Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching

Shiqing Zhang, Shiliang Zhang, *Member, IEEE,* Tiejun Huang, *Senior Member, IEEE,* and Wen Gao, *Fellow, IEEE*

***Abstract*—Speech emotion recognition is challenging because of the affective gap between the subjective emotions and low-level features. Integrating multi-level feature learning and model training, Deep Convolutional Neural Networks (DCNN) has exhibited remarkable success in bridging the semantic gap in visual tasks like image classification, object detection. This paper explores how to utilize a DCNN to bridge the affective gap in speech signals. To this end, we firstly extract three channels of log Mel-spectrograms (static, delta and delta-delta) similar to the RGB image representation as the DCNN input. Then the AlexNet DCNN model pre-trained on the large ImageNet dataset is employed to learn high-level feature representations on each segment divided from an utterance. The learned segment-level features are aggregated by a Discriminant Temporal Pyramid Matching (DTPM) strategy. DTPM combines temporal pyramid matching and optimal Lp-norm pooling to form a global utterance-level feature representation, followed by the linear Support Vector Machines (SVM) for emotion classification. Experimental results on four public datasets, *i.e.*, EMO-DB, RML, eNTERFACE05 and BAUM-1s, show the promising performance of our DCNN model and the DTPM strategy. Another interesting finding is that the DCNN model pre-trained for image applications performs reasonably good in affective speech feature extraction. A further fine-tuning on the target emotional speech datasets substantially promotes the recognition performance.**

***Index Terms*—Speech Emotion Recognition, Feature Learning, Deep Convolutional Neural Network, Discriminant Temporal Pyramid Matching, Lp-norm Pooling**

I. Introduction

Speech signals, as one of the most natural media of human communication, not only carry the explicit linguistic contents but also contain the implicit paralinguistic information about the speakers. During the last two decades, enormous efforts have been devoted to developing methods for automatically identifying human emotions from speech signals, which is called speech emotion recognition. At present, speech emotion recognition has become an attractive research topic in signal processing, pattern recognition, artificial intelligence, and so on, due to its importance in human-machine interactions [1], [2].

S.Q. Zhang, S.L. Zhang, T.J. Huang, and W. Gao are with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, China.
E-mail: {tzczsq, slzhang.jdl, tjhuang, wgao}@pku.edu.cn, all authors are corresponding authors.

S.Q. Zhang is also with the Institute of Intelligent Information Processing, Taizhou University, China.

Manuscript received xxxx, 2017; revised xxxx, 2017.

Feature extraction is a critical step to bridge the affective gap between speech signals and the subjective emotions. So far, a variety of hand-designed features have been used for speech emotion recognition [3], [4], [5]. However, these hand-designed features are usually low-level, they may hence not be discriminative enough to depict the subjective emotions. It is needed to develop automatic feature learning algorithms to extract high-level affective feature representations for speech emotion recognition.

To address this issue, the newly-emerged deep learning techniques [6] provide a possible solution. Among them, two typical deep leaning methods are Deep Neural Networks (DNN) [6], and Deep Convolutional Neural Networks (DCNN) [7]. Here, a DCNN is taken as a deep extension of the conventional Convolutional Neural Networks (CNN) [8]. Recently, deep learning techniques have been employed to automatically learn high-level feature representations from low-level data in tasks like speech recognition [9], image classification and understanding [7], [10], object detection [11]. As far as speech emotion recognition is concerned, one of the early-used deep learning methods is the DNN method. For instance, in [12], [13] a DNN is used to learn high-level feature representations from the extracted low-level acoustic features for emotion classification.

In recent years, several works [14], [15], [16], [17] have successfully employed CNNs for feature learning in speech signal processing. In [14], a 1-layer CNN is adopted to obtain promising performance for speech recognition. In [15], [16], the authors also employ a 1-layer CNN trained with a Sparse Auto-encoder (SAE) to extract affective features for speech emotion recognition. Recently, Trigeorgis *et al.*, [17] presents an end-to-end speech emotion recognition system by combining a 2-layer CNN with a Long Short-Term Memory (LSTM) [18]. Note that they employ 1-D convolution, such as frequency convolution [14], [15], [16] or time convolution [17], rather than 2-D convolution widely used in DCNN models [7], [10]. Additionally, these used 1-layer or 2-layer CNNs are much shallower compared with the deep structures in DCNN models [7], [10]. Accordingly, they may could not effectively learn affective features discriminative enough to distinguish the subjective emotions.

It has been recently found that, with deep multi-level convolutional and pooling layers, DCNNs usually exhibit much better performance than the shallow CNNs in computer vision [19], [20]. This is reasonable because the deep structures of DCNNs can effectively model the hierarchical architecture

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2766843, IEEE Transactions on Multimedia

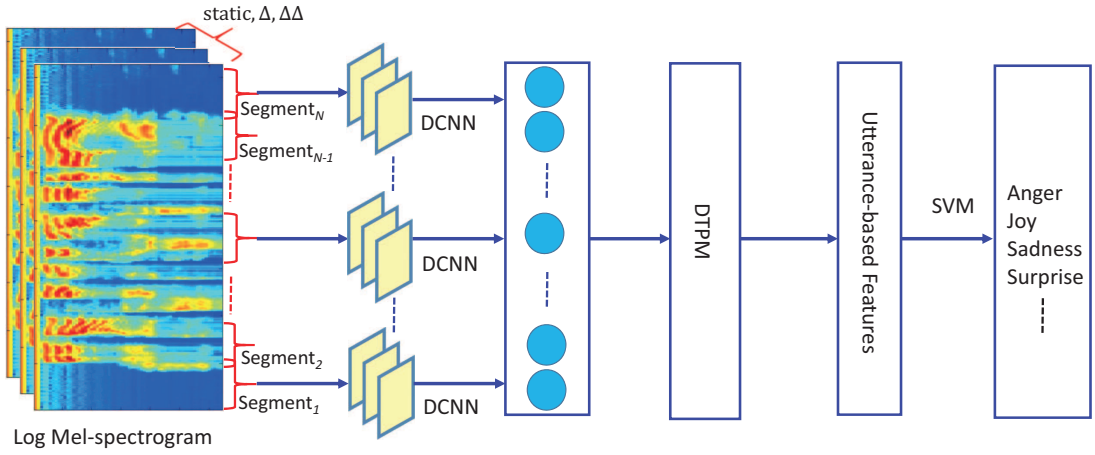

Fig. 1. An overview of the proposed speech emotion recognition framework using DCNNs and DTPM: (1) Three channels of log Mel-spectrograms (static, delta and delta-delta) are extracted and divided into N overlapping segments as the DCNN input. (2) A DCNN model is employed for automatic feature learning on each segment to generate segment-level features. (3) A DTPM scheme is designed to concatenate the learned segment-level features to form a global utterance-level feature representation. (4) With utterance-level features, a linear SVM classifier is employed to predict utterance-level emotions.

of information processing in the primate visual perception system [7], [10]. Motivated by the promising performance of deep models, this work aims to employ DCNNs to develop an effective speech emotion recognition system.

The success of DCNNs in visual tasks motivates us to test DCNNs in speech emotion recognition. To achieve this, three issues need to be addressed. First, a proper speech representation should be designed as the DCNN input. Previous works [14], [15], [16], [17] have employed 1-D speech signals as the CNN inputs, and 1-D convolution is adopted for CNNs. Compared with 1-D convolution, 2-D convolution involves more parameters to capture more detailed temporal-frequency correlations, thus is potential to present stronger feature learning ability. Therefore, it is important to convert 1-D speech signals into suitable 2-D representations as the DCNN input. Second, most existing emotional speech datasets [3], [4], [5] contain limited numbers of samples. They are not sufficient enough to train deep models having a large amount of parameters. Finally, speech signals may have variant time of duration but the DCNN models require fixed input size. It is hence easier to design the DCNN models for speech segments with a fixed length, rather than for the global utterance. Therefore, proper pooling strategies are needed to generate a global utterance-level feature representation based on the segment-level features learned by DCNNs.

In this paper, we use deep features learned by DCNNs [7] and propose a Discriminant Temporal Pyramid Matching (DTPM) algorithm to pool deep features for speech emotion recognition. As illustrated in Fig. 1, three channels of log Mel-spectrograms (static, delta and delta-delta) are extracted as the DCNN input. The DCNN models are trained to produce deep features for each segment. The DTPM pools the learned segment-level features into a global utterance-level feature representation, followed by the linear SVM emotion classifier. Extensive experiments on four public datasets, *i.e.*, the Berlin dataset of German emotional speech (EMO-DB) [21], the RML audio-visual dataset [22], the eNTERFACE05 audio-visual dataset [23], and the BAUM-1s dataset [24], demonstrate the promising performance of our proposed method.

The main contributions of this paper can be summarized as:

- We propose to use three channels of log Mel-spectrograms generated from the original 1-D utterances as the DCNN input. This input is similar to the RGB image representation, thus makes it possible to use existing DCNNs pre-trained on image datasets for affective feature extraction.
- The proposed DTPM strategy combines temporal pyramid matching and optimal Lp-norm pooling to generate a discriminative utterance-level feature representation from segment-level features learned by DCNNs.
- We find that the DCNN model pre-trained for image applications performs reasonably good in affective feature extraction. A further fine-tuning on target speech emotion recognition tasks substantially promotes the recognition performance.

The rest of this paper is structured as follows. The related works are reviewed in Section II. Section III describes our DCNN model for affective feature extraction. Section IV presents the details of our DTPM scheme. Section V describes and analyzes the experimental results. Section VI provides discussions, followed by the conclusions in Section VII.

II. Related Work

Generally, feature extraction and emotion classification are two key steps in speech emotion recognition. In this section, we first briefly review emotion classifiers and then focus on feature extraction since it is more relevant to our work.

