to every sensitivity vector $(\partial \phi^{(o)}_j(y))/(\partial y)$ associated with the jth non-discriminant feature $\phi^{(o)}_j$. This penalty component is denoted as $\mathcal{J}_{orth}$

$$\mathcal{J}_{orth} = \sum_{i,j} \left(\frac{\partial F^{(e)}_i(y)}{\partial y} \cdot \frac{\partial F^{(o)}_j(y)}{\partial y} \right)^2. \tag{14}$$

Since a salient feature for SER is usually a sensitive feature for reconstruction error or discrimination error, the features responding strongly to this property tend to be more important. We measure the saliency for each input as the sum of its weight saliency. Specifically, the saliency for input i is defined as

$$\begin{aligned}S_i &= \sum_{k \in \varphi(i)} Saliency(w_k) \\ &= \frac{1}{2} \sum_{k \in \varphi(i)} \frac{\partial^2 MSE}{\partial w_k^2} w_k^2\end{aligned} \tag{15}$$

where $\varphi(i)$ is the set of weights connected input i and w_k is the kth weight. In (15), MSE denotes the mean squared error. For features in $\phi^{(e)}(y)$, both the reconstruction and discrimination errors are taken into consideration, while for features in $\phi^{(o)}(y)$, only the reconstruction error is considered. The cost function is given as

$$\begin{aligned}\mathcal{J}_{SAL} &= -\frac{1}{2} \sum \frac{\partial^2 \|y - \hat{y}\|^2}{\partial w_k^2} w_k^2 \\ &= -\frac{1}{2} \sum_{k \in \phi^{(e)}(y)} \frac{\partial^2 L_{DISC}(z, \hat{z})}{\partial w_k^2} w_k^2\end{aligned} \tag{16}$$

where the first term encourages salient features in $\phi^{(e)}(y)$ and $\phi^{(o)}(y)$ to reduce reconstruction error, and the second term encourages affect-salient features in $\phi^{(e)}(y)$ to reduce discriminative error. This objective is achieved through weight suppression during training.

Putting all the components of the loss function together we get

$$\mathcal{L}(\theta) = \sum_{y=F(x)} L(y, g(h(y))) + \sum_{(x,z)\in S} L_{DISC}(z,\hat{z}) + \lambda_2 \mathcal{J}_{SAL} + \lambda_3 \mathcal{J}_{orth}. \tag{17}$$

The coefficients λ_2 and λ_3 weigh the contribution of the saliency penalty and the orthogonality penalty to the overall loss function, respectively, and they are empirically set as $\lambda_2 = 1$ and $\lambda_3 = 2$.

The set of parameters involved in SDFA is $\theta = (U, A, \delta, \beta)$. To train the network, we applied a common network training technique, back-propagation. After each sample passes the network, the error is calculated based on the loss function in (17) and weights are updated to minimize the error. All samples are cycled through until convergence.

IV. Experiments and Results

A. Datasets and Experimental Setup

Our affect-salient feature learning method was evaluated on four public emotional speech databases with different languages. The first one is Surrey Audio-Visual Expressed Emotion (SAVEE) Database [12], which contains emotional speech utterances covering seven emotions (i.e., anger, disgust, fear, happiness, sadness, surprise, and neutral) deliberately displayed by four English speakers. The sampling rate is 44.1 kHz. The second is Berlin Emotional Database (Emo-DB) [3], which also includes seven emotions (i.e., anger, disgust, fear, joy, sadness, boredom and neutral) displayed by ten German actors. The sample rate is 48 kHz. The third is Danish Emotional Speech database (DES) [7], which includes emotional speech utterances covering five emotions (i.e., anger, joy, surprise, sadness and neutral) displayed by four Danish actors. The sample rate is 48 kHz. The last is Mandarin Emotional Speech database (MES) [9], which contains five emotions (i.e., anger, joy, surprise, sadness and disgust) displayed by seven Mandarin actors. The sample rate is 11.025 kHz.

Except for the experiments conducted with language variance (Section IV-C4), we handle each database separately. That is, features are learned and evaluated using the samples from the same emotional speech database. Specifically, we first split the data into the training set and the testing set. In the unsupervised feature learning stage, we train kernels using one third of randomly selected data in the training dataset of each database. The labels are removed and not used in this stage. In the second stage (SDFA), we train with all the speech utterances in the training dataset.

