Speech Emotion Classification Using Attention-Based LSTM

Yue Xie, *Student Member, IEEE*, Ruiyu Liang, *Member, IEEE*, Zhenlin Liang, Chengwei Huang, *Member, IEEE*, Cairong Zou, and Björn Schuller, *Fellow, IEEE*

***Abstract*—Automatic speech emotion recognition has been a research hotspot in the field of human–computer interaction over the past decade. However, due to the lack of research on the inherent temporal relationship of the speech waveform, the current recognition accuracy needs improvement. To make full use of the difference of emotional saturation between time frames, a novel method is proposed for speech recognition using frame-level speech features combined with attention-based long short-term memory (LSTM) recurrent neural networks. Frame-level speech features were extracted from waveform to replace traditional statistical features, which could preserve the timing relations in the original speech through the sequence of frames. To distinguish emotional saturation in different frames, two improvement strategies are proposed for LSTM based on the attention mechanism: first, the algorithm reduces the computational complexity by modifying the forgetting gate of traditional LSTM without sacrificing performance and second, in the final output of the LSTM, an attention mechanism is applied to both the time and the feature dimension to obtain the information related to the task, rather than using the output from the last iteration of the traditional algorithm. Extensive experiments on the CASIA, eNTERFACE, and GEMEP emotion corpora demonstrate that the performance of the proposed approach is able to outperform the state-of-the-art algorithms reported to date.**

***Index Terms*—Speech emotion, frame-level features, LSTM, attention mechanism.**

I. Introduction

SPEECH emotion recognition (SER) has great practical value in human-computer interaction [1]–[4] and a range of further applications. To realize emotion classification based on speech, substantial research work was invested in machine learning algorithms, such as support vector machines [5]–[7], Bayesian classifiers [8], [9] and K nearest neighbors [10], [11]. In recent years, deep learning has been widely used for automatic speech emotion recognition. Deng [12] used semi-supervised learning with auto-encoders and a small amount of emotional label data for SER. Neumann [13] and Wöllmer [14] applied convolutional neural network and LSTM to SER respectively.

Although the above algorithms have been successfully applied in emotion recognition, most of the traditional machine learning algorithms and deep learning networks (such as auto-encoder and convolutional neural networks) can only accept data with fixed dimensions as input. This appears contradictory for utterance-level emotion recognition with a variable length of speech. To solve this problem, first, the most popular methods [15]–[18] extract emotion-related features (called frame-level features in this paper) from short-term speech frames, and then static statistical functions(e.g., mean, variance, maximum, or a linear regression coefficient) are applied to frame-level features, and the results are concatenated into a vector with a fixed dimension to represent the complete speech waveform. Although these features with fixed dimensions satisfy the requirements of model input, the speech features processed by statistical analysis lose the temporal information in the original speech. Another solution to this contradiction is to design a model that can accept variable length features. For example, the Long Short-term Memory (LSTM) structure is proposed by Schmidhuber [19] for recurrent neural networks (RNN). This method provides feasibility for processing temporal sequences with a variable length such as speech.

In recent years, to strengthen the ability of LSTM to process data in specific tasks, many improvements for the internal structure of LSTM were proposed. Schmidhuber [19] proposed a peephole connection by using the historical cell status as input information to enhance the ability to learn historical information. Yao [20] controls the flow of data between memory cells by introducing a depth gate to connect memory cells between layers. However, these improved LSTM variants enhance memory information at the expense of computational complexity. In addition, in many LSTM applications [14], [21]–[23], the output of the last time of LSTM is often selected as the input to the next model (because other models can only accept inputs with fixed dimensions). However, the speech is mostly silent at the end in the speech emotion recognition task, and there is almost no

Manuscript received November 2, 2018; revised April 14, 2019 and June 25, 2019; accepted June 27, 2019. Date of publication July 1, 2019; date of current version August 1, 2019. This work was supported in part by the National Natural Science Foundation of China under Grants 61871213, 61673108, and 61571106, in part by Six Talent Peaks Project in Jiangsu Province under Grant 2016-DZXX-023, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20161517. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tuomas Virtanen. *(Corresponding author: Ruiyu Liang.)*

Y. Xie, Z. Liang, and C. Zou are with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: 230169046@seu.edu.cn; zhenlinliang1@163.com; cr_zou@seu.edu.cn).

R. Liang is with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China, and also with the School of Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, China (e-mail: lly1711@163.com).

C. Huang is with the Sugon (Nanjing) Institute of Chinese Academy of Sciences Co. Ltd., Nanjing 211106, China (e-mail: huangcwx@126.com).

B. Schuller is with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg 86159, Germany, and also with the Group on Language, Audio, and Music, Imperial College London, London SW7 2AZ, U.K. (e-mail: bjoern.schuller@imperial.ac.uk).

Digital Object Identifier 10.1109/TASLP.2019.2925934

emotional information. Therefore, the emotional information will be weakened at the last moment. How to effectively use the LSTM output at all times (rather than a single last moment) is the key to improving the performance of speech emotion recognition. To solve the above issues, an improved LSTM model is proposed for the speech emotion recognition task. First, the model employs frame-level speech features as input. The dimension of feature changes with the actual speech length, and the temporal information in the original speech is preserved by the sequences between frames. Thus, it is more suitable for the input of LSTM with the ability to handle variable length sequences. Second, in order for the memory cells of LSTM to utilize the critical information in the historical state efficiently, an attention gate is proposed as an alternative to the forgetting gate in the traditional LSTM. This improvement not only reduces the computational complexity of the LSTM but also optimizes the emotion recognition performance. In addition, the emotional saturation is different among the time segments of speech (silent fragments contain less emotional information), and various speech features differ in abilities to distinguish emotions [24]. Therefore, it is feasible to distinguish the differences by weight coefficients to make full use of emotional information and improve emotion recognition performance. Hence, for the particularity of speech emotion recognition, this paper proposes a weighting method based on the attention mechanism for the output of LSTM, which has been successfully applied in the field of image processing [25]–[27]. This weighting operation not only acts on the time dimension but also on the feature dimension. Finally, the performance of the model is verified on the CASIA, eNTERFACE, and GEMEP corpus.

In summary, the main contributions of this paper to speech emotion recognition research are as follows:

- An attention gate for LSTM is proposed to address the problem that most machine learning algorithms only accept fixed-dimensional data as input and cannot handle time series effectively due to a lack of memory ability, which optimizes the forgetting gate in traditional LSTM, enabling memory cells to use historical information more efficiently and simultaneously reducing the computational complexity of LSTM.
- A weighting method based on the attention mechanism for the output on the time and feature dimensions is proposed to distinguish the differences of emotional saturation among speech time segments and the abilities of different features to distinguish emotions.

II. Related Works

With the successful application of LSTM in natural language processing [28]–[30], it also has been introduced into speech emotion recognition. Wöllmer [14] first applied LSTM to continuous emotion recognition and extracted 4843 features for each utterance as the input of LSTM. In his further work [15], the static features were used as the input of bidirectional LSTM (BLSTM) to predict the emotional expression of a spoken utterance. In the term of features, the temporal information in speech is not fully utilized because of the global statistics ignoring the temporal structure of speech [31]. In order to enhance the features, the time window is fed frame by frame into a recurrent layer in [32], and the experiments on emotion classification got a better result. In the earlier works, frame-level features were directly used for SER [31], [33]–[36], which preserve the temporal information through the sequences among frames. It is well known that LSTM is adept at processing sequential data, so frame-level features are more suitable for its input than that with statistics.

