Speech Emotion Recognition using Context-Aware Dilated Convolution Network

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*—Deep learning-based speech emotion recognition has been applied for social living assistance, health monitoring, authentication, and other human-to-machine interaction applications. Because of the ubiquitous nature of the applications, computationally efficient and robust speech emotion recognition models are required. The nature of the speech signal requires tracking of time steps, analyzing long-term dependencies and the contexts of the utterances as well as the spatial cues. Recurrent neural networks like long short-term memory and gated recurrent units coupled with attention mechanisms are often used to consider long-term dependencies and context in the speech signal. However, they do not take care of the spatial cues that may exist in the speech signal. Moreover, the operation of most of these systems is sequential which causes slow convergence, and sluggish training. Therefore, we propose a model that employs dilated convolutions layers in combination with hybrid attention mechanisms. The model uses multi-head attention to extract the global context in the feature representations which are fed into the bidirectional long short-term memory configured with self-attention to further handle the context and long-term dependencies. The model uses spectral and voice quality features extracted from the raw speech signals as input. The proposed model achieves comparable performance in terms of F1 score and accuracy. The proposed model's performance is also presented in terms of confusion matrices.**

***Index Terms*—context-aware emotion recognition, multi-head attention, dilated convolution**

I. Introduction

Speech emotion recognition (SER) is an affective computing domain that involves the detection and classification of emotion states from the speech signal or its extracted features. SER is applied in social living assistance machines, health monitoring, authentication systems, and interactive robots, etc. These applications of SER can be deployed in resource-constrained devices that can be used any time anywhere. Recently, SER systems were proposed for remote monitoring of senior citizens at smart homes [1], [2]. The progressive improvement of the performance of these systems using deep learning techniques is one of the major reasons for their enhancement.

Due to the need of tracking sequential time steps and long-term dependencies between the features in time series studies, recurrent neural networks (RNNs) like long short-term memory (LSTM) [3] are the ultimate deep learning methods used in most speech recognition studies. They are used in combination with different attention mechanisms [4], [5] and [6] so as to take into consideration the context of the inputs and feature representations. To consider long-term dependencies, the models proposed in [7], [8] and [9] use the LSTM and/or the bidirectional LSTM (BiLSTM). In [10] the bidirectional gated recurrent unit (BiGRU) is used to model the long-term dependencies for SER. However as suggested in [11] though RNNs alone achieve promising results, they encounter problems in convergence, sluggish training that uses a lot of memory resources due to the sequential manner in which they operate. In order to reduce these challenges [12], [13] and [14] proposed models that use dilated convolution layers for speech emotion recognition. However, the existing models compute long-term dependencies and the global context relationships between the features using attention mechanisms that operate sequentially which increases the computational cost and are slow in training. Some of them also do not consider spatial cues of the signal that depict different emotional states. Moreover, most of these models use raw signals as input and learn emotional cues and their relationships during training in an end-to-end approach. It is observed from the results in [12] and [13] that these models find difficulties in discriminating high arousal emotions especially happy and angry.

In this paper, we propose a model that uses self and multi-head attention in a hybrid manner together with dilated convolutions and bidirectional long short-term memory to achieve comparable performance. The model computes global contextualized dependencies between features in a parallel manner using multi-head attention. It further computes the global context and long-term dependencies by the use of self-attention configured in a stack of BiLSTM layers. The use of dilated convolution layers improves the receptive field with less increase in the number of parameters compared to the number of layers. The model uses spectral and voice quality features extracted from the raw speech signals as input since models that use raw signals tend to confuse happy and angry emotions or neutral and sad emotions.

The contribution of this paper is threefold;

- We propose a speech emotion recognition model that uses context-aware dilated convolution, multi-head attention and self-attention configured stack of BiLSTM layers.
- We validate the proposed model on North American,

German and Urdu languages using spectral and voice quality features.
- The performance of the proposed model is also compared with existing approaches that use raw signals in an end-to-end approach.

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

II. The Proposed Model

The architecture of the proposed model named context-aware dilated convolution network (CADCN) is shown in Fig. 1. The main objective of this model is to learn contextualized speech emotion cues with long-term dependencies using dilated convolutions and hybrid attention mechanisms coupled with BiLSTM. The main components of the proposed model are discussed in this section.