In our experiments, we first convert the time-domain signals into spectrograms. The spectrogram has a 20 ms window size with a 10 ms overlap. The spectrogram was further processed using principal component analysis (PCA) whitening (with 60 components) to reduce its dimensionality. Thus, the data we feed into the unsupervised feature learning network consist of 60 channels of one-dimensional vectors of length n_w. We then pre-train the unsupervised auto-encoder patch-wise to get the kernels which are used later to perform convolution. The unsupervised auto-encoder is trained with patches extracted randomly at different locations, the size of which matches that of the convolutional kernels being learned (6×60 and 10×60). Based on the kernels, we form 600 local invariant features with kernel size 6×60 and 360 features with kernel size 10×60. Then, SDFA is used to extract 300 affect-salient features and 300 non-discriminative features in which the former ones (across all the frames in the window) are fed into a linear SVM for the final emotion classification on each emotional speech database.

We evaluate affect-salient features based on the classification accuracy and compare it with several other well-established feature representations: spectrogram representation (“RAW” features), TEO [37], acoustic features extracted in [23] (A1) and [8] (A2), and local invariant features (LIF). We also compare features obtained in our system for two different cases: 1) with and without affect-salient penalty [the third term in (17)] and 2) with and without orthogonality penalty [the fourth term in (17)]. More specifically, the “RAW” feature contains statistics, i.e., mean, max, min, and standard deviation, computed for each channel over all frames of the spectrogram. TEO features were quantified using seven statistics (i.e., the mean, median, minimum, maximum, standard deviation, range, and inter-quartile) across all frames of an emotional speech utterance (see [37] for details). The acoustic feature set (A1) in [23] contains 101 widely-used emotional speech features for SER such as pitch-related features, energy-related features, speech rate, and MFCC, while the acoustic emotion feature vectors (A2) in [8] contains 6552 features extracted by the openEAR toolkit. LIF is the local invariant features learned by unsupervised feature learning. Finally, we denote the features learned by SDFA without saliency penalty and without orthogonality penalty as SDFA (no_s) and SDFA (no_or), respectively. The features learned by SDFA with all the penalty terms in (17) are denoted as SDFA.

These feature representations are first evaluated for SER in four public emotional speech databases. Except for the speaker-independent experiments reported in Table I, recognition accuracy on all the expressions of each database are reported using five-fold cross-validation with a SVM classifier. The speaker-independent experiments reported in Table I are conducted with two-fold cross-validation. Here, the test speaker’s utterances are excluded in the unsupervised training of CNN. To determine the parameters C and σ for SVM, we randomly selected 100 speech utterances from the dataset of SAVEE to form an independent validation set, based on which a grid search is performed in the range of $[-1, 10]$ and $[-1, 1.5]$ for C and σ, respectively. The pair of parameters ($C = 5$ and $\sigma = 0.5$) that gives the best results on the validation set is chosen as the one used for cross-validation. Besides overall classification accuracy, we also evaluate robustness of our method with respect to the common disturbing

TABLE I
SER ACCURACY AND STANDARD DEVIATION ON THE FOUR PUBLIC EMOTIONAL SPEECH DATABASES WITH SPEAKER VARIATION (SINGLE SPEAKER, SPEAKER-DEPENDENT, AND SPEAKER-INDEPENDENT). THE RECOGNITION ACCURACY IS REPORTED IN %, AND THE HIGHEST ONE IS HIGHLIGHTED IN BOLD. IN THE TABLE, SPEAKER-DEP AND SPEAKER-INDEP DENOTE SPEAKER-DEPENDENT AND SPEAKER-INDEPENDENT, RESPECTIVELY

speaker	database	RAW	TEO	A1	A2	LIF	SDFA(no_or)	SDFA(no_s)	SDFA
single speaker	SAVEE	31.4±1.56	68.3±1.45	72.4±1.23	87.3±4.22	82.9±0.75	85.0±0.69	88.6±0.67	**89.7**±0.35
	Emo-DB	40.5±1.23	77.2±1.41	85.8±1.09	90.5±3.97	88.7±0.99	92.7±0.73	90.5±0.82	**93.7**±0.28
	DES	38.9±1.43	69.1±1.35	75.6±1.15	87.9±3.47	84.3±0.81	90.0±1.84	88.9±0.53	**90.8**±0.26
	MES	39.1±1.36	70.7±1.19	77.1±1.03	87.5±2.52	84.1±0.75	84.9±1.87	87.1±0.63	**90.2**±0.29
speaker-dep	SAVEE	29.4±1.64	58.8±1.47	65.0±1.21	70.8±4.64	69.4±0.79	**86.7**±3.75	73.3±0.66	75.4±0.42
	Emo-DB	37.2±1.45	66.4±1.48	72.3±1.15	83.2±0.43	80.5±0.81	74.7±1.7	86.4±0.63	**88.3**±0.31
	DES	34.6±1.42	65.3±1.41	71.1±1.25	76.9±0.83	76.8±0.93	78.9±0.57	79.9±0.58	**82.1**±0.45
	MES	35.5±1.53	60.6±1.37	67.8±1.38	76.4±0.77	72.5±0.98	77.2±2.10	76.3±0.61	**79.3**±0.50
speaker-indep	SAVEE	26.7±1.83	51.3±1.65	59.5±1.25	65.0±1.42	62.6±0.80	63.3±0.94	67.0±0.71	**73.6**±0.51
	Emo-DB	35.9±1.62	61.2±1.35	69.3±1.12	74.5±1.56	79.5±0.73	81.4±0.61	82.9±0.69	**85.2**±0.45
	DES	32.6±1.72	59.3±1.58	66.8±1.37	70.8±0.94	72.5±0.83	74.6±0.85	76.1±0.72	**79.9**±0.53
	MES	30.7±1.77	57.5±1.66	64.5±1.39	74.9±0.17	71.0±0.86	76.1±1.76	75.9±0.74	**78.3**±0.61