A. Emotion Classifier

For emotion classification various machine learning algorithms have been utilized to constitute a good classifier to distinguish the underlying emotion categories. Early emotion classifiers contain K-Nearest-Neighbor (KNN) [25] and Artificial Neural Network (ANN) [26]. Then, a number of statistical

pattern recognition approaches, such as Gaussian Mixture Model (GMM) [27], Hidden Markov Models (HMM) [28], and SVM [29], are widely adopted for speech emotion recognition. Recently, some advanced classifiers based on sparse representation [30], [31] have also been studied. Nevertheless, each classifier has its own advantages and disadvantages. To integrate the merits of different classifiers, ensembles of multiple classifiers have been investigated for speech emotion recognition [32], [33].

B. Feature Extraction

Affective speech features widely used for emotion recognition can be roughly divided into four categories: 1) acoustic features [34], [35], 2) language features, such as lexical information [36], [37], 3) context information, such as subject, gender, culture influences [38], [39], 4) hybrid features [36], [40], such as the integration of two or three features above-mentioned.

Acoustic features, as one of the most popular affective features, mainly contain prosody features, voice quality features, and spectral features [34], [35]. Pitch, loudness, and duration are commonly used as prosody features [41], since they express the stress and intonation patterns of spoken language. Voice quality features, as the characteristic auditory colouring of an individual voice, have been shown to be discriminative in expressing positive or negative emotions [42]. The widely used voice quality features are the first three formants (F1, F2, F3), spectral energy distribution, harmonics-to-noise-ratio, pitch irregularity (jitter), amplitude irregularity (shimmer), and so on. Combining prosody features and voice quality features shows better performance than using prosody features alone [43], [44]. In recent years, glottal features [45] and voice source parameters [46] have been used as more advanced voice quality features for speech emotion recognition. The third typical acoustic features are spectral features, computed from the short-term power spectrum of sound, such as Linear Prediction Cepstral Coefficients (LPCC), Log Frequency Power Coefficients (LFPC) and Mel-frequency Cepstral Coefficients (MFCC). Among them, MFCC is the most popular spectral feature, since it is able to model the human auditory perception system. In recent years, modulation spectral features [47] from an auditory-inspired long-term spectro-temporal representation, and weighted spectral features [48] based on local Hu moments, have also been studied. In addition, the newly-developed Geneva minimalistic acoustic parameter set (GeMAPS) [5], such as frequency, energy, spectral related features, has shown promising performance in speech emotion recognition.

Language features, which are computed based on the verbal contents of speech, are another important representation conveying emotion information. Note that, language features are usually combined with acoustic features for speech emotion recognition [37], [36]. In [37], language features are extracted with the bag of n-gram and character n-gram approaches. Then the linguistic features are combined with acoustic features to predict dimensional emotions in a 3-D continuous space. In [36], by computing the weight of every word, a four-dimensional emotion lexicon for four emotion classes, *i.e.*, anger, joy, sadness and neutral, are obtained. Then, integrating these feature representations via early fusion and late fusion is employed for speech emotion recognition.

Context information has also been investigated in recent literatures [38], [39] for emotion recognition. In [38], the authors present a context analysis of subject and text on speech emotion recognition, and find that gender-based context information enhances recognition performance. In [39], the influences of cultural information on speech emotion recognition are explored. The authors claim that intra-cultural and multi-cultural emotion recognition paradigms give better performance than cross-cultural recognition.

Note that, since these hand-designed features mentioned above are low-level, they may not be discriminative enough to identify the subjective emotions. To tackle this issue, it may be feasible to employ deep learning techniques to automatically learn high-level affective features for speech emotion recognition.

III. DCNNs for Affective Feature Extraction

To utilize DCNNs in speech emotion recognition, three problems should be addressed. First, the DCNN input should be properly computed from 1-D speech signals. Second, DCNN's training requires a large amount of labeled data. Third, a feature pooling strategy is required to generate the global utterance-level feature representation from the DCNN outputs on local segments. In this section, we present the details of how the first two problems are addressed.

Fig. 2 illustrates the framework for affective feature extraction. From the original 1-D utterance, we first extract the static 2-D log Mel-spectrogram and then reorganize it into three channels of log Mel-spectrograms (static, delta and delta-delta). For data augmentation, the log Mel-spectrogram extracted from an utterance is divided into a certain number of overlapping segments as the DCNN input. More details about data augmentation can be found in Section V-B. Then the AlexNet DCNN model [7] pre-trained on the large-scale ImageNet dataset is employed to perform fine-tuning tasks for affective feature extraction. We present more details of the two steps in the following two sections.

A. Generation of DCNN Input

Because of the limited training data of speech emotion recognition, it is not possible to directly train a robust deep model. Motivated by the promising performance of available DCNN models, we propose to first initialize deep models with available DCNN models like AlexNet [7], then fine-tune it for transfer learning on target emotional datasets. Because available DCNN models take 2-D or 3-D images as inputs, we transform the raw 1-D speech spectrogram into 3-D array as the DCNN input.

In recent years, Abdel-Hamid *et al.*, [14] adopt the extracted log Mel-spectrogram and organize it into a 2-D array as the CNN input with a shallow 1-layer structure for speech recognition. Specifically, for each frame with a context window of 15 frames and 40 Mel-filter banks, they construct 45 (*i.e.*, 15×3) 1-D feature maps with size 40×45. Then,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2766843, IEEE Transactions on Multimedia

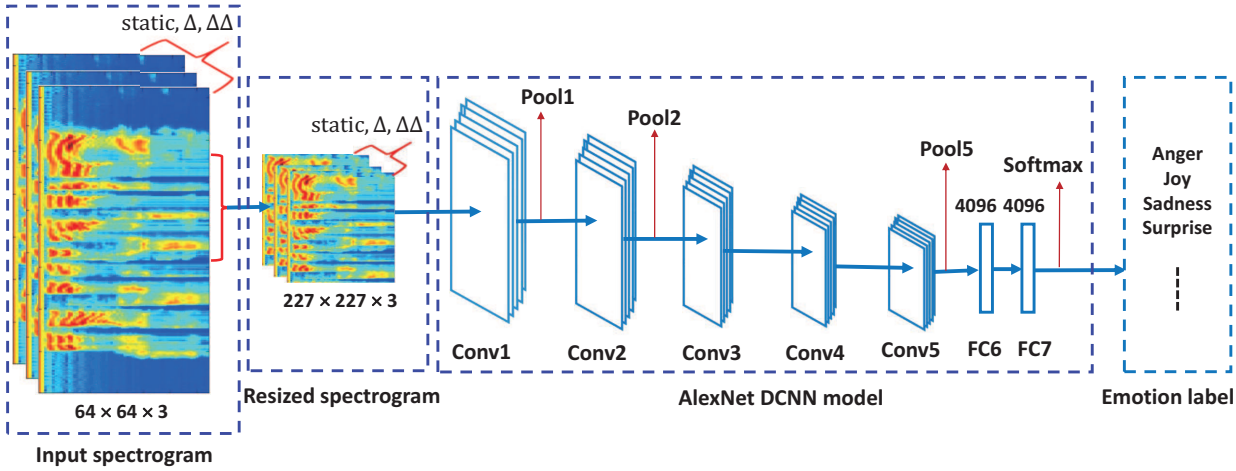

Fig. 2. The flowchart of our DCNN model for affective feature extraction. Three channels of log Mel-spectrograms with size $64 \times 64 \times 3$ (static, delta and delta-delta) are firstly produced, and then are resized to $227 \times 227 \times 3$ as the DCNN input. The DCNN model is first initialized with the AlexNet [7], then is fine-tuned on target emotional datasets. The 4096-D FC7 output is finally used as the segment-level affective features.

the 1-D convolutional kernel is applied along the frequency axis. However, speech emotion recognition using DCNNs is different from speech recognition in [14]. First, 1-D convolution operation along the frequency axis could not capture the temporal information, which is important for emotion recognition. Second, the divided segments with 15 frames (about 165 ms) used for speech recognition, are too short to distinguish emotions, since it has been found that only a speech segment length of more than 250 ms presents sufficient information for identifying emotions [49], [50].

To address these two issues, from the raw 1-D speech signals we generate the following overlapping Mel-spectrogram segments (abbreviated as $\mathrm{Mel\_SS}$) as the DCNN input

$$\mathrm{Mel\_SS} \in \mathbb{R}^{F \times T \times C}, \tag{1}$$

where F is the number of Mel-filter banks, T is the segment length corresponding to the frame number in a context window, and C ($C = 1, 2, 3$) represents the number of channels of Mel-spectrogram. Note that $C = 1$ denotes one channel of Mel-spectrogram, *i.e.*, the original static spectrogram, $C = 2$ denotes the static and delta coefficients of Mel-spectrograms, and $C = 3$ represents three channels of Mel-spectrograms including the static, delta and delta-delta coefficients of Mel-spectrogram.

As an example described in Fig. 2, we extract $\mathrm{Mel\_SS}$ with size $64 \times 64 \times 3$ ($F = 64, T = 64, C = 3$) as the input of DCNN. This kind of three channels of spectrograms is analogous to the RGB image representation of visual data. In detail, for an utterance we adopt 64 Mel-filter banks from 20 to 8000 Hz to obtain the whole log Mel-spectrogram using a 25ms Hamming window size with 10ms overlapping. Then, a context window of 64 frames is applied to the whole log Mel-spectrogram to extract the static 2-D Mel-spectrogram segments with size 64×64. A frame shift size of 30 frames is used to produce such overlapping segments of Mel-spectrogram. Each segment hence includes a context window of 64 frames and its length is $10\mathrm{ms} \times 63 + 25\mathrm{ms} = 655\mathrm{ms}$. In this case, the segment length is about 2.6 times longer than the suggested length of 250 ms in [49], [50], and conveys sufficient clues for emotion recognition.

Note that we set F as 64 because the input height-width ratio of our DCNN model is 1:1. Besides, F is usually set to be relatively large values for the usage of CNNs. For example, F is set to 40 in speech recognition [14] and 60 in speech emotion recognition [16], respectively. Therefore, it is reasonable to set F as 64 in this work.