In addition to the input of LSTM, the output of conventional LSTM should be improved. In most applications of LSTM [21]–[23], the output of the last moment in LSTM is selected as the input to the next model (since other models only accept inputs with a fixed dimension, while the dimension of LSTM's output is the same as the input of which the real dimension is not uniform). This can lead to imperfect use of LSTM output information at other historical moments; Specifically, the accumulative information of LSTM at the last moment is lossy because the time span of long-term dependencies is not infinite [37], [38]. In the emotion classification task, Keren [32] introduced a pooling operation of convolutional neural networks to the output of LSTM. Mirsamadi [39] proposed an attention mechanism for computing weights for frames with an attention parameter vector. Due to the memory capacity of LSTM, the accumulated information is the most abundant in the output of the last time. Therefore, the output of the last time is often taken as the final output of the LSTM (In both this study and [39], this method could recognize the emotion). Theoretically, the last time of LSTM networks should obtain a large weight. Therefore, this study takes the output of the last time as a reference to ensure that it can obtain a large weight by using the attention mechanism. Moreover, considering the difference of distinguishing ability between speech features, the attention mechanism is also applied to the feature dimension of the LSTM's output.

Not only can the attention mechanism be used to optimize the output of the LSTM, but it can also be used for updating memory cells. Some research has investigated how to update memory cells. Tao [40] applied the attention mechanism to update cell states of LSTM, who focused on the information between cells and considered more previous cell states. In the term of computation, Bradbury [41] presented quasi-recurrent neural networks for neural sequence modeling that allowed the output to depend on the overall order of elements in the sequence and had faster speed than the conventional LSTM at train and test time. Cho [42] proposed the gated recurrent unit (GRU), which combines the input and forgetting gates into an update gate and mixes the cell state and the hidden layer state, thus simplifying the calculation of LSTM. Greff [22] introduced a Coupled LSTM that uses only one gate to control the effects of historical cell states and candidate cell states on current cell status, simplifying the calculation of candidate cell state weights. Unlike the above works, this study focuses on the computation of the inside of cell and modifies the forgetting gate of LSTM with self-attention algorithm [43]. Therefore, the computation of the forgetting gate is different from that of the previous LSTM. Since the self-attention algorithm is only related to the historical cell state itself, regardless of the current input and

TABLE I
FRAME-LEVEL SPEECH FEATURES

Features	Description
voiceProb	Voicing probability
HNR	Log harmonics-to-noise ratio
F0	Pitch frequency
F0raw	Raw F0 candidate without threshold in unvoiced segments
F0env	F0 envelope
jitterLocal	The AAD[1] between consecutive periods
jitterDDP	The AAD between consecutive differences between consecutive periods
shimmerLocal	The AAD between the interpolated peak amplitudes of consecutive periods
harmonicERMS	Harmonic component RMS[2] energy
noiseERMS	Noise component RMS energy
pcm_loudness_sma	Loudness
pcm_loudness_sma_de	Delta regression of loudness
mfcc_sma[0]-[14]	Mel-Frequency Cepstral Coefficients
mfcc_sma_de[0]-[14]	delta regression of mfcc
pcm_Mag[0]-[25]	Mel Spectral
logMelFreqBand[0]-[7]	log Mel frequency bands
lpcCoeff[0]-[7]	Linear predictive coding coefficients
lspFreq[0]-[7]	Line spectral pair frequency
pcm_zcr	Zero-crossing rate

[1]Average Absolute Difference
[2]Root Mean Square

historical hidden layer states, the computational complexity can be reduced.

III. FRAME-LEVEL SPEECH FEATURES

The ComParE openSMILE features proposed by Schuller *et al.* is most widely used in speech emotion recognition [12], [44], of which one [17] has a dimension of 6373 features based on the extraction of Low-Level Descriptors (LLD, such as zero-crossing-rate, root mean square frame energy, pitch frequency and Mel-frequency cepstral coefficients 1–12), adding their deltas, and applying statistical functions. Based on the openSMILE ComParE features, frame-level speech features (i.e., the features without statistical functionals.) are directly used for emotion classification, which are shown in Table I. The basic reasons are: (1) the fixed-length feature calculation of the statistical functional loses much information from the original speech, such as time information. (2) Hinton [45] believed that deep learning has the ability to automatically learn feature changes, and can learn deep features related to tasks from the underlying speech features. Thus, the frame level feature appears more suitable as an input to the deep learning network suggested herein.

The openSMILE ComParE feature set uses the harmonic to noise ratio (HNR) that is the ratio of the harmonic energy (harmonicERMS) and the glottal noise energy (noiseERMS) as one of the characteristics:

$$\mathrm{HNR} = 10log_{10}\left\{\sum_{n=1}^{N} g^2(n) / \sum_{n=1}^{N} n^2(n)\right\}, \qquad (1)$$

where $g(n)$ and $n(n)$ represent glottal harmonic signals and noise signals, respectively. N is the length of speech.

However, HNR blurs the differences between different emotion categories due to the presence of divisions in the ratio. Conversely, the harmonicERMS and noiseERMS features can preserve the differences between the emotional categories. At the same time, some work such as [7] confirmed that the harmonic information of speech can be used to distinguish emotion categories in the CASIA and EMODB databases. Research [46] also shows that glottis waves contain certain emotional information. Thus, glottal harmonic energy and glottal noise energy are separately extracted as an emotional feature to reflect the glottal closure state.

To visualize the impact of features on classification, a number of samples (X-axis) are taken from the CASIA [47], eNTERFACE [48] and GEMEP corpus [49], then the mean values of the three features of these samples are calculated over the time frames. As Fig. 1 illustrates, on the eNTERFACE corpus, the discrimination among emotions was obvious in the harmonicERMS and noiseERMS contour, which was severely reduced in the HNR (the ratio of the harmonicERMS and noiseERMS) contour due to the division operation. Similarly, on the CASIA corpus, the difference in emotions in the HNR dimension is smaller than for the harmonicERMS and noiseERMS dimensions. In addition, the anger emotion is at a higher level on these corpora (this is also the commonality between the above corpus) among all categories of emotions, and hence it is relatively easier to distinguish from other emotions. On the CASIA corpus, the neutral emotion is at the lowest level in the three feature contour, so it is also relatively easier to distinguish. On the eNTERFACE corpus, the sad emotion, which is at the lowest level, theoretically has the stronger distinguishability, while the disgust, fear and surprise emotions, which overlap with each other, may be difficult to distinguish. On the GEMEP corpus, the contours of all emotions overlap with each other, for which one of the possible reasons is some emotion portrayals with non-semantic short text 'aaa'. Therefore, the emotions' distinguishability of GEMEP is lower than that on the other two corpus, which indicates that the average recognition rate on GEMEP would be lowest. In summary, different features have different capabilities to distinguish emotions in different databases.