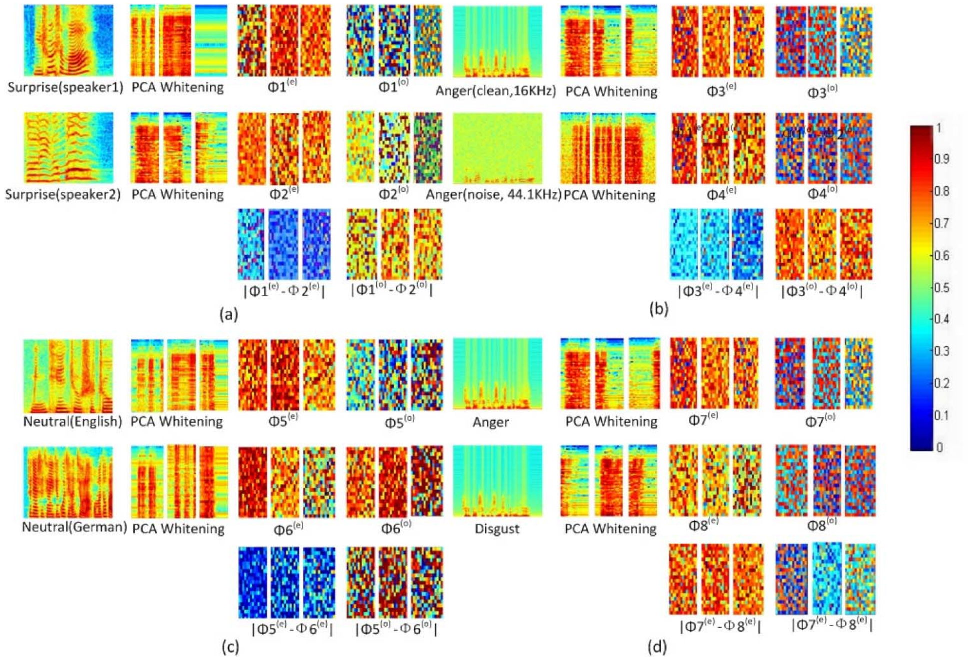

Fig. 3. Visualization of the features learned by SDFA. (a) Features with speaker variation, (b) features with environmental distortion, (c) features with language variation, and (d) features with emotion variation. In each panel, the first column provides the raw spectrogram, the second column shows the results after PCA whitening, and the third and the fourth column visualize affect-salient features and non-discriminative features, respectively. The difference images with various conditions (e.g., emotion, speaker, language, and environmental factors) are provided in the last row.

factors in SER, i.e., the speaker variation, environment distortion, and language variation.

B. Feature Visualization

In order to provide intuitive understanding of the learned features by SDFA, we first visualize several examples obtained with emotion, speaker, language, and environmental variations. Specifically, Fig. 3(a) shows the features associated with the same emotion but different speakers. The first column provides the raw spectrogram of sample speech fragments of "Surprise" for two different speakers. The raw signal is preprocessed by PCA whitening (the second column) in order to reduce its dimensionality. In the third and fourth columns, we visualize affect-salient features and non-discriminative features learned for the two speakers in the first and the second row, respectively. In addition, the difference between the features in the first and the second row is shown in the last row of the corresponding column. We first observe that the affect-salient features $\phi^{(e)}$ contain a higher energy (a larger value) in the entire spectrogram than the non-discriminative feature $\phi^{(o)}$. As the weights with larger values are considered as more important features for SER, this indicates that we achieved our goal of disentangling emotions from other noisy factors. More importantly, when two utterances have the same emotion but from different speakers, the affect-salient feature changes much less than the non-discriminative one, as shown clearly by the difference images in the last row.

