Performance Comparison of Raw Speech Signals and Handcrafted Features in Deep Learning-Based Emotion Recognition Models

Samuel Kakuba
Graduate School of Electronics and Electrical Engineering
Kyungpook National University
Daegu, Republic of Korea
2021327392@knu.ac.kr

Dong Seog Han
School of Electronics and Electrical Engineering
Kyungpook National University
Daegu, Republic of Korea
dshan@knu.ac.kr

***Abstract*—The advent of deep learning has seen its application in end-to-end speech emotion recognition systems that are later used in real life. These deep learning systems use either raw speech signals or handcrafted features as input. Commendable performance in terms of accuracy and sometimes F1 score has been reported in SER studies that use these features without the analysis of the robustness of these models in terms of the individual emotion class confusion ratio. In this paper, we carried out comparative experiments to ascertain the robustness of deep learning-based models that use raw signals, separate handcrafted speech features and in combination. In the experiments in which we used speech features, we extracted spectral and voice quality features from the raw speech signals as input to the model. We found out that though the accuracy and F1 score in all experiments are commendable, the robustness in terms of confusion ratio is worse with raw signals and best with a combination of handcrafted features.**

***Index Terms*—emotion recognition, raw speech signals, speech features**

I. Introduction

Emotion recognition (ER) is an affective computing domain that involves the inference of one's emotional state observed from changes in speech, face, gestures and physiological activities in the body. The advent of deep learning has seen the use of this domain in a number of applications like social living assistance robots, health monitoring, authentication systems, and interactive robots. Speech emotion recognition involves inference of emotions from the speech signal. Because of their benefits, end to end deep learning systems have been deployed in speech emotion recognition using raw signals [1], [2] and [3].

The works in [4] and [5] employed end-to-end deep learning systems on raw signals using one-dimensional convolutions as well as mel spectrograms using two-dimensional convolution layers in combination with long short-term memory (LSTM) and convolutional LSTM (convLSTM) respectively. They suggested the use of local feature learning blocks (LFLB) at the lower layers for local feature extraction and global feature learning blocks (GFLB) at the deeper layers for global feature extraction. These applications of SER can be deployed in ubiquitous computing devices that can be used anytime anywhere.

The authors in [1], [2] and [3] proposed models that use raw signals as input to end-to-end deep learning-based models to learn emotional cues and their relationships for speech emotion recognition (SER). These models compute long-term dependencies and the global context and relationships between the features using attention mechanisms [8], [9] and [10]. Though these models obtain commendable accuracies they are not robust in discrimination of high arousal emotions especially happy and angry. Commendable results in SER research have also been achieved by deep learning-based models that use handcrafted features like mel frequency cepstral coefficients (MFCCs) and mel spectrograms. In [11] a convolutional neural network with attention for speech emotion recognition using MFCCs was implemented. Mel spectrograms were used as inputs to the two-dimensional model suggested in [5]. In [12], a combination of MFCCs, mel spectrograms and chroma grams were used in a deep learning-based model for SER. However, though commendable accuracies have been reported, there has been no specific evaluation of the confusion error (ratio) of these models for the different categories of emotions they classify which would give researchers a direction for the improvement of the robustness of the SER models especially in real-life scenarios.

In this paper, we use a simple deep learning-based model to evaluate the robustness of raw signals, spectral and voice quality features extracted from the raw speech signals for speech emotion recognition. We report on the robustness of the model in terms of the confusion ratio (error) of the different emotion classes and suggest a way forward.

The contribution of this paper is twofold;

- We use a simple speech emotion recognition model to evaluate the robustness of deep learning-based models that

use either raw signals, spectral and voice quality features or their combination.
- We also suggest possible ways in which the robustness can be improved in SER studies in order to be effectively used in real-life applications.

The rest of the paper is organized as follows: the methods are presented in Section II. The results and discussion are presented in Section III. Section IV presents the conclusion.

II. Methods

To carry out the experimental study, we used the framework shown in Fig. 1. It consists of four convolution layers each with pooling layers where necessary for local feature extraction and self-attention configured bidirectional layer of 64 units that was used for global feature learning before the feature representations are fed into a dense layer and a subsequent softmax layer for classification. The self-attention mechanism was configured in order to further consider long-term dependencies and the global context of the speech representations. In order to be consistent with earlier research, we used the exponential linear unit (ELU) as the activation function.

A. Experiments

We carried out five experiments to evaluate the robustness of the model. We used the publicly available German dataset of Berlin (EMODB) [13]. The experiments involved separate use of raw signals, MFCCs, mel spectrograms, a combination of MFCCs and mel spectrograms (mel) and a combination of MFCCs, mel spectrograms and chroma grams which is represented as "All" in Table I. It should be noted that we evaluated chroma grams alone and never obtained commendable results. In this paper, we considered happiness, sadness, neutral and anger as emotional states. To carry out the experiments, we used Keras 2.8.0 API, TensorFlow 2.6 as the back-end with python programming, and Nvidia GeForce RTX 2080 super graphics processing unit (GPU).

B. Feature Extraction

Raw Signals: For each speech signal, a sequence length of 16000 samples was considered. Speech signals with a shorter sequence length were padded while the longer ones were truncated. We used a band pass filter to consider frequencies between 300 and 3500 Hz which we considered to have pertinent cues for emotion recognition. In order to bring the signal amplitude to a target level, we used root mean square normalization in python.

