Learning Salient Segments for Speech Emotion Recognition Using Attentive Temporal Pooling

XIAOHAN XIA, DONGMEI JIANG, AND HICHEM SAHLI, (Member, IEEE)

1 NPU-VUB Joint AVSP Research Laboratory, School of Computer Science, Northwestern Polytechnical University (NPU), Xi’an 710072, China
2 Shaanxi Key Laboratory on Speech and Image Information Processing, Northwestern Polytechnical University (NPU), Xi’an 710072, China
3 Peng Cheng Laboratory, Shenzhen 518055, China
4 VUB-NPU Joint AVSP Laboratory, Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), 1050 Brussels, Belgium
5 Interuniversity Microelectronics Centre (IMEC), 3001 Heverlee, Belgium

Corresponding author: Dongmei Jiang (jiangdm@nwpu.edu.cn)

This work was supported in part by the Shaanxi Provincial Key Research and Development Program under Grant 2017KW-ZD-14, in part by the 111 Project under Grant B16039, and in part by the Vrije Universiteit Brussel Interdisciplinary Research Program through the EMO-App Project.

ABSTRACT In the temporal process of expressing the emotions, some intervals embed more salient emotion information than others. In this paper, by introducing an attentive temporal pooling module into the deep neural network (DNN) architecture, we present a simple but effective speech emotion recognition (SER) framework, which is able to automatically highlight the emotionally salient segments while suppressing the influence of less relevant ones. For an input speech utterance, the extracted feature sequence of hand-crafted low-level descriptors (LLDs) are evenly split into several overlapping temporal segments, and the segment-level features are computed by performing functionals on the LLDs of each segment. These segment-level features are then input into a DNN model outputting the emotion probabilities as well as the more condensed representation of each segment. An attentive temporal pooling module, consisting of an auxiliary DNN and a Gaussian Mixture Model (GMM), is proposed to learn the emotional saliency weights of different temporal segments from the condensed representations, which are then assigned to the segment-level emotion probabilities for the final utterance-level prediction. Notably, the attentive temporal pooling module and the DNN architecture for feature abstraction can be jointly trained using only the utterance-level labels, while without any frame-level or segment-level supervisory information. Experimental results on the three public released emotion datasets RML, EMO-DB, and IEMOCAP show that the proposed framework obtains state-of-the-art performance on SER.

INDEX TERMS Attentive temporal pooling, deep neural networks, hand-crafted audio features, speech emotion recognition.

I. INTRODUCTION

Over the past two decades, automatic speech emotion recognition (SER) has attracted numerous research interests from both industry and academia [1]. Two major directions have been investigated: dimensional affect recognition and categorical emotion recognition [2]. Categorical emotion recognition is more commonly used under the hypothesis that there exist a small number of basic emotions, such as happy, angry, sad, surprise, fear and disgust, that can be recognized universally [3]. In this paper, we focus on categorical SER - a time series classification problem where the goal is to predict the emotion class given a speech utterance.

One of the most important issues for SER is how to model the temporal dynamics of emotions embedded in speech. Especially, during a speech utterance, there may exist some intervals that are silent and do not carry any emotion information, while other segments strongly express emotions. In building emotion recognition models, it is crucial to automatically learn and pay more attention to the emotionally salient parts while ignoring the other parts such as the silent or neutral segments in speech.

Various models or frameworks have been proposed for SER, including utterance-wise, frame-wise, or segment-wise
recognition methods. For the utterance-wise methods, normally frame-level hand-crafted Low-Level Descriptors (LLDs) are firstly extracted [4], then the global feature vectors are computed as statistics of the LLDs over the whole utterance [4], [5], which are used to train classifiers such as Support Vector Machines (SVM) or Extreme Learning Machines (ELM) [6], [7]. Recently, a promising trend is to utilize deep convolutional neural networks (CNN) to learn discriminative feature representations or classify emotions. In [8], Yenigalla et al. adopted a set of four parallel convolutions with different kernel sizes to extract features from spectrograms, after which two fully-connected layers were used to perform classification. Nevertheless, the input speech signal has to be trimmed or padded into a fixed-length segment of 6 seconds, which may lead to information redundancy or loss. In the above utterance-wise recognition models, the temporal dynamics of emotion are not considered.

For frame-wise SER, various dynamic models, such as Hidden Markov Models (HMM) [9] and Recurrent Neural Networks (RNN) especially those with Long Short-Term Memory (LSTM) [10], have been adopted to capture the temporal dynamics of emotions. They face a crucial problem on how to aggregate over frame-level LLDs to match a single emotion label for each utterance. Various methods have been proposed to solve this problem. Some works accumulated the temporal emotion information to the last frame in RNN and then used the activations of the final hidden layer of the last frame for classification [11]. On the other hand, some works assigned the same emotion label to each frame within the utterance, and trained the RNN in a frame-wise manner by back-propagating cross-entropy errors from each frame [12]. Alternatively, in some works [13], mean pooling was performed to calculate the average of hidden representations of all frames as embedding features, which were then input into a classification model for SER. For such methods, the information along the frames are evenly considered, no matter they are silence, emotionally salient region, or speech segments that do not contain any emotion information. In [14], Mirsamadi et al. proposed a weighted pooling strategy to adaptively assign weights to different segments. Finally, the segment-level emotion probabilities can aggregate into the utterance-level prediction through linear combination with these weights. It should be noted that this attentive temporal pooling module is introduced to learn the emotional saliency weights of different temporal segments. Specifically, the attentive temporal pooling module consists of an auxiliary DNN and a Gaussian Mixture Model (GMM), where the auxiliary DNN learns the parameters of GMM from the condensed segment-level features, and the GMM functions as a weights generator to adaptively assign weights to different segments. The main contributions of the paper can be summarized as follows:

1) A segment-wise SER framework is proposed, which integrates the DNN with an attentive temporal pooling model consisting of an auxiliary DNN and a GMM. It provides a simple architecture that effectively learns emotionally salient regions over the long-term speech signal, and adaptively generates the weight for each temporal segment. Through jointly back-propagating the attentive temporal pooling module and the DNN architecture, more attention might be given to the emotionally salient segments.

2) Through evenly splitting a speech utterance into a fixed number of overlapping temporal segments, and inputting the functional features of the segments into the DNN model, it avoids transforming the audio into spectrograms and use CNN models [8] which requires the confidences in each segment, then apply a second-stage classifier to form the final decision for the utterance [16], [17]. Zhang et al. [18] first computed the log Mel-spectrogram for one utterance and split it into overlapping spectrogram segments, then each segment was fed into a fine-tuned AlexNet [19] model obtaining a deep feature vector. Average pooling was applied on the features of all segments to obtain the utterance-level features and finally a linear SVM classifier was employed for emotion classification. Whereas their work achieved state-of-the-art performance, the average pooling assigned equal weights to each segment while ignoring the emotionally salience in speech.