IV. ATTENTION-BASED LSTM

The attention mechanism was first applied to the field of image processing [25]–[27], with very good results. The core idea is that the human brain's attention to the whole picture is not balanced, and there is a certain weight distinction. Inspired by this phenomenon, this paper introduces the self-attention mechanism into the forgetting gate calculation of LSTM to reduce the model calculation with the premise of ensuring the performance of the model. At the same time, the frame-level speech features used in emotion recognition include not only time information but also feature-level information. These different characteristics may have different degrees of influence on the final classification performance. For this reason, feature level information is also multiplied by the attention weighting coefficients to improve the final performance of the model.

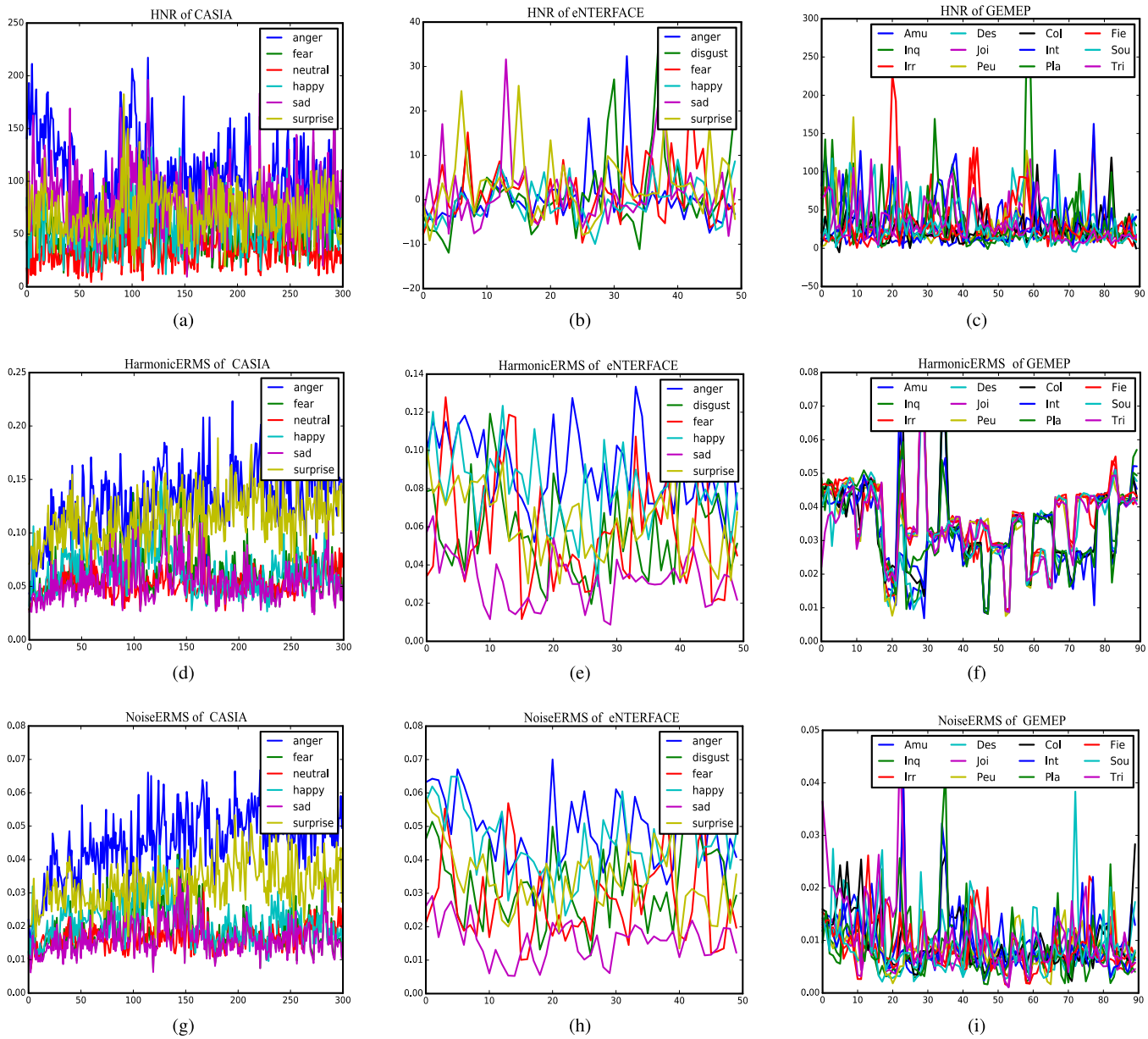

Fig. 1. Features analyses; Fig. 1a, Fig. 1b and Fig. 1c are HNR statistical analysis; On the CASIA corpus, the mean of the anger emotion is at the highest level, and the neutral emotion is at the lowest level. Therefore, it is easier to distinguish them from the 6 emotions. However, this feature has poorer distinguishability on both eNTERFACE and GEMEP corpus. Fig. 1d, Fig. 1e and Fig. 1f are harmonicERMS statistical analysis; Fig. 1g, Fig. 1h and Fig. 1i are noiseERMS statistical analysis; On the CASIA, the anger emotion and the surprise emotion have the stronger distinguishability, which are at the highest and second highest levels respectively and with a relatively small overlap with each other. On the eNTERFACE, the anger emotion and the sad emotion have the stronger distinguishability, which are at two different extremes respectively. On the GEMEP corpus, the contours of all emotions overlap with each other.

A. *Attention Gate*

The forgetting gate of the LSTM cell is used to determine what information should be discarded in the cell state at the previous moment and participate directly in updating the cell state. In the original LSTM proposed by Hochreiter [19], the update algorithm of the cell state is related to the hidden layer output at the previous moment and the input at the current moment. Furthermore, they added a peephole connection and took the cell state of the previous moment as a parameter to update the current state.

The forget gate calculation formula is shown in Eq. (2):

$$f_t = \sigma(W_f \times [C_{t-1}, h_{t-1}, x_t] + b_f) \tag{2}$$

The cell state update formula is as shown in Eq. (3)–(5):

$$i_t = \sigma(W_i \times [C_{t-1}, h_{t-1}, x_t] + b_i) \tag{3}$$

$$\tilde{C}_t = \tanh(W_C \times [C_{t-1}, h_{t-1}, x_t] + b_C) \tag{4}$$

$$C_t = f_t \bullet C_{t-1} + i_t \bullet \tilde{C}_t \tag{5}$$

where C_{t-1} and h_{t-1} are the cell state and hidden layer output at the previous moment, respectively. x_t is the input at the current moment. $\tilde{C}_t$ is the candidate value for updating cell state. W_f, W_i and W_C are the weights of forgetting gate, input gate and candidate cell receptively, and b_f, b_i and b_C are their biases receptively. i_t is the weight coefficients of $\tilde{C}_t$. $\bullet$ is the Hadamard product. σ is the logistic sigmoid function.

$$\sigma(x) = \frac{1}{1+e^x} \tag{6}$$

In Coupled LSTM [22], i_t is obtained by $(1-f_t)$,which means that the forget gate determines the weight coefficient of both the previous and current information. The cell state updating formula is modified to equation (7):

$$C_t = f_t \bullet C_{t-1} + (1-f_t) \bullet \tilde{C}_t \tag{7}$$

As seen from Eq. (7), the forget gate f_t essentially updates the current state of the cell by calculating the new and old cell states' weighted summation. In light of the weighting coefficient, this paper proposes a method using the self-attention mechanism [43] to obtain key information about the cell's own state by training the parameters of the self-attention model to update the new cell state. In this paper, we refer to this as the attention gate. Its formula is indicated in the following equation (8).

$$f_t = \sigma(V_f \times \tanh(W_f \times C_{t-1})) \tag{8}$$

where $\mathrm{V}_f \in R^{N\times N}$ and $\mathrm{W}_f \in R^{N\times N}$ are the parameters to be trained and N is the number of hidden units. Compared with (2), the dimension of the weight parameter W_f is reduced because Eq. (8) has no h_{t-1} and x_t, that means the number of parameters to be trained is fewer. This greatly reduces the number of training calculations. The forget gate of LSTM is calculated at each moment, so the reduction of the amount of calculation of the forget gate can greatly improve the effectiveness of LSTM model training. The experimental results show that the combination of Eq. (7) and Eq. (8) to update the cell state does not affect the performance of the final LSTM mode.