Fig. 3(b) shows the features associated with the same emotion "Anger" but under different environmental factors (clean vs. noisy background with different sampling rates). Again, we

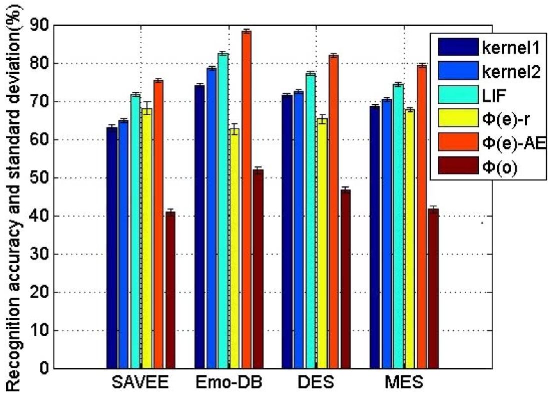

Fig. 4. Average recognition accuracy and standard deviation (%) on the four public emotional speech emotion databases with features learned in different stages of CNN. In the figure, kernel1 and kernel2 denote the kernel with the size of 10×60 and 6×60, respectively.

observe that changes on $\phi^{(e)}$ is much smaller than on $\phi^{(o)}$, which indicates that affect-salient features are robust to environment distortion. Fig. 3(c) shows the features associated with the emotion "Neutral" but with different languages (English vs. German). Clearly, our method is also robust to the language variation. Finally, Fig. 3(d) shows the features obtained with two different emotions "Anger" and "Disgust" while all other factors are kept as the same. Obviously, we get the opposite results this time: the affect-salient feature $\phi^{(e)}$ has a larger change than the non-discriminative one.

C. Performance Evaluation

Accuracy and Convergence on Public Emotional Speech Databases: In this section, we report the recognition accuracy on the four public emotional speech databases using 5-fold cross-validation based on features learned in different stages of CNN: single non-convolutional one-layer kernels (Kernel(10 × 60), Kernel (6 × 60)), LIF, affect-salient features $\phi^{(e)}$ learned with random kernels ($\phi(e) - r$), affect-salient features $\phi^{(e)}$ learned with auto-encoder ($\phi(e) - AE$), and non-discriminative features $\phi^{(o)}$. In all cases, SVM classifier is used for the emotion classification. The results are shown in Fig. 4. These results clearly show that each successive layer in CNN helps to disentangle discriminative features, yielding better classification performance. Notice that on all the databases, the accuracy obtained by affect-salient features is much higher than that obtained by the non-discriminative features, both are obtained in the last layer of CNN. Furthermore, on all the databases, the accuracy of affect-salient features with auto-encoder ($\phi(e) - AE$) is much higher than that obtained by using random kernels. Our experiments also show that auto-encoder can help speed up the convergence of CNN training. Fig. 5 compares the training error of CNN on Emo-DB with respect to the training epoch for ($\phi(e) - r$) and ($\phi(e) - AE$) using the same learning rate. Clearly, the network with the auto-encoder converges faster (and achieves a lower error) than that with random kernels.

Robustness to Speaker Variation: A major source of variability in SER is the variance across speakers. The performance of SER methods degenerates greatly if the speakers in the training set are not the same as those in the testing set. In this section, we further evaluate the features learned by SDFA by comparing it with other well-established feature representations with respect to speaker variance.

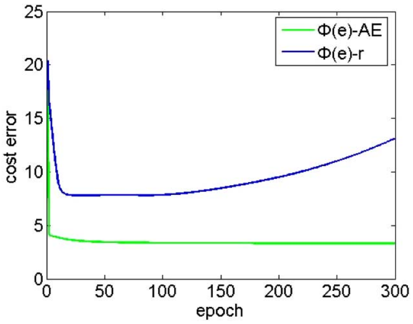

Fig. 5. Convergence speed on Emo-DB with the same learning rate, auto-encoder ($\phi(e) - AE$) versus random kernels ($\phi(e) - r$).

In Table I, single speaker means that the emotional speech utterances come from the same speaker both in the training and the testing set. Speaker-dep means that the person whose emotional speech utterances appear in the testing set also has speech utterances that were used to train the model. Finally, speaker-indep means that the person whose emotional speech utterances appear in the testing set did not have any speech utterance that was used to train the model. The average accuracy and the standard deviation for RAW, TEO, A1 [23], A2 [8], LIF, SDFA(no_or), SDFA(no_s), and SDFA on four public emotional speech databases are reported.