In speech recognition, the first and second temporal derivatives on the extracted acoustic features such as MFCC, are widely used as additional features. Similarly, after extracting the static 2-D Mel-spectrogram, we also calculate the first order and second order regression coefficients along the time axis as the delta and delta-delta coefficients of Mel-spectrogram. In this way, we organize the 1-D speech signals into three channels of Mel-spectrogram segments, *i.e.*, $\mathrm{Mel\_SS}$ with size $64 \times 64 \times 3$ (three channels: static, delta and delta-delta) as the DCNN input. Then, 2-D convolution operation along the frequency axis and time axis can be performed for DCNN's training on this input.

When using the AlexNet DCNN model [7] for affective feature extraction, we have to resize the spectrogram $64 \times 64 \times 3$ into $227 \times 227 \times 3$, which is the input size of AlexNet. Since the extracted three channels of Mel-spectrograms can be regarded as the RGB image representation, we perform the resize operation with bilinear interpolation, which is commonly used for image resizing. Note that, the number of channels of Mel-spectrogram C and the segment length T may have an important impact on the learned deep features. Therefore, we will investigate their effects on the recognition accuracy in experiments.

B. DCNN Architecture

As shown in Fig. 2, our DCNN model includes five convolutional layers, three of which are followed by max-pooling layers, and two fully-connected layers. The last fully-

connected layer consists of 4096 units, giving a 4096-D feature representation. It can be observed that this structure is identical to the one of AlexNet [7], which is trained on the large-scale ImageNet dataset. The initial parameters of this DCNN model can thus be copied from the AlexNet, making this DCNN model easier to train on speech emotion recognition tasks. In the followings, we introduce the computations and principles of convolutional layer, pooling layer and fully-connected layer, respectively.

Convolutional layer: A convolutional layer employs a set of convolutional filters to extract multiple local patterns at each local region in the input space, and produces many feature maps. This can be denoted as

$$(h_k)_{ij} = (W_k \otimes q)_{ij} + b_k, \tag{2}$$

where $(h_k)_{ij}$ denotes the (i, j) element of the k-th output feature map, q represents the input feature maps, W_k and b_k denotes the k-th filter and bias, respectively. The symbol $\otimes$ represents 2-D spatial convolution operation.

Pooling layer: After each convolutional layer, a pooling layer may be used. The pooling layer aims to down-sample the obtained feature maps from the previous convolutional layers and produces a single output from local regions of convolution feature maps. Two widely used pooling operators are max-pooling and average-pooling. A max-pooling or average-pooling layer produces a lower resolution version of convolution layer activations by taking the maximum or average filter activation from different positions within a specified window.

Fully-connected layer: This layer integrates the outputs from previous layers to yield the final feature representations for classification or regression. The activation function is a sigmoid or tanh function. The output of fully-connected layers is computed by

$$x_k = \sum\nolimits_l W_{kl} q_l + b_k, \tag{3}$$

where x_k denotes the k-th output neuron, q_l denotes the l-th input neuron, W_{kl} represents the weight value connecting q_l with x_k, and b_k denotes the bias term of y_k.

Since fully-connected layers can be taken as convolutional layers with a kernel size of 1×1, Eq. (3) can be reformulated as

$$(x_k)_{1,1} = (W_k \otimes q)_{1,1} + b_k. \tag{4}$$

For DCNN's training, Stochastic Gradient Descent (SGD) is commonly employed with parameters like the batch size of examples, the momentum value (*e.g.*, 0.9), and the weight decay value (*e.g.*, 0.0005). In this case, the weight w is updated by

$$\begin{aligned} v_{i+1} &= 0.9 \cdot v_i - 0.0005 \cdot \eta \cdot w_i - \eta \cdot \langle \tfrac{\partial L}{\partial w} |w_i \rangle_{D_i}, \\ w_{i+1} &\Leftarrow w_i + v_{i+1}, \end{aligned} \tag{5}$$

where v denotes the momentum variable, η is the learning rate, i is the iteration number index, and $\langle \frac{\partial L}{\partial w} |w_i \rangle_{D_i}$ is the mean of derivatives of the i-th batch D_i. The network hence can be updated by back-prorogation. More details of DCNN's training can be found in [7].

In our DCNN's training, we first initialize the network with parameters in the AlexNet, then fine-tune the network in emotion classification tasks, which uses the Mel_SS with size $227 \times 227 \times 3$ as input and multiple emotion classes as output. Note that, the number of classes used in the AlexNet model is 1000, but in our emotion classification tasks, the number of emotion categories is 6 or 7. Therefore, our used DCNN model differs from the AlexNet in the last two layers, where our model predicts 6 or 7 emotion categories.

After fine-tuning the AlexNet model, we take the output of its FC7 layer as the segment-level affective features x. Given N overlapping Mel-spectrogram segments as the inputs of the DCNN model, we can obtain a segment-level feature representation $X{=}(x_1, x_2, \cdots, x_N) \in \mathbb{R}^{d \times N}$ with feature dimensionality $d = 4096$. This representation X hence is used as the input of the following DTPM algorithm to produce the global utterance-level features for emotion classification.

IV. DTPM for Utterance-level Feature Representation

Because of the unfixed time of duration for speech utterances, the above-mentioned segment-level features X have a variant number of segments. This unfixed dimensionality makes such segment-level features not directly useable for emotion recognition. Therefore, we proceed to convert the segment-level features into an utterance-level feature representation with fixed dimensionality. This process, which is also called as feature pooling, is widely used in computer vision to convert the local features into the global features for image classification and retrieval.

There are two types of widely-used pooling strategies, *i.e.*, average-pooling and max-pooling, which compute the averaged values and max values on each dimension, respectively. Note that, different pooling strategies are suited for different types of features, *e.g.*, max-pooling is suited for sparse features. It is difficult to decide which pooling strategy is optimal for our segment-level affective features. Moreover, most of pooling strategies discard the temporal clues of speech signals, which might be important to distinguish emotions.

Our DTPM is motivated to simultaneously embed the temporal clues and find the optimal pooling strategy. It is partially inspired by the Spatial Pyramid Matching (SPM) [51], which embeds the spatial clues during feature pooling for image classification. In SPM, an image is first divided into regions at different scales, then feature pooling is conducted on each region. The final feature is hence the concatenation of the pooled features at each scale. Similarly, we also divide the segment-level features X into non-overlapping sub-blocks along the time axis at different scales, then conduct feature pooling on each sub-block. The final concatenated feature thus integrates the temporal clues at different scales. The details will be presented in Section IV-A.

To acquire the optimal pooling strategy, we formulate the feature pooling as

$$f^p(X) = \left(\frac{1}{N} \sum_{j=1}^{N} |x_j|^p \right)^{\frac{1}{p}}, \tag{6}$$

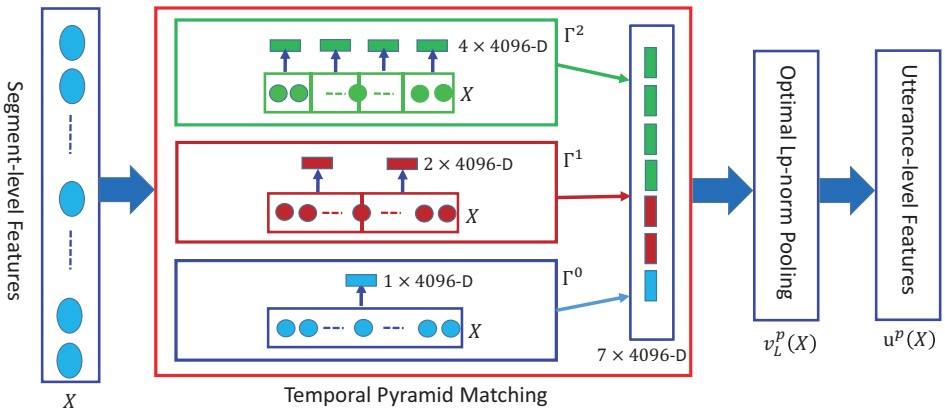

Fig. 3. The framework of Discriminant Temporal Pyramid Matching (DTPM).

where $f^p(X)$ denotes the acquired feature after pooling operation, N is the number of segment features, and p controls the pooling strategy. *E.g.*, $p = 1$ corresponds to average-pooling, whereas $p = \infty$ corresponds to max-pooling. To testify the advantages of the optimal pooling strategy, we compare it with average-pooling and max-pooling in the latter experiments, as shown in Section VI.

From Eq. (6), it can be observed that the parameter p decides the performance of the pooling. In recent years, it has been found in [52], [53] that the p value has an important impact on the image classification accuracy. Therefore, we proceed to acquire an optimal p for our affective features. More details will be presented in Section IV-B.

In a word, given the segment-level features X, DTPM aggregates X to produce the global utterance-level features $v_L^p(X)$ with the optimal pooling parameter p and pyramid level $\ell = 0, 1, 2, \cdots, L$. For example, Fig.3 presents the framework of DTPM. The original segment-level features X are divided at three scales with $\ell = 0, 1, 2$. The final feature is generated by concatenating the features at each scale with the optimal pooling parameter p.

A. Temporal Pyramid Matching

Temporal Pyramid Matching (TPM) first divides the segment-level features X at multiple levels. Specifically, $X=(x_1, x_2, \cdots, x_N) \in R^{d\times N}$ is equally divided into 2^ℓ successive non-overlapping sub-blocks along the time axis at different levels with $\ell = 0, 1, 2, \cdots, L$. For the ℓ-th level, this can be expressed as

$$X = (X_1, X_2, \cdots, X_m), \tag{7}$$

where $m=2^\ell$, $\ell = 0, 1, 2, \cdots, L$.

For a sub-block $X_m = (x_1, x_2, \cdots, x_n) \in R^{d\times n}$ with n segments, we use the pooling strategy in Eq. (6) to produce fixed-length d-dimension feature representation $f^p(X_m)$, *i.e.*,

$$f^p(X_m)= \left(\frac{1}{n} \sum_{j=1}^{n} |x_j|^p \right)^{\frac{1}{p}}, \tag{8}$$

where $p = 1$ corresponds to average-pooling, whereas $p = \infty$ corresponds to max-pooling. The optimal p will be computed in Section IV-B.