B. Output

The length of feature of frame-level speech varies with the number of speech frames, and classical LSTM can learn deep features with fixed length from the variable-length frame-level speech features by selecting the output of the last moment. The output of the LSTM model proposed by Gers [50] is as shown in Eq. (9).

$$o_t = \sigma(W_o \times [C_t, h_{t-1}, x_t] + b_o) \tag{9}$$

where W_o and b_o are weights and biases of the output gate. Traditional LSTM selects the last moment of output (denoted as $o_{max\_time} \in R^{B\times 1\times N}$. B and represent the size of batch. 1 means the last time step) as the input to full connection layers (or another model that requires fixed length data as the input). o_t is the output at the t-th step. Combined with the characteristics of frame-level speech features, this paper proposes a method of attention weighting for output of all time steps $\mathrm{o}_{all\_time} \in R^{B\times M\times N}$ (M is the number of time steps) on the time dimension and feature dimension simultaneously, and then combines the weighted results together as the final output.

1) Attention on Time Dimension: Since the degree of emotional saturation in each frame is not uniform – that means the contribution of each frame to the final emotional recognition is different – the degree of contribution can be expressed by the weight coefficients of the frames. In [51], the weight coefficients are calculated by the output of the encoder and the current input of the decoder based on the attention mechanism. Mirsamadi [39] also proposed an attention mechanism for computing weights for frames with an attention parameter vector u, as described in Eq. (10).

$$\alpha_{\mathrm{t}} = \frac{\exp(u^H y_t)}{\sum_{\tau=1}^{T} \exp(u^H y_\tau)} \tag{10}$$

where α_{t} is the weight for the output at t-th time step y_t. H denotes the transpose operator. Due to the memory ability of LSTM, the accumulated information is the most abundant in the output of the last moment. Therefore, the output of the last moment is often taken as the final output of the LSTM (In both this study and [39], this method could recognize the emotion). Theoretically, the last moment of LSTM networks should obtain a large weight. Therefore, this study takes the output of the last moment as a reference to ensure that it can obtain a large weight by using attention mechanism. Finally, the weight coefficients are applied to o_{all_time} on the time dimension and summed up in the time dimension as an output. The relevant calculation formula is:

$$s_T = \mathrm{softmax}(o_{max\_time} \times (o_{all\_time} \times w_t)^H) \tag{11}$$

$$output_T = s_T \times o_{all\_time} \tag{12}$$

where $w_t \in R^{N\times N}$ is the weight for training. $s_T \in R^{B\times 1\times M}$ represents the attention weight coefficients on the time dimension. The M corresponds to the frames. $output_T$ should has the dimension of $[B, 1, N]$, which could be reshaped as the dimension $[B, N]$ and used as the input of full connection layer.

2) Attention on Feature Dimension: It is well known that it is difficult to use single features to accomplish multicategory classification tasks, so multiple features often must be combined to accomplish these tasks. However, the distinguishability of each feature to the target task is not the same. To express the difference among features, an attention weighting is calculated on the feature dimension:

$$s_F = \mathrm{softmax}(\tanh(o_{all\_time} \times w_F) \times v_F) \tag{13}$$

$$output_F = \sum_{time} s_F \bullet o_{all\_time} \tag{14}$$

where $v_F, w_F \in R^{N\times N}$ are trainable parameters of self-attention algorithm. N is not only the number of hidden units, but also represents a new feature space. $s_F \in R^{B\times M\times N}$ could be obtained of which the value is different from each other in the N-axis, that means, it could reflect the difference among features in the new space. In Eq. (14), the summation is operated on the time frames, of which the aim is to calculate the statistical functions of features over the time dimension. If each frame gets the

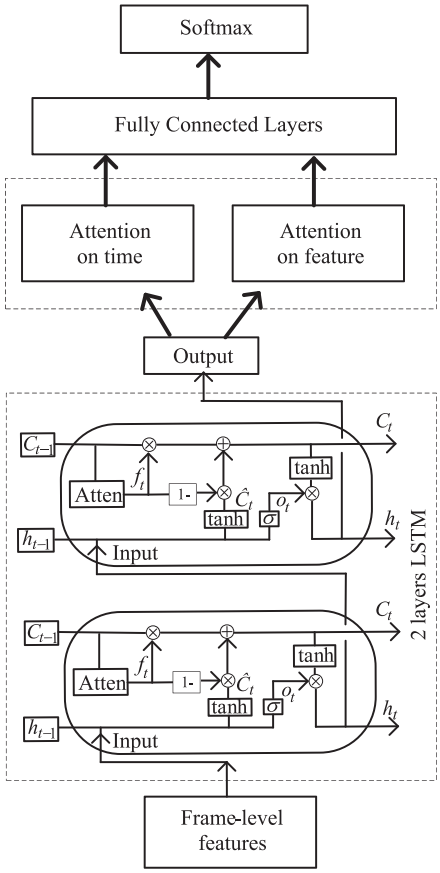

Fig. 2. Model architecture. The model takes the frame-level speech feature as input and obtains the output which corresponding time of each frame through the 2-layer LSTM. The LSTM's internal forgetting gate has been replaced by attention gate. In order to distinguish the difference of emotion in time and features, the model performs weighting operation on the output of LSTM on time dimension and feature dimension respectively, and takes the two weighted results as the input of full connection layers. Finally, the output of the softmax layer is the result of emotion recognition.

same values, the mean of feature can be obtained by summing on the time dimension. Therefore, $output_F \in R^{B\times 1\times N}$ is like the statistical value of the feature in the time dimension.

Finally, $[output_T, output_F]$ is used as the input of full connection layers, rather than the output o_{max_time} corresponding to the last moment of o_{all_time}, as shown in Fig. 2. This new LSTM output considers differences in both time levels and feature levels. It can enhance key information and weaken secondary information, thereby improving the ability to represent features.

