Recurrent Neural Networks With Conformer for Speech Emotion Recognition

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

[1]South China Normal University
[2]Guangdong Polytechnic Normal University

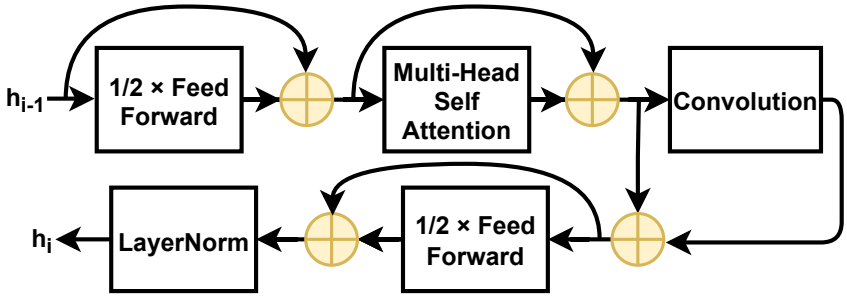
October 24, 2022

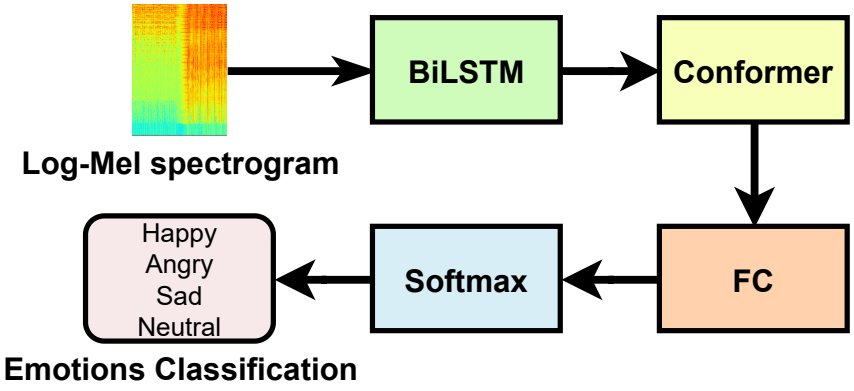
Abstract

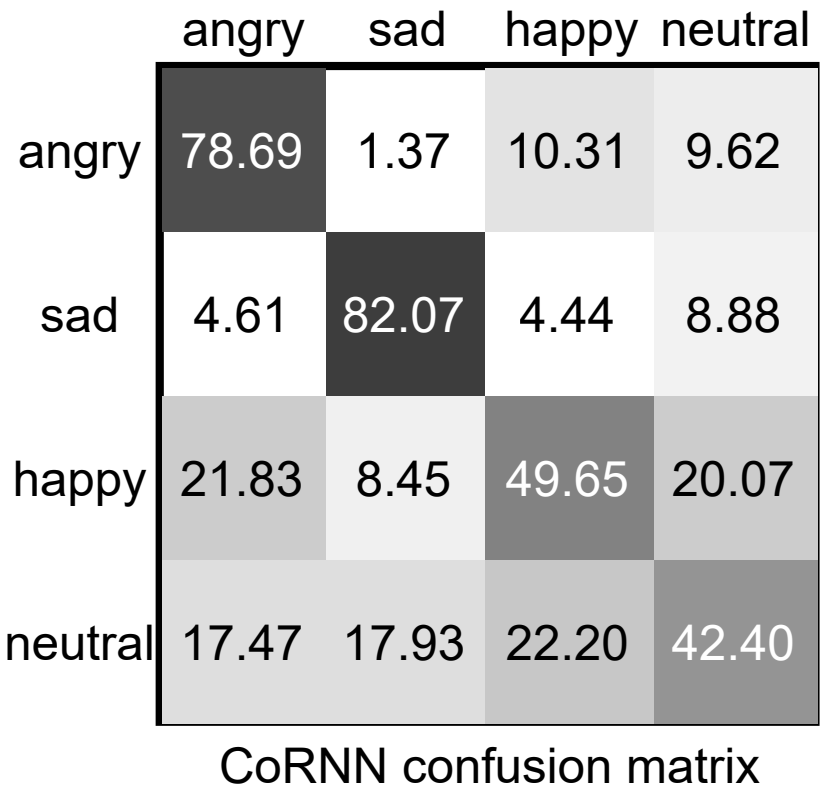
Speech emotion recognition plays an important role in many applications, but the task is challenging due to various factors such as background noise, different speaker speech characteristics, etc. The well known speech emotion recognition system ACRNN uses CNN to extract local features of speech signals and attention mechanism focuses on the parts with prominent emotions. However, it has no ability to capture long-term global information and it also has no ability to jointly attend to the information from different representation subspaces at different positions because only one single attention module is used. In order to settle out the drawbacks of ACRNN, CoRNN is proposed in this letter by applying Conformer to replace the modules of CNN and attention module. The experimental results on IEMOCAP dataset demonstrate the unweighted average recall of the proposed CoRNN can achieve 65.53%, which improves 0.79% comparing with ACRNN.