The results clearly show that the learned features (i.e., LIF, SDFA(no_or), SDFA(no_s), and SDFA) generally outperform the baseline ones (RAW, TEO and A1) and achieve comparable results with the well-established acoustic feature set A2. SDFA gets the highest accuracy in all the cases (1% to 6% higher than A2) except that it is second to SDFA(no_or) for the case of speaker-dep on SAVEE. Notice that in the case of speaker-independent, that is, the speakers in training set are mismatched with those in the testing set, SDFA has the smallest standard deviation on most of the databases (SAVEE, Emo-DB and DES), which indicates that SDFA is able to learn a feature representation that is salient to emotion while being robust to speaker variation.

Robustness to Environment Distortion: In SER, there are often cases where significant mismatch between training and test data persists. Environmental factors, e.g., ambient noise, reverberation, microphone type and capture device, are common sources of such mismatch. In this section, we evaluate the extent to which invariance can be obtained with respect to distortions caused by the environment.

In Table II, noise means that the utterances in the test set are corrupted by the Gaussian noise of 20 dB SNR. Channel means we do not use a consistent sampling rate between the training and test samples. Specifically, the sampling frequency of the utterances in the test set of four public emotional speech databases is 16 kHz, which is different from that of the samples in the training set.

Table II compares the accuracy and the standard deviation for RAW, TEO, A1 [23], A2 [8], LIF, SDFA(no_or), SDFA(no_s),

TABLE II
SER Accuracy and Standard Deviation on Four Public Emotional Speech Database With Environmental Distortion (Noise, Channel, and $\mathrm{noise} + \mathrm{channel}$). The Recognition Accuracy is Reported in %, and the Highest One is Highlighted in Bold. In the Table, ED Denotes Environment Distortion

ED	database	RAW	TEO	A1	A2	LIF	SDFA(no_or)	SDFA(no_s)	SDFA
none(clean)	SAVEE	29.4±1.64	58.8±1.47	65.0±1.21	70.8±4.64	69.4±0.79	**86.7**±3.75	73.3±0.66	75.4±0.42
	Emo-DB	37.2±1.45	66.4±1.48	72.3±1.15	83.2±0.43	80.5±0.81	74.7±1.7	86.4±0.63	**88.3**±0.31
	DES	34.6±1.42	65.3±1.41	71.1±1.25	76.9±0.83	76.8±0.93	78.9±0.57	79.9±0.58	**82.1**±0.45
	MES	35.5±1.53	60.6±1.37	67.8±1.38	76.4±0.77	72.5±0.98	77.2±2.10	76.3±0.61	**79.3**±0.50
noise	SAVEE	16.2±2.02	38.5±1.93	49.5±1.96	54.8±1.26	53.6±0.88	55.7±0.89	53.8±0.91	**62.2**±0.56
	Emo-DB	28.3±1.98	47.6±1.99	55.6±1.72	72.3±1.45	68.2±0.83	74.6±0.71	76.8±0.63	**80.1**±0.53
	DES	26.7±1.96	40.6±2.01	53.3±1.68	66.2±1.27	67.8±0.91	75.1±0.98	73.2±0.80	**77.9**±0.55
	MES	28.6±1.92	43.2±1.65	50.2±1.73	62.5±1.30	64.1±1.09	69.3±1.06	70.8±0.75	**72.7**±0.70
channel	SAVEE	16.7±1.97	48.7±1.77	56.6±1.89	62.9±1.73	58.2±0.93	63.2±1.15	58.7±0.68	**69.0**±0.58
	Emo-DB	34.3±1.82	59.9±1.77	68.6±1.65	75.7±1.45	76.1±0.92	80.5±0.79	82.3±0.81	**84.5**±0.49
	DES	31.2±1.56	57.2±1.63	67.8±1.44	74.2±1.27	72.7±1.05	75.2±0.94	76.6±0.68	**77.9**±0.57
	MES	32.3±1.62	53.7±1.44	61.7±1.55	69.3±1.40	67.5±1.01	69.5±1.13	70.2±0.80	**74.8**±0.77
noise+channel	SAVEE	14.5±2.54	36.2±2.35	47.5±2.88	53.9±2.33	49.7±1.94	52.3±0.94	53.6±1.75	**60.7**±0.91
	Emo-DB	29.2±2.35	45.4±2.28	51.7±2.73	69.7±2.55	65.9±1.77	76.6±0.87	71.1±1.55	**78.3**±0.89
	DES	27.5±1.93	38.9±2.47	50.1±2.08	64.6±2.09	63.2±1.49	75.0±1.03	70.4±1.38	**75.8**±0.73
	MES	29.4±2.03	41.4±1.97	47.7±1.99	60.3±1.99	61.9±1.95	**71.1**±0.63	67.7±1.21	69.9±0.97

