CNN-n-GRU: end-to-end speech emotion recognition from raw waveform signal using CNNs and gated recurrent unit networks

Alaa Nfissi[1, 2, 4], Wassim Bouachir[2, 4], Nizar Bouguila[1], and Brian L. Mishara[3, 4]
[1]Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada
[2]Department of Science and Technology, TÉLUQ University, Montreal, Canada
[3]Psychology Department, University of Quebec at Montreal, Montreal, Canada
[4] CRISE Research Centre, Montreal, Canada

***Abstract*—We present CNN-n-GRU, a new end-to-end (E2E) architecture built of an n-layer convolutional neural network (CNN) followed sequentially by an n-layer Gated Recurrent Unit (GRU) for speech emotion recognition. CNNs and RNNs both exhibited promising outcomes when fed raw waveform voice inputs. This inspired our idea to combine them into a single model to maximise their potential. Instead of using hand-crafted features or spectrograms, we train CNNs to recognise low-level speech representations from raw waveform, which allows the network to capture relevant narrow-band emotion characteristics. On the other hand, RNNs (GRUs in our case) can learn temporal characteristics, allowing the network to better capture the signal's time-distributed features. Because a CNN can generate multiple levels of representation abstraction, we exploit early layers to extract high-level features, then to supply the appropriate input to subsequent RNN layers in order to aggregate long-term dependencies. By taking advantage of both CNNs and GRUs in a single model, the proposed architecture has important advantages over other models from the literature. The proposed model was evaluated using the TESS dataset and compared to state-of-the-art methods. Our experimental results demonstrate that the proposed model is more accurate than traditional classification approaches for speech emotion recognition.**

***Index Terms*—Speech emotion recognition, CNN, RNN, GRU, Signal processing, Waveform signal.**

I. INTRODUCTION

In this study, we are interested in speech emotion recognition (SER), which is an approach for detecting feelings and emotions from speech signal. It is the process of extracting the speaker's emotional state from speech signal's para-linguistic (PROSODIC) aspects without necessarily understanding the language [1], which is a relatively recent area of study that has experienced tremendous growth since the turn of the century.

Based on CNN and RNN models, we propose a new type of architecture that contributes in solving these issues according to two aspects. On one hand, setting convolutional layers at the beginning of the network is an effective technique to minimise input dimensionality, which can greatly simplify the training procedure. On the other hand, a deep CNN, may be used to extract high-level properties, which are then transferred to an RNN for final time aggregation.

Classification models can be divided into two categories, including traditional machine learning classifiers and deep learning classifiers. Previous works have also used hybrid approaches that combine both categories (see Fig. 1). In [2] and [3], the authors' starting point was the extraction of handcrafted features from speech signal and the generation of the Mel frequency cepstral coefficients (MFCC) spectrogram. However, this could not be an appropriate representation of speech because it is based on the mel scale, a perceptual scale of pitches judged by listeners to be equal in distance from one another. As a result, there is no assurance that they are ideal for all speech-related activities since many characteristics may be lost. Following that, [2] used a CNN as the basic component of the model to extract the relevant speech features, while taking into account the locally distributed data and ignoring the time dependency of the voice, whereas [3] attempted to avoid this weakness by using Bidirectional Long Short-Term Memory (BLSTM) to consider time and attention mechanism to focus on the most important parts of the spectrogram that emphasise the emotions. In [4], the authors used MFCC with no spectrogram, since the conventional SVM model was considered to classify emotions. In [5], another principle was employed to address the temporal representation of the speech, which consists of a 3D-CNN after computing deltas and delta-deltas for log-Mels attention CRNN. However the limitation of this study is still the speech representation that was used to feed the model. In [6], the authors attempted to address the data's high dimentionality by using Principal Component Analysis (PCA) as a first approach to represent the data, loses some of the data's information that may be relevant. For the second approach, which is based on spectrograms, the authors still have data representation issues. To address this problem, they applied the VGG16 network to the spectrogram as an image. However, this does not take into consideration the time dependency in the speech signal. In our work, we used raw waveform instead of going through any sort of hand-crafted feature extraction or spectrograms, which is to our knowledge applied for the first time on TESS dataset [7]. By using this conception, we take into consideration both local and time-distributed features of the speech for emotion recognition.