The generated feature $f^p(X_m)$ at different scales encodes different temporal clues. For example, compared with the feature at 0-th level, the pooled feature at the second level embeds more refined temporal clues. We thus aggregate the pooling results on all sub-blocks at different levels into the final global utterance-level feature representation.

Let $\Gamma^\ell(X)=(f^p(X_1), f^p(X_2), \cdots, f^p(X_m))$ denote the concatenated feature of X at pyramid level ℓ. Then we can get the global utterance-level features $v_L^p(X)$ of TPM by means of concatenating all $\Gamma^\ell(X)$ at different pyramid level, *i.e.*,

$$v_L^p(X) = (\tfrac{1}{2^L}\Gamma^0, \tfrac{1}{2^L}\Gamma^1, \tfrac{1}{2^{L-1}}\Gamma^2, \cdots, \tfrac{1}{2}\Gamma^L), \tag{9}$$

where $\Gamma^\ell(X)$ is abbreviated as Γ^ℓ. In the final utterance-level features, we set higher weights for the features on higher levels, which embeds more refined temporal clues. This is also similar to the weighting strategy in SPM [51].

B. Optimal Lp-norm Pooling

To improve the discriminative power of $v_L^p(X)$, we employ the class separability criteria according to the Marginal Fisher Analysis (MFA) [54] to learn the optimal Lp-norm pooling. Let $u^p(X)$ denote the final utterance-level features after optimal Lp-norm pooling, then we get

$$u^p(X) = \alpha^T v_L^p(X), \tag{10}$$

where α is a diagonal matrix used to weight $v_L^p(X)$, making $v_L^p(X)$ discriminant. In the following, $v_L^p(X)$ is abbreviated as v^p, and $u^p(X)$ is abbreviated as u^p. Therefore, the task is acquiring the optimal α and p to make the final feature u^p as discriminative as possible.

To optimize both α and p simultaneously, the objective function should maximize the inter-class separability while minimize the inner-class separability. This induces our objective function, *i.e.*,

$$\alpha^*, p^* = \arg\max_{\alpha,p} \Omega(\alpha, p) := \frac{\alpha^T S_b(p) \alpha}{\alpha^T S_\omega(p) \alpha}, \tag{11}$$

where $S_b(p)$ represents the inter-class separability, $S_\omega(p)$ represents the inner-class separability. They are computed by

$$\begin{aligned} S_b(p) &= \sum_i \sum_{j \in N_k^-(i)} (v_i^p - v_j^p)(v_i^p - v_j^p)^T, \\ S_\omega(p) &= \sum_i \sum_{j \in N_k^+(i)} (v_i^p - v_j^p)(v_i^p - v_j^p)^T, \end{aligned} \tag{12}$$

where $N_k^-(i)$ denotes the index set for k nearest neighbors of the pooling data v_i^p from different classes, and $N_k^+(i)$ represents the k nearest neighbors of the pooling data v_i^p from the same classes.

Eq. (11) can be solved by optimizing α and p alternatively. When fixing p, this objection function is transformed into the classical Linear Discriminant Analysis (LDA) [55], [56] problem. In this case, $S_b(p)$ and $S_\omega(p)$, represent the between-class scatter matrix and the within-class scatter matrix, respectively. Therefore, the optional solution α^* can be obtained with the closed-form solution for a fixed p:

$$\begin{aligned} \alpha^* &= \arg\max_\alpha \lambda, \\ s.t. S_b\alpha &= \lambda S_\omega \alpha. \end{aligned} \tag{13}$$

The diagonal vector for the optional solution α^* is the eigenvector corresponding to the largest eigenevalue $\lambda_{\max}$.

When fixing α, the optimizing problem in Eq. (11) has no closed-form solution. Nevertheless, it can be solved with a gradient descent process in an iterative way. Specifically, with a fixed α, we can get

$$\begin{aligned} \widetilde{S}_b(p) &= \alpha^T S_b(p)\alpha = \sum_i \sum_{j \in N_k^-(i)} (\mathrm{u}_i^p - \mathrm{u}_j^p)^2, \\ \widetilde{S}_\omega(p) &= \alpha^T S_\omega(p)\alpha = \sum_i \sum_{j \in N_k^+(i)} (\mathrm{u}_i^p - \mathrm{u}_j^p)^2. \end{aligned} \tag{14}$$

The partial derivatives of $\widetilde{S}_b(p)$ and $\widetilde{S}_\omega(p)$ related to p are then computed by

$$\begin{aligned} \tfrac{\partial \widetilde{S}_b}{\partial p} &= 2\sum_i \sum_{j \in N_k^-(i)} (\mathrm{u}_i^p - \mathrm{u}_j^p)\alpha^T(\beta_i - \beta_j), \\ \tfrac{\partial \widetilde{S}_\omega}{\partial p} &= 2\sum_i \sum_{j \in N_k^+(i)} (\mathrm{u}_i^p - \mathrm{u}_j^p)\alpha^T(\beta_i - \beta_j), \end{aligned} \tag{15}$$

where β denotes the Hadamard product $\beta = v^p \circ \ln v$. Then we can get the partial derivative of Eq. (11) with respect to p:

$$\nabla p = \frac{\partial}{\partial p}\Omega(\alpha, p) = \frac{1}{\widetilde{S}_\omega^2}\left(\frac{\partial \widetilde{S}_\omega}{\partial p}\widetilde{S}_b - \frac{\partial \widetilde{S}_b}{\partial p}\widetilde{S}_\omega\right). \tag{16}$$

The p value can be updated along the gradient direction with a step size γ, *i.e.*,

$$p^{(t+1)} = p^{(t)} + \gamma \cdot \nabla p, \tag{17}$$

where the superscript t denotes the t-th iteration. In our implementation, the iteration stops if the number of iterations exceeds the permitted number N_{iter}. After acquiring the final feature representation $\mathrm{u}^p(X)$, we use it for emotional classification with classifiers like SVM.

Our training strategy divides the utterances into segments. This enlarges the training set for DCNNs, but is potential to make emotion recognition on each segment more difficult if the segment is too short. We have carefully set the length of each segment to 655ms, which is about 2.6 times longer than the suggested 250ms for emotion recognition in [49], [50]. Therefore, each segment should preserve sufficient clues for emotion recognition. To conduct utterance-level emotion recognition, we generate utterance-level features with the DTPM, which aggregates segment-level features at different scales with Lp-norm pooling. DTPM is inspired by the Spatial Pyramid Matching (SPM) [51] commonly used in image classification. SPM aggregates low-level features from image patches to form a global feature discriminative to high-level semantics. Similar to SPM, DTPM is potential of learning a discriminative utterance-level feature from local segment-level features. In the following section, we will testy the validity of this training strategy.

V. Experiments

A. Datasets

We test the proposed method on four public datasets, including the Berlin dataset of German emotional speech (EMO-DB) [21], the RML audio-visual dataset [22], the eNTERFACE05 audio-visual dataset [23], and the BAUM-1s audio-visual dataset [24].

EMO-DB: The acted EMO-DB speech corpus [21] contains 535 emotional utterances with seven different acted emotions: anger, joy, sadness, neutral, boredom, disgust and fear. Ten professional native German-speaking actors (five female and five male) are asked to simulate these emotions, giving 10 German utterances (five short and five long sentences) which are able to be used in everyday communication. These actors are required to read these predefined sentences in the targeted seven emotions. The recordings in this dataset are taken in an anechoic chamber with high-quality recording equipment and produced at a sampling rate of 16 kHz with a 16-bit resolution and mono channel. The audio files are on average around 3 seconds long. A human perception test with other 20 subjects is conducted to evaluate the quality of the recorded data.

RML: The acted RML audio-visual dataset [22], collected from Ryerson Multimedia Research Lab, Ryerson University, contains 720 utterances of eight subjects from different gender and culture, in six different speaking languages. It consists of six emotions: anger, disgust, fear, joy, sadness, and surprise. The samples were recorded at a sampling rate of 44,100 Hz with a 16-bit resolution and mono channel. The audio files are on average around 5 seconds long. To ensure the context independency of speech samples, more than ten reference sentences for each emotion are presented. At least two participants who do not know the corresponding language are employed in human perception test to evaluate whether the correct emotion is expressed.

eNTERFACE05: The eNTERFACE05 [23] is an induced audio-visual emotion dataset with six basic emotions, *i.e.*, anger, disgust, fear, joy, sadness, and surprise. 42 subjects from 14 different nationalities are included. Each subject is asked to listen to six successive short stories, each of which is used to induce a particular emotion. Two experts are employed to evaluate whether the reaction expresses the intended emotions

in an unambiguous way. The speech utterances are pulled from video files of the subjects speaking in English. The sampling rate is 48 kHz for audio. The audio files are on average around 3 seconds long. Overall, the eNTERFACE05 dataset contains 1290 utterances.

BAUM-1s: The spontaneous BAUM-1s [24] audio-visual dataset contains eight emotions (joy, anger, sadness, disgust, fear, surprise, boredom and contempt), and four mental states (unsure, thinking, concentrating and bothered). It has 1222 utterances collected from 31 Turkish subjects, 17 of which are female. Emotion elicitation using video clips is employed to get spontaneous audio-visual expressions. Each utterance is given an emotion label by using a majority voting over the five annotators. The audio files have a sampling rate of 48 kHz, and the average time of duration is around 3 seconds. As done in [22], [23], this work aims to identify six basic emotions (joy, anger, sadness, disgust, fear, surprise), giving 521 utterances in total for experiments. Note that, BAUM-1s, is a latest audio-visual emotional data set released in 2016. Moreover, BAUM-1s records spontaneous emotions rather than acted emotions, thus defines a more challenging emotion recognition problem than the aforementioned datasets like EMO-DB and eNTERFACE05. Therefore, BAUM-1s is a reasonable and challenging testset.