TABLE III
SER Accuracy and Standard Deviation on Four Public Emotional Speech Databases With Language Variance. The Recognition Accuracy is Reported in %, and the Highest One Is Highlighted in Bold. In the Table, Fe-Database Denotes the Database Used to Learn the Features, and Cl-Database Denotes the Database Used to Perform and Evaluate SER

Fe-database	Cl-database	RAW	TEO	A1	A2	LIF	SDFA(no_or)	SDFA(no_s)	SDFA
SAVEE	Emo-DB	25.5±1.97	40.1±1.55	55.3±1.82	64.7±1.88	63.3±1.45	66.2±1.27	68.3±0.83	**71.8**±0.65
Emo-DB	SAVEE	18.6±2.33	32.7±1.89	43.2±2.05	51.3±1.72	50.6±1.67	56.1±1.38	53.8±1.15	**57.2**±1.11
DES	MES	20.1±2.25	29.3±2.23	45.6±1.98	51.1±1.93	52.1±1.66	57.4±1.59	56.3±1.41	**60.4**±1.07
MES	DES	18.9±2.40	28.7±2.21	40.5±1.89	48.9±1.85	49.3±1.78	54.5±1.03	54.2±1.66	**57.8**±1.43

and SDFA on the four emotional speech databases, respectively. Clearly, among all the methods SDFA achieves the highest and the most stable accuracy (the smallest standard deviation) in 14 out of 16 cases, and is the second best in the remaining two cases. In addition, the learned features (i.e., LIF, SDFA(no_or), SDFA(no_s), and SDFA) generally outperform the baseline ones (RAW, TEO and A1).

Robustness to Language Variance: The goal of this experiment is to evaluate whether the features learned by SDFA can achieve competitive performance with respect to language variance. We divide the four public emotional speech databases into two groups: one group is SAVEE and Emo-DB, and the other one is DES and MES. We conduct language variance experiment in each group separately as the databases in one group have the similar categories of emotions. Specifically, in each group, the affect- salient features $\phi^{(e)}$ are learned by SDFA using one database, but the SVM classifier is trained and tested by all the samples in the other database, after which we exchange these two databases and conduct the experiment again. In this experiment, we only use the emotion categories included in both databases of each group. Specifically, in the group of SAVEE and Emo-DB, the common emotion categories (anger, disgust, fear, happiness, sadness and neutral) are used, and in the group of DES and MES, only anger, joy, surprise and sadness are used.

Table III compares the average recognition accuracy and standard deviation of the eight feature extraction methods on each experiment using 5-fold cross-validation. Even though the features are learned from a database with a different language, competitive classification performance is still achieved by SDFA. Clearly, SDFA achieves the highest accuracy and the smallest standard deviation compared with all other feature representations for all the cases.

V. Conclusion and Future Work

Learning salient, discriminative features is an important research issue for SER. The main contribution of this work is two-fold. First, we introduce feature learning to SER, in which the optimal feature set for SER are learned automatically by CNN through two-stage training: SAE and SDFA. Second, in SDFA, we propose a novel objective function that encourages the feature saliency, orthogonality, and discrimination. Consequently, our method can disentangle affect-salient features from other noisy factors such as speakers and language. Experimental results on public emotional speech databases show superior performance of the learned features with respect to speaker variation, environment distortion, and language variation when compared with several well-established feature representations.

In this paper, our focus is on the recognition of prototypic expressions of several basic emotions based on displayed emotional utterances in laboratory settings. Subtle, continuous, and context-specific interpretations of affective utterances recorded in naturalistic and real-world settings are clearly more important and more difficult research problems. Feature learning, as an advanced technique to learn a transformation of raw inputs to a representation that can be effectively exploited by a classifier, is well-suited for addressing these challenges. In the future, we plan to extend the proposed method in this paper and evaluate its performance on naturalistic speech data.

References

[1] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowiea, E. Douglas-Cowie, and C. Cox, "ASR for emotional speech: Clarifying the issues and enhancing performance," *Neural Netw.*, vol. 18, pp. 437–444, 2005.

[2] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[3] F. Burkhardt, A. Paeschke, M. Rolfes, and W. S. *et al.*, "A database of German Emotional Speech," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 1517–1520.

[4] G. Davood, S. Mansour, N. Alireza, and G. Sahar, "Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network," *Neural Comput. Appl.*, vol. 21, no. 8, pp. 2115–2126, 2012.

[5] 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. 39th IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 4818– 4822.