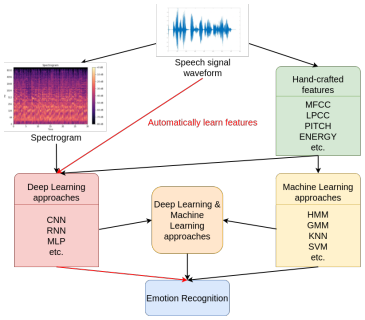

Fig. 1. SER approaches

II. PROPOSED METHOD

In order to ensure robust and invariant representations, weight sharing, local filters, and pooling are important techniques that often useful within CNN architectures for processing raw speech samples. Moreover, GRUs have also the potential to ensure effeicient representations, since they process data using their gate principles and memory capabilities, and depict it in a time-distributed manner. The basic concept behind our strategy is to combine a 1-D CNN with a GRU into a single model to take advantage of their respective capabilities. This model will be applied directly to the speech signal's raw waveform, without manual feature manipulation, in order to ensure relevant feature extraction within a robust end-to-end (E2E) model. This is illustrated by the path with red arrows in Fig. 1.

A. Motivation

The CNN component of our design is motivated by the work [1] on the ambient sound detection challenge using CNNs. To simulate the band-pass filter, it employs a relatively tiny receptive field in the convolutional layers, while a large receptive field in the first layer is determined based on the audio sampling rate. To optimize representation learning in the convolutional layers, our CNN component layers are completely convolutional, with no dropout, and can be applied to audio signal with varying duration. We solve the difficulty of training very deep models while keeping the computation cost low by using batch normalization and a careful design of down-sampling layers.

As human beings, we only keep the information that we believe is sufficient to assess a situation or an event, while the rest simply fades away from memory. This is the driving force underlying the selection of GRUs. They were developed as a solution to the problem of short-term memory. They have internal mechanisms known as gates that allow them to govern information flow. These gates can learn which data in a sequence should be kept or discarded. This allows to convey important information down the lengthy chain of sequences in order to generate predictions. [8].

B. Method

The major components of our CNN's portion architecture are the convolutional layers, which accept as input a 1D vector. They also allow to process time-series waveforms encoded as a long 1D vector as input instead of hand-tuned features or specifically generated spectrograms. To decrease the cost of computation in the rest of the network, we drastically lowered the temporal resolution in the top two layers, by utilising massive convolutional and max-pooling strides with a massive receptive field. Following that, we apply a batch normalisation to its output before passing it to an activation function, which in our instance is Leaky ReLU since it lacks zero-slope sections, overcomes the "dying ReLU" problem and speeds up training, based on the evidence that maintaining the "mean activation" at zero speeds up training [9]. Following the first convolutional layer, we implement a series of convolutional layers with smaller receptive fields to reduce the number of parameters in the model, followed by a single max-pooling layer (see Fig. 2).

Thus, to ensure that the CNN and GRU networks are properly linked, we use a single global average pooling layer that averages the activation throughout the temporal dimension to reduce each feature map to a single float. Then, we use this as an input to a fully connected layer, which will reduce the data by half to lower computation cost in the n-layer GRU component. This, in our case, is an important component of the model since it is the piece that will contribute the final time aggregation for time-dispersed data features representation due to the gates concept that controls the memory functioning. After the GRU layers, we apply another fully connected layer to aggregate the data and get an output equal to the number of classes, on which we apply softmax for final classification (see Fig. 2).