V. Experiments and Results Discussion

To showcase the performance of the suggested approach, we chose three different popular databases to avoid observations based on single corpus evaluation. The CASIA [47], eNTERFACE [48] and GEMEP corpus [49] are used for experiments. The CASIA is an emotion corpus introduced by the Institute of Automation, Chinese Academy of Sciences that contains 6 categories of emotion (i.e., anger, fear, happy, neutral, sad and surprise). The corpus contains 7200 speech samples recorded from 4 speakers (2 males and 2 females), of which 1000 samples are randomly selected as the test set. The eNTERFACE is an audio and video emotion corpus in English, recorded from 43 speakers from 14 countries, and classifies samples based on the following 6 emotions: anger, disgust, fear, happy, sad and surprise. Only the audio data from this corpus is used in this work. 1260 valid speech samples are obtained for the emotion recognition study, of which 260 samples are used as the test set. GEMEP is a French-content corpus with 18 speech emotional categories and 1260 utterance samples. We choose 12 categories of emotions (amusement (amu), anxiety (inq), despair (des), hot anger (col), joy (joi), panic fear (peu), interest (int), irritation (irr), pleasure (pla), pride (fie), relief (sou), sadness (tri); NOTE: the abbreviations come from French) in our experiments as in [8], [52]. Those are totally 1080 samples by ten speakers belonging to the chosen categories, where 200 samples are randomly selected as the test set. As the current knowledge to authors, the recognition accuracy based on speech is less than 90% on CASIA[53], [54], 80% on eNTERFACE [8], [44] and 50% on GEMEP [8], [55].

TABLE II
Parameters in Network

Parameters	Values
Input	[128, timestep, 93]
Learning Rate	0.0001/0.001
Hidden units (the first LSTM)	512
Hidden units (the second LSTM)	256
Hidden units (full connection layer)	[512, 128]/[256, 128]
Hidden units (output)	[128, 6]/[128, 12]

The proposed models, including the LSTM based on attention-weighting in the time dimension (LSTM-T), the LSTM based on attention-weighting on the feature dimension (LSTM-F), the LSTM based on the modified forget gate with the attention mechanism (LSTM-at), the LSTM based on attention-weighting on both time and feature dimensions (LSTM-TF) and its variant of the forget gate (LSTM-TF-at),consist of 2 LSTM layers, and the settings of relevant parameters are given in Table II. The input has the dimension of $[128, timestep, 93]$, where 128 is the size of batch, $timestep$ is the number of frames and 93 is the number of features exacted from speech. In order to compare the time complexity, these parameters are the same on all corpus without screening. Only the learning rate is adjusted according to the stability of the convergence on the training set. The initial learning rate is 0.0001 on CASIA, and 0.001 on both eNTERFACE and GEMEP. The dimension of output is determined by the number of emotion in the corpus (CASIA and eNTERFACE have 6 categories, while GEMEP has 12 categories). Since two attention-weighting operations are performed on the output matrix of LSTM, and the results are concatenated in the form of $[output_T, output_F]$ as the input of the subsequent fully connected layer, the cell number of the full connection layer is doubled. The full connection layer parameter $[256, 128]$ in Table II corresponds to the network based on traditional LSTMs, while the $[512, 128]$ is the parameter setting of the LSTM networks based on the attention mechanism for the time and feature

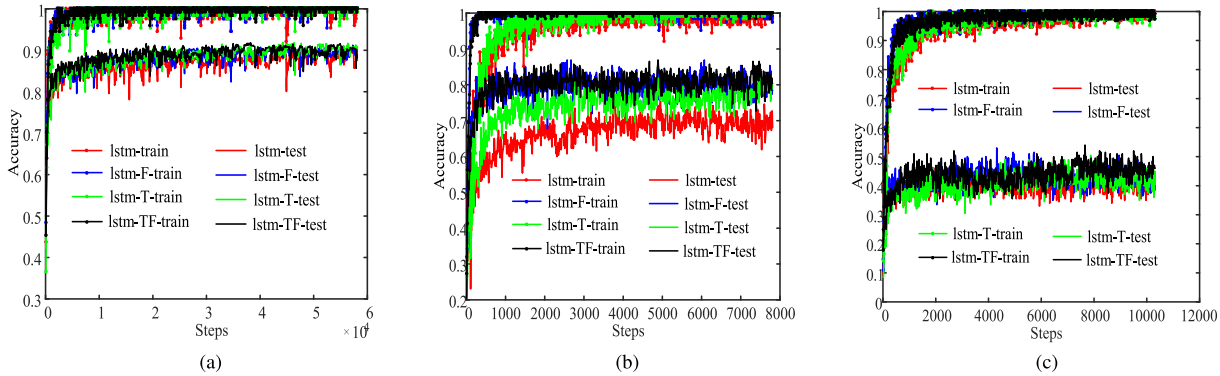

Fig. 3. Convergence curves of the models.

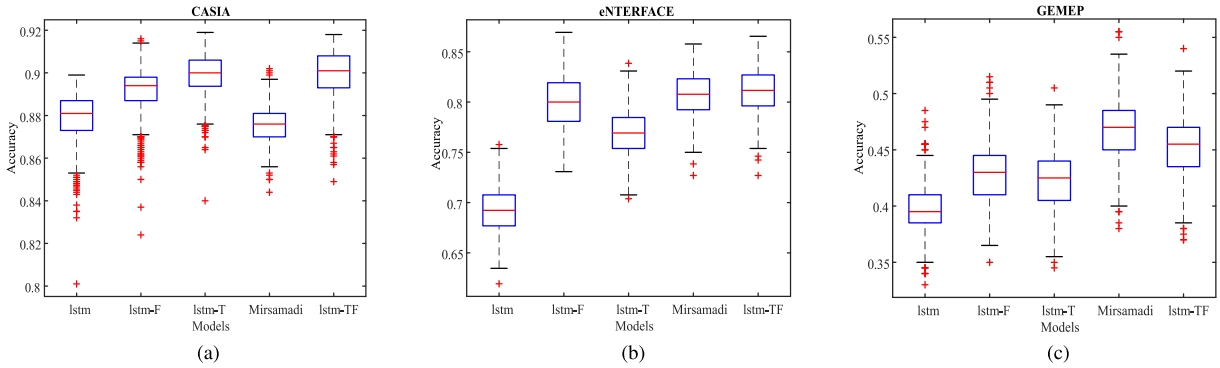

Fig. 4. Stability of models on test set.

dimension. Other parameters remain invariant to ensure the validity of the experiment results.

A. *Attention Mechanism for LSTM's Output*

Attention-weighting is implemented on the output of LSTM to emphasize the key information relevant to the task. To demonstrate its validity, comparisons among the normal LSTM, LSTM-T, LSTM-F, and LSTM-TF were conducted. Fig. 3 depicts the convergence of models during training to explain the speed of convergence under the training steps. On all mentioned corpora, the speed of LSTM-TF (black curves) is obviously faster than that of LSTM, which implies that attention mechanism on output could speed up the model mining the information related to the emotion classification. Fig. 4 shows the stability of models on the test set, where the vertical axis reports the unweighted average recall (UAR), the height of rectangular box denotes the stability of model and the red lines are the stable UAR. As shown in the figures, the LSTM-TF achieve higher stable UAR and ensure the stability of model similar to the traditional LSTM. Even though the attention mechanism is only applied to a single dimension (LSTM-F or LSTM-T), the performance still is improved. On CASIA, the height of rectangular box is smaller than that on the others, that means it has better stability because of the distinguishing features.

As to the time dimension, the lstm-T is compared with [39]. On the eNTERFACE and GEMEP corpora, the [39] outperforms the lstm-T method. However, lstm-T achieves better accuracy on CASIA corpus that contains about six times as much data as the other two corpora. In Eq. (11), w_t is a matrix instead of a vector u in Eq. (10) used in [39], that means the number of parameters of the proposed method (lstm-T) is more than that of [39]. Therefore, it needs more data to train and works better on large data sets.