In this paper, we propose a segment-wise SER framework with deep neural networks and attentive temporal pooling, which employs a self-attention mechanism to automatically highlight the emotionally salient segments while suppressing the influence of less relevant ones during back-propagation. For a speech utterance, we first extract frame-level hand-crafted LLD features, then evenly split the feature sequence into several overlapping segments, and compute functional features on the LLDs of each segment. For each segment, the functional features are input into a DNN model to obtain the condensed features through nonlinear transformation and predict segment-level emotion probabilities. An attentive temporal pooling module is introduced to learn the emotional saliency weights of different temporal segments. Specifically, the attentive temporal pooling module consists of an auxiliary DNN and a Gaussian Mixture Model (GMM), where the auxiliary DNN learns the parameters of GMM from the condensed segment-level features, and the GMM functions as a weights generator to adaptively assign weights to different segments. Finally, the segment-level emotion probabilities can aggregate into the utterance-level prediction through linear combination with these weights. It should be noted that this attentive temporal pooling module and the segment-level DNN architecture are jointly trained through back-propagation, therefore no frame-level or segment-level emotion labels are needed for training.
The related work are reviewed in section II. Section III aggregate over frame-level LLDs to match the single emotion applying RNNs either as utterance representation learners level features [23]. Recently, there is a growing trend towards been used to learn the distribution of the frame based low-dynamics. Traditional dynamic models, such as HMMs, have been widely adopted in SER. One of the first works is [21], where a DNN model was applied on utterance-level statistics to improve the performance. Experiments on nine databases showed a significant improvement over the baselines with SVMs. In [22], the fully convolutional neural network (FCN) was adopted to handle the variable-length spectrograms for different utterances. Then, an attention layer was added on top of the FCN model to focus on emotionally salient regions of the spectrogram followed by a softmax layer to perform classification. Experiments showed superior results on the benchmark dataset.

In the following subsections, we will introduce each module in details.

B. FRAME-WISE SER
Frame-wise approaches focus on modeling the temporal dynamics. Traditional dynamic models, such as HMMs, have been used to learn the distribution of the frame based low-level features [23]. Recently, there is a growing trend towards applying RNNs either as utterance representation learners or emotion classifiers, where a crucial problem is how to aggregate over frame-level LLDs to match the single emotion label for each utterance. Mirsamadi et al. [14] analyzed four RNN architectures: 1) assigning the same emotion label to each frame within the utterance, and training the RNN in a frame-wise manner by computing the cross-entropy error at all frames; 2) using only the final hidden representation at the last frame of an utterance as the representation; 3) performing mean-pooling to calculate the average of hidden representations of all frames as the long-term aggregation; 4) adopting a weighted-pooling strategy to generate a weighted sum instead of mean-pooling, where the weights are determined from an additional attention mechanism based model. Experiments indicated that the attention-based weighted pooling strategy outperformed all other RNN architectures by focusing on emotional salient frames. More recently, some studies have investigated the application of the CNNs in SER. In [24], the authors set all the utterances to a fixed length through padding or cutting the speech, and extracted diverse hand-crafted LLDs as input to an attention-based CNN (ACNN) with multi-view learning objective function. Liu et al. [25] used a CNN with global k-max pooling over the temporal frame-level LLDs, and experiments showed an improvement over the ACNN-based method.

C. SEGMENT-WISE SER
Segment-wise approaches usually start from splitting the utterance into several fixed-length segments. In [16], hand-crafted features were firstly extracted at the frame level, and segment-level features were constructed by stacking the features of consecutive frames, which were used to train a DNN model as the segment-level emotion recognizer. Then, utterance-level features were computed from the statistics of the segment-level probabilities, which were fed into an ELM model for the final emotion classification. In [26], the authors used a fine-tuned AlexNet [19] model to learn segment-level discriminative features from the log Mel-spectrogram, and proposed a discriminant temporal pyramid matching strategy to aggregate the learned segment-level features into a global utterance-level feature representation, based on which a linear SVM was employed to predict utterance-level emotion. Chen et al. [27] first split each utterance into equal-length segments of 3 seconds and computed the log-Mels with first and second derivatives to obtain a 3D feature representation for each segment. Next, the 3D attention-based convolutional LSTM model was employed to generate discriminative features, after which a fully connected layer was used for segment-level classification. Finally, the whole utterance prediction was obtained through max pooling over each segment. Nevertheless, in both [26] and [27], the authors used the emotion label of the whole utterance as the label of each segment for training the models, whether the segment contains emotion information or not.

Our work differs from the above segment-wise frameworks by its novel adoption of the joint training of a standard DNN model and an attentive temporal pooling module, which enables efficient emotional salience learning over the long-term temporal structure using only the utterance-level supervision information, while without any assumptions on either frame-level or segment-level labels.

III. THE PROPOSED SEGMENT-WISE SER MODEL WITH ATTENTIVE TEMPORAL POOLING
With the goal of learning the emotionally salient regions within an utterance, we propose a novel SER framework by integrating an attentive temporal pooling module into a standard DNN architecture, as illustrated in Fig. 1. The input speech signal is first processed to extract LLD features,
which are then evenly split into several overlapping segments. Segment-level global features are computed over these temporal segments and DNNs are adopted to condense these high-dimensional features and predict the segment-level emotion probabilities. The attentive temporal pooling module is designed to automatically learn the emotional saliency weights of different segments with self-attention mechanism. Specifically, auxiliary DNNs take the condensed features from the features generation DNNs to learn the parameters of a GMM. The GMM acts as a generator of weights, representing emotional saliency assigned to the temporal segments of each utterance. Finally, the learned weights are used to aggregate the segment-level estimations into the utterance-level prediction.

### A. HAND-CRAFTED AUDIO FEATURES

1) FRAME-LEVEL LOW-LEVEL DESCRIPTOR (LLD) FEATURES

In this work, we construct a brute-force feature set to make most of the domain knowledge in SER. Specifically, we leverage the experience from both the eGeMAPS set [4] and the INTERSPEECH Challenges feature sets [28]–[31] to cover more acoustic parameters that shall be potentially useful in representing emotion information.

