Group No: 231

Group Member Names:

1. ASHIQUE ZZAMAN (2021sc04612)

2. AMRITESH KUMAR DAS (2021sc04432)

3. BISHNU CHARAN SINHA (2021sc04431)

4. RAKSHANDA KAUL (2021sc04406)

Domain: Speech Recognition

Comparison Table:

	PAPER 1	**PAPER 2**	**PAPER 3**
Title of the paper	Emotion recognition from speech using convolutional neural network with recurrent neural network architecture	Convolutional Neural Network (CNN) Based Speech-Emotion Recognition	Speech Emotion Recognition Using Deep Learning Hybrid Models
Authors	Saikat Basu, Jaybrata Chakraborty, Md. Aftabuddin	Alif Bin Abdul Qayyum, Asiful Arefeen, Celia Shahnaz	Jamsher Bhanbhro, Shahnawaz Talpur, Asif Aziz Memon
Year of publication	2017	2019	2022
Network used / No of layers	CNN and LSTM	1D CNN and Dense layer	2D CNN (4 sequential time distributed Conv2D block) and LSTM
Depth of the network	In CNN, used three convolution layer having 32, 16, 8 filter respectively. In LSTM network, provided two hidden layer with 50 nodes in first layer and 20 nodes in second layer	7 1D convolutional layers each followed by a batch normalization layer and a max pooling layer with a pool size of 2, except for the last convolutional layer. The last layer of the model is a dense layer with 7 nodes	In first block, channel 1 was used as an input, and channel 16 for output with padding & stride one and kernel size 3. In second block channels, 16 were used as input, and channel 32 for output with padding & stride one and kernel size 3 The other two blocks also have the same configuration as the second block. After the above configuration, the model uses whatever is found after flattening LSTM is combined with the Linear SoftMax Layer.
How is the network helping the overall task? **eg: feature engg or classification or regression or all**	13 MFCC (Mel Frequency Cepstral Coefficient) are used with 13 velocity and 13 acceleration component as features and a CNN (Convolution Neural Network) and LSTM (Long Short Term Memory) based approach for classification. Use of CNN-LSTM model for recognition of emotion from speech is a effective step towards designing a generic emotion recognition system. Although the size of data set is not so large the performance of the proposed model is promising enough. Some normalized input data or use of Bidirectional LSTM instead of LSTM can lead to more better solution.	Research on speech-emotion recognition exploiting concurrent machine learning techniques has been on the peak for some time. Numerous techniques like Recurrent Neural Network (RNN), Deep Neural Network (DNN), spectral feature extraction and many more have been applied on different datasets. This paper presents a unique Convolutional Neural Network (CNN) based speech-emotion recognition system. A model is developed and fed with raw speech from specific dataset for training, classification and testing purposes with the help of high end GPU. This work will be influential in developing conversational and social robots and allocating all the nuances of their sentiments.	Speech Emotion Recognition (SER) has been essential to Human-Computer Interaction (HCI) and other complex speech processing systems over the past decade. Initially, the Mel spectrogram's temporal features are trained using a combination of stacked Convolutional Neural Networks (CNN) & Long-term short memory (LSTM). The said model performs well. For enhancing the speech, samples are initially preprocessed using data improvement and dataset balancing techniques. CNN + LSTM delivers benefits and has higher accuracy rates due to repeated learning and deep feature extractions. The model is more complex, demands significant training, and yields more accurate predictions than the other SERs.
Loss function used	Categorical cross entropy	Categorical Crossentropy	Cross entropy loss
Evaluation / Performance metric used	Training accuracy reached at 96% and test accuracy reached at 80%	Generates accuracy of 83.61%	Generates accuracy above 93.9% for the model on mentioned data set when classifying emotions into one of eight categories.
Name of Dataset used. **If a public dataset, provide the URL.**	Berlin Emotional Speech (EmoDB) dataset URL: http://www.emodb.bilderbar.info/download/	Surrey Audio-Visual Expressed Emotion (SAVEE) URL: http://kahlan.eps.surrey.ac.uk/savee/Database.html	Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDNESS) dataset URL: https://zenodo.org/record/1188976

Conclusion:

Per information gathered from the above-mentioned papers, all 3 papers focussed on Speech Recognition using various deep learning algorithms. Below are few conclusions drawn from each paper –

Paper 1:

1. Dataset size is not large, but the use of a CNN-LSTM model for emotion recognition from speech is effective.
2. Suggests that using normalized input data or Bidirectional LSTM instead of LSTM could lead to better results.
3. Recommends using a larger dataset for improved training outcomes.

Paper 2:

1. Presents a neural network-based speech-emotion recognition method that can classify 7 types of emotions with high accuracy.
2. Emphasizes the user-friendliness and interpretability of the method.
3. Suggests that this method has significant potential for various applications and can be expanded to classify more emotions.

Paper 3:

1. Describes a complex model that requires significant training but yields highly accurate predictions for speech-emotion recognition (SER).
2. Attributes the precision of performance to a balanced dataset, augmentation techniques, and careful handling of convolution blocks and layers.
3. Concludes that using the best hyperparameters with the Mel Spectrogram of the input signal can yield excellent SER results.

Given the details, especially the performance matrices with respect to Learning curve and Confusion matrix, if we look at the learning curve of **Paper 1** and **Paper 3**, there is a significant gap between Train and Test curve showing the model did not do much learning and is capable of further learning.

In comparison, model presented in **Paper 2** seems to have done good as far as learning curve is concerned and also had good Accuracy and don't seem to be overfitted or underfitted. Moreover, model in **Paper 2** shows uniform performance across emotion classes in the confusion matrix indicates that the model is consistent in recognizing different emotions. This could be important if we need balanced performance across various emotional states.

So, for Part B of the task, implementing the methodologies from **Paper 2** might offer a good learning curve and more accurate results with the deep learning techniques used.