Bangla Speech Emotion Recognition Using Transfer Learning and Convolutional Neural Networks

Ahmed Al Jawad and Faisal Hossain
Department of Electrical & Computer Engineering
North South University
Dhaka 1229, Bangladesh
ahmed.jawad01@northsouth.edu; faisal.hossain03@northsouth.edu

***Abstract*—In this paper, speech emotion recognition with deep learning based approach is introduced. Implementations consist of convolutional neural networks (CNN) and residual neural network (ResNet) to classify emotions from the Bangla speech corpus SUBESCO using spectral features - Mel Frequency Cepstral Coefficients (MFCCs) and derivatitves of MFCCs. The final results of the study concluded with ResNet50 being trained on the RAVDESS dataset and fine-tuned on SUBESCO to achieve an accuracy of 71.14% in correctly identifying the 7 unique emotions embedded in the dataset.**

***Index Terms*—Deep CNN, RAVDESS, SUBESCO, Bangla SER, MFCC, Transfer Learning**

I. Introduction

Understanding and extracting human emotions has been a significant research topic for many years. Apart from visual representations such as facial expressions, it is considered that speech has a strong correlation with human emotion. Speech emotion recognition is the process of classifying the emotional state of a speaker, and it emerges as a key research topic in the domain of Human Computer Interaction (HCI). Speech emotion recognition systems are applicable to many applications such as human-computer interaction, education, entertainment industries, automotive, and security systems. Recognizing human emotion from audio samples remains a challenge and various machine learning approaches were taken by researchers. However, choosing the optimal features for emotion recognition is a difficult task and there are several features used in speech emotion recognition, with MFCCs being the most widely used feature [1]. In this study, an SER system for Bangla utterances is designed using convolutional neural networks and by applying transfer learning. The datasets, input feature, model architectures used as well as experimental setups for this study are discussed later in this paper in their respective sections.

II. Literature Review

Speech emotion recognition (SER) can be leveraged to enhance a wide range of applications. In recent years, the significance of SER has increased drastically, causing more researches to be performed within this domain. The classifiers used in speech emotion recognition have used handcrafted features with traditional machine learning approaches. They do not yield a high-accuracy result due to the uncertainty of effective feature selection [2]. Although, with recent deep learning approaches, higher performance can be observed. Some of the fundamental deep learning techniques used for SER are Convolutional Neural Networks (CNN), Deep Boltzmann Machine (DBM), Recurrent Neural Network (RNN), Recursive Neural Network (RNN), Auto Encoder (AE), and Long Short-Term Memory (LSTM) [3]. Generally, spectrograms generated from audio are used as the input to the models. In the work of Dias et al. [4], they proposed a one-dimensional deep convolutional neural networks approach using five spectral features such as Mel-frequency Cepstral Coefficients (MFCCs), Mel-scaled spectrogram, Chromagram, Spectral contrast feature, and Tonnetz representation. Their proposed framework claims that it has outperformed most of the previously known state-of-the-art models and reported further insights for achieving higher accuracy.

For features, deep learning algorithms use unstructured audio representations like the spectrogram and MFCCs. Sandra et al. [5] developed a DeepSpectrum system that extracts deep picture-based descriptors from Mel-spectrograms, allowing knowledge to be transferred from the image domain to audio recognition. They also used openSMILE, an open-source audio extraction tool, to extract the INTERSPEECH ComParE challenge 2013 feature set, comprising about 6373 features. Stochastic Gradient Descent (SGD) is used to compare these two feature extraction algorithms by minimizing logistic regression or modified Huber losses. Their experiments demonstrate that using DenseNet-121 on Deep Spectrum gives the best accuracy of 62.43 % on the test set among ten different pre-trained CNN. Considering data scarcity is a significant concern in SER, some research has lately included data augmentation in this domain to alleviate the problem. For instance, [6] shows the use of vocal tract length perturbation (VTLP), augmenting speech datasets by transforming spectrograms, using a random linear warping along the frequency dimension, and oversampling the minor classes. They achieved a 1.1 percent increase in unweighted accuracy and a 0.7 percent increase in weighted accuracy for four different classes of emotions utilizing the VLTP method on CNN with Bi-LSTM architecture on the IEMOCAP dataset, compared to the baseline dataset. Implementing transfer learning is another way to improve performance. For speech emotion recognition, [7] employed a resnet34 model with a statistics pooling layer and spectrogram augmentation. To reduce overfitting

