EMOTION RECOGNITION IN SPEECH

By - Mahrukh Shabbir Hasan
Supervised by – Carolyn Deveruex

Introduction:

In daily interpersonal human relationships and their comprehension of one another, emotion plays a crucial role. If this level of understanding could be extended to human-machine interaction, it would prove to be a huge progress in technology. The goal of this project is to conduct a comparative study of the accuracy, predictive quality and training/testing speed of the four models used in this project—Convolutional Neural Networks (CNN), Multilayer Perceptron (MLP), Support Vector Machines (SVM), and Long-Short Term Memory (LSTM) on the same dataset, in order to detect emotion in speech.

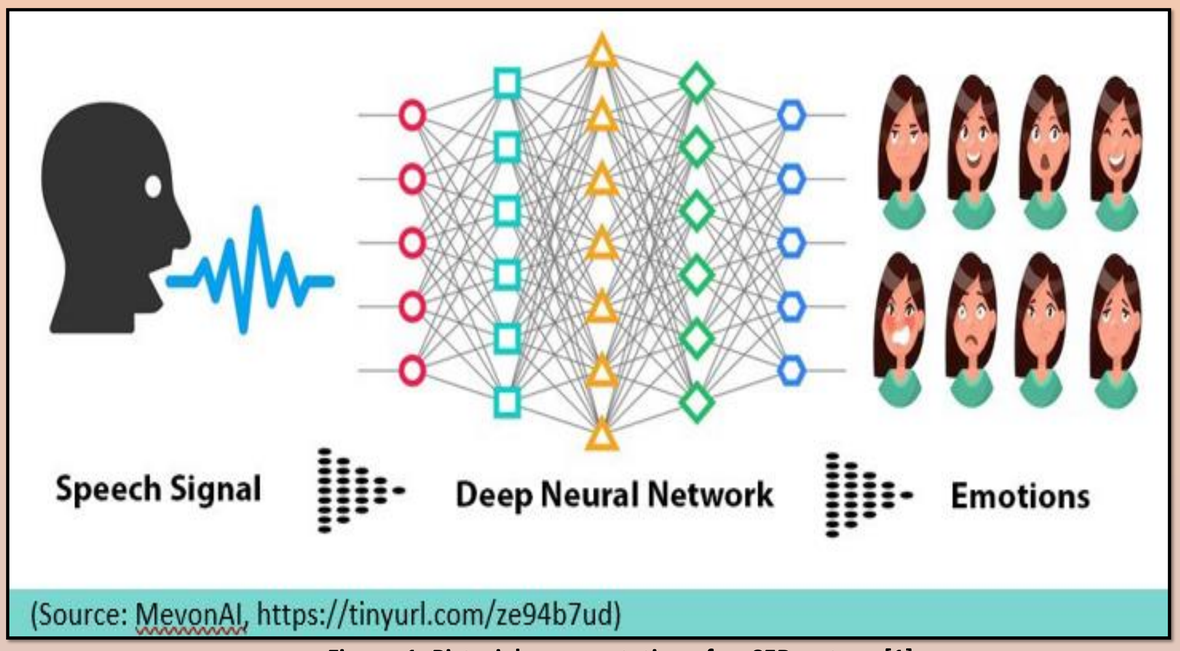

Figure 1: Pictorial representation of an SER system [1]

Papers	Classifiers	Dataset Used	Results
Speech Emotion Recognition: Methods and Cases Study. (Kerkeni et al., 2018)[2]	Multivariate Linear regression, SVM, RNN	Berlin and Spanish Emotional database	Highest results for each classifier: MLR – Spanish – 82.41%, Berlin – 75% SVM – Spanish – 77.63%, Berlin – 63.30% RNN- Spanish – 90.05%, Berlin – 69.55%
Speech emotion recognition using hidden Markov models. (Nwe, Foo, and De Silva, 2003)[3]	LFPC to represent speech signals and Hidden Markov Models were used as classifiers	A user defined dataset, with voices of Burmese and mandarin speaking people. A total of 720 voice recordings.	An average accuracy of 77.1%, 89% when emotions were identified individually
Evaluating deep learning architectures for Speech Emotion Recognition. (Fayek, Lech and Cavedon, 2017) [4]	DNN – CNN + LSTM	IEMOCAP	Best accuracy achieved was 64.78%. Preprocessing was done with hamming window and log FFT.
Automatic speech emotion recognition using recurrent neural networks with local attention. (Mirsamadi et al., 2017) [5]	RNN	IEMOCAP	61.8% recognition rate with raw features 63.5% recognition rate with LLD features.
Speech emotion recognition with deep convolutional neural networks (Issa, Fatih Demirci and Yazici, 2020) [6]	CNN	RAVDESS, EMODB, IEMOCAP	RAVDESS - 71.61% EMODB – 86.1% IEMOCAP – 64.30%

