Title: **End-to-End Speech Emotion Recognition Using Deep Neural Networks (ICASSP 2018)**

Author: Panagiotis Tzirakis, Jiehao Zhang, Bjorn W. Schuller

Summary: This paper is focused on continues emotion recognition from speech. Different from previous Long-Short-Term Memory architectures, the authors present a new a method combining CNN and RNN. The proposed framework first processes raw speech signals with tree convolutional layers, then stacked with two layers of LSTM on the output. After that, the 16 Hz input data is processed to 25 Hz labels. Experiments among different methods prove the effectiveness of the proposed method.

Good Things about This Paper

1. This is a new hybrid CRNN architecture, where successive convolutional layers learn features at different levels and a LSTM network learns the temporal relationships between those features.
2. The authors also mentioned the methodology on how the pooling size should be selected on the kernel size of the convolutional layer in part 3, which is considered to be ignored by the former researchers.
3. Use concordance correlation coefficient (CCC) as loss function instead of mean square error (MSE) in the network training process.

Major Comments

The proposed method is considered to be better because it creatively combined convolutional layers and recurrent layers, taking the advantages of both architectures to automatically extract features from raw data and utilize the context information among different fragments at the same time. However, as there are some defects in the evaluation stages, the good performance of the described models may not be so promising as the author described.

First, it is better to distinguish the effects of introducing CNN and RNN separately, for readers to understand which part contributes to which kind of performance improvement. For example, as the authors mentioned in section 5.3 and Fig.2, the performance on the valence dimension is not as good as on the arousal. There is no further discussion of this issue, but it is important because valence is associated with the acoustic features learned by the CNN part. One possible cause for the underperformance may be the CNN is not as useful as it is thought to be. More discussions about this among different models may be added to the article.

Another improvement may be specifying the significance of this method. The author mentioned in the last but one paragraph that their creation of network architecture is inspired by the way convolutional speech features are computed. However, as one of the baseline models, [1] has a very similar CRNN architecture as the proposed model, with only the hyperparameters differ.

[1] G. Trigeorgis *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)*, Shanghai, 2016, pp. 5200-5204.

Minor Comments

1. Equation (1) should be ended with full stop('.') instead of comma(',').
2. The performance comparison tables are not described in detail, e.g. no units.