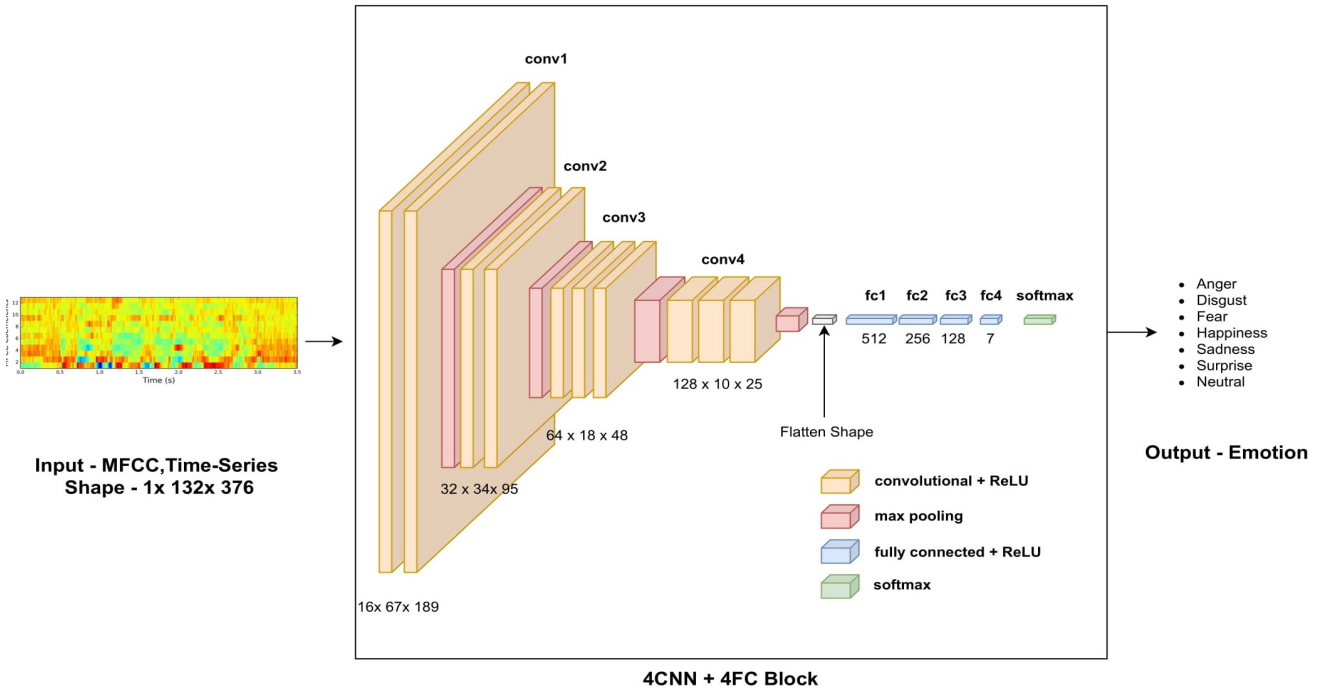

Fig. 1. CNN Block Diagram

and increase the number of samples, they used random time-frequency masks on spectrograms. Their experiment illustrates several comparisons between transfer learning, augmentation, statistics pooling, and merging all of the architecture, with the combined system outperforming the previous experiments significantly. They also tested their suggested work's competitive performance, comparing it to numerous past research and claiming state-of-the-art performance in multiple settings.

Most of these deep learning models consist of high computational complexity, containing millions of parameters. As a result, they are not preferable for implementations in embedded systems. Average values of MFCCs are taken to reduce the dimensionality. Panuwit et al. [8] have derived the delta (differential coefficient) from MFCCs, and the delta-delta coefficient (acceleration coefficient) from the delta. These coefficient values are added with the MFCC to form a feature vector, and the mean value of MFCC was calculated and converted into unit variance and standardized. They have concluded using their mean value of MFCCs consist of significantly a smaller number of parameters and gives higher accuracy on their ANN when compared to normal MFCCs. Their results show better performance than state-of-the-art methods, with an accuracy rating of 87.8% on the EMO-DB dataset and 82.3% on the RAVDESS dataset.

In contrast, we propose using a Bangla Emotional Speech Corpus (SUBESCO) [9] to conduct experiments. We plan to use transfer learning techniques and data augmentation methodologies to train and evaluate the corpus to build multiple SER models. Furthermore, we anticipate that our research would help more advanced research on the Bangla speech corpus.

III. Proposed Methodology

For our experiments, we have used the SUBESCO dataset and the RAVDESS dataset, both of which were available for public use. From these experiments, we chose the best performing model. Data preprocessing, datasets, and experimental setup are discussed in the following ensuing subsections.