Hosted file

`ELECTRONICS LETTERS5.7z` available at https://authorea.com/users/516951/articles/591624-recurrent-neural-networks-with-conformer-for-speech-emotion-recognition







CoRNN confusion matrix

Recurrent Neural Networks With Conformer for Speech Emotion Recognition

Chenjing Sun, Jichen Yang, Xin Huang and Xianhua Hou

Speech emotion recognition plays an important role in many applications, but the task is challenging due to various factors such as background noise, different speaker speech characteristics, etc. The well known speech emotion recognition system ACRNN uses CNN to extract local features of speech signals and attention mechanism focuses on the parts with prominent emotions. However, it has no ability to capture long-term global information and it also has no ability to jointly attend to the information from different representation subspaces at different positions because only one single attention module is used. In order to settle out the drawbacks of ACRNN, CoRNN is proposed in this letter by applying Conformer to replace the modules of CNN and attention module. The experimental results on IEMOCAP dataset demonstrate the unweighted average recall of the proposed CoRNN can achieve 65.53%, which improves 0.79% comparing with ACRNN.

Introduction: Speech emotion recognition (SER) is widely used in network teaching, smart home, emotion conversion, expressive speech synthesis and other fields, which has important research value [1]. Convolutional Recurrent Neural Network (CRNN) was firstly proposed on raw audio samples for SER [2]. 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 [3]. 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 has been widely used in many fields such as expressive speech synthesis [4]. It has two drawbacks, one is that it has no ability to capture 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 has no ability to jointly attend to the information from different representation subspaces at different positions.

CoRNN: In order to settle down the two issues in ACRNN, in this letter, a method of CoRNN is proposed by using Conformer [5] to modify ACRNN. Further speaking, Conformer is used to replace the modules of CNN and attention in ACRNN. The reasons are as follows:

- Conformer has the ability to capture global and local feature at the same time. The reason behind this is that Conformer is mainly composed of Transformer [6] and CNN, wherein Transformer can capture global information while CNN can be used to capture local feature. The diagram of Conformer can be found in Fig. 1.
- Conformer is able to jointly attend to the information from different representation subspaces at different positions because it has multi-head attention module. 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 structure of the CoRNN is shown in Fig. 2. From Fig. 2, it can be found that there are totally four modules in CoRNN, which are BiLSTM, Conformer, fully connected layer (FC) and softmax. 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.

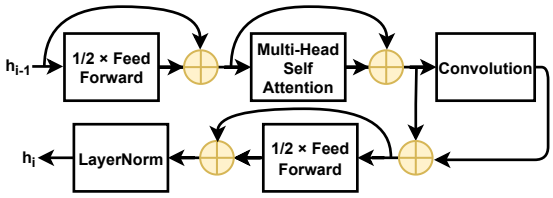


Fig. 1 *The architecture of Conformer Block.*

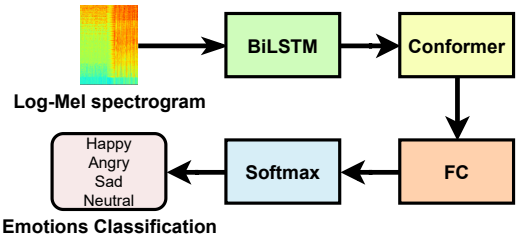


Fig. 2 *Schematic diagram of CoRNN architecture for SER.*

BiLSTM

Since BiLSTM is suitable for processing temporal sequences, it performs well in natural language processing and has been introduced into SER. In this paper, BiLSTM [7] is used. 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.

Conformer

When extracting the emotional representation of the speaker, the focus is on extracting the local and global features of the speech.

To extract sentiment features more efficiently, we use the Conformer structure that can model both global and local features, as shown in Fig. 2. Conformer is a combination of CNN and Transformer, so it can well capture the local and global features of speech. 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 [8].

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 (FNN) with half-step residual connection, which are located before the MHSA module and after the Conv module. Such a structure can yield better results compared to a single FFN [9].

Database and experimental setup: Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [10] is used to evaluate the proposed CoRNN. Following [3], four types of emotions, which are happy, neutral, angry and sad, are selected from the improvisation version of the database and there are 2280 utterances in it. 10-fold Cross Validation is used in our evaluation.

In addition, the openEAR toolkit [11] is used to extract traditional speech emotion feature log-Mel spectrogram as input [12] in this work. The log-Mel spectrogram of every utterance with 3s length by truncating or padding before entering the CoRNN.

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 model that has been trained for 250 epochs is saved as the final model. Due to the uneven distribution of labels, Unweighted Average Recall (UAR) [13] is used as metric to evaluate the performance of CoRNN, which can avoid overfitting of the model to a certain category.

Experimental results and analysis: Table 1 shows the experimental results (UAR(%)) using CoRNN on IEMOCAP dataset. 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

Confusion matrix analysis: In order to further compare and analyze the experimental results, confusion matrix is used here. Fig. 3 shows the confusion matrix of the CoRNN.

By observing the confusion matrix of CoRNN's experimental results in the figure, it can be found that the model has a good effect on the recognition for the emotions of angry 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.