A. Dilated Convolution Block (DCB)

The dilated convolution block consists of two dilated convolution layers of dilation rates 1 and 2. The dilated convolution (DC) layers are used to provide temporal sequence modeling while covering a large receptive field. Dilated convolutions spread filters by skipping values in the input sequence in specified predetermined steps. In so doing dilated convolution layers provide a large receptive field with a few layers. In so doing the filters are applied over a large receptive area without an increase in the kernel inputs and the number of parameters. This, therefore, ensures temporal dependencies. The output feature representations of the dilated convolution block are fed into the Multi-head attention mechanism block that is made up of eight heads.

B. Contextualized Block (CB)

The other part of the model that aids contextualized learning consists of multi-head attention mechanism of eight heads together with three BiLSTM layers that employ self-attention. Multi-head attention mechanism operates the eight self-attention heads in parallel which improves its performance, reduces complexity and training time. The position encoding is handled by the use of the triangle positional embeddings used in [6]. The encoding result is passed through linear models to obtain the query Q, key K, and value V that would later be used in the scaled dot product computation to obtain similarities between feature representations. For linear projection, we use feed-forward neural networks (FFN). The concatenated output of the Multi-head attention is fed into three BiLSTM layers configured with self-attention mechanism to further consider long-term dependencies and the global context.

III. Experiments

In this section, we present the experiments carried out to validate the model on three datasets of different languages. 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).

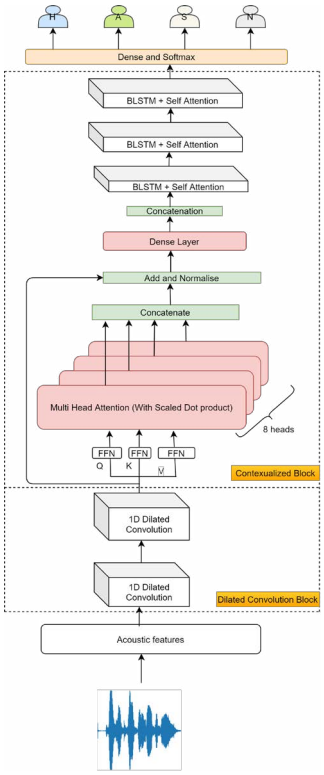

Fig. 1. A framework of the proposed Context-Aware Dilated Convolution Network (CADCN) that uses dilated convolution, multi-head attention and BiLSTM that is configured with self-attention.

A. Datasets

To study the performance of the proposed model we used the German dataset of Berlin (EMODB) [15] for the German language, Ryerson audio-visual database of emotional speech and song (RAVDESS) [16] for the English language and the Urdu language dataset [18]. For all the datasets, we consider happiness, sadness, neutral and anger as emotional states.

B. Feature Extraction

For each of the datasets, we extracted spectral and voice quality features using Librosa 0.9.2 and used them as input to the proposed model. We considered features that can depict loudness, pitch and quality of sound. The spectral low-level descriptors of sound extracted from the speech signal were mel frequency cepstral coefficients (MFCCs) and chroma grams. For voice quality, we extracted mel spectrograms. The mean value of these features extracted from each frame was calculated and concatenated to form one-dimensional inputs that represent the emotional knowledge of that frame that is fed into the model.

IV. Results and Discussion

In this section, we present the results and discussion of the performance of the proposed model on speech emotion recognition. We also discuss the significance of the dilated convolution layers in combination with multi-head attention

mechanism and stacked BiLSTM layers configured with self-attention for speech emotion recognition.

A. Results

Table I shows the results in terms of accuracy (A) and F1 score (F1) obtained by the proposed model in comparison with existing approaches. The proposed model was validated on EMODB, Urdu, and RAVDESS datasets. The results show the significance of the contextualized dilated convolution model. We observe from the confusion matrices shown in Fig. 2 to Fig. 4 that the model is good at discriminating emotions that belong to the same arousal dimension plane which is a challenge to the existing models that use dilated convolutions.

TABLE I
PERFORMANCE COMPARISON OF THE PROPOSED MODEL WITH THE EXISTING APPROACHES

Model	Datasets	Inputs	A(%)	F1(%)
[12]	EMODB	Signal	83.82	-
	URDU	Signal	-	-
	RAVDESS	Signal	-	-
[13]	EMODB	Signal	90.01	90.0
	URDU	Signal	-	-
	RAVDESS	Signal	-	-
Ours	EMODB	Features	96.36	96.43
	URDU	Features	95.31	95.31
	RAVDESS	Features	88.96	86.86

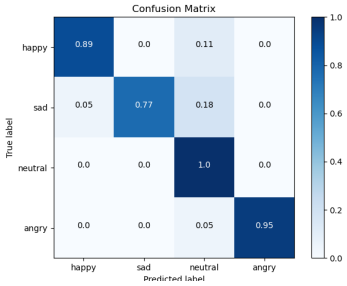

Fig. 2. Confusion Matrix for proposed model when tested on Urdu language testing datasets.