B. Experimental Setup

1) Details of DCNN Training: Each of the four emotional datasets contains a limited number of samples. It is thus desirable to generate more samples for DCNN's training. To address this issue, we directly split an utterance into a certain number of overlapping segments. Each of the segments is labeled with the utterance emotion category for DCNN's training. In this case, the number of training samples is decided by the overlap length (a frame shift size) between two adjacent segments, *i.e.*, smaller overlap results in a larger number of training samples. However, as suggested in [50], the overlap length should be larger than 250 ms in speech emotion recognition. Therefore, we set the overlap length as 30 frames, which is about $10\text{ms} \times 29 + 25\text{ms} = 315\text{ms}$. As a result, when extracting Mel-spectrogram segments with size $64 \times 64 \times 3$, we can significantly augment the size of training data, *i.e.*, from 535 utterances to 11,842 segments for the EMO-DB dataset, from 720 utterances to 11,316 segments for the RML dataset, from 1290 utterances to 16,186 segments for the eNTERFACE05 dataset, and 521 utterances to 6368 segments for the BAUM-1s dataset, respectively.

Note that, segmenting an utterance into small segments, was widely used for discrete emotion classification, as in [13], [57], [58]. Although it is not necessarily true that the emotion labels in all segments divided from an utterance are equivalent to that of the whole utterance, we can still employ DCNNs to learn effective segment-level features from the segment-level emotions, which can be utilized to predict utterance-level emotions.

The structure of the used DCNN model [7] is presented in Fig. 2. The DCNN model is trained with mini-batch size of 30, Stochastic Gradient Descent (SGD) with a momentum of 0.9, and a learning rate of 0.001. The maximum number of epochs is set as 300. We perform DCNNs on the MATLAB2014 platform with the MatConvNet package [59], which is a MATLAB toolbox implementing CNNs for computer vision applications. One NVIDIA GTX TITAN X GPU with a 12GB memory is used to train DCNNs with a GPU mode. We employ the LIBSVM package [60] with the linear kernel function and the one-versus-one strategy for multi-class classification. When implementing optimal Lp-norm pooling, we set the number of permitted iteration $N_{iter} = 50$, and the number of nearest neighbors $k = 20$, as done in [52].

It is noted that the used DCNN model called AlexNet, is firstly reported in [7] with input size of $224 \times 224 \times 3$. However, in many practical implementations such as imagenet-caffe-alex, available at http://www.vlfeat.org/matconvnet/pretrained/, researchers commonly use input size $227 \times 227 \times 3$ rather than $224 \times 224 \times 3$.

2) Evaluation Methods: As suggested in [61], test-runs are implemented by using a speaker-independent Leave-One-Speaker-Out (LOSO) or Leave-One-Speakers-Group-Out (LOSGO) cross-validation strategy, which are usually adopted in most real applications. Specifically, for the EMO-DB and RML datasets, we employ the LOSO scheme. For the eNTERFACE05 and BAUM-1s datasets, we use the LOSGO scheme with five speakers group, similar to [24]. Note that, we adopt the speaker-independent test-runs, which is more realistic and challenging than the speaker-dependent test-runs. Therefore, we only compare with works also using the same setting and wont compare with works like [58] that report speaker-dependent results. The Weighted Average Recall (WAR), also known as the standard accuracy, is reported to evaluate the performance of speech emotion recognition. Here, WAR denotes the recognition rates of individual classes weighted by the class distribution.

We evaluate the performance of two methods, *i.e.*, *DCNN-Average*, and *DCNN-DTPM*. The details of these two methods are described below.

DCNN-Average also uses DCNNs as feature extractor. After extracting features on each Mel-spectrogram segment with DCNNs, the conventional average-pooling is employed over all the segments to produce the final fixed-length global utterance-level features. Then the linear SVM classifier is adopted for emotion identification. Therefore, we compare our method to DCNN-Average to show the validity of the proposed DTPM.

DCNN-DTPM is our proposed method described in Fig.3.

C. Experimental Results and Analysis

We use Mel-spectrogram segments with size $\text{Mel\_SS} \in \mathbb{R}^{F \times T \times C}$ as the DCNN input, where F is the number of Mel-filter banks commonly set as 64, T is the number of frames in each segment, and C represents the number of channels of Mel-spectrogram. The parameters C and T largely affects the amount of affective cues DCNNs could perceive. In this part, we first investigate the effects of C and the validity of our DCNN's training strategy. Then we will validate the effects

TABLE II
SPEAKER-INDEPENDENT RECOGNITION ACCURACY (%) USING THE ALEXNET WITHOUT FINE-TUNING AS A FEATURE EXTRACTOR. L^* DENOTES THE VALUE OF L CORRESPONDING TO THE BEST PERFORMANCE.

Dataset	EMO-DB	RML	eNTERFACE05	BAUM-1s
DCNN-Average	72.35	59.46	51.33	36.10
DCNN-DTPM (L^*)	76.27 (3)	62.40 (3)	56.08 (2)	38.42 (2)

TABLE III
SPEAKER-INDEPENDENT RECOGNITION ACCURACY (%) USING THE FINE-TUNED ALEXNET AS A FEATURE EXTRACTOR. L^* DENOTES THE VALUE OF L CORRESPONDING TO THE BEST PERFORMANCE.

Dataset	EMO-DB	RML	eNTERFACE05	BAUM-1s
DCNN-Average	82.65	66.17	72.80	42.26
DCNN-DTPM (L^*)	87.31 (2)	69.70 (2)	76.56 (2)	44.61 (2)

of T on the recognition accuracy and compare to the state-of-the-arts.

1) Effects of the number of channels in Mel-spectrogram: To investigate the effects of the number of channels, *i.e.*, C, we use a simplified DCNN model for feature extraction. This DCNN model contains five layers (Conv1-Pool1-Conv2-Pool2-Conv3-Conv4-FC5) and finally generates a 600-D feature representation. Specifically, the size of the input is $64\times64\times C$, the first three convolutional layers (Conv1, Conv2, Conv3) have 128 kernels of size 5×5 with a stride of 1. The fourth convolution layer (Conv4) has 256 kernels of size 4×4 with a stride of 1. We adopt average-pooling for the pooling layers. Pooling size of 3×3 with a stride of 3 is used for Pool1, and 2×2 with a stride of 2 is used for Pool2. The fully-connected layer in FC5 has 600 neurons, giving a 600-D feature representation. For the DCNN inputs with different C, we change the number of input channel of this DCNN model.

Table I presents performance comparisons with different values of C. Note that, for DCNN-DTPM, we test different pyramid levels with $L = 1$, 2, and 3, respectively. We present the best performance as well as the corresponding L in Table I. From the results, we can make the following two observations.

First, setting $C = 3$ shows the best performance at most cases, and constantly outperforms the case when $C = 1$. This indicates that the first order and second order derivatives of 2-D Mel-spectrogram segments preserve helpful cues for emotion recognition. The fact that, $C = 3$ slightly outperforms $C = 2$ indicates that further introducing higher order of derivatives may not significantly boost the performance. Nevertheless, $C = 3$ results in an input similar to the RGB image representation. Accordingly, we set $C = 3$ in our following experiments.

Second, DCNN-DTPM clearly outperforms DCNN-average on four datasets. It is also clear that dividing the segment-level features into multiple levels, *i.e.*, setting L larger than 1, improves the performance of DCNN-DTPM. This demonstrates the advantages of our DTPM over the conventional average-pooling strategy when coding the local segment-level features.

TABLE IV
THE BEST RECOGNITION ACCURACY (%) AND CORRESPONDING T USING THE FINE-TUNED ALEXNET ON FOUR DATASETS. L^* DENOTES THE VALUE OF L CORRESPONDING TO THE BEST PERFORMANCE.

Fine-tuning	EMO-DB	RML	eNTERFACE05	BAUM-1s
Segment length	$T = 64$	$T = 220$	$T = 80$	$T = 64$
DCNN-DTPM (L^*)	87.31 (2)	75.34 (3)	79.25 (2)	44.61 (2)

2) The performance of DCNN pre-trained on ImageNet: The above experiment suggests $C = 3$, corresponding to a DCNN input similar to the RGB image representation. Such input can be directly processed by available DCNNs pre-trained on large-scale image datasets. In this experiment, we first directly use the original AlexNet [7] to extract affective features. Then, we fine-tune the AlexNet on the target emotion recognition tasks and test the performance of the fine-tuned model. Note that, to use the Alexnet we resize $64\times64\times3$ spectrogram to $227\times227\times3$ with bilinear interpolation.

Table II gives the recognition performance obtained by the AlexNet without fine-tuning. It can be observed that, the AlexNet shows reasonably good performance, *e.g.*, on the RML dataset it gives performance close to the results in Table I obtained with the simplified DCNN model. This demonstrates that, although the AlexNet is trained on an independent image dataset, it also extracts discriminative affective features from emotional speech datasets with our DCNN input.

We further show the performance of the fine-tuned AlexNet in Table III. It is easy to observe that the fine-tuning procedure significantly boosts the discriminative power of the extracted features. After fine-tuning, the best performance of DCNN-DTPM comes up to 87.31%, 69.70%, 76.56%, and 44.61%, respectively on four datasets. Note that the recognition performance on the spontaneous BAUM-1s dataset is much lower than the obtained performance on other three emotional datasets. This shows that the spontaneous emotions are more difficult to be identified well than the acted and induced emotions. It is also clear that the fine-tuned AlexNet significantly outperforms the simplified DCNN in Table I, which is trained directly on the target datasets. This indicates the advantages of our DCNN's training strategy, *i.e.*, using available models trained on image datasets to initialize our DCNNs for fine-tuning. Moreover, the experimental results also show the validity of our generated DCNN input.

The comparisons among Table I, Table II, and Table III clearly show the advantages of our training strategy, *i.e.*, initialize with the AlexNet, then fine-tune on the target emotional speech datasets. The reason why the AlexNet helps emotion recognition might be because we convert the audio signals into an image-like representation, as well as the deep structure and huge training data of the AlexNet.

3) Effects of the segment length: The segment length T decides the duration of audio signals the DCNN model processes. It hence may largely affect the discriminative power of the extracted affective features. We thus show the effects of T on the emotion recognition performance.