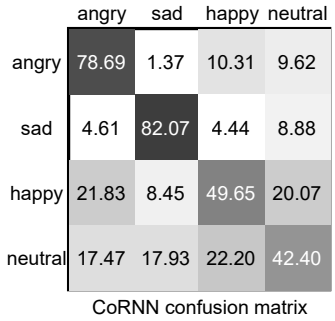


Fig. 3 *Confusion matrix of CoRNN with the UAR of 65.53% on the IEMOCAP dataset.*

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 2 shows the comparison between CoRNN and other existing systems on the IEMOCAP dataset in terms of UAR. In which,

- **Raw Speech + CRNN** [14]: Taking raw speech as input, a parallel CNN is used to capture long-term and short-term interactions from the raw speech. The extracted features are input into a classification module composed of CNN and LSTM, high-level features are captured by convolutional layers, and long-term temporal modeling is performed by LSTM layers.
- **3-D log-Mel features + ACRNN** [3]: The three-dimensional Mel spectrogram is extracted as the input to CRNN, the CRNN model is used to learn high-level feature representations of speech segments, and the attention model is used to score the importance of a series of high-level representations to the final emotion representation.

Table 2: 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

First, we compare our system with the CRNN system. Unlike the system in this letter, 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.

Conclusion: In order to improve the performance of speech emotion recognition, this letter modifies ACRNN. The CoRNN model is proposed that uses the Conformer module to replace the CNN and attention modules in the original model. At the same time, the BiLSTM network is combined to achieve the goal of extracting more comprehensive emotional representation from various aspects. The experimental results on the IEMOCAP dataset show that the effect of using CoRNN for speech emotion recognition is better than that of using ACRNN. In addition, our system also outperforms some previous systems.

Acknowledgment: This work was supported by NSFC(62001173, 62171188). The author gratefully acknowledges the support of 2022 Guangdong Hong Kong-Macao Greater Bay Area Exchange Programs of South China Normal University (SCNU).

Chenjing Sun, Xin Huang and Xianhua Hou are with the School of Electronics and Information Engineering, South China Normal University, Foshan, China and SCNU Qingyuan Institute of Science and Technology Innovation Co., Ltd., Qingyuan 511517, China.
Jichen Yang is with the School of Cyberspace Security, Guangdong Polytechnic Normal University, Guangzhou, China.

E-mail: (Jichen Yang) nisonyoung@163.com, (Xin Huang) huangxin@m.scnu.edu.cn.

References

1 S. Zhong, B. Yu, and H. Zhang.: 'Exploration of an Independent Training Framework for Speech Emotion Recognition', *IEEE Access*, 2020, **8**, pp. 222533-222543
2 G. Trigeorgis, F. Ringeval, R. Brueckner, et al.: 'Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network', *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2016, pp. 5200-5204
3 M. Chen, X. He, J. Yang, and H. Zhang.: '3-D Convolutional recurrent neural networks with attention model for speech emotion recognition', *IEEE Signal Processing Letters*, 2018, **25**(10), pp. 1440-1444
4 R. Liu, B. Sisman, G. Gao, and H. Li.: 'Expressive TTS training with frame and style reconstruction loss', *IEEE Transactions on Audio, Speech and Language Processing*, 2021, **29**, pp. 1806-1818
5 A. Gulati, J. Qin, C. C. Chiu, et al.: 'Conformer: Convolution-Augmented transformer for speech recognition', *Proc. Interspeech*, 2020, pp. 5036-5040
6 A. Vaswani, N. Shazeer, N. Parmar, et al.: 'Attention is all you need', *Advances in Neural Information Processing Systems (NIPS 2017)*, 2017, pp. 5998-6008
7 A. Graves, J. Schmidhuber.: 'Framewise phoneme classification with bidirectional LSTM and other neural network architectures', *Neural networks*, 2005, **18**(5-6), pp. 602-610
8 Z. Dai, Z. Yang, Y. Yang, et al.: 'Transformer-XL: Attentive language models beyond a fixed-length context', *arXiv preprint arXiv:1901.02860*, 2019
9 Y. Lu, Z. Li, D. He, et al.: 'Understanding and improving transformer from a multi-particle dynamic system point of view', *arXiv preprint arXiv:1906.02762*, 2019
10 C. Busso, M. Bulut, C. C. Lee, et al.: 'IEMOCAP: Interactive emotional dyadic motion capture database', *Language resources and evaluation*, 2008, **42**, pp. 335-359
11 F. Eyben, M. Wöllmer, and B. Schuller.: 'OpenEAR-Introducing the Munich open-source emotion and affect recognition toolkit', *2009 3rd international conference on affective computing and intelligent interaction and workshops*, 2009, pp. 1-6
12 S. Latif, R. Rana, S. Khalifa, et al.: 'Survey of deep representation learning for speech emotion recognition', *IEEE Transactions on Affective Computing*, 2021, pp. 1-1
13 B. Schuller, A. Batliner, S. Steidl, and D. Seppi.: 'Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge', *Speech communication*, 2011, **64**, pp. 1062-1087
14 S. Latif, R. Rana, S. Khalifa, et al.: 'Direct modelling of speech emotion from raw speech', *arXiv preprint arXiv:1904.03833*, 2019