[6] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. 5th Biannu. Humaine Assoc. Conf. Affective Comput. Intell. Interaction*, Geneva, Switzerland, 2013, pp. 511–516.

[7] I. Engberg, A. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a Danish emotional speech database," in *Proc. 5th Eur. Conf. Speech Commun. Technol.*, Rhodes, Greece, Sep. 1997, pp. 1695–169.

[8] F. Eyben, M. Wollmer, and B. Schuller, "openEAR: Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in *Proc. Affective Comput. Intell. Interaction*, Amsterdam, The Netherlands, 2009, 1996, pp. 576–581 .

[9] X. Mao and L. Chen, "Speech emotion recognition based on parametric filter and fractal dimension," *IEICE Trans. Inf. Syst.*, vol. E93–D, no. 8, pp. 2324–2326, 2010.

[10] Z. Guo and Z. Wang, "An unsupervised hierarchical feature learning framework for one-shot image recognition," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 621–632, Apr. 2013.

[11] E. Guven and P. Bock, "Speech emotion recognition using a backward context," in *Proc. IEEE 39th Appl. Imagery Pattern Recog. Workshop*, Washington, D.C., USA, Oct. 2010, pp. 1–5.

[12] S. Haq, P. J. B. Jackson, and J. D. Edge, "Speaker-dependent audio-visual emotion recognition," in *Proc. Int. Conf. Auditory-Visual Speech Process.*, Norwich, U.K., Sep. 2009, pp. 53–58.

[13] H. Hu, M. X. Xu, and W. Wu, "Fusion of global statistical and segmental spectral features for speech emotion recognition," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, vol. 2, pp. 1013–1016.

[14] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proc. 22nd ACM Int. Conf. Multimedia*, Orlando, FL, USA, 2014.

[15] S. G. Koolagudi and S. R. Krothapalli, "Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features," *Int. J. Speech Technol.*, vol. 15, no. 4, pp. 495–511, 2012.

[16] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech using source, system, and prosodic features," *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 265–289, 2012.

[17] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in *Proc. Automac. Speech Recog. Understand.*, Olomouc, Czech Republic, 2013, pp. 216–221.

[18] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[19] H. Lee, Y. Largman, P. Pham, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. Advances Neural Inform. Process. Syst,*, 2009, pp. 1096–1104.

[20] I. Luengo, E. Navas, and I. Hernandez, "Feature analysis and evaluation for automatic emotion identification in speech," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 490–501, Oct. 2010.

[21] M. Lugger and B. Yang, "Combining classifiers with diverse feature sets for robust speaker independent emotion recognition," in *Proc. Eur. Signal Process. Conf.*, Trivandrum, Kerala, India, Dec. 2009, pp. 1225–1229.

[22] S. Mansour, B. Mahdi, and G. Davood, "Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method," *Neural Comput. Appl.*, vol. 23, no. 1, pp. 215–227, 2013.

[23] Q. R. Mao and Y. Z. Zhan, "Speech emotion recognition method based on improved decision tree and layered feature selection," *Int. J. Humanoid Robot.*, vol. 7, no. 2, pp. 245–261, 2010.

[24] E. A. Moataz, K. M. S. , and K. Fakhri, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recog.*, vol. 44, no. 3, pp. 572–587, 2011.

[25] D. Morrison, R. Wang, and L. D. Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Commun.*, vol. 49, no. 2, pp. 98–112, 2007.

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

[27] M. Pantic, A. Nijholt, A. Pentland, and T. S. Huang, "Human-centred intelligent human-computer interaction (hci^2): How far are we from attaining it?," *Int. J. Autonomous and Adaptive Commun. Syst,*, vol. 1, no. 2, pp. 168–187, 2008.

[28] T. L. Pao, Y. T. Chen, J. H. Yeh, Y. M. Cheng, and Y. Y. Lin, "A comparative study of different weighting schemes on KNN-based emotion recognition in Mandarin speech," *Advanced Intell. Comput. Theories and Appl. With Aspects of Theoretical and Methodological*, vol. 4681, pp. 997–1005, 2007.

[29] S. Ramakrishnan and I. M. M. E. Emary, "Speech emotion recognition approaches in human-computer interaction," *Telecommun. Syst,*, vol. 52, no. 3, pp. 1467–1478, 2013.

[30] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion recognition from speech using global and local prosodic features," *Int. J. Speech Technol.*, vol. 16, no. 2, pp. 143–160, 2012.

[31] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *Proc. Eur. Conf. Comput. Vision*, Florence, Italy, Oct. 2012, vol. 7577, pp. 808–822.

[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[33] E. M. Schmidt and Y. E. Kim, "Learning emotion-based acoustic features with deep belief networks," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, 2011, pp. 65–68.