The length of the shortest utterance is 1.23 second long

TABLE I
SPEAKER-INDEPENDENT ACCURACY (%) COMPARISONS BY SETTING DIFFERENT VALUES OF C USING A SIMPLIFIED DCNN MODEL. THE SIZE OF THE SPECTROGRAM IS $64 \times 64 \times C$. L^* DENOTES THE VALUE OF L CORRESPONDING TO THE BEST PERFORMANCE OF DCNN-DTPM.

Dataset	EMO-DB			RML			eNTERFACE05			BAUM-1s		
C	1	2	3	1	2	3	1	2	3	1	2	3
DCNN-Average	73.86	77.44	78.92	59.25	61.84	61.19	62.33	65.01	66.42	36.49	38.21	38.62
DCNN-DTPM (L^*)	77.03 (3)	82.69 (2)	**83.53** (3)	61.48 (2)	**64.88** (3)	64.21 (3)	65.95 (1)	69.88 (2)	**70.25** (2)	38.74 (2)	39.05 (3)	**40.57** (2)

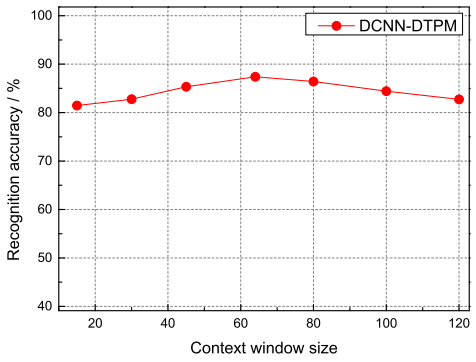

Fig. 4. The effects of T on the EMO-DB dataset.

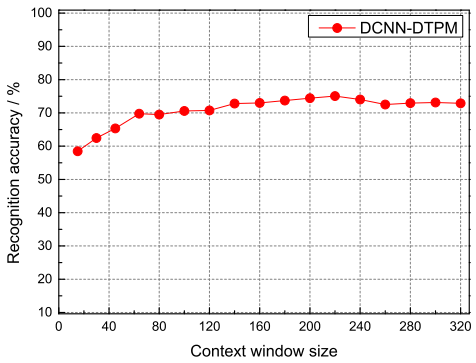

Fig. 5. The effects of T on the RML dataset.

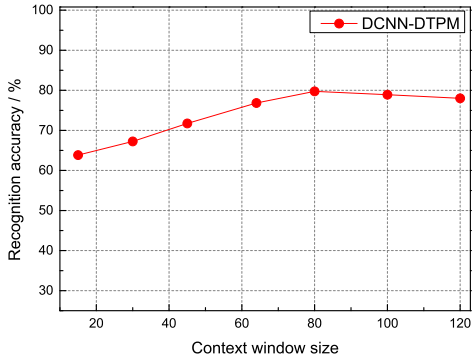

Fig. 6. The effects of T on the eNTERFACE05 dataset.

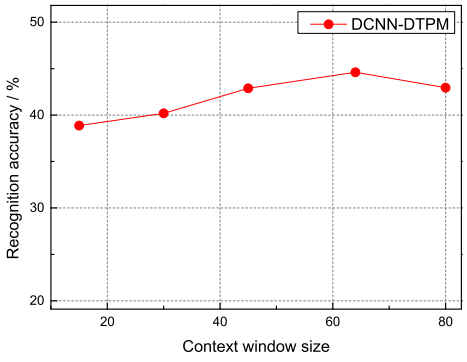

Fig. 7. The effects of T on the BAUM-1s dataset.

on the EMO-DB dataset, and 1.12 second long on the eNTERFACE05 dataset. Accordingly, for the EMO-DB dataset and the eNTERFACE05 dataset, we test T ranges in [15, 30, 45, 64, 80, 100, 120], where $T = 120$ corresponds to about 1.22 second, which is close to the length of the shortest utterance. The length of the shortest utterance is 3.27 seconds long on the RML dataset. We thus test T ranges in [15, 30, 45, 64, 80, 100, 120, 140, $\cdots$, 320] on the RML dataset. On the BAUM-1s dataset, we test T ranges in [15, 30, 45, 64, 80], since the length of the shortest utterance is 0.768 seconds long. For some certain utterances shorter than T, we simply repeat the first frame and last frame in an utterance so that the length of this utterance equals to T. Note that for $T = 15$, as a benchmark used in speech recognition, the overlap length of Mel-spectrogram segments is 15 frames, whereas for $T \geq 30$ the overlap length is 30 frames. All spectrograms with different T are resized to be $227 \times 227 \times 3$ with bilinear interpolation as the input of DCNN. Fig.4, Fig.5, Fig.6, and Fig.7 show the effects of T on four datasets. Table IV presents the best performance and the optimal T on four datasets. From the experimental results, we can draw two conclusions.

First, it can be observed that larger T is helpful for better performance. However, too large T does not constantly improve the performance. Table IV shows that the best performance on four datasets are 87.31%, 75.34%, 79.25%, and 44.61%, respectively. The corresponding optimal T on four datasets are 64, 220, 80, and 64, respectively. This may be because setting larger T decreases the number of generated training samples for DCNNs. Therefore, DCNN-DTPM does not always improve the performance with the increase of the

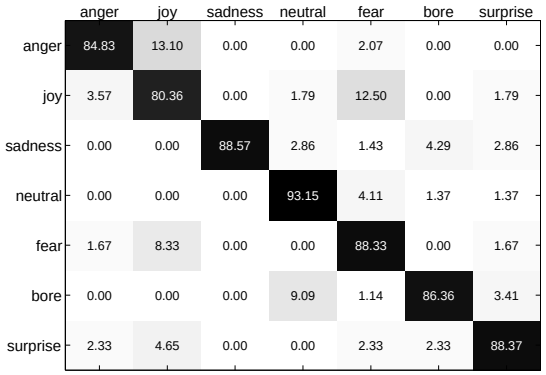

Fig. 8. Confusion matrix of DCNN-DTPM with an average accuracy of 87.31% on the EMO-DB dataset.

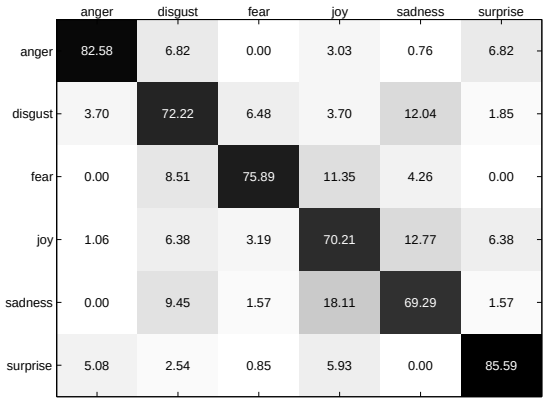

Fig. 9. Confusion matrix of DCNN-DTPM with an average accuracy of 75.34% on the RML dataset.

segment length.

Second, the four curves shows that the recognition performance of DCNN-DTPM remains stable when T is larger than 64. Setting $T = 64$ generally gives promising performance on four datasets. This might be because the DTPM also considers the temporal clues, thus makes the algorithm more robust to T. It is also interesting to observe that segment length of 15 frames, *i.e.*, $T = 15$, widely used for speech recognition [14], does not get promising emotion recognition performance. This might be because $T = 15$ is too short to provide sufficient temporal cues for distinguishing emotions.

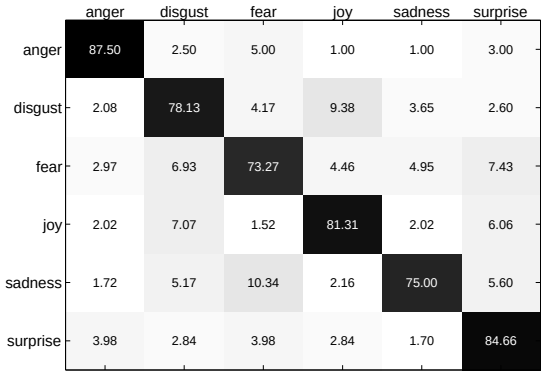

Fig. 10. Confusion matrix of DCNN-DTPM with an average accuracy of 79.25% on the eNTERFACE05 dataset.

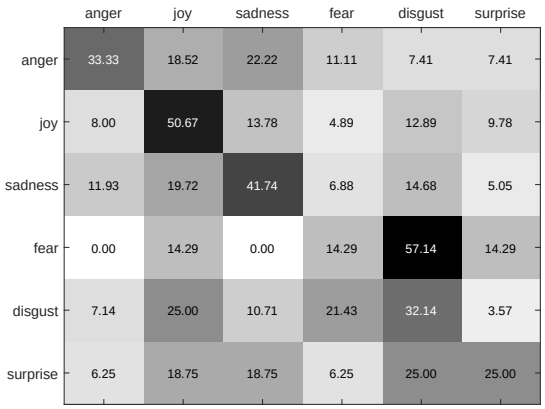

Fig. 11. Confusion matrix of DCNN-DTPM with an average accuracy of 44.61% on the BAUM-1s dataset.

TABLE V
COMPARISONS OF RECOGNITION ACCURACY (%) WITH STATE-OF-THE-ART WORKS. HERE, "FEATURES" DENOTES THE USED AFFECTIVE FEATURES IN THOSE WORKS.