Given an utterance, 121 LLD features including energy, spectral and voicing related features are firstly extracted frame by frame, with a sliding Hamming window of size as 25 ms and shift as 10 ms. Table 1 lists these LLD features, where the number on the right of each row represents the dimension of the corresponding feature.

| 21 Energy related LLD |
|-----------------------|
| Sum of auditory spectrum (loudness) | 1 |
| RMS Energy | 1 |
| log. Energy | 1 |
| Zero-Crossing Rate | 1 |
| Intensity | 1 |
| LP gain | 1 |
| F0 envelope | 1 |
| L1 norm length | 2 |
| Chroma for 12 semitones | 12 |

| 83 Spectral related LLD |
|------------------------|
| Alpha ratio (50-1000 Hz / 1-5 kHz) | 1 |
| Spectral Slope (0-500 Hz, 0.5-1.5 kHz) | 2 |
| Hammarberg index | 1 |
| RASTA-filt. aud. spect. bds. 1-26 (0-8 kHz) | 26 |
| Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9 | 4 |
| MFCC 0-14 | 15 |
| PLP cepstral coefficients 1-5 | 5 |
| LPC coefficients | 11 |
| LSP frequency 0-7 | 8 |
| Spectral energy 250-650 Hz, 1-4 kHz | 2 |
| Spectral Flux, Centroid, Entropy, Variance, Skewness, Kurtosis, Harmonicity, Flatness | 8 |

| 17 Voicing related LLD |
|-----------------------|
| F0 (SHS & Viterbi smoothing) | 1 |
| Formants 1, 2, 3 (freq., bandwidth, ampl.) | 9 |
| Harmonic difference H1-H2, H1-A3 | 2 |
| log. HNR | 1 |
| Jitter (local) | 1 |
| Jitter (DPD) | 1 |
| Shimmer (local) | 1 |
| Probability Of voicing | 1 |

2) TEMPORAL SEGMENTATION AND SEGMENT-LEVEL FUNCTIONAL FEATURES

Given an utterance \( U \), we split it into \( N \) segments \( \{ U_1, U_2, \ldots, U_N \} \) with equal length. Each segment can be represented as \( U_i = \{ x_{i1}, \ldots, x_{ij}, \ldots, x_{iM} \} \), where \( x_{ij} \in \mathbb{R}^{121} \) \( (i = 1, \ldots, N; j = 1, 2, \ldots, M) \) is the 121-dimensional LLD feature vector and \( M \) is the number of frames within the segment. Note that, we fix the number of segments \( N \) for all utterances, and the number of frames \( M \)
TABLE 2. Functionals over several consecutive frames within one segment. The number on the right represents the dimension of the corresponding functionals.

| 29 Statistical functionals |  |
|---------------------------|--|
| Arithmetic mean           | 1 |
| Max, min                  | 2 |
| Variance, standard deviation, skewness, kurtosis | 4 |
| Quartiles 1-3             | 3 |
| Up-level time 25%, 50%, 75% | 3 |
| Rise/Fall time            | 2 |
| Number of peaks           | 1 |
| Mean of peak distances    | 1 |
| Mean value of peaks       | 1 |
| peakMeanMeanDist          | 1 |
| Sample position           | 1 |
| Number of segments        | 1 |
| Mean/max/min length of segments | 3 |
| Norm                      | 1 |
| Linear regression coefficient a (slope) | 1 |
| Linear regression coefficient b (offset) | 1 |
| Linear regression quadratic/linear error | 2 |

in each segment may vary in different utterances. Following the conclusion in [15] that only a segment longer than 250 ms can contain enough emotional information, in this work we set the shortest segment length to 300 ms, corresponding to 30 speech frames. Therefore for an utterance of length $L$, if $L \geq 15 \times (N + 1)$, then we split it into $N$ segments each containing $M = \text{floor}[2L/(N + 1)]$ frames, with the shift of $S = \text{floor}[L/(N + 1)]$ frames. Otherwise, we set $M = 30$, and the shift $S = \text{floor}[(L - 30)/(N - 1)]$.

In this way, the framework can adaptively adjust the length and shift of the segments to make sure that the utterance is segmented into $N$ equal segments whose length is not shorter than 30 frames (300 ms). This simple but effective temporal segmentation enables the architecture to work on the untrimmed or unpadded utterances to preserve the original information to the utmost.

For each segment $U_n$, 29 statistical functionals, as listed in Table 2, are performed over the 121 frame-level LLDs and their corresponding delta coefficients (except for the 25 descriptors related to Intensity, Chroma, LPC coefficients and LP gain), yielding $(121 + 96) \times 29 = 6293$ features. In addition, the 29 functionals are also applied to the second order derivations of 21 descriptors on PLP cepstral coefficients, MFCC and logarithm energy, resulting in another $21 \times 29 = 609$ parameters. Finally, a 6902-dimensional feature vector is obtained for each segment. Therefore, each utterance of $N$ segments can be represented by a temporal feature sequence $\{V_1, \ldots, V_n, \ldots, V_N\}$, where $V_n \in \mathbb{R}^{6902}$ is the segment-level global feature vector.

### B. SEGMENT-LEVEL EMOTION PROBABILITY ESTIMATION

For each segment $n$, the segment-level feature vector $V_n$ is input into a DNN architecture for feature abstraction and segment-level emotion probability estimation.

With the DNN, the condensed high-level feature vector can be computed as:

$$Y_n = \mathcal{F}_1(V_n; W_1); \quad n = 1, 2, \ldots, N \quad (1)$$

where $\mathcal{F}_1$ is the function representing the non-linear transformation in a standard DNN with parameters $W_1$. The rectified linear unit (ReLU) is adopted as activation function. Normally, an additional softmax layer is used to perform the $C$-way classification, where $C$ is the number of emotion categories.

$$S_n = \frac{e^{Y_n}}{\sum_{i=1}^{C} e^{Y_{ni}}}; \quad n = 1, 2, \ldots, N \quad (2)$$

Here the softmax score $S_n$ is the normalized log-probability of emotion confidences for the $n^{th}$ segment. Note that, the DNNs of all the segments share parameters in our framework to alleviate over-fitting.

### C. ATTENTIVE TEMPORAL POOLING

After obtaining the segment-level emotion estimation results, the common approach is to perform average pooling over these segment-level prediction scores for each emotion class respectively. However, it is not advisable to view every part in an utterance with equal contributions and the mean operation will distort the final recognition performance. One alternative approach is linear weighting, i.e., to perform an element-wise weighted linear combination on the prediction scores from each segment. Although this aggregation method is expected to learn weights of different phases of an emotion category, it is data-independent without considering the difference between utterances.