A. Dataset

SUBESCO is the largest emotional audio corpus for Bangla language. It consists of seven different emotional states in the 7000 utterances, performed by 20 native speakers which is also gender balanced. The corpus contains the six fundamental emotions anger, disgust, fear, happiness, sadness, and surprise along with neutral emotion, all in equal proportions. SUBESCO is made available free of charge and is available from the open access library Zenodo.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is an English audio-visual database, which was performed by 12 male and 12 female actors, in a native North American accent. For our work, the audio-emotional speech portion was used which consists of 1440 utterances containing the emotions calm, disgust, happy, sad, angry, fearful, surprise, and neutral expressions. RAVDESS has gained its popularity for SER studies among researchers, and it is free of charge and available from Zenodo.

B. Audio Preprocessing

Preprocessing the data is very crucial because the performance achieved during supervised learning is strongly dependent on the data which is fed to the classifier. When generating the csv file consisting of the necessary target output

labels, neutral emotion was dropped to make the dataset balanced, since there are significantly lesser number of neutral expressions. Audio preprocessing were performed using the Torchaudio library which comes with Pytorch framework. All the audio files were sampled at 48KHz and the length of the audio were fixed at 4 seconds. For RAVDESS audio files, the duration is 3 seconds long and right-padding was applied.

C. Feature Extraction

Human vocal elements, such as tongue, teeth and vocal cord affects the sound and makes it discrete for each individual. The resulting voice consists of a shape which exhibits in the envelope of the short-time power spectrum, which is represented by MFCCs. Using the Torchaudio library, 20 and 44 MFCCs were extracted from the audio samples along with their derivatives, for different experiments. Higher number of MFCCs may be chosen to capture more information [10]. Therefore, in applications like SER, this would allow more number of parameters to be used. The MFCCs, delta, and delta-deltas were concatenated and normalized before using as model input.

D. Model Architecture

We experimented with custom four layer CNN designs and other pretrained CNN architectures to fit the objective of audio-based emotion classification. Four Convolutional layers follow a Max-pooling layer, making up the customized CNN architecture. The kernel size for executing the convolutional operation in each layer was set at (3×3) with the 2-pixel zero padding and 1-pixel stride. The convolutional layers' kernel numbers were set to 16, 32, 64, and 128. For non-linearity, the ReLU activation function was introduced. We flatten the output from the CNN layer and input them in four fully connected layer with 512,256,128 and 7 respectively. A softmax layer was used to distribute prediction probabilities among those seven different emotions.Figure 1 For next experiment, we used ResNet-18 and ResNet-50 pretrained models and finetuned them. Moreover, we tried out transfer learning. We finetuned the models, first on the RAVDESS dataset. We then used the trained models to further train and test on the SUBESCO dataset.

E. Evaluation Metrics

Audio emotion identification is a classification problem, and essential metrics for model performance evaluation include accuracy, precision, recall, and the f1-score. We chose accuracy metrics to highlight our best experiment outcomes because our class distribution is balanced across the datasets.

IV. Experimental Setup

The experiments were conducted in Google Colaboratory and Pytorch framework was used to implement our work. Multiple experiments were carried out on the two datasets. The datasets for both SUBESCO and RAVDESS were split into 70% for training, 20% for testing, and 10% for validation. The corresponding splits were also gender and actor balanced for SUBESCO to make sure each split has a good gender ratio balance as well as distinct speaker for each spilt. Emotion recognition is independent of gender or speaker, so the splitting was carried in this manner. As mentioned earlier, for each of the experiments a feature vector was extracted which consisted of MFCCs, Deltas, and Delta-Deltas, concatenated. The model architectures for these experiments are custom 4-layer CNN with 3 fully connected layer, and pre-trained Resnet18 and Resnet50 [11]. Different number of MFCCs were also experimented, however the results differed slightly. As there are 7 classes of emotions, cross-entropy loss [12] was used as the criterion and for the optimizer, we chose Adam [13]. Training was carried out for 100 epochs with a learning rate of 10^{-4}, weight decay of 10^{-6} and batch-size of 32.

V. Result Analysis

As previously stated, our goal is to demonstrate various models and setups, compare them, and then select the best model based on the evaluation metrics. On the RAVDESS dataset, we achieved validation accuracy of 73.77 % and model inference accuracy of 62.77 % in our first experiment with 4CNN+4FC. However, ResNet-50 outperformed our previous result, with a validation set accuracy of 77.05 % and a testing set accuracy of 76.60 %. So far, this has been our best model on the RAVDESS dataset. The experiment on the SUBESCO dataset was then carried out using the same architecture of 4CNN+4FC, but the initial results were poor with 50.29% testing accuracy. With training on the ResNet-18, SUBESCO achieved competitive results with RAVDESS, with 74.57 % validation and 69.93 % testing accuracy.