Table 1: Relevant papers and their details

Background:

Many papers have been published with research about different models and techniques that could be employed in SER. But the challenge of achieving the right balance of accuracy vs model complexity is still a topic of research. Although there are many papers regarding this topic, few of the most relevant papers are listed in table 1. Papers from 2017-2020, focused more on deep learning models but (Nwe. et al)[3], written in 2003, used Hidden Markov Models (HMM). CNN and RNN being the most common models in many papers, showed various levels of accuracy implying that datasets and feature extraction play an key role in SER.

Methodology:

A dataframe of 10,242 audios from CREMA-D[7] and TESS[8] datasets was built. Then, audio files were pre-processed to extract features using MFCCs (Mel Freq. Cepstrum Coeff.). Then, the data was augmented to make the models robust and was split into train and test data in the ratio of 70% & 30% respectively. CNN, SVM, LSTM and MLP models were trained on the train data and were tested and compared, on test dataset, in terms of accuracy, predictive quality and their train and test speed. (Fig.2)

Feature extraction was based on underlying principle that audios can be analyzed in amplitude, frequency and time domain. (Fig.3)

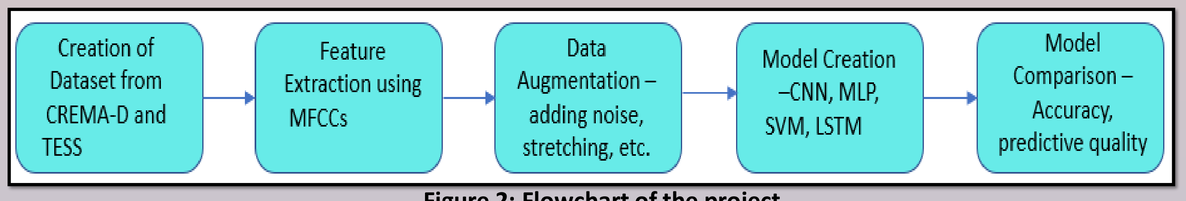

Figure 2: Flowchart of the project

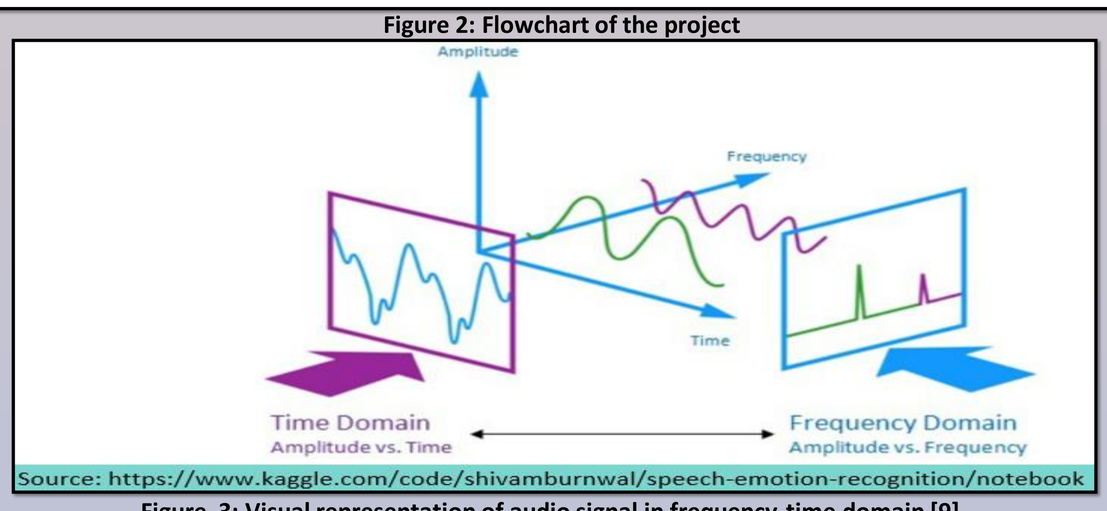

Figure 3: Visual representation of audio signal in frequency-time domain [9]