In this paper, referring to the deep adaptive temporal pooling strategy proposed in [32] by combining weight generators and an auxiliary CNN as a self-attention module, and the idea in [33] where a GMM is represented as the last layer of a DNN architecture jointly optimized with all previous layers for speech recognition, we propose an attentive temporal pooling module, as illustrated in Fig. 2, which contains two components: a GMM as the weights generator that models the distribution of the weights of different segments, and an auxiliary DNN to learn the parameters of the GMM. The auxiliary DNN with the input from the condensed features of all segments is firstly trained to learn the parameters of GMM, which are then used by the weights generator to sample the weights of the different temporal segments representing the emotional salience in each utterance. It is worth noting that the probability density function of a GMM model is differentiable with respect to its parameters, which is a superb benefit for back-propagating the gradients to jointly update the whole framework.

Formally, let us consider the auxiliary DNN as $\mathcal{F}_2$, representing the non-linear transformation with parameters $W_2$ between the condensed high-level features $Y$ extracted from the segment-level DNN and the parameter set $\Theta$ of GMM,
As illustrated in Equation (3),
\[
\Theta = F_{2}(Y; W_{2})
\]
where the output $\Theta \in \mathbb{R}^{K \times 3}$ parameterizes a mixture of $K$ Gaussian distributions. It is worth mentioning that the parameters of GMM are learned from the condensed segment-level features, thus the GMM is adaptively parameterized in a data-dependent manner which could better learn the emotional salient regions.

With the parameters of GMM, we can sample the emotional saliency weight for each segment. Given the index of the temporal segment $i_{n} (n = 1, 2, \cdots, N)$, the corresponding weight $w_{n}$ is sampled as follows. Note that, the $\{i_{n}\} (n = 1, 2, \cdots, N)$ values are first normalized into $[0, 1]$ before feeding them into the GMM model.

\[
w_{n} = \mathcal{H}(i_{n}; \Theta) = \sum_{k=1}^{K} \pi_{k} N(i_{n} \mid \mu_{k}, \sigma_{k}^{2}) = \sum_{k=1}^{K} \frac{\pi_{k}}{\sqrt{2\pi}\sigma_{k}^{2}} e^{-\frac{(i_{n}-\mu_{k})^{2}}{2\sigma_{k}^{2}}}; \quad n = 1, 2, \cdots, N
\]

where $\pi_{k}$ is mixture coefficient subject to $\sum_{k=1}^{K} \pi_{k} = 1$, $\mu_{k}$ and $\sigma_{k}$ is the mean and variance of the $k^{th}$ GMM component.

Within this formulation, we can calculate the gradients of $w_{n}$ with respect to $\pi_{k}$, $\mu_{k}$, and $\sigma_{k}$, respectively, as follows [32]:

\[
\frac{\partial w_{n}}{\partial \pi_{k}} = N(i_{n} \mid \mu_{k}, \sigma_{k}^{2})
\]

\[
\frac{\partial w_{n}}{\partial \mu_{k}} = \frac{i_{n} - \mu_{k}}{\sigma_{k}^{2}} \pi_{k} N(i_{n} \mid \mu_{k}, \sigma_{k}^{2})
\]

\[
\frac{\partial w_{n}}{\partial \sigma_{k}} = \frac{1}{\sigma_{k}} \left[ \frac{(i_{n} - \mu_{k})^{2}}{\sigma_{k}^{2}} - 1 \right] \pi_{k} N(i_{n} \mid \mu_{k}, \sigma_{k}^{2})
\]

Through these gradient formulas, we can update the GMM parameters $\Theta$ using back-propagation together with the DNN parameters $W_{1}$ and $W_{2}$.

D. UTTERANCE-LEVEL ESTIMATION AGGREGATION

Once the weights of the temporal segments are generated from (4), the weighted average is performed on the segment-level softmax scores in (2) to perform the utterance-level estimation aggregation. After that, an additional softmax layer is adopted on the normalized aggregated vector to get the final prediction.

Overall, there are three advantages of introducing the attentive temporal pooling into the standard DNN model: 1) The attentive pooling module adaptively generates weights for temporal segments, which is able to model the emotional saliency over the long-term temporal structure. In this way, the segments which are more emotion-relevant will be magnified and the non-relevant ones will be simultaneously suppressed. Therefore, the salient segments could contribute more to the final utterance-level prediction. 2) Since the attentive pooling module is based on the high-level representations from the DNN model for segment-level emotion estimation, the DNN parameters are expected to be updated with different coefficients according to the temporal weights, which takes full advantage of extra back-propagation information to help the training process of DNNs. 3) Only the utterance-level supervisory information is needed in the whole training process, while without any assumptions of segment-level labels. This avoids the confusing information when simply labeling every segment as the same emotion with the utterance.

IV. EXPERIMENTS AND ANALYSIS

A. EMOTION DATASETS

In this paper, to evaluate the performance of the proposed speech emotion recognition method, we conduct experiments on three well-known and publicly-available emotion datasets, i.e. the RML dataset [34], the Berlin Emotional Speech Database [35] and the Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [36].

1) RML

The RML dataset [34] contains 720 audio-visual emotional samples acted by eight subjects with different languages, accents, and cultural backgrounds. Specifically, six different speaking languages, i.e., English, Mandarin, Urdu, Punjabi, Persian and Italian, are included, together with different accents of English and Chinese. The basic six emotions are expressed: Anger, Disgust, Fear, Happiness, Sadness, and Surprise. More than ten different sentences were provided for each emotional class to ensure the context independency of the speech data.

2) EMO-DB

The Berlin emotional speech database [35], also known as EMO-DB, is recorded by ten professional native German-speaking actors (five female and five male). Five short and five long sentences are elaborately constructed for...
the actors to speak in seven target emotions, namely Anger, Fear, Joy, Sadness, Disgust, Boredom, and Neutral, resulting in 535 utterances.

3) IEMOCAP
The IEMOCAP database [36] consists of about 12 hours audio-video data of ten actors over 5 dyadic affective interactions (sessions), in each of which one female and one male are involved in. Both scripted and improvised scenarios are involved in each session. Following previous works [16], [22], [37], we only consider four emotional categories, namely Happy, Sad, Angry, and Neutral from the improvised speech data.

B. SETUP OF THE EXPERIMENTS
The hand-crafted acoustic features are extracted with the open-source toolkit openSMILE [38] using the configuration files provided in the 2.3.0 release version. Z-normalization is performed to alleviate the impact of different feature scales, where the mean and standard deviation are calculated only on the training data and then used on all data, respectively in each fold of cross-validation experiment on all datasets. It should be noted that, although the previous work [39] has revealed that per-speaker z-normalization can improve the performance, we do not adopt it given the assumption that speaker identity information is not available in our study.