Speech Features: We extracted spectral and voice quality features using Librosa 0.9.2. We considered features that can depict loudness, pitch and quality of sound. We extracted MFCCs and chroma grams as spectral features and mel spectrograms as voice quality features. The mean value of these features extracted from each frame was calculated and was separately used as input to the model in the first experiments. In the other two experiments, a combination of either MFCCs and mel spectrograms or MFCCs, chroma grams and mel spectrograms were used as input to the model after concatenation.

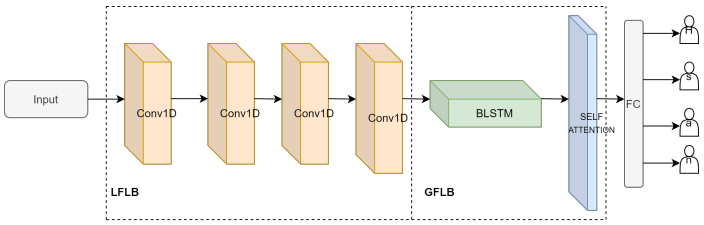

Fig. 1. The framework used in our experiments.

III. Results and Discussion

A. Results

Table I shows the experimental results in terms of accuracy (A) and F1 score (F1) obtained by the model. We also present comparative results of the robustness of the model in terms of the confusion ratio of the different classes of emotions for each input. CH, CS, CA and CN are confusion ratios for happy, sad, angry and neutral respectively. The results show that for all the inputs, the accuracy and F1 score can be commendable however the robustness especially in terms of the confusion error between high arousal dimension emotions needs to be given more attention.

TABLE I
Performance Comparison of the Model on Raw Signals and Extracted Features

Input	A(%)	F1(%)	CH(%)	CS(%)	CA(%)	CN(%)
Raw signal	81.18	80.52	07	92	99	71
Mel	80.0	78.91	73	79	67	60
MFCCs	85.46	85.71	40	93	88	87
Mel & MFCCs	94.55	95.48	40	100	100	93
All	89.09	87.77	53	71	88	87

B. Discussion

The experimental results show that models that use raw signals can achieve a commendable accuracy and F1 score however, they are not robust in terms of discriminating the high arousal emotion states of happy and angry. This is partly because the emotional cues of happy and angry are similar in terms of emotional dimension. Therefore, robust models that aid complex speech signal processing are required if they are to be used in SER systems. The experiments also show that mel spectrograms which depict voice quality cues in a speech signal are quite robust for happy and sad however the model still does not perform well especially for the neutral and angry emotions that tend to be confused with all the other emotions. On the other hand, MFCCs can be used by models if the interest is to achieve robustness for sad, angry and neutral however, the models that use them still confuse happy and other emotions

especially anger with which they belong to the same plane. Moreover, a combination of MFCCs and mel spectrograms improves the robustness results further to as high as 100% for sad and angry but the confusion ratio for happy remains the same. A combination of MFCCs, mel spectrograms and chroma grams that takes the pitch of sound into consideration improves the confusion ratio of happy but there is need for its robustness for the other emotions compared to the model that uses a combination of MFCCs and mel spectrograms. These results show that, in terms of the robustness of deep learning-based SER systems, models that use a combination of features perform better than those that either use a single kind of features or those that use raw signals. It should however, be noted that for all the inputs, the accuracy and F1 scores are commendable which further suggests that accuracy and F1 score are not the best metrics for SER studies especially for deployment in real-life situations.

IV. Conclusion

In this paper, we carried out an experimental evaluation of the robustness of deep learning-based SER models in terms of accuracy, F1 score and confusion ratio. We found out that the models perform well in terms of accuracy and F1 score when subjected to all the kinds of inputs considered in the experiments. They are however not robust enough in discriminating emotions that belong to the same dimensional plane especially when raw speech signals are used. The robustness in terms of confusion ratio improves for handcrafted feature inputs. The experiments show that the robustness improves even further when the handcrafted features are combined. However, it is still important to investigate the performance of complex deep learning-based models that can be robust enough to be deployed in real-time situations.

Acknowledgment

This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2022-2020-0-01808) supervised by the Institute of Information & Communications Technology Planning & Evaluation (IITP).

References

[1] S. K. Pandey, H. S. Shekhawat, and S. Prasanna, "Emotion recognition from raw speech using wavenet," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019, pp. 1292–1297.

[2] S. Kwon *et al.*, "Mlt-dnet: Speech emotion recognition using 1d dilated cnn based on multi-learning trick approach," *Expert Systems with Applications*, vol. 167, p. 114177, 2021.

[3] D. Tang, P. Kuppens, L. Geurts, and T. van Waterschoot, "End-to-end speech emotion recognition using a novel context-stacking dilated convolution neural network," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–16, 2021.

[4] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical signal processing and control*, vol. 47, pp. 312–323, 2019.

[5] S. Kwon, "Clstm: Deep feature-based speech emotion recognition using the hierarchical convlstm network," *Mathematics*, vol. 8, no. 12, p. 2133, 2020.

[6] X. Wu and Q. Zhang, "Intelligent aging home control method and system for internet of things emotion recognition," *Frontiers in Psychology*, vol. 13, 2022.

[7] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, "Real-time speech emotion analysis for smart home assistants," *IEEE Transactions on Consumer Electronics*, vol. 67, pp. 68–76, 2021.

[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[9] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] M. Xu, F. Zhang, and S. U. Khan, "Improve accuracy of speech emotion recognition with attention head fusion," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, 2020, pp. 1058–1064.

[12] S. Tripathi and H. Beigi, "Multi-modal emotion recognition on iemocap with neural networks," *arXiv preprint arXiv:1804.05788*, 2018.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.