Model	Accuracy	Training Time	Test time	Predictive quality
CNN	84	90mins	9sec	Very good
MLP	84.76	7 mins	0.25s	Very good
RNN-LSTM	88.88	75 mins	2.95 sec	Great
SVM	72.9	4.2 mins	1 min	Average

Table 2: Comparison of performance of each model

	precision	recall	f1-score	support
angry	0.96	0.92	0.94	1127
disgust	0.81	0.88	0.85	1182
fear	0.91	0.86	0.89	1175
happy	0.88	0.90	0.89	1172
neutral	0.88	0.87	0.87	1052
sad	0.88	0.88	0.88	1204
surprise	1.00	0.99	1.00	258
accuracy			0.89	7170
macro avg	0.90	0.90	0.90	7170
weighted avg	0.89	0.89	0.89	7170

----accuracy score 88.88423988842399 ----

Figure 4: Confusion matrix of LSTM

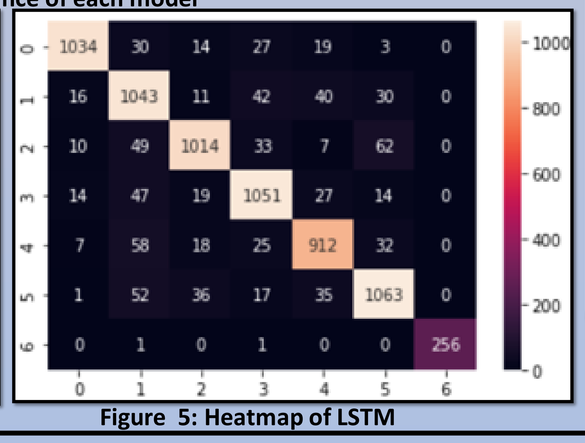

Figure 5: Heatmap of LSTM

Results Achieved So Far:

Table 2 depicts the comparison of models with different parameters. LSTM has the highest accuracy so far of 88.8% with a predictive quality better than the rest of the models. MLP has an accuracy of 84.76%, almost that of LSTM, but has the best testing time of 0.25sec.

Fig. 4 shows the precision with which the LSTM detects each emotion. Anger, fear and surprise are emotions with more than 90% precision. The heatmap of LSTM classifier, presenting the correlation of each emotion with itself as well as other emotions is shown in fig. 5. The more the correlation, the lighter the color.

Applications:

- Call Centres
- Medical Studies
- Games
- AI interface
- Web-based E learnings
- Fraud detection
- Emotion levels of MPs
- NGOs
- Interface with Siri/Alexa
- Emotion in audio surveillance
- Smartphone interface instead of emojis

References:

[1] GitHub - SuyashMore/MevonAI-Speech-Emotion-Recognition: Identify the emotion of multiple speakers in an Audio Segment
[2] Kerkeni, L., Serrestou, Y., Mbarki, M, et. all, 2018. Speech Emotion Recognition: Methods and Cases Study. Proceedings of the 10th International Conference on Agents and Artificial Intelligence,.[176]
[3] Nwe, T., Foo, S. and De Silva, L., 2003. Speech emotion recognition using hidden Markov models. Speech Communication, 41(4), pp.603-623.
[4] Fayek, H., Lech, M. and Cavedon, L., 2017. Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks, 92, pp.60-68.
[5] Mirsamadi, S., Barsoum, E. and Zhang, C., 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),.
[6] Issa, D., Fatih Demirci, M. and Yazici, A., 2020. Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, p.101894.
[7] https://www.kaggle.com/datasets/ejlok1/cremad
[8] https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess
[9] https://www.kaggle.com/code/shivamburnwal/speech-emotion-recognition/notebook