All the networks in this paper have been implemented using the open-source deep learning framework PyTorch [40] on one NVIDIA 1080 Ti GPU with 12 GB memory. As a common setting, the mini-batch Stochastic Gradient Descent (SGD) algorithm is utilized to learn the network parameters, where the batch size is set to 16 and momentum set to 0.9. During the training phase, the maximum number of epochs is set to 200 and the early-stopping strategy is adopted to deal with over-fitting. The rectified linear unit (ReLU) is used as activation function and the categorical cross-entropy as cost function. To deal with the problem of covariate shift, speed up the convergence of training and alleviate the over-fitting problem, Batch Normalization [41] is adopted in the DNN model. For the temporal segmentation, the number of segments within each utterance has been empirically set to 5.

Following the suggestion in [42], test-runs are performed by employing a speaker-independent Leave-One-Speaker-Out (LOSO) or Leave-One-Group-Out (LOSGO) cross-validation strategy, and the final performance metric is the average of these results. Specifically, the LOSO strategy is adopted on the RML and EMO-DB datasets, and the LOSGO strategy is adopted on the IEMOCAP database with five sessions. Both the weighted accuracy (WA) and unweighted accuracy (UA) are reported to evaluate the performance, as done in other related studies [16], [22], [37]. The WA is the percentage of correct classification decisions over the test set and the UA is the average of accuracies over each emotion categories. In this work, we consider the weighted accuracy as the primary measurement, since it is self-weighted from the usability standpoint [43].

C. EXPERIMENTAL RESULTS
In the following, we provide detailed experimental analysis on the proposed framework, to demonstrate the importance of learning salient regions via attentive temporal pooling. Finally, we compare the performance of our method with the state of the art works on the three SER benchmarks.

1) INSIGHT INTO THE EMOTIONAL SALIENT REGIONS
We first consider an utterance labeled as “happy” from the EMO-DB dataset as an example to illustrate the emotional salient region, as well as how the attentive temporal pooling strategy highlights the salient region and further benefits the final classification. Figure 3 shows the confidence scores of the seven emotions for each segment, as well as the final confidence score of the utterance, where in (a) and (b) are the results from average pooling, while in (c) and (d) are those from the attentive temporal pooling. As we can see from Figure 3 (a) and (b), a wrong label “angry” is predicted through average pooling where each segment is assigned the same weight to aggregate the segment-level softmax scores into the utterance-level score. Among all the five segments, the third one obtains the highest probability on the ground-truth emotion class “happy”, while others have much more confidence in “angry” class. The average pooling method simply considers every segment with equal contribution, therefore even though the third segment predicts the correct emotion, the whole utterance is misclassified as “angry” due to the averaging over all softmax scores.

In contrast, through attentive temporal pooling, the utterance is correctly classified as “happy” as shown in Figure 3 (c) and (d). We can see that: 1) The largest weight is assigned to the salient segment, i.e., the third segment, which is the only one that predicts the correct label in the segment level prediction with average pooling. Therefore the learned weights for different segments are able to highlight the salient segments and restrain the irrelevant ones and therefore boost correct emotion classification. 2) Joint training of the whole framework contributes to the segment-level emotion probability prediction. This can be seen from the comparison between Figure 3 (a) and (c), where the softmax scores have been improved on all the five segments through attentive temporal pooling, among which the fourth segment also obtains the correct prediction.

2) EVALUATION OF THE POOLING FUNCTION
a: BASELINE
We first conduct baseline experiments using the unsegmented (utterance-level) global features with deep neural networks. That is to apply the functionals of Table 2 to all the LLDs within the whole utterance without temporal segmentation. We evaluate diverse DNN architectures consisting of different layers \(N_{\text{layers}} \in \{2, 3, 4\}\) with various number of hidden nodes \(N_{\text{nodes}} \in \{128, 256, 512, 1024\}\) in each layer. The learning rate is set to \(1e-3\) and the whole training procedure...
FIGURE 3. Average pooling (1st row) vs attentive temporal pooling (2nd row) for a "happy" utterance from EMO-DB. In (a) and (b), the utterance is misclassified as "angry" through average pooling. In (c) and (d), the utterance is correctly classified as "happy" through attentive temporal pooling.

TABLE 3. Results of Utterance-level Global Features with Deep Neural Networks (%).

| Dataset | Best DNN architecture | ACC | MAR |
|---------|-----------------------|-----|-----|
| RML     | (128, 128, 128, 128) | 67.72 | 66.65 |
| EMO-DB  | (512, 512, 512)      | 83.51 | 77.66 |
| IEMOCAP | (256, 256, 256, 256) | 65.23 | 59.12 |

stops at 200 epochs. The best results and the corresponding architectures on each dataset are listed in Table 3.

b: AVERAGE POOLING
To investigate the performance of temporal segmentation, we first conduct experiments using average pooling strategy. For each utterance, we first split it into $N$ consecutive segments with a certain overlap following the description in Section III-A2, where $N$ is set to 5 empirically. The 6902 dimension features of each segment are first input into a DNN for high-level feature abstraction and initial emotion classification, then average pooling is adopted to aggregate these segment-level predictions into the utterance-level scores.

Table 4 lists the results of segment-level global features with DNNs and average pooling. Note that, the structure of DNNs are selected after evaluating the best architecture considering different layers ($N_{\text{layers}} \in \{2, 3, 4\}$) with various number of hidden nodes ($N_{\text{nodes}} \in \{128, 256, 512, 1024\}$) in each layer, and a learning rate set to $1e^{-3}$. Analyzing Table 3 and 4, one can see that the temporal segmentation together with the average pooling has improved the baseline performance, which demonstrates the effectiveness of temporal segmentation.

c: ATTENTIVE TEMPORAL POOLING
For the experiments using the proposed SER framework with attentive temporal pooling, to accelerate the convergence of training, we utilize the pre-trained DNN model with the segment-level global features and average pooling as the backbone network, i.e., the $F_1$ in (1). For the auxiliary DNN $F_2$ as illustrated in (3), we randomly initialize it and select the best architecture from different numbers of hidden layers ($N_{\text{layers}} \in \{2, 3, 4\}$) and various hidden nodes ($N_{\text{nodes}} \in \{16, 32, 64, 128, 256\}$) in each layer. For the GMM components, we set the number of Gaussian components $K$.
as 3 by default and the effect of different values of $K$ will be discussed later. It is worth mentioning that the attentive temporal pooling module and the backbone DNN architecture are jointly trained using only the utterance-level labels, and the learning rate of attentive temporal pooling module is set to $1e-2$ whereas that of the pre-trained backbone network is set to $1e-5$. Note that all the segment-level DNNs share parameters, this allows us to carry out fair comparison with the baseline method using utterance-level DNNs without temporal segmentation and the segment-level DNNs with average pooling strategy.