Datasets	Refs.	Features	WAR	UAR
EMO-DB	[61]	Prosody, MFCC	85.60	84.60
	[12]	Prosody, MFCC	81.90	79.10
	[5]	ComParE set	N/A	86.00
	[62]	AVEC-2013 set	N/A	86.10
	Ours	DCNNs	**87.31**	**86.30**
RML	[63]	Prosody	51.04	N/A
	[64]	PNCC	58.33	N/A
	Ours	DCNNs	**75.34**	**75.20**
eNTERFACE05	[61]	Prosody, MFCC	72.40	72.50
	[12]	Prosody, MFCC	61.10	61.10
	[24]	MFCC, RASTA-PLP	72.95	N/A
	[62]	ComParE set	N/A	80.50
	Ours	DCNNs	**79.25**	**79.40**
BAUM-1s	[24]	MFCC,RASTA-PLP	29.41	N/A
	Ours	DCNNs	**44.61**	**44.03**

To further investigate the recognition accuracy, we present the confusion matrix corresponding to the results of DCNN-DTPM in Table IV. Fig.8 shows that on the EMO-DB dataset, "neutral" is identified with the highest accuracy of 93.15%, and the other six emotions are classified with accuracies higher than 80%. Fig.9 indicates that only two emotions, *i.e.*, "anger" and "surprise", are distinguished with accuracies higher than 82% on the RML dataset. On the eNTERFACE05 dataset, "anger", "joy" and "surprise" can be recognized with accuracies of 87.50%, 81.31%, 84.66%, respectively, as shown in Fig.10. Fig.11 indicates that on the BAUM-1s dataset "joy" and "sadness", are classified with accuracies of 50.67%, 41.74%, respectively, whereas the other four emotions are identified with accuracies lower than 40%. The low recognition accuracies on the BAUM-1s dataset demonstrate the difficulty in recognizing spontaneous emotions.

4) Comparisons with the state-of-the-art results: We compare our method with some previous works on four public datasets in Table V. We compare with these works because they also use the speaker-independent LOSO or LOSGO test-runs, which are more reasonable than the speaker-dependent test-runs used in [22]. Note that some pervious works [5], [62] also employ Unweighted Averaged recall (UAR), which is used to better reflect unbalance among classes, as the

evaluate measures of recognition performance, although we have presented the common WAR for performance evaluation. Accordingly, we present both WAR and UAR on these four datasets for a fair comparison.

From Table V, we can see that our method is very competitive to the state-of-the-art results. Specially, on the EMO-DB dataset our method performs best, compared with [5], [12], [61], [62]. On the RML dataset, our method gives much better performance than [63], [64]. On the eNTERFACE05 dataset, our method obviously outperforms [12], [61], [24], and presents a little lower performance than [62]. On the BAUM-1s dataset, our method also clearly outperforms [24], *i.e.*, our 44.61% vs. 29.41% of [24] in term of WAR. Therefore, although the BAUM-1s is a relatively small dataset, it defines a challenging emotion recognition problem and also validates the advantages of the proposed algorithm. Note that in [61], the authors employ 6552 LLD acoustic features such as prosody and MFCC for emotion classification. This shows the advantages of our learned affective features using DCNNs. [12] also uses a DNN to learn discriminative features. Different from our work, [12] learns features from 6552 LLD acoustic features, rather than from the raw speech signals or the spectrogram. This thus clearly shows the advantages of our DCNN model, *i.e.*, using three channels of spectrograms as input and coding raw DCNN features with DTPM to get the final feature representation. [62] reports the best performance of by using the large AVEC-2013 feature set [65] on the EMO-DB dataset, and the large ComParE feature set [66] on the eNTERFACE05 dataset.

Our experimental results show that our method gets impressive recognition accuracies in comparison with the state-of-the-art works. For example, we report an UAR accuracy of 86.30% on the EMO-DB dataset, on which outperforms all the three compared works, *i.e.*, 79.1% by [12], 84.6% by [61], 86.0% by [5] and 86.1% by [62]. As far as we know, this is an early work using DCNNs pre-trained on image domain for emotion recognition. The success of this work guarantees further investigation in this direction. These distinctive characteristics distinguish our work from existing efforts on speech emotion recognition.

VI. Discussions

The pyramid level L controls the number of levels in DTPM, thus may affect the recognition performance. In our experiments, we investigate the effects of L with a value range between 1 and 3. We do not use $L \geq 4$, since the resulted feature dimensionality is too large. As shown in the above experimental results, $L = 2$ or $L = 3$ generally gets the optimal results. This indicates that dividing the Mel-spectrogram into multiple levels, *i.e.*, $L \geq 2$, helps to improve the performance. It also can be inferred that our algorithm is not quite sensitive to L, and setting $L = 2$ or $L = 3$ is a reasonable option at most cases.

To verify the effectiveness of our Lp-norm pooling, we compare it with two commonly used pooling methods, *i.e.*, average-pooling and max-pooling, in Table VI. This is conducted by modifying the value of p in DTPM, *e.g.*, $p = 1$ corresponds to average-pooling, whereas $p = \infty$ corresponds to max-pooling. It can be seen from Table VI that our Lp-norm pooling performs better than the other two pooling methods. It also can be seen that, it is hard to decide which pooling strategy performs better for a specific task with experience. *E.g.*, max-pooling performs better than average-pooling on the RML and eNTERFACE05 datasets, but average-pooling performs better on the EMO-DB and BAUM-1s datasets. This thus shows the necessarily of pooling strategy learning.

Since the Mel-spectrogram domain is represented as a 2-D matrix, it is natural to utilize CNNs to learn emotion information. To this end, it is straightforward to train a deep model on 64×64 spectrogram data. However, Table I and Table IV indicate that directly using 64×64 features to train a deep model obtains lower performance than our fine-tuned AlexNet. The reason might be the limited training data of speech emotion recognition. This motivates us to use the pre-trained AlexNet, which is already trained with millions of images and shows reasonably good performance in emotion feature extraction as shown in Table II. Therefore, we initialize a deep model with the same structure and parameters of the AlexNet and fine-tune it on target emotional datasets. Experimental results in Table II and Table IV have shown the effectiveness of the pre-trained AlexNet as well as our fine-tuned deep model.

It is a challenging problem to collect and annotate large numbers of utterances for emotion classification due to the difficulty of emotion annotation. At present, on existing small emotional speech datasets, it is a good choice to fine-tune pre-trained deep models. As shown in our experiments, fine-tuned the AlexNet pre-trained on the ImageNet works well on speech emotion recognition tasks. The reason why the AlexNet helps emotion recognition might be because we convert the audio signals into an image-like representation as well as the strong feature learning ability of the AlexNet, *e.g.*, higher-level convolutions gradually deduce semantics from larger receptive fields. The extracted three channels of Mel-spectrograms are analogous to the RGB image representation. This representation makes it feasible to first generate meaningful low-level time-frequency features with low-level 2-D convolutions, then deduce more discriminative features with higher-levels of convolutions. Besides, three channels of Mel-spectrograms may characterize emotions as certain shapes and structures, which are thus able to be effectively perceived by the AlexNet pre-trained on the image domain.

The proposed method is based on the AlexNet. Similar to the AlexNet for ImageNet large-scale classification, our method is capable of learning on million-scale training data with the commonly used GPU, *e.g.*, NVIDIA TITAN X. It is thus also interesting to retrain deep models on larger emotional speech datasets than the used EMO-DB, eNTERFACE05, and BAUM-1s in our future work.

VII. Conclusions and Future Work

This paper is motivated by how to employ DCNNs for automatic feature learning on speech emotion recognition tasks. We present a new method combining DCNNs with DTPM for

TABLE VI
RECOGNITION ACCURACY (%) COMPARISON OF THREE POOLING METHODS IN DTPM USING $64 \times 64 \times 3$ MEL-SPECTROGRAM AND $L = 2$ ON THREE DATASETS. p^* DENOTES THE MEAN OPTIMAL VALUES OF p IN LOSO OR LOSGO TEST-RUNS.

Feature pooling	EMO-DB	RML	eNTERFACE05	BAUM-1s
Average	83.28	60.73	71.08	41.94
Max	82.64	63.48	72.75	40.26
Ours (p^*)	**87.31**(1.12)	**69.70**(1.50)	**76.56**(1.58)	**44.61**(0.21)

automatic affective feature learning. A DCNN is used to learn discriminative segment-level features from three channels of log Mel-spectrograms similar to the RGB image representation. DTPM is designed to aggregate the learned segment-level features into the global utterance-level feature representation for emotion recognition. Extensive experiments on four data sets show that our method can yield promising performance in comparison with the state-of-the-arts. In addition, we also find that with our generated DCNN input, DCNN models pre-trained on the large-scale ImageNet data could be leveraged in speech affective feature extraction. This makes DCNN's training with a limited amount of annotated speech data easier. The success of this work warranties further investigation on using deep learning in speech emotion recognition.

Although this paper focuses on discrete emotion recognition, it is interesting to explore the effectiveness of deep features in continuous dimension emotion recognition on datasets like SEMAINE [67], RECOLA [68] and JESTKOD [69]. Note that this work focuses on global utterance-level emotion classification and proposes the algorithm accordingly, *i.e.*, first uses DCNNs to extract segment-level feature, then aggregates segment-level features with DTPM to form a global feature, and finally performs emotion classification with the linear SVM. Therefore, this algorithm is still not capable to deal with continuous dimensional emotion recognition. To tackle this problem, one possible way is to consider extra temporal cues and combine CNN and LSTM [17], which is commonly used to select and accumulate frame-level features for video categorization. This will be one of our future works. Moreover, there are many open issues that still need to be further studied to make emotion recognition work well in real-life settings. For example, as show in Table V, it is more difficult for our model to recognize the spontaneous emotions. It is also necessary to take the personality into consideration because different persons may have different ways to express emotions. Additionally, it is also interesting to apply our proposed method for affective analysis of music video [70].

VIII. ACKNOWLEDGEMENTS

This work was supported in part by National Science Foundation of China (NSFC) and Zhejiang Provincial National Science Foundation of China under Grant No. 61572050, LY16F020011, 91538111, 61620106009, and the National 1000 Youth Talents Plan.