To quantitatively analyze these models' recognition accuracy for different emotions in each test set, the highest UAR of each model is selected for comparing. Table III, Table IV and Table V are recognition results for CASIA , eNTERFACE and GEMEP, respectively. As seen from the tables, the anger emotion shows higher recognition rates than other emotions in both CASIA and eNTERFACE corpus, consistent with the results of the feature analyses (see Fig. 1). The hot anger emotion reaches the highest recall rate among all the emotions on the GEMEP corpus. For the CASIA corpus, the anger and neutral emotions whose feature levels are, respectively, highest and lowest, also show higher distinguishability. The LSTM-TF models improve the recall of all emotions except anger, but the overall performance is only improved by 2% compared with normal LSTM, possibly because the baseline of recognition performance on the CASIA corpus is high, and hence room for improvement is limited. However, despite the low baseline of GEMEP, the UAR is raised only 53% from 48%, as the overlap of emotion (see Fig. 1c, Fig. 1f and Fig. 1i) is severely unrecognizable. On the eNTERFACE corpus, the results obtained by the LSTM-F model are

TABLE III
RESULTS OF CASIA

Models	Anger	Fear	Happy	Neutral	Sad	Surprise	UAR
LSTM	91.3%	87.1%	84.9%	95.6%	86.3%	94.8%	90.0%
LSTM-T	93.6%	91.8%	87.9%	93.0%	93.2%	91.9%	91.9%
LSTM-F	95.9%	90.1%	86.7%	95.6%	91.9%	90.8%	91.8%
LSTM-TF	87.2%	89.5%	87.3%	98.7%	95.0%	94.8%	92.0%
LSTM-at	91.9%	85.4%	86.7%	94.9%	91.3%	93.6%	90.6%
LSTM-TF-at	95.9%	88.9%	87.9%	97.5%	92.6%	94.2%	92.8%

TABLE IV
RESULTS OF ENTERFACE

Models	Anger	Disgust	Fear	Happy	Sad	Surprise	UAR
LSTM	88.4%	64.3%	76.6%	83.8%	68.9%	73.9%	75.8%
LSTM-T	93.0%	81.0%	72.3%	91.9%	82.2%	84.8%	83.8%
LSTM-F	95.4%	76.2%	78.7%	97.3%	100%	78.3%	87.3%
LSTM-TF	88.4%	85.7%	80.9%	97.3%	86.7%	84.8%	86.9%
LSTM-at	81.4%	71.4%	76.6%	94.6%	77.8%	89.1%	81.5%
LSTM-TF-at	90.7%	97.6%	76.6%	97.3%	88.9%	89.1%	89.6%

TABLE V
RESULTS OF GEMEP

Models	Amu	Inq	Irr	Des	Joi	Peu	col	Int	Pla	Fie	Sou	Tri	UAR
LSTM	57.1%	36.4%	34.6%	30.8%	35.0%	50.0%	87.5%	46.7%	70.8%	46.7%	30.8%	46.7%	48.5%
LSTM-T	64.3%	9.1%	50.0%	23.1%	40.0%	77.8%	87.5%	46.7%	45.8%	40.0%	69.2%	73.3%	53.0%
LSTM-F	64.3%	18.2%	38.5%	38.5%	55.0%	61.1%	81.3%	53.3%	37.5%	40.0%	61.5%	60.0%	50.5%
LSTM-TF	71.4%	36.4%	46.2%	46.2%	45.0%	55.6%	81.3%	60.0%	50.0%	33.3%	69.2%	60.0%	54.0%
LSTM-at	42.9%	27.3%	42.3%	46.2%	40.0%	50.0%	68.8%	53.3%	62.5%	46.7%	61.5%	66.7%	51.0%
LSTM-TF-at	64.3%	27.3%	57.7%	7.7%	65.0%	77.8%	93.8%	53.3%	37.5%	53.3%	53.9%	80.0%	57.0%

basically consistent with the feature analyses, i.e., the recognition rate of the sad emotion is highest, and those for the disgust, fear and surprise emotions are relatively lower. Compared with improvement on CASIA and GEMEP, the UAR of LSTM-TF increases obviously with 11.1% on eNTERFACE, indicating that the attention-weighted deep features emphasize key emotional information.

TABLE VI
P-VALUES OF LEFT-TAILED T-TEST WITH 0.05 SIGNIFICANCE LEVEL

Models	CASIA	eNTERFACE	GEMEP
(LSTM,LSTM-TF)	9.26e-65	4.94e-308	1.81e-111
(LSTM, LSTM-at)	0.052	2.88e-118	1.89e-75
(LSTM, LSTM-TF-at)	1.30e-120	0	0

B. Attention Gate

To verify that the modified forget-gate based on attention-mechanism can effectively reduce training time under the premise of ensuring system performance, contrast experiments are performed on two groups of experiments in this paper. One experiment is between the LSTM-at and the traditional LSTM model, and the other is between the LSTM-TF model and LSTM-TF-at. As shown in Fig. 5 which depicts the training time under the same training steps on the corresponding corpus. These four models are trained on the CASIA with 1200 epochs, eNTERFACE with 1000 epochs and GEMEP with 1500 epochs. In other words, the models performed the same iterations on the same database. As seen from the figures, the training time of each model with the same number of training steps was different. The time cost of the LSTM model based on the attention-gate is less than for the model that is not modified. Comparing the training time on these corpora, CASIA required more time, and the training time differences between LSTM-at and LSTM (3.5h), LSTM-TF-at and LSTM-TF (1h) are larger than that on eNTERFACE (0.8h and 0.9h) and GEMEP (0.7h on both experiments). This indicates that longer training time is correlated with a more prominent advantage of LSTM based on attention-gates.

In terms of computational complexity, the GRU has less training time than the proposed attention-based forgetting gate (lstm-at), as shown in Fig. 5. However, the lstm-at achieves better performance than GRU on CASIA corpus that has more samples than eNTERFACE and GEMEP. Although GRU has low computational complexity, its performance is not good in large data sets. The similar conclusion was drown by Britz [56]. In his work, LSTM cells consistently outperformed GRU cells. Therefore, the lstm-at reduces computational complexity without sacrificing performance.

In terms of updating cell states, Tao [40] applied the attention mechanism to update cell states of LSTM, who focused on the information between cells. However, this study pays attention

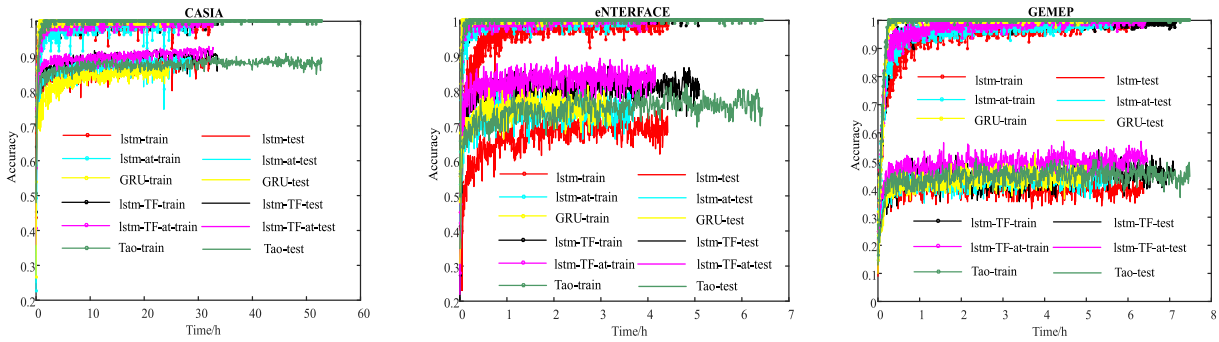

Fig. 5. Training time of models with the same steps.