The results are summarized in Table 5 together with the corresponding DNN architectures, where BBNet denotes the pre-trained backbone DNN for feature extraction and AUNet denotes the auxiliary DNN in the attentive temporal pooling module. Compared to the results of average pooling in Table 4, one can see that the attentive temporal pooling module boosts the recognition accuracy from 70.3% to 73.15% on RML dataset, from 83.82% to 85.73% on EMO-DB dataset, and from 67.51% to 69.37% on IEMOCAP dataset.

For the attentive temporal pooling module, the crucial parameter is the number of Gaussian components $K$ besides the architecture of the auxiliary DNN. To investigate the effect of different numbers of Gaussian components, we vary $K$ from 1 to 4 and evaluate the performance using the same network architecture. The results are summarized in Fig. 4. As we can see, for all the three datasets, the best performances are obtained with $K$ equals to 3. When further increasing the number of Gaussian components, the performance drops drastically.

d: ANALYSIS OF CONFUSION MATRICES

To further investigate the recognition performance, the confusion matrices corresponding to the results of segment-level global features with DNN and the attentive temporal pooling in Table 5 are presented in Fig. 5, for RML, EMO-DB and IEMOCAP dataset, respectively. As we can see from Fig. 5 (a), on the RML dataset, most samples belonging to “anger” and “surprise” are correctly recognized and the accuracies of both categories are higher than 84%, whereas for “fear” and “sandness”, these two emotions are prone to be recognized as other emotions and obtain relatively low accuracy. On the EMO-DB dataset, as shown in Fig. 5 (b), the recognition accuracy for “anger”, “sadness” and “boredom” reaches 93.7%, 93.55% and 91.36%, respectively, which are very promising performance. On the contrary, “joy” obtains a relatively low accuracy and is recognized as “anger”. This may be due to the fact that “joy” and “anger” have similar activation level in the activation-valence space, which is also proved on the IEMOCAP dataset as shown in Fig. 5 (c) where most samples in “angry” category are wrongly recognized as “happy”.

For comparison, the confusion matrices of the results in Table 4 using DNN and average pooling are also presented in Fig. 6, for RML, EMO-DB and IEMOCAP dataset, respectively. Through the comparison of Fig. 5 (a) and Fig. 6 (a), one can notice that the recognition accuracies of “happiness” and “fear” have been improved a lot through attentive temporal pooling. On the EMO-DB dataset, as shown in Fig. 5 (b) and Fig. 6 (b), the accuracy of “joy” has been significantly improved from 54.93% using averaging pooling to 70.42% using attentive temporal pooling, which demonstrates the attentive temporal pooling helps a lot to distinguish “joy” from “anger”. It is similar on the IEMOCAP dataset, where the accuracy of “happy” has improved by 4.32% through our proposed method using DNN and attentive temporal pooling.

3) COMPARISON WITH THE STATE OF THE ART

We compare our proposed SER framework with some state of the art works on the three datasets, results are summarized in Table 6. It can be seen that our proposed method obtains very competitive performance compared to the state-of-the-art results, showing the advantage of integrating the attentive temporal pooling into the DNN model. Specifically, on both RML dataset and EMO-DB dataset, our proposed framework obtains the highest weighted accuracy and unweighted accuracy. On the IEMOCAP dataset, it also achieves a competitive WA of 69.37% and an UA of 60.66%, which are higher than most state-of-the-art results but slightly lower than those of [22].

TABLE 5. Results of Segment-level global features with Deep Neural Networks and Attentive Temporal Pooling (%). BBNet denotes the pre-trained backbone DNN and AUNet denotes the auxiliary DNN in the attentive temporal pooling module. The number of Gaussian components is set to 3 for all the results in this table.

| Dataset | Best DNN architecture | ACC | MAR |
|---------|-----------------------|-----|-----|
| RML     | (256, 256, 256)       | 73.15 | 71.90 |
| EMO-DB  | (256, 256, 256)       | 85.73 | 82.86 |
| IEMOCAP | (256, 256, 256)       | 69.37 | 60.66 |

FIGURE 4. The effect of different numbers of Gaussian components ($K$) in the attentive temporal pooling module, on the RML, EMO-DB and IEMOCAP dataset, respectively. Using the DNN architectures as listed in Table 5.
The work in [16] was similar with our work, it combines hand-crafted features and deep learning models. Nevertheless, it didn’t consider the attention mechanism. In [46], the authors first split each utterance into 3-second long segments, which were used to train a convolutional LSTM model, and then the prediction for the whole utterance was obtained by averaging the posterior probabilities of those segments. The average pooling over segment-level estimation did not consider the importance of different segments. However, our proposed framework can attentively learn the emotional salient regions and therefore contribute to a better performance.

Attention mechanisms have been adopted in several works such as [24], [27], [47] and [22]. Different from our work, [24] investigated learning salient frames through an attentive CNN with multi-view learning objective function. In [27], the attention model was adopted during the segment-level classification and then max pooling was performed for the utterance-level aggregation, while our proposed method directly applies the attention mechanism at the utterance-level and therefore can better locate and focus on the salient parts for the whole utterance. In [47], two attention mechanisms were investigated to learn the emotionally relevant frames based on the BLSTM-CTC framework. One is the component attention, the other is quantum attention. In this work we explore the effectiveness of attention mechanism at the segment level. In [22], the attention mechanism was adopted on top of the fully convolutional neural network to focus on the time-frequency regions of the input spectrogram which were more emotion-relevant. The superiority of FCN model dealing with variable-length input together with the attention mechanism significantly boosted the model performance. Although our proposed method obtains a slightly lower accuracy than that in [22], our whole framework is much simpler and efficient, which is more suitable for applications in real life scenarios. More recently, deep generative models such as Variational Auto-encoders (VAE) [37] and Generative Adversarial Nets (GAN) [45] were investigated to learn the input data distribution and generate powerful features for SER. Nevertheless, these complex systems did not achieve better results, and our proposed method is very simple but works well.
TABLE 6. Comparison with the state-of-the-art results on the RML, EMO-DB and IEMOCAP dataset, respectively. Note that the word “features” is denoted as “Fea.” for short, “utt-Fea.” denotes the utterance-level global features and “seg-Fea.” for the segment-level global features.