[34] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Quebec, Canada, May 2004, vol. 1, pp. 577–580.

[35] S. Steidl, "Automatic classification of emotion related user states in spontaneous children's speech," Ph. D. dissertation, Comput. Sci. Dept., Univ. Erlangen-Nuremberg, Erlangen, Germany, 2009.

[36] A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *IEEE Int. Conf. Acoustics, Speech Signal Process.*, Prague, Czech Republic, 2011, pp. 5688–5691.

[37] R. Sun and E. M. Ii, "Investigating glottal parameters and Teager energy operators in emotion recognition," *Affective Comput. Intell. Interaction, Lecture Notes Comput. Sci.*, vol. 6975, pp. 425–434, 2011.

[38] Y. Sungrack and C. D. Yoo, "Loss-scaled large-margin Gaussian mixture models for speech emotion classification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 585–598, Feb. 2011.

[39] A. Tawari and M. M. Trivedi, "Speech emotion analysis: Exploring the role of context," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 502–509, Oct. 2010.

[40] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features and methods," *Speech Commun.*, vol. 48, no. 9, pp. 1162–1181, 2006.

[41] M. Wollmer, A. Metallinou, N. Katsamanis, B. Schuller, and S. Narayanan, "Analyzing the memory of BLSTM neural networks for enhanced emotion classification in dyadic spoken interactions," in *Proc. 37th Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, 2012, pp. 4157–4160.

[42] C. H. Wu and W. B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Trans. Affective Comput.*, vol. 2, no. 1, pp. 10–21, Jan.-Jun. 2011.

[43] D. Yu, M. L. Seltzer, J. Li, J. T. Huang, and S. Frank, "Feature learning in deep neural networks - Studies on speech recognition tasks," in *Proc. 1st Int. Conf. Learn. Representations*, Scottsdale, AZ, USA, 2013, pp. 1–9.

[44] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[45] Z. Zeng, J. Tu, M. Liu, and T. Huang, "Audio-visual affect recognition," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 424–428, Feb. 2007.

[46] S. Q. Zhang, "Emotion recognition in Chinese natural speech by combining prosody and voice quality features," in *Proc. 5th Int. Symp. Neural Netw.*, Beijing, China, Sep. 2008, pp. 457–464.

[47] W. Zhang, X. Zhang, and Y. Sun, "Based on EEMD-HHT marginal spectrum of speech emotion recognition," in *Proc. Int. Conf. Comput., Meas., Control Sensor Netw.*, Taiyuan, China, Jul. 2012, pp. 91–94.

Qirong Mao (M'12) received the M.S. and Ph.D. degrees in computer application technology from Jiangsu University, Zhenjiang, China, in 2002 and 2009, respectively.

She is currently an Associate Professor with the School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China. Her research interests include affective computing, pattern recognition, and multimedia analysis. She has published over 30 technical articles, some of them in premium journals and conferences such as *ACM Multimedia*.

Ming Dong (S'01–A'01–M'02) received the B.S. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 1995, and the Ph.D. degree in electrical engineering from the University of Cincinnati, OH, USA, in 2001.

He is currently an Associate Professor of Computer Science and the Director of the Machine Vision and Pattern Recognition Laboratory, Wayne State University, Detroit, MI, USA. His research interests include pattern recognition, data mining, and multimedia analysis. He has published over 90 technical articles, many in premium journals and conferences such as IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Visualization and Computer Graphics, IEEE Transactions on Neural Networks, IEEE Transactions on Computers, IEEE Transactions on Knowledge and Data Engineering, IEEE International Conference on Data Mining, IEEE Conference on Computer Vision and Pattern Recognition, *ACM Multimedia*, and *WWW*.

Dr. Dong was an Associate Editor of IEEE Transactions on Neural Networks and *Pattern Analysis and Applications* (Springer), and was on the editorial board of the *International Journal of Semantic Web and Information Systems*. He also serves as a program committee member for many related conferences.

Zhengwei Huang received the B.S. degree in computer science and technology from Jiangsu University, Zhenjiang, China, in 2012, and is currently an M.S. candidate in computer application technology with the School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China.

His research interests include affect computing and pattern recognition.

Yongzhao Zhan received the B.S. degree in computer science and technology from Fuzhou University, Fujian, China, in 1984, and the Ph.D. degree in computer science and technology from Nanjing University, Nanjing, China, in 2000.

He is currently a Professor and the Dean of the School of Computer Science and Communication Engineering with Jiangsu University, Zhenjiang, China. His research interests include multimedia analysis and pattern recognition. He has published over 60 technical articles.