TABLE I
Accuracy Comparison of Dataset and Model

Dataset	Model	Validation Accuracy	Testing Accuracy	Total Parameters
RAVDESS	4CNN+4FC	73.77%	62.77%	16,4646,791
	ResNet - 50	77.05%	**76.60%**	23,516,167
SUBESCO	4CNN+4FC	56.71%	50.29%	16,4646,791
	ResNet- 18	74.57%	69.93%	11,173,895
	RAVDESS ResNet - 50	77.29%	**71.14%**	23,516,167

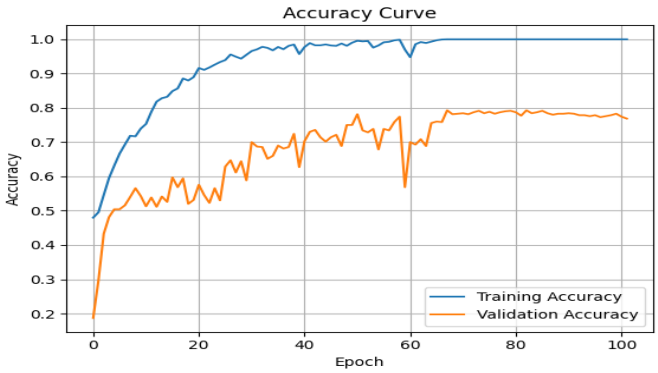

Fig. 2. SUBESCO accuracy plot on best model

Our best accuracy on SUBESCO were achived using transfer learning on the RAVDESS dataset earlier. We trained the Resnet-50 model which outperformed all other experiment on the RAVDESS dataset, on the SUBESCO dataset. This experiment achieved best result on the dataset with a 77.29% validation accuracy with a 71.14% testing accuracy. Training history on the SUBESCO for our best result is shown in figure 2 and 3. Summarization of our comparitive and best results can be found on table I.

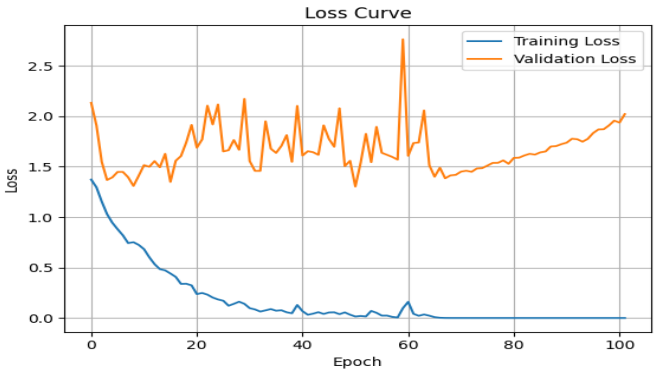

Fig. 3. SUBESCO loss plot on best model

VI. Conclusion

In this study, a SER system for Bangla audio was formulated using SUBESCO and RAVDESS datasets. We conducted a series of experiments, from which the best performing model resulted with an accuracy of 76% on RAVDESS dataset and 71.14% on SUBESCO dataset after fine-tuning. Further work can be done to achieve higher performance by applying further preprocessing to reduce noise and applying data augmentation techniques. As there are relatively less research performed on domain of Bangla audio SER, our study aims to provide contribution to it.

References

[1] D. Rana and A. Jain, "Effect of windowing on the calculation of mfcc statistical parameter for different gender in hindi speech," *International Journal of Computer Applications*, vol. 98, no. 8, 2014.

[2] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, 2014.

[3] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[4] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[5] S. Ottl, S. Amiriparian, M. Gerczuk, V. Karas, and B. Schuller, "Group-level speech emotion recognition utilising deep spectrum features," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 821–826, 2020.

[6] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "Cnn+ lstm architecture for speech emotion recognition with data augmentation," *arXiv preprint arXiv:1802.05630*, 2018.

[7] S. Padi, S. O. Sadjadi, R. D. Sriram, and D. Manocha, "Improved speech emotion recognition using transfer learning and spectrogram augmentation," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 645–652, 2021.

[8] P. Nantasri, E. Phaisangittisagul, J. Karnjana, S. Boonkla, S. Keerativittayanun, A. Rugchatjaroen, S. Usanavasin, and T. Shinozaki, "A light-weight artificial neural network for speech emotion recognition using average values of mfccs and their derivatives," in *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 41–44, IEEE, 2020.

[9] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla," *Plos one*, vol. 16, no. 4, p. e0250173, 2021.

[10] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using mfcc features and support vector machine," in *Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia*, vol. 2, pp. 556–561, 2007.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[12] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.