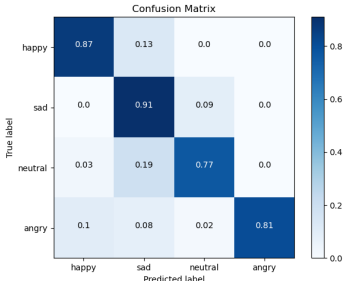

Fig. 3. Confusion Matrix for proposed model when tested on RAVDESS testing datasets.

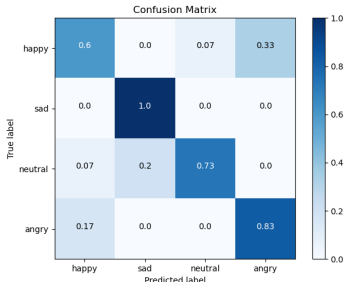

Fig. 4. Confusion Matrix for proposed model when tested on EMODB testing datasets.

B. Discussion

The results of the proposed model show that it achieves an accuracy of 96.36% and an F1 score of 96.36% on the EMODB dataset. The minimum performance of accuracy of 88.96% and F1 score of 86.86% is registered when the model is tested on the RAVDESS dataset. The model obtained an accuracy and F1 score of 95.31% on the Urdu language dataset. In comparison with the existing models, we observe that the proposed model improves the accuracy when validated on the EMODB dataset in a range of 6.35% to 12.54% and 6.43% of F1 score. It is however observed that the performance is best when the model is validated on non-English datasets. On the whole, the parallelism through the use of dilated convolution layers and multi-head attention coupled with BiLSTM that is configured with self-attention mechanism improves the performance of deep learning-based speech emotion recognition models.

V. CONCLUSION

In this paper, we proposed a context-aware speech emotion recognition model that uses dilated convolution layers, multi-head attention and a stack of BiLSTM layers configured with self-attention. It learns contextualized speech emotion cues with long-term dependencies from spectral and voice quality features of a speech signal using dilated convolution layers, multi-head attention and a stack of self-attention enabled BiLSTM layers. It is also observed that the model discriminates well the discrete emotions in the same arousal plane. However, it is still important to investigate the impact of the model on each emotion in real-time.

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

[2] 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.

[3] S. Hochreiter, "Ja1 4 rgen schmidhuber (1997)."long short-term memory"," *Neural Computation*, vol. 9, no. 8.
[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
[5] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.
[6] 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.
[7] Y. Yu and Y.-J. Kim, "Attention-lstm-attention model for speech emotion recognition and analysis of iemocap database," *Electronics*, vol. 9, no. 5, p. 713, 2020.
[8] Y. Xie, R. Liang, Z. Liang, and L. Zhao, "Attention-based dense lstm for speech emotion recognition," *IEICE TRANSACTIONS on Information and Systems*, vol. 102, no. 7, pp. 1426–1429, 2019.
[9] H. Zhang, H. Huang, and H. Han, "Attention-based convolution skip bidirectional long short-term memory network for speech emotion recognition," *IEEE Access*, vol. 9, pp. 5332–5342, 2020.
[10] Z. Zhu, W. Dai, Y. Hu, and J. Li, "Speech emotion recognition model based on bi-gru and focal loss," *Pattern Recognition Letters*, vol. 140, pp. 358–365, 2020.
[11] R. A. Hamad, M. Kimura, L. Yang, W. L. Woo, and B. Wei, "Dilated causal convolution with multi-head self attention for sensor human activity recognition," *Neural Computing and Applications*, vol. 33, no. 20, pp. 13 705–13 722, 2021.
[12] 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.
[13] 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.
[14] 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.
[15] 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.
[16] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.
[17] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.
[18] M. U. Arshad, M. F. Bashir, A. Majeed, W. Shahzad, and M. O. Beg, "Corpus for emotion detection on roman urdu," in *2019 22nd International Multitopic Conference (INMIC)*. IEEE, 2019, pp. 1–6.