Improved Speech Emotion Recognition Using LAM and CTC

Lingyuan Meng[1], Zhe Sun[1], Yang Liu[1(✉)], Zhen Zhao[1], and Yongwei Li[2]

[1] School of Information Science and Technology, Qingdao Universicity of Science and Technology, Qingdao 266061, China
yangliu@qust.edu.cn

[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100089, China

Abstract. Time sequence based speech emotion recognition methods are difficult to distinguish between emotional and non-emotional frames of speech, and cannot calculate the amount of emotional information carried by emotional frames. In this paper, we propose a speech emotion recognition method using Local Attention Mechanism (LAM) and Connectionist Temporal Classification (CTC) to deal with these issues. First, we extract the Variational Gammatone Cepstral Coefficients (VGFCC) emotional feature from the speech as the input of LAM-CTC shared encoder. Second, CTC layer performs automatic hard alignment, which allows the network to have the largest activation value at the emotional key frame of the voice. LAM layer learns different degrees on the emotional auxiliary frame. Finally, BP neural network is used to integrate the decoding outputs of CTC layer and LAM layer to obtain emotion prediction results. Evaluation on IEMOCAP shows that the proposed model outperformed the state-of-the-art methods with a UAR of 68.5% and an WAR of 68.1% respectively.

Keywords: Speech emotion recognition · Attention · CTC · VGFCC · IEMOCAP

1 Introduction

As the fundamental research of affective computing, the study of speech emotion recognition has attracted great attention in recent years. Traditional machine learning approaches, such as Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Kernel Regression, Maximum Likelihood Classification (MLC), and Support Vector Machine (SVM) are widely adopted for emotion recognition using extracted features in previous works [1–3]. However, traditional speech emotion recognition methods are difficult to implement when encountering large training datasets.

Supported by The Natural Science Foundation of Shandong Province (No. ZR2020 QF007).

J. Xiong et al. (Eds.): MobiMedia 2021, LNICST 394, pp. 748–756, 2021.
https://doi.org/10.1007/978-3-030-89814-4_55

With the development of deep learning technology, deep neural networks (DNN) could not only easily handle large-scale training data, but also learn the deep level characteristics of speech emotion. For example, George trained a convolution recurrent neural network (CRNN) to perform continuous emotion prediction [4]. Zhang used spectrogram training based on full convolutional neural networks to convert the sequence transformation problem into image recognition problem [5]. Keren et al. used deep residual network (ResNet) to enhance speech to improve the performance of speech emotion recognition [6]. The intensity of speech emotions changes continuously over time, however, most DNN cannot distinguish between emotional and non-emotional frames in the speech, thus leading to performance degradation.

Graves et al. proposed the Connectionist temporal classification (CTC) algorithm [7], which can learn the error of the entire samples so that the network can automatically converge to the point where the emotional characteristics are most obvious [8,9]. However, the CTC algorithm only considers whether the current frame belongs to an emotional frame, while neglecting the difference of the amount of emotional information contained in each frame. The attention mechanism (AM [10] calculates the relative weights of the emotional features in the speech signal and each time domain, and selects the time domain signal with larger weight for recognition, so as to ensure that key information will not be lost [11]. Chen proposed a fusion model of AM and CTC in which the CTC module uses the emotional semantic coding sequence output by the decoder to calculate the loss [12], and the AM module encodes the emotional semantics output by the encoder. However, AM is global attention and does not consider the time order of the sequence.

In this paper, we propose a speech emotion recognition method based on Local Attention Mechanism (LAM) and CTC. First, we use SVM to extract the Variational Gammatone Cepstral Coefficients (VGFCC) features of the input speech, which are then used as the input of shared encoder; Second, the CTC layer performs back propagation using cross-entropy error for training and aligns the key emotional frames in the speech. Meanwhile, the LAM layer calculates the context relevance for the matching degree and the encoder output by attention mechanism algorithm, and different levels of information are extracted from different emotional frames for learning. Finally, the outputs of the CTC and LAM layers are integrated through BP neural network.

The main contributions of this paper include three aspects:

1) We apply local attention mechanism to reduce the interference of irrelevant speech frames on the current speech frame weight calculation;
2) By optimizing the decoder network structure and using supervised learning to merge the results of the CTC layer and LAM layer, the proposed model can fully learn the emotional features from speech;
3) Experimental results on benchmark dataset IEMOCAP demonstrate that our model gains an absolute improvement of 0.5%–0.8% over state-of-the-art strategies.

2 Proposed Method

The proposed framework as shown in Fig. 1 mainly consists: VGFCC feature extraction, CTC-Attention shared encoder, LAM layer, CTC layer, decoder and output layer.

2.1 VGFCC Feature Extraction

VGFCC considers the nonlinear and non-stationary characteristics of speech signals. First, set the maximum length of the voice *x(n)* to 7.5 s, cut the longer voice into 7.5 s, and fill the shorter voice with zeros. Second, perform pre-emphasis, framing, and windowing on *x(n)* (Hamming window) processing, and the sampling rate is set 16000 Hz. Third, *x(n)* is decomposed into K intrinsic modal function (IMF) components and all IMF components are subjected to fast Fourier transform (FFT) to obtain the spectral amplitude. Next, all the spectrum amplitudes are modulus squared and summed to obtain the energy spectrum of the signal and the energy spectrum is filtered by the Gammatone filter. Finally, the discrete cosine transform is performed on the filtered result to obtain the VGFCC coefficient.

2.2 Encoding Module

CTC-Attention Shared Encoder: CTC layer and LAM layer share the same encoder which is a double hidden layer LSTM, since LSTM can selectively affect the state of the neural network at each moment through a special gate structure. The output of each LSTM unit is used as the input of the CTC layer and LAM layer.

LAM Layer: Following the Encoding model, a structured LAM network aggregates information from the LSTM hidden states H^{lstm} and produces a fixed-length vector **Conv** as the encoding of the emotion feature.

Given hidden states H^{lstm} as input, first determine the number of speech frames T, and then determine the size of the optional range K.

When T is less than K, the model will use the global attention mechanism. The network computes a vector of attentional weights c and the weighted sum of the hidden states h^{attn} as follows:

$$\mathrm{c} = \mathrm{softmax}\left(w_{s2}\tanh\left(W_{sl}{H^{lstm}}^{T}\right)\right) \tag{1}$$

$$h^{\mathrm{attn}} = \sum_{i=1}^{T} cH^{lstm}, \tag{2}$$

where W_{s1} and w_{s2} are trainable parameters.

When T is larger than K, the model uses a sliding window to limit the calculation range of the vector of attentional weights, and slides the window according to the current calculated position. As shown in Fig. 2, X_n is the hidden

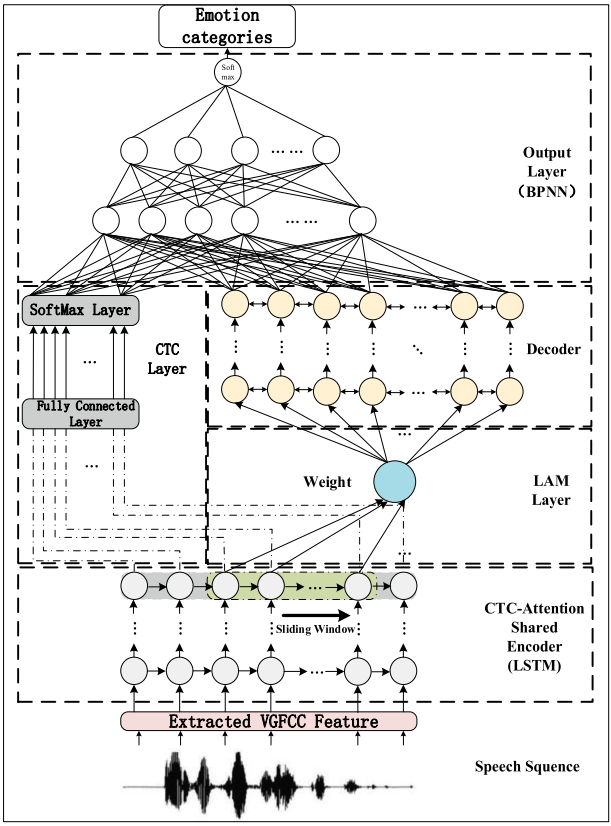

Fig. 1. Overall structure of the proposed framework.

layer output of the encoder neuron, and the current calculation position u is in the middle of the sliding window. If the lower limit of the sliding window is less than position 1 of X_1, start from position 1. If the upper limit of the sliding window is greater than the last position of the sequence, it ends with the last position. The length of the sliding window D is set to $K/2$. Note that the calculation formula of the weight vector c_{slide} and the hidden layer state weighted sum h^{attn} is:

$$c_{\text{slide}} = \text{softmax}\left(w_{s2}\tanh\left(W_{sl}H^{lstm^K}\right)\right) \tag{3}$$

$$h^{\text{attn}} = \sum_{i=u-D}^{u+D} c_{\text{slide}}\, H^{lstm} \tag{4}$$

CTC Layer: Define the intermediate label sequence $\pi = (\pi_1, \ldots \pi_T)$. Assumingis y' the expansion after adding the separator for y. Define a many-to-one mapping

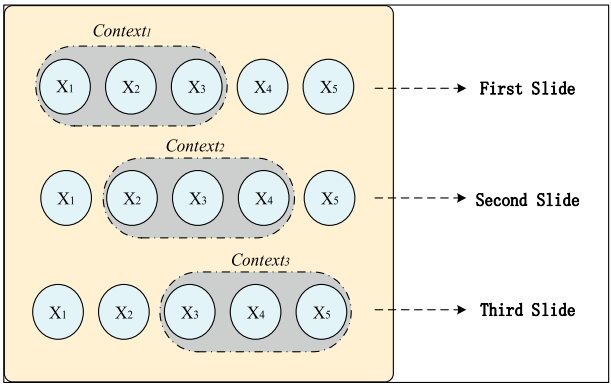

Fig. 2. The procedure of LAM.

as follow:

$$B : L' \rightarrow L^{\leq T} \tag{5}$$

where $L^{\leq T}$ is the output set of π of possible intermediate label sequences to obtain the probability of output label $P(y \mid x)$:

$$P(y \mid x) = \sum_{\pi \in B^{-1}(y')} p(\pi \mid x) \tag{6}$$

At each time step t of the input sequence x, the corresponding output pi_t must be calculated. Assuming that the output sequence between each time step is conditional and independent, $P(y \mid x)$ is the probability of a single label in the intermediate label sequence. (7) can be obtained as follow:

$$P(\pi \mid x) \approx \prod_{t=1}^{T} P\left(\pi_1 \mid \pi_1, \pi_2, \ldots, \pi_{t-1}, x\right), \forall \pi \in L' \tag{7}$$

The value of the negative log probability of the CTC loss function can be defined as follow:

$$L_{CTC}(S) = -\sum_{(x,y') \in S} \ln \sum_{\pi \in B^{-1}(y')} \prod_{t=1}^{T} q_t\left(\pi_t\right), \forall \pi \in L' \tag{8}$$

The CTC algorithm uses the HMM forward and backward algorithm to improve the calculation speed:

$$p_{CTC}(y \mid x) = \sum_{t=1}^{T} \sum_{u=1}^{|y'|} \frac{\alpha_t(u)\beta_t(u)}{q_t\left(\pi_t\right)} \tag{9}$$

where $\alpha_t(u)$ is the forward probability of the u-th label at time step t, and $\beta_t(u)$ is the backward probability of the u-th label at time step t.

2.3 Decoder

LSTM can only process the speech sequence in order, while the BLSTM could process the global speech sequence. Therefore, we select the BLSTM with double hidden layers as the decoder unit.

2.4 Output Layer

The output sequence of CTC and LAM is fused through double layers BP neural network. The gradient descent method is used for learning until the mean square error value reaches the threshold or the maximum number of iterations.

3 Experiment and Analysis

3.1 Database Description

We perform experiments on the public database, Interactive Emotional Dyadic Motion Capture (IEMOCAP) released by the School of Engineering of the University of Southern California. We use 12 h of IEMOCAP audio data which are divided into short sentences and labeled by three experts for discrete emotion categories. The labeled categories are "Angry", "Excited", "Happy", "Neutral" and "Sad". Since the emotions of "Excited" and "Happy" are very similar, both data are merged into the "Happy" category, and then these speech data are used as the emotion dataset for experiment which include 5531 utterances: "Happy" (1636), "Angry" (1103), "Sad" (1084), and "Neutral" (1708).

3.2 Experimental Setup

The experiment used Keras as the development framework and was completed on GeForce1080-Ti graphics card and Ubuntu 18.04 LST system. We convert the speech signal to 16 kHz and use 16bit to quantize the speech signal. We choose cross entropy as the cost function, Adadelta as the optimizer, and relu as the activation. The number of neurons selected for the encoder and decoder is 256, the learning rate is 0.01, the number of training cycles is 500, and the batch size model adjustment parameter is 64. 5531 pieces of speech emotion data are randomly divided into 80% training set and 20% testing set. The data set has a total of 4 emotion categories, and the proportion in the entire training/test set is still the same as the entire corpus, and the parameters are shown in the numbers above.

3.3 Comparative Experiments

In order to quantitatively evaluate the performance of the proposed model, the classification results of comparative experiments are provided in Table 1. The average training time, weighted average recall (WAR) and unweighted average recall (UAR) of the model are used as evaluation criterions.

Table 1. Results of comparative experiments. Note: Bold front denotes the best performance.

Model	UAR [%]	WAR [%]	Average training time [min]
LSTM	58.2	57.1	**76**
LSTM-CTC	63.7	62.9	109
LSTM-selfattention	66.1	65.8	114
AttRNN-RNN	67.6	67.5	105
LAM-CTC	**68.5**	**68.1**	85

We selected four classic models for comparison, including LSTM model [11], LSTM-selfattention model [11], LSTM-CTC model[12] and AttRNN-RNN model [12].

As shown in Table 1, the proposed model LAM-CTC is superior to other classic models in terms of UAR and WAR. Compared with the AttRNN-RNN model which performs best among the four models, the UAR and WAR of the proposed model increase by 0.9% and 0.6%, respectively. LAM effectively improves the tightness of the emotional context of certain long sentences, and resolve the problem that the global attention cannot perform emotional weight calculation for local speech, thereby ignoring the change of local emotions over time. In addition, when using BLSTM as the unit of the decoder, the current calculation vector contains more post information compared to only using RNN.

The average training time of proposed model is 85 min which only costs 78% of that of the LSTM-selfattention model and 84.7% of that of the AttRNN-RNN model. Since the global attention mechanism needs to calculate the weight of each frame at once when calculating the sample weight, LAM only needs to calculate the weight of the frame in the current window which slides backward overtime. At the same time, LAM calculates fewer parameters which could effectively reduce training time.

3.4 Ablation Experiments

The effectiveness of the LAM and CTC modules was verified through ablation experiments. The settings of the ablation experiments were as follows: (1) Replace LAM with global attention to verify the contribution of LAM to the performance (Attention-CTC). (2) Remove the CTC from the LAM-CTC model to verify the contribution of CTC to the performance (LAM). The comparison results of the ablation experiments are shown in Table 2.

As is shown in Table 2, the UAR and WAR of proposed model increase by 0.7% and 1.2% respectively compared with Attention-CTC because LAM could improve the tightness of the context of speech emotion. Compared with LAM, the UAR and WAR of the proposed model increase by 2.2% and 5%, respectively. By fusing the CTC module which could fully learn the key emotional frames of the speech, the proposed model can learn more about the emotional features in the

Table 2. Results of ablation experiments. Note: Bold front denotes the best performance.

Model	UAR [%]	WAR [%]	Average training time [min]
Attention-CTC	67.8	66.9	118
LAM	66.3	63.1	**67**
LAM-CTC	**68.5**	**68.1**	85

emotional speech. The training time of LAM-CTC only consumes 73.5% of that of the Attention-CTC model, while the LAM model only takes 67 min. Because LAM only needs to calculate the weight of the frame in the current window when calculating the context vector weight, and the attention window slides backward overtime. At the same time, LAM calculates fewer parameters, so using LAM can reduce training time. The results of ablation experiments show that both LAM and CTC could improve the performance of speech emotion recognition.

4 Conclusion

In this paper, we propose a deep learning model that integrates LAM and CTC for speech emotion recognition. Our method uses LAM to match the attention model according to the length of the speech sequence, and changes the calculation range of the context vector on the long sentence for modeling. CTC is utilized to perform alignment, which allows the network to have the largest activation value at the emotional key frame of the voice. The ablation experiment shows the effectiveness of reducing the interference of irrelevant speech frames on the current weight calculation. Compared with state-of-the-art models, the UAR and WAR achieves 68.5% and 68.1% with absolute increments more than 0.9% and 0.6%.

Acknowledgement. This work is supported by the Natural Science Foundation of Shandong Province (No. ZR2020QF007).

References

1. Schuller, B., Rigoll G., Lang, M.: Hidden Markov model-based speech emotion recognition. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP 2003), pp. II–1 (2003). https://doi.org/10.1109/ICASSP.2003.1202279
2. Dong, F., Zhang, G., Huang, Y., Liu, H.: Speech emotion recognition based on multi-output GMM and SVM. In: 2010 Chinese Conference on Pattern Recognition (CCPR), pp. 1–4 (2010). https://doi.org/10.1109/CCPR.2010.5659255
3. Caihua, C.: Research on multi-modal mandarin speech emotion recognition based on SVM. In: 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), pp. 173–176 (2019). https://doi.org/10.1109/ICPICS47731.2019.8942545

4. Trigeorgis, G., et al.: Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5200–5204 (2016). https://doi.org/10.1109/ICASSP.2016.7472669
5. Zhang, Y., Du, J., Wang, Z., Zhang, J., Tu, Y.: Attention based fully convolutional network for speech emotion recognition. In: 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1771–1775 (2018). https://doi.org/10.23919/APSIPA.2018.8659587
6. Ariav, I., Cohen, I.: An end-to-end multimodal voice activity detection using WaveNet encoder and residual networks. IEEE J. Sel. Top. Signal Process. **13**(2), 265–274 (2019). https://doi.org/10.1109/JSTSP.2019.2901195
7. Graves, A., Metze, F.: A first attempt at polyphonic sound event detection using connectionist temporal classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2986–2990 (2017). https://doi.org/10.1109/ICASSP.2017.7952704
8. Miao, Y., Gowayyed, M., Metze, F.: EESEN: end-to-end speech recognition using deep RNN models and WFST-based decoding. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 167–174 (2015). https://doi.org/10.1109/ASRU.2015.7404790
9. Shan, C., et al.: Investigating end-to-end speech recognition for Mandarin-English code-switching. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6056–6060 (2019) https://doi.org/10.1109/ICASSP.2019.8682850
10. Su, J., Zeng, J., Xie, J., Wen, H., Yin, Y., Liu, Y.: Exploring discriminative word-level domain contexts for multi-domain neural machine translation. IEEE Trans. Pattern Anal. Mach. Intell. **43**(5), 1530–1545 (2021). https://doi.org/10.1109/TPAMI.2019.2954406
11. Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., Schuller, B.: Speech emotion classification using attention-based LSTM. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(11), 1675–1685 (2019). https://doi.org/10.1109/TASLP.2019.2925934
12. Chen, X.: Research on Speech Emotion Recognition Method Based on Time Series Deep Learning Model. Harbin Institute of Technology (2018)