TABLE VII
FEATURES AND MODELS ON THE CASIA

Models	Anger	Fear	Happy	Neutral	Sad	Surprise	UAR
ComPARE+SVM	87.4%	77.9%	84.1%	95.8%	88.5%	90.4%	87.4%
LSTM-TF-at	95.9%	88.9%	87.9%	97.5%	92.6%	94.2%	92.8%

TABLE VIII
FEATURES AND MODELS ON THE ENTERFACE

Models	Anger	Disgust	Fear	Happy	Sad	Surprise	UAR
ComPARE+SVM	74.4%	38.1%	34.0%	64.9%	80.0%	45.7%	55.8%
LSTM-TF-at	90.7%	97.6%	76.6%	97.3%	88.9%	89.1%	89.6%

to the inside of cell and modifies the forgetting gate of LSTM. Therefore, the computation of the forgetting gate is different from that of the previous LSTM and that in [40]. On all mentioned corpus, the performance achieved by lstm-at is similar to that of Tao. However, the later needs more time to train because of calculating more previous cell states, as shown in Fig. 5. In this study, lstm-at focuses on the internal calculation of cell state and takes both computational complexity and performance into consideration.

To quantitatively analyze the LSTM model based on the attention-gate in terms of identifying performance, the best recognition performance of each model was analyzed, as shown in Table III, Table IV and Table V. The LSTM based on the attention-gate reduced the matrix operations inside the model, and there is no negative impact on the UAR of all databases, in fact, even improved performance was sometimes observed. Compared to the baseline of traditional LSTM, the UAR of LSTM-at was improved by approximately 0.6% 5.7% and 2.5% on the CASIA, eNTERFACE and GEMEP corpus, respectively; compared with the LSTM-TF, the UAR of LSTM-TF-at was improved approximately by 0.8%, 2.7% and 3% on the CASIA, eNTERFACE and GEMEP corpus, respectively.

In summary, LSTM-TF-at enhances emotion-related information and significantly improves the UAR by introducing an attention mechanism into the time and feature dimension, as shown in Table VI that depicts the P-value with the left-tailed T-test between LSTM and the improved ones. However, on CASIA corpus, the improvement of LSTM-at is not significant with 0.052 P-value because of the high baseline. In addition, the forget-gate modified by the attention mechanism is designed to reduce the computational complexity of the model, accelerate the model convergence speed and shorten training time while ensuring the performance. The significance is even more pronounced by combining LSTM-TF and LSTM-at because of the smaller P-values obtained by LSTM-TF-at. Therefore, this model has obvious advantages in emotion classification.

C. Feature Comparison

Since the feature set used in this paper is modified on the basis of openSMILE ComParE features [16], the two feature sets – the original one and the modified one – were also compared. However, because the final openSMILE ComParE feature set after functional application to the LLDs is a one-dimensional feature vector, the internal data does not have a timing relationship and is not suitable as an input of the LSTM model presented for categorical emotion recognition. Therefore, the openSMILE ComParE functional feature set is combined with the traditional machine learning algorithm SVM to serve as a comparison baseline. The results are shown in Table VII, Table VIII and Table IX, which correspond to the CASIA, eNTERFACE and GEMEP databases, respectively.

As shown in the tables, the UAR obtained with the proposed method is improved by 5.4%, 33.8% and 17.0% on CASIA, eNTERFACE and GEMEP, respectively. Especially on CASIA and eNTERFACE, the recalls have increased for each category

TABLE IX
FEATURES AND MODELS ON THE GEMEP

Models	Amu	Inq	Irr	Des	Joi	Peu	col	Int	Pla	Fie	Sou	Tri	UAR
ComPARE+SVM	71.4%	18.2%	7.7%	38.5%	15.0%	50.0%	81.2%	40.0%	37.5%	20.0%	53.8%	73.3%	40.0%
LSTM-TF-at	64.3%	27.3%	57.7%	7.7%	65.0%	77.8%	93.8%	53.3%	37.5%	53.3%	53.9%	80.0%	57.0%

of emotion, which indicates the advantage of the LSTM-TF-at with the frame-level features.

VI. CONCLUSION

In this work, an improved attention-based LSTM is proposed for emotion classification. The attention mechanism is introduced into both the forget gate and the output of LSTM. The improved gate is only related to the historical cell state, and is independent of the current input, which could reduce the computational complexity. The experiments demonstrate that the new attention-gate can also improve the recognition rate. Moreover, due to the consideration of emotional saturation of different time segments and the ability of different features to distinguish emotions by applying attention mechanism into the time and feature dimension of LSTM's output, the proposed model (LSTM-TF-at) could achieve better performance than the others, especially on large data sets. On small data sets, LSTM-TF-at and Mirsamadi's model have similar UAR, but the latter has an advantage in algorithm complexity. In addition, compared with classical SVM classifier, the recalls increase for each category of emotion with LSTM-TF-at on both CASIA and eNTERFACE corpus.

Future work includes the following aspects. First, although the proposed method is effective on the classification task, it has much sense to the continuous emotion. Thus, the improved LSTM for continuous emotion recognition should be studied. Second, this algorithm takes the ability of different features to distinguish emotions into consideration. Hence, in our future research, it would be used for feature filtering. In addition, following our research, this attention-based LSTM is expected to be conducted in more applications.

REFERENCES

[1] R. Cowie *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] C. N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *Artif. Intell. Rev.*, vol. 43, no. 2, 2015.

[3] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[4] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Hoboken, NJ, USA: Wiley, 2013.

[5] B. Schuller *et al.*, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 119–131, Jul. /Dec. 2011.

[6] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker independent speech emotion recognition by ensemble classification," *Proc. IEEE Int. Conf. Multimedia Expo.*, 2005, pp. 864–867.

[7] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using fourier parameters," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 69–75, Jan.–Mar. 2015.

[8] X. Xu *et al.*, "A two-dimensional framework of multiple kernel subspace learning for recognizing emotion in speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1436–1449, Jul. 2017.

[9] A. álvarez *et al.*, "Feature subset selection based on evolutionary algorithms for automatic emotion recognition in spoken Spanish and standard basque language," in *Proc. Int. Conf. Text, Speech, Dialogue*, 2006, pp. 565–572.

[10] D. Morrison, R. Wang, and L. C. D. Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Commun.*, vol. 49, pp. 98–112, 2007.

[11] T. L. Pao, W. Y. Liao, Y. T. Chen, and J. H. Yeh, "Comparison of several classifiers for emotion recognition from noisy mandarin speech," in *Proc. Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, 2007, pp. 23–26.

[12] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Semi-supervised autoencoders for speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 31–43, Jan. 2018.

[13] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *18th Annu. Conf. Int. Speech Commun. Assoc., INTERSPEECH*, pp. 1263–1267, 2017.