III. EXPERIMENTS AND RESULTS

A. Dataset

In this work we used **TESS** [7]. Toronto emotional speech set (TESS) is essentially a dataset that contains a total of 2800 stimuli in which two actresses, a young female and an older female, said a set of 200 target words in the carrier phrase "Say the word ____,". Recordings were taken of the set expressing each of seven moods (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral) [10].

B. Experimental setup

We first ensured that our signal has a sampling rate of 16KHz and is mono-channel to standardise our experimental data format. The dataset is segmented as follows: 80% for training, 10% for validation, and 10% for testing. We applied a Grid search using Asynchronous Successive Halving algorithm (ASHA) to find the appropriate hyperparameters, due to its high performance [11].

We examined four CNN-n-GRU architectures with n = 3, 5, 11, and 18. Each model is run for 100 epochs until it converges using Adam [12] without using any pretrained model (i.e. the weights of each model are started from scratch). The receptive field of our first CNN layer is equal to *(sampling rate / 100)*, which is in our case *160* to cover a 10-millisecond time span, to be comparable to the window size for many MFCC

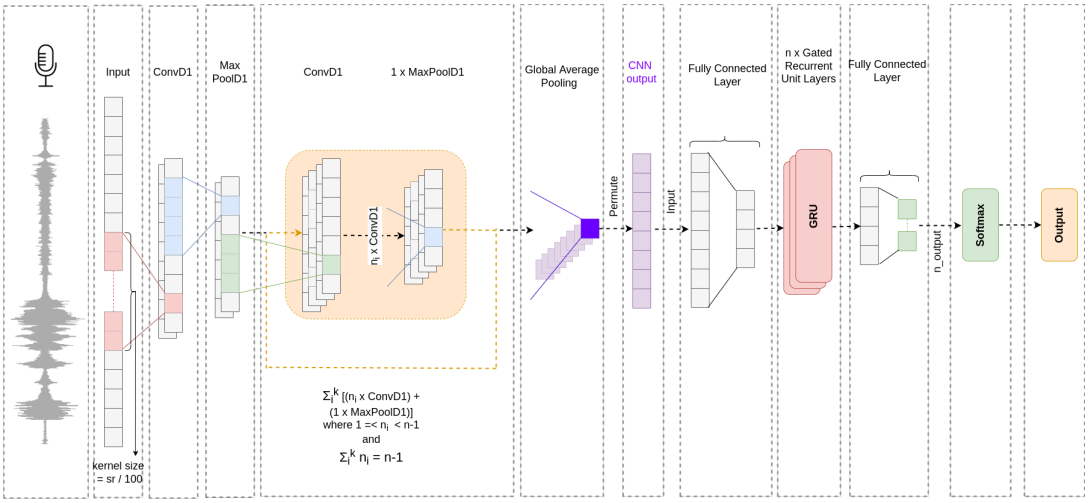

Fig. 2. Layer-wise CNN-n-GRU model architecture

computations. Our trials were run on a distributed multi-node system powered by four Nvidia Tesla T4 GPUs. Since TESS dataset is balanced we used the accuracy and F1-score as evaluation metrics for our model [13].

C. Results

Table I presents the results of other state-of-the-art algorithms compared to our proposed method, where Table II shows the outcome of the four tested architectures of our model. These results show that our model performs well, where the best architecture of our model was CNN-18-GRU that reached 99.2% in accuracy and a F1-score of 99%, outperforming the state-of-the-art methods. The detailed findings derived from the CNN-18-GRU model are presented in Table III, fig. 3, and the confusion matrix in fig. 4.

TABLE I
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE TESS DATASET

Related works		
TESS	**Test Accuracy**	**F1-score**
Iqbal et al. [4]	97%	NA
Aggarwal, Apeksha, et al. [6]	97.6%	97%
Dupuis et al. [14]	82%	NA
Praseetha et al. [15]	95.8%	NA
Huang et al. [16]	85%	NA
Choudhary et al. [17]	97.1%	96%
Kapoor et al. [18]	97.5%	97.4%
Krishnan et al. [19]	93.3%	NA
Our Method: CNN-18-GRU	99.2%	99%

TABLE II
CNN-N-GRU PERFORMANCE ON THE TESS DATASET

Our Results				
TESS	**Precision**	**Recall**	**F1-score**	**Test Accuracy**
CNN-3-GRU	98.8%	98.8%	98.8%	98.9%
CNN-5-GRU	96%	98%	97%	99%
CNN-11-GRU	64%	62.5%	63%	67%
CNN-18-GRU	98.7%	98.5%	99%	99.2%

Table III presents the classification report, including precision, recall, and F1-score of CNN-18-GRU. It's shown that the highest F1-score was obtained for the classes 'fear', 'neutral' and 'sad', while the lowest F1-score was achieved for the class 'surprised'.

TABLE III
CLASS-WISE CNN-18-GRU PERFORMANCE ON THE TESS DATASET

Emotion	**Precision**	**Recall**	**F1-score**	**Support**
angry	100%	97.5%	98.7%	40
disgust	97.6%	100%	98.8%	40
fear	100%	100%	100%	40
happy	100%	97.5%	98.7%	40
neutral	100%	100%	100%	40
sad	100%	100%	100%	40
surprised	94.9%	97.9%	96.4%	32
accuracy			99.2%	272
macro avg	98.7%	98.5%	99%	272
weighted avg	98.7%	98.5%	99%	272

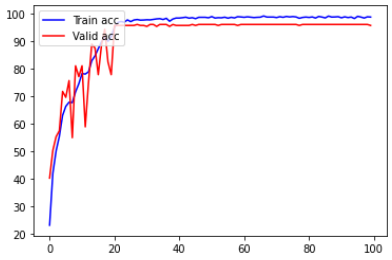

Fig. 3. CNN-18-GRU training and validation accuracy over epochs

Fig. 3 illustrates the training and validation accuracy considering 100 epochs in model CNN-18-GRU.

D. Result analysis and interpretation

We can observe that performance highly depends on the depth of the network. Generally, the models performed well, except for CNN-11-GRU (67% accuracy and 63% F1-score,

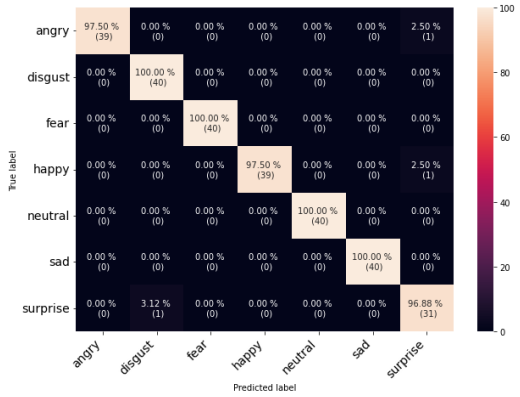

Fig. 4. CNN-18-GRU model confusion matrix

see Table II). To validate our method, we compared it to several state-of-the-art methods. While they all performed well, data representation choices, as well as the ability to generalise regarding the variety of speech characteristics, remained an issue with most state-of-the-art methods. This is because they all use heuristically hand-crafted features, such as MFCC and spectrograms, which we aimed to eliminate as a step in the process of speech emotion recognition.

One of the limitations of this study is the use of the TESS dataset. In fact, the reliability of this actor-based dataset in comparison with everyday human speech is limited by several factors. This includes the brevity of the content, the intensity of emotional expression, the fact that only female voices were used, and the limited demographic variation of participating people.

As shown in fig. 3, CNN-18-GRU converged quickly (around the 20th epoch) and it did not take a lot of time for training compared to other techniques that need more than 100 epochs to train. Given our balanced classes, we can see in table III that the emotions 'fear', 'neutral', 'disgust', and 'sad' were correctly predicted. However, the F1-score for 'disgust' was 98.8%, and 97.5% for both 'mad' and 'happy', which were confused with 'surprise'. This suggests that the two classes may share certain acoustic characteristics. We can see from the confusion matrix (fig. 4) that an correct classification score of 96.88% was achieved for the class 'surprise', while we observe a confusion of 3.12% with 'disgust'. We can also notice from table III that the F1-score of 'surprise' is 96.4%, which indicates that it is relatively more prone to be misclassified than the other emotions. These results highlight the importance of providing the model with the ability to learn by itself the appropriate local and time-distributed features.

IV. CONCLUSION

We presented a new deep learning model called CNN-n-GRU for end-to-end speech emotion recognition from acoustic waveform data. We employ a wide receptive field in the first convolutional layer, which acts as an acoustic features extractor. The rest of the network uses narrow receptive fields followed by the multi-layer GRU neural networks, that contributes in the time-distributed features aggregation. Our proposed technique outperforms state-of-the-art methods in terms of accuracy and F1-score, as it captures local and time distributed features. In this way, we take advantage of both CNNs and RNNs capabilities. As future work, we first plan to improve the first layer filters in order to enhance our model's generalizability. Second we aim to test our model on more realistic datasets, with an emphasis on the end-to-end approach that eliminates the need for explicit feature extraction.

REFERENCES

[1] A. A. T. Garcia, C. A. R. Garcia, L. Villasenor-Pineda, and O. Mendoza-Montoya, *Biosignal Processing and Classification Using Computational Learning and Intelligence: Principles, Algorithms, and Applications*, pp. 307–326. Academic Press, 2021.

[2] A. Aftab, A. Morsali, S. Ghaemmaghami, and B. Champagne, "Lightsernet: A lightweight fully convolutional neural network for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6912–6916, IEEE, 2022.

[3] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning.," in *Interspeech*, pp. 2803–2807, 2019.

[4] M. Z. Iqbal, "Mfcc and machine learning based speech emotion recognition over tess and iemocap datasets," 2020.

[5] 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*, vol. 25, no. 10, pp. 1440–1444, 2018.

[6] A. Aggarwal, A. Srivastava, A. Agarwal, N. Chahal, D. Singh, A. A. Alnuaim, A. Alhadlaq, and H.-N. Lee, "Two-way feature extraction for speech emotion recognition using deep learning," *Sensors*, vol. 22, no. 6, p. 2378, 2022.

[7] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020.

[8] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014.

[9] A. L. Maas, A. Y. Hannun, A. Y. Ng, *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, p. 3, Citeseer, 2013.

[10] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (tess)," *Scholars Portal Dataverse*, vol. 1, 2020.

[11] L. Li, K. Jamieson, A. Rostamizadeh, E. Gonina, J. Ben-Tzur, M. Hardt, B. Recht, and A. Talwalkar, "A system for massively parallel hyperparameter tuning," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 230–246, 2020.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[13] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[14] K. Dupuis and M. K. Pichora-Fuller, "Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set," *Canadian Acoustics*, vol. 39, no. 3, pp. 182–183, 2011.

[15] V. Praseetha and S. Vadivel, "Deep learning models for speech emotion recognition," *Journal of Computer Science*, vol. 14, no. 11, pp. 1577–1587, 2018.

[16] A. Huang and P. Bao, "Human vocal sentiment analysis," 2019.

[17] R. R. Choudhary, G. Meena, and K. K. Mohbey, "Speech emotion based sentiment recognition using deep neural networks," in *Journal of Physics: Conference Series*, vol. 2236, p. 012003, IOP Publishing, 2022.

[18] S. Kapoor and T. Kumar, "Fusing traditionally extracted features with deep learned features from the speech spectrogram for anger and stress detection using convolution neural network," *Multimedia Tools and Applications*, pp. 1–22, 2022.

[19] P. T. Krishnan, A. N. Joseph Raj, and V. Rajangam, "Emotion classification from speech signal based on empirical mode decomposition and non-linear features," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1919–1934, 2021.