REFERENCES

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

[2] S. Ramakrishnan and I. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, no. 3, pp. 1467–1478, 2013. 1

[3] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011. 1, 2

[4] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015. 1, 2

[5] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "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. 1, 2, 3, 11, 12

[6] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006. 1

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105. 1, 2, 3, 4, 5, 8, 9

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 1

[9] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012. 1

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 346–361. 1, 2

[11] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, Boston, USA, 2015, pp. 5325–5334. 1

[12] A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, 2011, pp. 5688–5691. 1, 11, 12

[13] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proceedings of INTERSPEECH*, Singapore, 2014, pp. 223–227. 1, 8

[14] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014. 1, 2, 3, 4, 11

[15] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proceedings of the ACM International Conference on Multimedia*, NewYork, USA, 2014, pp. 801–804. 1, 2

[16] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014. 1, 2, 4

[17] G. Trigeorgis, F. Ringeval, R. Bruckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *41st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5200–5204. 1, 2, 13

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 1

[19] D. C. Ciresan, U. Meier, J. Masci, L. Maria Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 22, no. 1, Barcelona, Spain, 2011, pp. 1237–1242. 1

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions,"

in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 1–9. 1

[21] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech." in *Interspeech*, vol. 5, Lisbon, Portugal, 2005, pp. 1517–1520. 2, 7

[22] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals*," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 936–946, 2008. 2, 7, 8, 11

[23] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops*, Atlanta, GA, USA, 2006, pp. 8–8. 2, 7, 8

[24] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: a spontaneous audio-visual face database of affective and mental states," *IEEE Transaction on Affective Computing*, 2016. 2, 7, 8, 11, 12

[25] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Fourth International Conference on Spoken Language (ICSLP'96)*, vol. 3, Philadelphia, PA, USA, 1996, pp. 1970–1973. 2

[26] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural computing & applications*, vol. 9, no. 4, pp. 290–296, 2000. 2

[27] D. Ververidis and C. Kotropoulos, "Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm," in *IEEE International Conference on Multimedia and Expo (ICME)*, Amsterdam, The Netherlands, 2005, pp. 1500–1503. 3

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

[29] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP'04)*, vol. 1, Montreal, Quebec, Canada, 2004, pp. I–577. 3

[30] X. Zhao, S. Zhang, and B. Lei, "Robust emotion recognition in noisy speech via sparse representation," *Neural Computing and Applications*, vol. 24, no. 7-8, pp. 1539–1553, 2014. 3

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

[32] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech communication*, vol. 49, no. 2, pp. 98–112, 2007. 3

[33] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Computer Speech & Language*, vol. 25, no. 3, pp. 556–570, 2011. 3

[34] I. Luengo, E. Navas, and I. Hernáez, "Feature analysis and evaluation for automatic emotion identification in speech," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 490–501, 2010. 3

[35] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69–75, 2015. 3

[36] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, 2015, pp. 4749–4753. 3

[37] B. Schuller, "Recognizing affect from linguistic information in 3D continuous space," *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 192–205, 2011. 3

[38] A. Tawari and M. M. Trivedi, "Speech emotion analysis: exploring the role of context," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 502–509, 2010. 3

[39] M. A. Quiros-Ramirez and T. Onisawa, "Considering cross-cultural context in the automatic recognition of emotions," *International Journal of Machine Learning and Cybernetics*, vol. 6, no. 1, pp. 119–127, 2015. 3

[40] H. Cao, A. Savran, R. Verma, and A. Nenkova, "Acoustic and lexical representations for affect prediction in spontaneous conversations," *Computer speech & language*, vol. 29, no. 1, pp. 203–217, 2015. 3

[41] V. A. Petrushin, "Emotion recognition in speech signal: experimental study, development, and application," in *6th International Conference on Spoken Language Processing*, Beijing, China, 2000, pp. 222–225. 3

[42] R. Tato, R. Santos, R. Kompe, and J. M. Pardo, "Emotional space improves emotion recognition." in *INTERSPEECH*, Denver,Colorado, 2002, pp. 2029–2032. 3

[43] M. Lugger and B. Yang, "The relevance of voice quality features in speaker independent emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Honolulu, HI, 2007, pp. 17–20. 3

[44] S. Zhang, "Emotion recognition in Chinese natural speech by combining prosody and voice quality features," in *Advances in Neural Networks-ISNN 2008*. Springer, 2008, pp. 457–464. 3

[45] A. I. Iliev, M. S. Scordilis, J. P. Papa, and A. X. Falcão, "Spoken emotion recognition through optimum-path forest classification using glottal features," *Computer Speech & Language*, vol. 24, no. 3, pp. 445–460, 2010. 3

[46] J. Sundberg, S. Patel, E. Björkner, and K. R. Scherer, "Interdependencies among voice source parameters in emotional speech," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 162–174, 2011. 3

[47] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011. 3

[48] Y. Sun, G. Wen, and J. Wang, "Weighted spectral features based on local Hu moments for speech emotion recognition," *Biomedical Signal Processing and Control*, vol. 18, pp. 80–90, 2015. 3

[49] E. M. Provost, "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, 2013, pp. 3682–3686. 4, 7

[50] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, 2013. 4, 7, 8

[51] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, New York, USA, 2006, pp. 2169–2178. 5, 6, 7

[52] J. Feng, B. Ni, Q. Tian, and S. Yan, "Geometric Lp-norm feature pooling for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, 2011, pp. 2609–2704. 6, 8

[53] C. Gulcehre, K. Cho, R. Pascanu, and Y. Bengio, "Learned-norm pooling for deep feedforward and recurrent neural networks," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 530–546. 6

[54] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007. 6

[55] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936. 7

[56] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic press, 2013. 7

[57] M. T. Shami and M. S. Kamel, "Segment-based approach to the recognition of emotions in speech," in *2015 IEEE International Conference on Multimedia and Expo (ICME)*, Amsterdam, Netherlands, 2005, pp. 4–7. 8

[58] B. W. Schuller and G. Rigoll, "Timing levels in segment-based speech emotion recognition." in *INTERSPEECH*, Pittsburgh, Pennsylvania, 2006, pp. 1818–1821. 8

[59] A. Vedaldi and K. Lenc, "MatConvNet-convolutional neural networks for MATLAB," *arXiv preprint arXiv:1412.4564*, 2014. 8

[60] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011. 8

[61] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: a benchmark comparison of performances," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU-2009)*, Merano, 2009, pp. 552–557. 8, 11, 12

[62] E. Florian, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer, 2016. 11, 12

[63] L. Gao, L. Qi, and L. Guan, "Information fusion based on kernel entropy component analysis in discriminative canonical correlation space with application to audio emotion recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 2817–2821. 11, 12

[64] N. E. D. Elmadany, Y. He, and L. Guan, "Multiview emotion recognition via multi-set locality preserving canonical correlation analysis," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, Montral, QC, Canada, 2016, pp. 590–593. 11, 12

[65] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013 - the continuous audio/visual emotion and depression recognition challenge," in *3rd ACM international workshop on Audio/Visual Emotion Challenge*, Barcelona, Spain, 2013, pp. 3–10. 12

[66] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *INTERSPEECH-2013*, Lyon, France, 2013, pp. 148–152. 12

[67] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012. 13

[68] F. Ringeval, A. Sonderegger, J. S. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Shanghai, China, 2013, pp. 1–8. 13

[69] E. Bozkurt, H. Khaki, S. Kececi, B. B. Turker, Y. Yemez, and E. Erzin, "JESTKOD database: dyadic interaction analysis," in *2015 23th Signal Processing and Communications Applications Conference (SIU)*, Malatya, Turkey, 2015, pp. 1374–1377. 13

[70] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian, "Affective visualization and retrieval for music video," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 510–522, 2010. 13

Shiqing Zhang received the Ph.D. degree at school of Communication and Information Engineering, University of Electronic Science and Technology of China, in 2012. Currently, he is a postdoctor with the School of Electronic Engineering and Computer Science, Peking University, Beijing, China, and also works as an associate professor of the Institute of Intelligent Information Processing, Taizhou University, China. His research interests include audio and image processing, affective computing and pattern recognition.

Shiliang Zhang is currently a tenure-track Assistant Professor in School of Electronic Engineering and Computer Science, Peking University. He received the Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences in 2012. He was a Postdoctoral Scientist in NEC Labs America and a Postdoctoral Research Fellow in University of Texas at San Antonio.

Dr. Zhang's research interests include large-scale image retrieval and computer vision for autonomous driving. He was awarded the National 1000 Youth Talents Plan of China, Outstanding Doctoral Dissertation Awards from both Chinese Academy of Sciences and Chinese Computer Federation (CCF), President Scholarship by Chinese Academy of Sciences, NEC Laboratories America Spot Recognition Award, and the Microsoft Research Fellowship. He has published over 30 papers in journals and conferences including IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, ACM Multimedia, and ICCV. He is the recipient of Top 10% Paper Award in IEEE MMSP 2011. His research is supported by the National 1000 Youth Talents Plan and Natural Science Foundation of China (NSFC).

Tiejun Huang (M'01−SM'12) is currently a Professor with the School of Electronic Engineering and Computer Science, Peking University, Beijing, China, where he is also the Director of the Institute for Digital Media Technology. He received the Ph.D. degree in pattern recognition and intelligent system from Huazhong (Central China) University of Science and Technology, Wuhan, China, in 1998, and the masters and bachelors degree in computer science from the Wuhan University of Technology, Wuhan, in 1995 and 1992, respectively. His research area includes video coding, image understanding, digital right management, and digital library. He has authored or co-authored over 100 peer-reviewed papers and three books. He is a member of the Board of Director for Digital Media Project, the Advisory Board of the IEEE Computing Society, and the Board of the Chinese Institute of Electronics.

Wen Gao (M'92−SM'05−F'09) received the Ph.D. degree in Electronics Engineering from the University of Tokyo, Tokyo, Japan, in 1991.

He is currently a Professor with the School of Electronic Engineering and Computer Science with Peking University, Beijing, China. Before joining Peking University, he was a Professor of Computer Science with the Harbin Institute of Technology, Harbin, China, from 1991 to 1995, and a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has authored five books and more than 600 technical articles in refereed journals and conference proceedings in image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interface, and bioinformatics.

Dr. Gao serves the editorial board for several journals, such as IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Multimedia, IEEE Transactions on Autonomous Mental Development, EURASIP Journal of Image Communications, and Journal of Visual Communication and Image Representation. He chaired a number of prestigious international conferences on multimedia and video signal processing, such as the IEEE ICME and ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations.