Speech Emotion Recognition Based on Recurrent Neural Networks with Conformer for Emotional Speech Synthesis

Xin Huang[1], Chenjing Sun[1], Jichen Yang[2(✉)], and Xianhua Hou[1]

[1] School of Electronics and Information Engineering, South China Normal University, Foshan 528225, China
[2] School of Cyber Security, Guangdong Polytechnic Normal University, Guangzhou 510665, China
nisonyoung@163.com

Abstract. Speech emotion recognition is the basis of emotional speech synthesis, a good speech emotion recognition system can learn more emotional expressions in speech and help in the synthesis of emotional speech. However, there are a number of issues that make the speech emotion recognition task difficult, including background noise and the distinct speech features of each speaker. The widely recognized speech emotion recognition system ACRNN extracts local features from speech signals using CNN, and its attention mechanism concentrates on the emotional content of the speech data. However, because only a single attention module is used, it is unable to simultaneously attend to the information from distinct representation subspaces at different locations, nor is it able to acquire long-term global information. The paper proposes CoRNN, which applies Conformer to replace CNN and attention module, with the purpose of overcoming the shortcomings of ACRNN. The experimental results on IEMOCAP dataset demonstrate that the unweighted average recall of the proposed CoRNN can achieve 65.53%, which improves 0.79% comparing with ACRNN.

Keywords: Speech emotion recognition · Emotional speech synthesis · Conformer

1 Introduction

The development of text-to-speech (TTS) technology makes the machine synthesized voice no longer cold and can better imitate human speech [1]. However, the current synthesized voice is still insufficient in emotional expressiveness, and it is necessary to further improve the system's emotional speech synthesis ability. Speech emotion recognition (SER) is the basis of emotional speech synthesis, and

J. Cai et al. (Eds.): SPNCE 2023, LNICST 525, pp. 268–277, 2025.
https://doi.org/10.1007/978-3-031-73699-5_19

can serve it, specifically in: the final synthesis stage of emotional speech synthesis system needs to provide the emotional embeddings of the desired emotion, and an excellent SER system can extract the emotional embeddings that contains more emotionally relevant information; for the voice input into the emotional speech synthesis system, the SER system can be used to first identify its emotional category, so as to eliminate the emotion-related information. Therefore, this paper focuses on the study of SER.

SER is widely used in network teaching, smart home, emotion conversion, expressive speech synthesis and other fields, which has important research value [2]. In recent years, people have conducted in-depth research on SER using acoustic features and a variety of machine learning (ML) and deep learning (DL) models. DL provides a variety of models for SER research to more accurately extract emotional states from speech. Deep neural network (DNN) models are often utilized to develop representations from low-level audio characteristics [3]. SER research usually uses convolutional neural networks (CNNs) and recurrent neural networks (RNNs) based on long short-term memory (LSTM) to extract local information in speech sequences. In CNN-based SER instances such as [4], CNN is often used to obtain time-frequency information derived from spectral features, while in LSTM-based instances such as [5], LSTM is used to focus on extracting sequence correlations of speech time series.

Convolutional Recurrent Neural Network (CRNN) was firstly proposed on raw audio samples for SER [6]. Then at the base of CRNN and attention mechanism, Chen et al. proposed a combination of attention model and convolutional recurrent neural network (ACRNN) for SER [7]. Because of its good performance, it has become a popular SER method to date.

The ACRNN mainly consists of three modules: CNN, Bidirectional Long Short-Term Memory (BiLSTM) and attention module, in which, CNN is used to extract local feature, BiLSTM plays the role of capturing contextual information and attention module is used to focus on emotion part. Though ACRNN is still extensively utilized in many domains, including expressive speech synthesis [8]. It has two drawbacks, one is that it is unable to extract long-term global information because it only makes use of CNN to capture local feature, the other is that only one single attention module in ACRNN is unable to attend to input from different representation subspaces at different positions simultaneously.

In order to settle down the two issues in ACRNN, in this paper, a method of CoRNN is proposed by using Conformer [9] to modify ACRNN. Further speaking, CNN and attention modules in ACRNN are swapped out for conformer. The reasons are as follows:

- Conformer has the ability to capture both global and local features simultaneously. The reason behind this is that Conformer is mainly composed of Transformer [10] and CNN, wherein Transformer can capture global information while CNN can be used to capture local feature.
- Conformer's multi-head attention module allows it to concurrently attend to data from different representation subspaces at different positions. Therefore, the model is more capable of sequence modeling for the relative dependency between features at different positions.

– Two half-step feed-forward layers are used in the Conformer, and the nonlinear activation function is introduced, so the nonlinear fitting ability and performance ability of the network can be improved.
– Conformer can extract better representation for emotion recognition.
– Conformer is effective in dealing with long-distance dependencies, and can make up for the problem that LSTM cannot deal with long-term dependencies.

The remaining parts of the paper are as follows. Section 2 details the architecture of the BiLSTM and Conformer based CoRNN. Section 3 introduces the database and experimental setup. Section 4 reports the studies on IEMOCAP dataset. Finally, Sect. 5 concludes the paper.

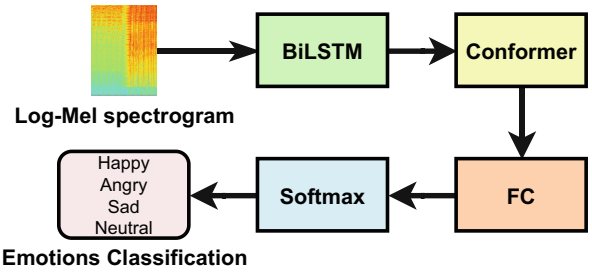

Fig. 1. Schematic diagram of CoRNN architecture for SER.

2 CoRNN

The structure of the CoRNN is shown in Fig. 1. From Fig. 1, it can be found that there are totally four modules in CoRNN, which are BiLSTM, Conformer, fully connected layer (FC) and softmax, in which the diagram of Conformer can be found in Fig. 2. The 3-D log-Mel spectrogram is fed into the BiLSTM network to capture the contextual features, and then the Conformer structure is connected to extract both local and global information of the sequence, and finally the classification is performed. Firstly, we introduce the role of each module briefly. BiLSTM is used to capture contextual information, Conformer is used to extract global and local features, FC plays the role of line transformation and Softmax is used to obtain the prediction probability of each emotion category of the input speech as the output. The emotion label corresponding to the dimension with the highest probability is the predicted emotion.

Next, the two principal modules in CoRNN, which are BiLSTM and Conformer, will be introduced in detail.

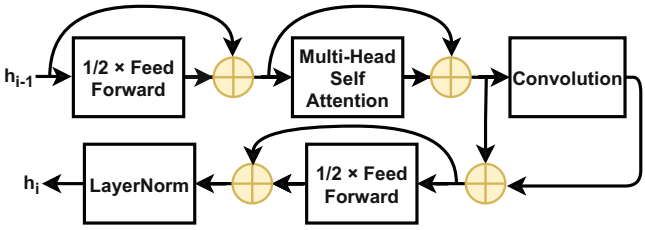

Fig. 2. The architecture of Conformer Block.

2.1 BiLSTM

RNNs introduce recurrent connections within layers, enabling parameters to be shared across time [11]. LSTM neural network is a variant of RNN. Relying on its unique mechanism, LSTM can process feature information with long interval distance, which is suitable for predicting temporal sequences. Therefore, it performs well in natural language processing and has been introduced into SER. In this paper, BiLSTM [12] is used. Compared with LSTM, BiLSTM combines the advantages of long short-term memory network and bidirectional recurrent neural network, which can better capture the temporal bidirectional context information of speech data, so it is more robust.

BiLSTM consists of two LSTM layers with opposite directions, which can simultaneously consider features from past and future timesteps, so it can capture the temporal bidirectional context information of speech data.

2.2 Conformer

When extracting the emotional representation of the speaker, the focus is on extracting the local and global features of the speech. Among them, global context modeling can increase the robustness of speaker feature extraction. CNN extracts features through local perception, so it can extract local features of speech well, but it is poor in capturing global features. While Transformer has good performance in capturing long-range global context dependencies, but has low local attention [13].

To extract emotional features more efficiently, we use the Conformer structure that can model both global and local features, as shown in Fig. 2. Transformer model is good at capturing global features but has low local attention, while CNN has a good ability to extract local features but is poor at capturing global features. Conformer is a combination of CNN and Transformer, which possesses the feature extraction capabilities of both, and thus can capture both local and global features of speech very well. The key components of the Conformer architecture include multi-head self-attention module (MHSA) and convolution module (Conv). The MHSA module can expand the ability of the model in sequence modeling of the relative dependency between features in different positions. Its relative position encoding module makes the model more robust to speech of different lengths [14].

The Conv module consists of Pointwise convolution layers, 1D Depthwise convolution layers, and also includes a BatchNorm layer to accelerate model convergence. The Conv module uses the local modeling ability of CNN to obtain the local features of sequences, which is the key to improve the performance of the model.

Different from the encoder of the Transformer model, the Conformer structure contains two feed-forward modules (FFN) with half-step residual connection, which are located before the MHSA module and after the Conv module. The addition of FFN introduces a nonlinear activation function that improves the nonlinear fitting ability of the network and improves the performance of the model. Such a structure can yield better results compared to a single FFN [15].

When the input feature h_{i-1} is fed into the Conformer architecture, the output feature h_i is generated as follows:

$$\widetilde{h_i} = h_{i-1} + \frac{1}{2}\mathrm{FNN}(h_{i-1}) \tag{1}$$

$$h'_i = \widetilde{h_i} + \mathrm{MHSA}(\widetilde{h_i}) \tag{2}$$

$$h''_i = h'_i + \mathrm{Conv}(h'_i) \tag{3}$$

$$h_i = \mathrm{LayerNorm}(h''_i + \frac{1}{2}\mathrm{FNN}(h''_i) \tag{4}$$

where $h_{i-1} \in R^{d\times T}$, $h_i \in R^{d\times T}$, d is the dimension of the Conformer, and T is the frame length.

3 Database and Experimental Setup

The suggested CoRNN is assessed using the Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [16]. The following four categories of emotions-happy, neutral, angry and sad-are taken from the database's improvisation version, which has 2280 utterances [7]. In our assessment, 10-fold Cross Validation is employed.

In addition, the openEAR toolkit [17] is used to extract traditional speech emotion feature log-Mel spectrogram [11] in this work. Then its delta and delta-delta features are computed based on the log-Mel. The log-Mel, delta, and delta-delta features are synthesized into a 3-D log-Mel set as input to minimize the effect of speaker differences. To facilitate batch training, the 3-D log-Mel spectrogram of every utterance with 3 s length by truncating or padding before entering the CoRNN. For the padding process in this study, we employ the zero-padding method.

Our work uses the Python platform to deploy experiments, and the network uses the Adam optimizer to optimize the classification cross entropy. In the network parameters, the batchsize for single training of the CoRNN model is

set to 40, the learning rate is set to 10^{-3}, and the overall dropout of the model is set to 0.2. The last model selected is the one that has been trained across 250 epochs. Owing to the non-uniform label distribution, Unweighted Average Recall (UAR)) [18] is employed as a performance statistic for CoRNN, helping to prevent the model from being overfit to a particular category.

4 Experimental Results and Analysis

The experimental findings (UAR(%)) on IEMOCAP dataset using CoRNN are displayed in Table 1. It can be seen that using the CoRNN model proposed in this paper, the UAR can achieve a result of 65.53%, which is a good recognition effect.

Table 1. Experimental results on IEMOCAP dataset using CoRNN in terms of UAR(%)

Models	UAR
CoRNN	65.53

4.1 Ablation Experiment of CoRNN

In order to better verify the effectiveness of the CoRNN, ablation experiments were carried out respectively:

- **CoRNN**: Extract the 3-D log-Mel feature of speech, and calculate its deltas, and delta-deltas to obtain a three-dimensional Mel spectrogram, which is input to the CoRNN for emotion classification.
- **w/o Conformer**: 3-D log-Mel spectrogram is only input into BiLSTM.
- **w/o BiLSTM**: 3-D log-Mel spectrogram is only input into Conformer.

Table 2 displays the outcomes of the experiment. The findings of the experiment show that the UAR value of the CoRNN model after deleting the Conformer network or BiLSTM network is smaller than the UAR value obtained by the original model, which means that the emotion recognition performance of the separate BiLSTM model and the separate Conformer model is not as good as the CoRNN model. An significant part of the CoRNN model recognition process involves the BiLSTM and Conformer networks.

Table 2. Ablation experiment results on IEMOCAP dataset using CoRNN in terms of UAR(%)

Models	UAR
CoRNN	65.53
w/o Conformer	63.68
w/o BiLSTM	59.76

4.2 Confusion Matrix Analysis

To compare and examine the experimental outcomes even more, confusion matrix is used here. Figure 3 shows the confusion matrix of the CoRNN.

By observing the confusion matrix of CoRNN's experimental results in the figure, it is discovered that the model significantly improves the ability to identify the emotions of anger and sad, but has a poor recognition effect on the emotions of happy and neutral, both of which are often recognized into each other. The reason may be that the small size of the happy category data, its feature acquisition is insufficient. The neutral category has not been able to capture its features well because of its own emotional factors are not prominent enough. These issues will be investigated in future work.

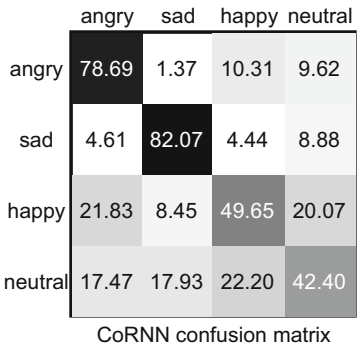

Fig. 3. Confusion matrix of CoRNN with the UAR of 65.53% on the IEMOCAP dataset

4.3 Comparison with the State-of-the-Art Systems

Here, we would like to compare the proposed CoRNN with other existing systems. To this end, Table 3 shows the comparison between CoRNN and other existing systems on the IEMOCAP dataset in terms of UAR. In which,

– **Raw Speech + CRNN** [19]: Taking raw speech as input, a parallel CNN is used to extract both long-term and short-term interactions from the raw

speech. The features that were captured are fed into a CNN and LSTM classification module. Convolutional layers pick up high-level information, while LSTM layers handle long-term temporal modeling.

– **3-D log-Mel features + ACRNN** [7]: The CRNN model is used to train high-level feature representations of speech segments, the attention model is employed to assess the value of a sequence of high-level representations to the resulting emotion representation, and the three-dimensional Mel spectrogram is obtained as the input.

Table 3. Comparison with the state-of-the-art systems on IEMOCAP dataset in terms of UAR(%)

Systems	Features	Models	UAR
1	Raw Speech	CRNN	60.23
2	3-D Log-Mel	ACRNN	64.74
Proposed	3-D Log-Mel	CoRNN	65.53

Our system is first contrasted with the CRNN system. Unlike the system in this paper, which uses hand-crafted acoustic features, the CRNN system uses CNN to extract features from raw speech, which are more general and contextual. However, the data size of the IEMOCAP dataset is too small to capture sufficiently accurate features. The experimental results show that the UAR value of the CoRNN system is increased by 5.3% compared with CRNN system. It reflects the effectiveness of the speech emotion recognition system proposed in this paper.

Second, our system is compared with the commonly used ACRNN systems, both of which take the 3-D log-Mel spectrum as input. Only CNN, BiLSTM and attention mechanism are used in the ACRNN system. The results show that the UAR of the CoRNN system is 0.79% higher than that of the ACRNN system, which indicates that the Conformer model has more advantages in the SER task than the ordinary attention model plus CNN.

To sum up, the proposed system based on CoRNN in this work outperforms other systems to a certain extent.

5 Conclusion

This paper modifies ACRNN to enhance SER's performance to better serve emotional speech synthesis. The CoRNN model is proposed that uses the Conformer module to replace the CNN and attention modules in the original model to improve the model's ability to extract global and local features. At the same time, the BiLSTM network is combined to achieve the goal of extracting more comprehensive emotional representation from various aspects. The effect of using

CoRNN for SER is better than that of using ACRNN, according to experimental results on the IEMOCAP dataset. In addition, the proposed system also outperforms some previous systems. Thus the proposed system can provide emotional embeddings containing more emotionally relevant information for emotional speech synthesis systems to improve the emotional expressiveness of synthesized speech.

Acknowledgments. The author gratefully acknowledges the support of NSFC (62001173, 62171188).

References

1. Lei, Y., Yang, S., Xie, L.: Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis. In: 2021 IEEE Spoken Language Technology Workshop (SLT), pp. 423–430 (2020)
2. Zhong, S., Yu, B., Zhang, H.: Exploration of an independent training framework for speech emotion recognition. IEEE Access **8**, 222533–222543 (2020)
3. Han, K., Yu, D., Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine. In: Interspeech 2014, pp. 223–227. Singapore, Malaysia (2014)
4. Zhang, S., Zhang, S.L., Huang, T., Gao, W.: Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. IEEE Trans. Multimedia **20**(6), 1576–1590 (2018)
5. Xie, Y., Liang, R., Liang, Z., et al.: Speech emotion classification using attention-based LSTM. IEEE/ACM Trans. Audio Speech Lang. Process. **27**, 1675–1685 (2019)
6. Trigeorgis, G., Ringeval, F., Brueckner, R., 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)
7. Chen, M., He, X., Yang, J., Zhang, H.: 3-D Convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Process. Lett. **25**(10), 1440–1444 (2018)
8. Liu, R., Sisman, B., Gao, G., Li, H.: Expressive TTS training with frame and style reconstruction loss. IEEE Trans. Audio Speech Lang. Process. **29**, 1806–1818 (2021)
9. Gulati, A., Qin, J., Chiu, C.C., et al.: Conformer: convolution-augmented transformer for speech recognition. In: Proceedings of Interspeech, pp. 5036–5040 (2020)
10. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (NIPS 2017), pp. 5998–6008 (2017)
11. Latif, S., Rana, R., Khalifa, S., et al.: Survey of deep representation learning for speech emotion recognition. IEEE Trans. Affect. Comput. **14**(2), 1634–1654 (2021)
12. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
13. Zhang, Y., et al.: MFA-conformer: multi-scale feature aggregation conformer for automatic speaker verification. In: Proceedings of Interspeech 2022, pp. 306–310 (2022)

14. Dai, Z., Yang, Z., Yang, Y., et al.: Transformer-XL: attentive language models beyond a fixed-length context. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2978–2988. Association for Computational Linguistics, Florence, Italy (2019)
15. Lu, Y., Li, Z., He, D., et al.: Understanding and improving transformer from a multi-particle dynamic system point of view, arXiv preprint arXiv:1906.02762 (2019)
16. Busso, C., Bulut, M., Lee, C.C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**, 335–359 (2008)
17. Eyben, F., Wöllmer, M., Schuller, B.: OpenEAR-introducing the Munich open-source emotion and affect recognition toolkit. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–6 (2009)
18. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun. **64**, 1062–1087 (2011)
19. Latif, S., Rana, R., Khalifa, S., et al.: Direct modelling of speech emotion from raw speech, arXiv preprint arXiv:1904.03833 (2019)