| Dataset | Reference | Method | WA (%) | UA (%) |
|---------|-----------|--------|--------|--------|
| RML     | [9]       | LLDo + HMM | 64.00 | N/A |
|         | [18]      | spectrogram + AlexNet | 66.17 | N/A |
|         | [44]      | hand-crafted Fea. + multi-level fusion | 68.57 | N/A |
|         | Proposed  | seg-Fea. + DNN + average pooling | 70.60 | 69.85 |
|         |           | seg-Fea. + DNN + attentive temporal pooling | 73.15 | 71.90 |
| EMO-DB  | [23]      | LLDo + DNN-HMM | 77.92 | N/A |
|         | [16, 27]  | LLDo + DNN-ELM | N/A | 71.56 |
|         | Proposed  | spectrogram + 3-D attention-based convolutional LSTM | N/A | 82.82 |
|         |           | utt-Fea. + DNN | 83.51 | 77.66 |
|         |           | seg-Fea. + DNN + average pooling | 85.82 | 78.33 |
|         |           | seg-Fea. + DNN + attentive temporal pooling | 85.73 | 82.86 |
| IEMOCAP | [10, 16]  | LLDo + DNN-ELM | 57.91 | 52.13 |
|         | [24]      | LLDo + attentive CNN | 62.11 | N/A |
|         | [10]      | LLDo + RNN-ELM | 62.85 | 63.89 |
|         | [37]      | LLDo + conditional VAE + LSTM | 64.93 | 62.81 |
|         | [45]      | hand-crafted Fea. + GAN + SVM | N/A | 60.29 |
|         | [46]      | spectrogram + CNN-LSTM | 68.80 | 59.40 |
|         | [47]      | spectrogram + attention-based BLSTM-CTC | 69.0 | 67.0 |
|         | [22]      | spectrogram + attention-based PCN | 70.40 | 63.90 |
|         | Proposed  | utt-Fea. + DNN | 65.23 | 59.12 |
|         |           | seg-Fea. + DNN + average pooling | 67.51 | 61.37 |
|         |           | seg-Fea. + DNN + attentive temporal pooling | 69.37 | 60.66 |

V. CONCLUSION
This paper is motivated by how to model the salient regions over the temporal speech signal for speech emotion recognition. We present a simple but effective framework by integrating an attentive temporal pooling module into a DNN to effectively learn emotionally salient regions and adaptively generate weight for each temporal segment. The DNN model is used to learn discriminative segment-level feature representations from the comprehensive hand-crafted features, and generate emotion estimation for each segment. The attentive temporal pooling module, consisting of an auxiliary DNN and a GMM, is introduced to learn the emotional saliency weights of different segments to aggregate the segment-level estimation into the final utterance-level prediction. The whole framework is jointly trained using only the utterance-level labels, without any assumptions of either the frame-level or segment-level supervisory information. Experimental results on three public emotion datasets indicate the superiority of our proposed method for SER. In the future work, we will investigate the relationship between adjacent segments within one utterance to better explore the rich contextual dependencies involved in emotional speech expression. In addition, we will extend the proposed framework with attentive temporal pooling to audio visual multi-modal emotion recognition to further improve the emotion recognition performance.

REFERENCES
[1] B. W. Schuller, “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” Commun. ACM, vol. 61, no. 5, pp. 90–99, Apr. 2018.
[2] 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.
[3] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognit., vol. 44, no. 3, pp. 572–587, Mar. 2011.
[4] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, “The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” IEEE Trans. Affect. Comput., vol. 7, no. 2, pp. 190–202, Apr. 2016.
[5] M. Schmitt, F. Ringeval, and B. W. Schuller, “At the border of acoustics and linguistics: Bag-of-Audio-Words for the recognition of emotions in speech,” in Proc. Interspeech, Sep. 2016, pp. 495–499.
[6] N. El Din Elmadany, Y. He, and L. Guan, “Multiview emotion recognition via multi-set locality preserving canonical correlation analysis,” in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), May 2016, pp. 590–593.
[7] J. Deng, N. Cummins, J. Han, X. Xu, Z. Ren, V. Pandit, Z. Zhang, and B. W. Schuller, “The university of passau open emotion recognition system for the multimodal emotion challenge,” in Proc. Chin. Conf. Pattern Recognit., 2016, pp. 652–666.
[8] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, “Speech emotion recognition using spectrogram and phoneme embedding,” in Proc. Interspeech, Sep. 2018, pp. 3688–3692.
[9] Y. Wang, L. Guan, and A. N. Venetsanopoulos, “Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition,” IEEE Trans. Multimedia, vol. 14, no. 3, pp. 597–607, Jun. 2012.
[10] J. Lee and I. Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,” in Proc. Interspeech, 2015, pp. 1537–1540.
[11] E. Tzinis and A. Potamianos, “Segment-based speech emotion recognition using recurrent neural networks,” in Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII), Oct. 2017, pp. 190–195.
[12] W. Han, H. Ruan, X. Chen, Z. Wang, H. Li, and B. W. Schuller, “Towards temporal modelling of categorical speech emotion recognition,” in Proc. Interspeech, Sep. 2018, pp. 932–936.
J. Huang, Y. Li, J. Tao, and Z. Lian, “Speech emotion recognition from variable-length inputs with triplet loss function,” in Proc. Interspeech, Sep. 2018, pp. 3673–3677.

S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 2227–2231.

E. M. Provost, “Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2013, pp. 3682–3686.

K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in Proc. Interspeech, 2014, pp. 2227–2227.

L. Guo, L. Wang, J. Deng, L. Zhang, and H. Guan, “A feature fusion method based on extreme learning machine for speech emotion recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 2666–2670.

S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, “Learning affective features with a hybrid deep model for Audio–Visual emotion recognition,” IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 10, pp. 3030–3043, Oct. 2018.

A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017.

D. Verderidis and C. Kotropoulos, “Emotional speech classification using Gaussian mixture models,” in Proc. IEEE Int. Symp. Circuits Syst., 2005, pp. 2871–2874.

A. Stuhlsatz, C. Meyer, F. Eyben, T. Ziekel, G. Meier, and B. W. Schuller, “Deep neural networks for acoustic emotion recognition: Raising the benchmarks,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2011, pp. 5688–5691.

Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, “Attention based fully convolutional network for speech emotion recognition,” in Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC), Nov. 2018, pp. 1711–1715.

L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli, “Hybrid deep neural network-hidden Markov model (DNN-HMM) based speech emotion recognition,” in Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact., Sep. 2013, pp. 312–317.

M. Neumann and N. T. Vu, “Attentional convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech,” in Proc. Interspeech, Aug. 2017, pp. 1263–1267.