[14] W. Martin *et al.*, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2008, pp. 597–600

[15] W. Martin, A. Metallinou, N. Katsamanis, B. Schuller, and S. Narayanan, "Analyzing the memory of BLSTM neural networks for enhanced emotion classification in dyadic spoken interactions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 4157–4160.

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

[17] B. Schuller *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 2794–2797.

[18] B. Schuller *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity and native language," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 2001–2005.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.

[20] K. Yao, T. Cohn, K. Vylomova, K. Duh, and C. Dyer, "Depth-gated recurrent neural networks," 2015, *arXiv:1508.03790*.

[21] T. N. Sainath and B. Li, "Modeling time-frequency patterns with LSTM vs. convolutional architectures for LVCSR tasks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 813–817.

[22] K. Greff, R. K. Srivastava, J. KoutnÍk, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[23] J. H. Yoo, "Large-scale video classification guided by batch normalized LSTM translator," 2017, *arXiv:1707.04045*.

[24] B. Vlasenko, B. Schuller, and A. Wendemuth, "Tendencies regarding the effect of emotional intensity in inter corpus phoneme-level speech emotion modelling," *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2016, pp. 1–6.

[25] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," *Comput. Vis. Pattern Recognit.*, vol. 40, pp. 842–850, 2014.

[26] F. Wang *et al.*, "Residual attention network for image classification," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6450–6458.

[27] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1245–1256, Jun. 2017.

[28] B. Athiwaratkun and J. W. Stokes, "Malware classification with LSTM and GRU language models and a character-level CNN," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 2482–2486.

[29] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," 2017, *arXiv:1708.02182*.

[30] W. Li and B. Mak, "Derivation of document vectors from adaptation of LSTM language model," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, Vol. 2, pp. 456–461.

[31] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden Markov models," in *Proc. 7th Eur. Conf. Speech Commun. Technol.*, 2001, pp. 2679–2682.

[32] G. Keren and B. Schuller, "Convolutional RNN: An enhanced model for extracting features from sequential data," in *Proc. Int. Joint Conf. Neural Netw.*, 2016, pp. 3412–3419.

[33] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* , 2003, vol. 2, pp. 1–4.

[34] Z. Inanoglu and R. Caneel, "Emotive alert: HMM-based emotion detection in voicemail messages," in *Proc. 10th Int. Conf. Intell. User Interfaces*, 2005, pp. 251–253.

[35] J. Wagner, T. Vogt, and Andre, "A Systematic Comparison of different HMM designs for Emotion Recognition from Acted and Spontaneous Speech," in *Affective Computing and Intelligent Interaction*, A. Paiva, R. Prada, and R. W. Picard, Eds., Berlin, Germany: Springer, 2007, pp. 114–125.

[36] T. Nwe, S. Foo, and L. D. Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, pp. 603–623, 2003.

[37] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[38] I. Sutskever, "Training recurrent neural networks," Ph.D. dissertation, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2013.

[39] 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.*, Mar. 2017, pp. 2227–2231.

[40] T. Fei, and G. Liu, "Advanced LSTM: A study about better time dependency modeling in emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 2906–2910.

[41] B. James *et al.*, "Quasi-recurrent neural networks," 2016, arXiv:1611.01576v2.

[42] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[43] T. Shen, J. Jiang, T. Zhou, S. Pan, G. Long, and C. Zhang, "Disan: Directional self-attention network for RnN/CNN-free language understanding," in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018, February 2, 2018 - February 7, 2018*, New Orleans, LA, USA, 2018, pp. 5446–5455.

[44] W. A. Jassim, R. Paramesran, and N. Harte, "Speech emotion classification using combined neurogram and INTERSPEECH 2010 paralinguistic challenge features," *IET Signal Process.*, vol. 11, no. 5, pp. 587–595, Jul. 2017.

[45] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted Boltzmann machines," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 5884–5887.

[46] E. Moore and M. Clements, "Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. I-101–104.

[47] J. T. F. L. M. Zhang and H. Jia, "Design of speech corpus for mandarin text to speech," in *Proc. Blizzard Challenge Workshop*, 2008, pp. 1–4.

[48] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE$'$05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops*, 2006, Art. no. 8.

[49] T. Bänziger, K. R. Scherer, "Introducing the Geneva multimodal emotion portrayal (GEMEP) corpus," *A Blueprint for Affective Computing A Sourcebook and Manual*, Oxford, U.K.: Oxford Univ. Press, 2010.

[50] F. A. Gers, J. Schmidhuber, "Recurrent nets that time and count[C]," *Proc. IEEE-INNS-ENNS Int. Joint Conf. Neural Netw., Neural Comput., New Challenges Perspectives New Millennium*, 2000, vol. 3, pp. 189–194.

[51] A. Graves, N. Jaitly, and A. R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2014, pp. 273–278.

[52] F. Eyben, F. Weninger, and B. Schuller, "Affect recognition in real-life acoustic conditions—A new perspective on feature selection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 2044–2048.

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

[54] Z. T. Liu, M. Wu, W. H. Cao, J. W. Mao, J. P. Xu, and G. Z. Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271–280, 2018.

[55] F. Eyben *et al.*, "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.–Jun. 2016.

[56] D. Britz, A. Goldie, M.-T. Luong, and Q. Le, "Massive exploration of neural machine translation architectures," 2017, *arXiv:1703.03906*.

Yue Xie is currently working toward a Doctoral degree at the school of information and communication engineering, Southeast University, Nanjing, China. His research interests include speech emotion recognition, deception detection, and deep learning.

Ruiyu Liang received the Ph.D. degree from Southeast University, Nanjing, China, in 2012. He is currently an Associate Professor with the Nanjing Institute of Technology, Nanjing, China. His research interests include speech signal processing and signal processing for hearing aids.

Zhenlin Liang is currently working toward a postgraduate degree at the school of information and communication engineering, Southeast University, Nanjing, China. His research interests include deception detection and machine learning.

Chengwei Huang received the Bachelor's and Ph.D. degrees from the Southeast University, Nanjing, China, in 2006 and 2013, respectively. He was an Associate Professor with Soochow University from 2013 to 2014. He started a robotics company as a partner and the General Manager in 2015 focusing on natural human-computer interaction. Since 2017, he has been a CTO with Big Data Technologies, Sugon (Nanjing) Institute of Chinese Academy of Sciences, Nanjing, China.

Cairong Zou received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the Southeast University, Nanjing, China in 1984, 1987, and 1991, respectively. He is currently a Professor with the Southeast University. In 1992, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada.

Björn Schuller received the Diploma in electrical engineering and information technology in 1999, the Doctoral degree in electrical engineering and information technology focusing on automatic speech and emotion recognition in 2006, and the Habilitation degree and the Adjunct Teaching Professorship in electrical engineering and information technology focusing on signal processing and machine intelligence in 2012 from the Technical University of Munich, Munich, Germany. He is currently a Full Professor of Artificial Intelligence, the Head of the Group on Language, Audio and Music, Imperial College London, London, U.K., a Full Professor with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, the co-founding CEO and the CSO of audEERING, Gilching, Germany, and an Associate with the Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland. He has authored/co-authored five books and more than 700 publications in peer-reviewed books, journals, and conference proceedings leading to more than 20 000 citations (h-index = 68). He is currently the President Emeritus of the Association for the Advancement of Affective Computing, an elected member of the IEEE Speech and Language Processing Technical Committee and a Senior Member of the ACM.