J. Liu, W. Han, H. Ruan, X. Chen, D. Jiang, and H. Li, “Learning salient features for speech emotion recognition using CNN,” in Proc. 1st Asian Conf. Affect. Comput. Intell. Interact. (ACHI Asia), May 2018, pp. 1–5.

S. Zhang, S. Zhang, T. Huang, and W. Gao, “Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching,” IEEE Trans. Multimedia, vol. 20, no. 6, pp. 1576–1590, Jun. 2018.

M. Chen, X. He, J. Yang, and H. Zhang, “3-D convolutional recurrent neural networks with attention for speech emotion recognition,” IEEE Signal Process. Lett., vol. 25, no. 10, pp. 1440–1444, Oct. 2018.

B. W. Schuller, S. Steidl, and A. Batliner, “The interspeech 2009 emotion challenge,” in Proc. Interspeech, 2009, pp. 312–315.

B. W. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, “The interspeech 2010 paralinguistic challenge,” in Proc. Interspeech, 2010, pp. 2794–2797.

B. W. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, and G. Mohammadi, “The interspeech 2012 speaker trait challenge,” in Proc. Interspeech, 2012, pp. 753–757.

B. W. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, and M. Mortillaro, “The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism,” in Proc. Interspeech, 2013, pp. 148–152.

S. Song, N.-M. Cheung, V. Chandrasekhar, and B. Mandal, “Deep adaptive temporal pooling for activity recognition,” in Proc. ACM Multimedia Conf. Multimedia Conf., 2018, pp. 1829–1837.

E. Variani, E. McDermott, and G. Heigold, “A Gaussian mixture model layer jointly optimized with discriminative features within a deep neural network architecture,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2015, pp. 4270–4274.

Y. Wang and L. Guan, “Recognizing human emotional state from audio-visual signals,” IEEE Trans. Multimedia, vol. 10, no. 5, pp. 936–946, Aug. 2008.

F. Burkhardt, A. Paeschke, M. Rolfs, W. F. Sendmeier, and B. Weiss, “A database of German emotional speech,” in Proc. Interspeech, 2005, pp. 1517–1520.

C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.-N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP. Interactive emotional dyadic motion capture database,” Lang. Resour. Eval., vol. 42, no. 4, pp. 335–359, Dec. 2008.

S. Latif, R. Rana, J. Qadir, and J. Epps, “Variational autoencoders for learning latent representations of speech emotion: A preliminary study,” in Proc. Interspeech, Sep. 2018, pp. 3107–3111.

F. Eyben, F. Weninger, F. Gross, and B. W. Schuller, “Recent developments in openSMILE, the Munich open-source multimedia feature extractor,” in Proc. 21st ACM Int. Conf. Multimedia, 2013, pp. 835–838.

C. Busso, A. Metallinou, and S. S. Narayanan, “Iterative feature normalization for emotional speech detection,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2011, pp. 5692–5695.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, “PyTorch: An imperative style, high-performance deep learning library,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 8026–8037.

S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. 32Nd Int. Conf. Int. Mach. Learn. (ICML), 2015, pp. 448–456.

B. W. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, “Acoustic emotion recognition: A benchmark comparison of performances,” in Proc. IEEE Workshop Autom. Speech Recognit. Understand., 2009, pp. 552–557.

I. J. Tashev, Z.-Q. Wang, and K. Godin, “Speech emotion recognition based on Gaussian mixture models and deep neural networks,” in Proc. Int. Theory Appl. Workshop (ITA), Feb. 2017, pp. 1–4.

C. S. Ooi, K. F. Seng, L.-M. Ang, and L. W. Chew, “A new approach of audio emotion recognition,” Expert Syst. Appl., vol. 41, no. 13, pp. 5858–5869, Oct. 2014.

S. Sahu, R. Gupta, and C. Espy-Wilson, “On enhancing speech emotion recognition using generative adversarial networks,” in Proc. Interspeech, Sep. 2018, pp. 3693–3697.

A. Satt, S. Rozenberg, and R. Hoory, “Efficient emotion recognition from speech using deep learning on spectrograms,” in Proc. Interspeech, Aug. 2017, pp. 1089–1093.

Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. W. Schuller, “Attention-enhanced connectionist temporal classification for discrete speech emotion recognition,” in Proc. Interspeech, Sep. 2019, pp. 206–210.

XIAOHAN XIA received the B.Eng. degree from Zhengzhou University, Henan, China, in 2012, and the M.Eng. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2015. She is currently pursuing the Ph.D. degree with the Shaanxi Key Laboratory on Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi’an, China. She was the author of the Award-Winning Paper from the Multimodal Emotion Recognition Challenge (MEC2016). Her main research interests include affective computing, speech emotion recognition, and facial expression analysis.
DONGMEI JIANG (Member, IEEE) received the B.Eng. and master’s degrees in automatic control and the Ph.D. degree in computer science and technology from Northwestern Polytechnical University (NPU), Xi’an, China, in 1994, 1997, and 2000, respectively. She was a Visiting Scholar with the Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Belgium, from 2001 to 2002, and from 2006 to 2007. Since 2005, she has been a NPU Team Coordinator with the Joint NPU-VUB Audio Visual Signal Processing (AVSP) Laboratory. She was also affiliated with NPU, where she was a Professor of computer science and technology, in 2010. She is the Corresponding Author of the Winner Paper from the Audio Visual+ Emotion Challenge (AVEC2015) and the AVEC2016 Depression Sub-challenge. Her research interests include multimodal affective computing, including emotion recognition from speech, facial expression, body gesture, and expressional facial animation synthesis.

HICHEM SAHLI (Member, IEEE) received the degree in mathematics and computer science, the D.E.A. degree in computer vision, and the Ph.D. degree in computer sciences from the Ecole Nationale Supérieure de Physique de Strasbourg, France. He was a Chargé de Recherche with the CAD and Robotics Department, Ecole des Mines de Paris. Since 1999, he has been a Professor with the Department of Electronics and Informatics (ETRO) and has been a Team-Coordinator with the Interuniversitair Micro-Elektronica Centrum vzw (IMEC). Since 2005, he has been a Coordinator with the AVSP Research Cluster, ETRO, and the Joint NPU-VUB AVSP Laboratory. Within ETRO-IRIS, he coordinates the research team in computer vision. His research team is dealing with variational and partial differential equations, stochastic models in computer vision, image and motion analysis, pattern recognition, and machine learning for object detection recognition and tracking. His main research interests include mathematical models and tools for the application of novel machine learning techniques for multimodal content analysis, automatic detection of affect in audio-visual speech, and body gestures. He serves as a Regular Reviewer for the major vision and multimedia conferences and journals.