Speech and Emotion Recognizer Using Deep Learning and Librosa

Vaijayanti Deshmukh[1], Kavisha Ghodasara[2], Jagrut Khetalpar[3], Prof. S. S. Bhong[4]

1,2,3 Students at Smt. Kashibai Navale College of Engineering, Pune
4 Professor at Smt. Kashibai Navale College of Engineering, Pune

---

Abstract - Speech and Emotion Recognition (SER) means to identify the underlying emotion by extracting the features of a human from his/her voice. Emotion recognition is a part of speech recognition. SER has many applications in today's world. Its applications can be found in healthcare, education, security, telecommunication industries, Human Computer Interaction (HCI), etc. Through this analysis it can be found whether a customer is unhappy or satisfied. SER is an open research area and there are many algorithms available which classify these voice signals and detect the emotions in it. Also what features influence the detection or change in a voice are yet uncertain. Here in this paper, techniques like Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) and a feature named Mel-frequency Cepstrum Coefficient (MFCC) are used. Various python libraries to extract the emotion out of the speech signal provided are utilized. Various datasets are used namely Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Surrey Audio-Visual Expressed Emotion (SAVEE).

Key Words: Speech Emotion Recognition, Convolutional Neural Networks, Deep Learning, Python, Librosa

1. INTRODUCTION

The main aim of this project is to correctly classify the emotions from a speech signal provided as input. The model consists of various stages. It undergoes preprocessing stage, feature extraction stage then classification, and at the last stage emotion is recognized as the output. Speech signals are given as input to the model. Feature extraction, feature enhancement and selection are performed using CNN and LSTM. Neural Network based classification is performed thus giving us the desired output which is the emotion that is to be predicted.

2. Literature Survey

In the field of Speech Emotion Recognition, numerous works on Emotion Recognition, Speech Recognition, Neural Networks and Machine Learning are available. In the work discussed in [1] Bi-directional LSTM is used along with CNN, as bi-directional LSTM used to hold the temporal data for part-of-speech (PoS) tagging and CNN to extract the potential features. The author used ADAM (Adaptive Moments Estimation) to observe the outcomes of different learning rates on the input parameter.

In [2] the author uses the Berlin database to observe the accuracy of a presented model under various features, namely, MFCC, prosodic, LSP and LPC (Linear Predictive Coding) features. They use a formerly trained CNN model, CNN ResNet34 as the proposed model. There are other three models, ANN (Artificial Neural Networks), kNN, GMM (Gaussian Mixture Modeling) that are used to make comparisons with regards to performance based on the different features as mentioned above. They perform ROC analysis under diverse features to conclude that the CNN model performs best on an average when measures for accuracy, sensitivity and specificity are taken into consideration.

In [3], a framework for Speech Emotion Recognition is used which includes RBFN (Radial Basis Function Network) similarity measurement for clusters. It is passed into the CNN model to extract prominent features, and further to the BiLSTM (deep bi-directional long short-term memory) for ultimately predicting the final state of the emotion. The author used various datasets such as IEMOCAP, RAVDESS and EMO-DB for improving the accuracy of the proposed model. [4] tells us about how the authors propose bagged ensemble learning for speech emotion recognition instead of single estimators. They use the RAVDESS, EMO-DB and IITKGP-SEHSC (Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus) datasets for this purpose. Feature extraction was performed for the audio files, and the algorithm proposed consists of bagging, which involves using several estimators to work on multiple varied subsets of the data at hand. This provides variability and reliability. The results on the three datasets are observed for three scenarios - MFCCs and Spectral Centroid features with feature selection, MFCCs and Spectral Centroid features

without feature selection and MFCCs with feature selection.

In [5] the author attempts to extract multiple features, other than MFCCs, to feed as an input to the CNN model. The result is an architecture that consists of extracting MFCC, chromagram, mel-scale spectrogram, Tonnetz representation, and spectral contrast features from the audio files. These features are then stacked and provided as input to the model. The accuracy of the model also depends on the order of stacking of the features. The RAVDESS, IEMOCAP and EMO-DB datasets are used to determine accuracy and it varies slightly for each dataset. The IEMOCAP and EMO-DB datasets are used in [6] to be used for the proposed model based on Dilated Convolutional Neural Network (DCNN). This enables extracting multiple salient features parallelly to predict the emotion of speech with a better accuracy.

In [7], a new approach is presented for SER as the previous methods discussed were unable to perform well in noisy conditions. In this paper the method that is used is modeling on the base of multi agent cognitive architecture. It involves multivariate analysis of signals. The model consists of various stages or levels where in the first layer of the architecture called previous recognition, is the system of agents detecting signal acoustics. The next level is subconscious recognition where the keywords from the previous layer are grouped into meaningful objects and actions. However, the main limitation of this method is that it cannot perform exceptionally well in a loud noisy environment. Various feature extraction algorithms to improve the speech emotion recognition rate are discussed in [8].

MFCC, Discrete Wavelet Transform (DWT), pitch, energy and Zero crossing rate (ZCR) algorithms are used for extracting features. Parameters like pitch, energy, standard deviation, skewness and kurtosis are taken into consideration. [9], [10] give us the reviews of various speech emotion recognition techniques using deep learning. These include various classifiers like Bayes, KNN, GMM, HMM, SVM, ELM and PCA. The techniques used are Deep Boltzmann Machine (DBM), Recurrent Neural Network (RNN), Recursive Neural Network (RvNN), Deep Belief Network (DBN), CNN, Auto Encoder (AE), etc. [11] The framework uses the dynamic entropy of quantitative Electroencephalogram (EEG) to extract continuous entropy values that change with time for EEG signals to achieve emotion recognition irrelevant to the subject. Paper presented a subject independent emotion recognition system that is suitable for real time operation and has the ability to recognize the actual emotional state. It is used to achieve good results but manual feature extraction and early level fusion leads to features to reduce redundancy. In [12] study introduces a method to design a curriculum for machine learning to maximize efficiency during the training process of DNNs for speech emotion recognition. The study proposed to quantify difficulty level of sentences by relying on results from pre-trained Model of training set. Important parameters are not investigated, including optimum number of difficulty bins and optimum number of epochs.

The above literature survey helped us understand several techniques and methods to implement a speech emotion recognition system, and to gather the merits and limitations of each method. This helped us make an informed decision on which model would help us best predict emotions from speech and would be most suited to our project.

3. PROPOSED ALGORITHM

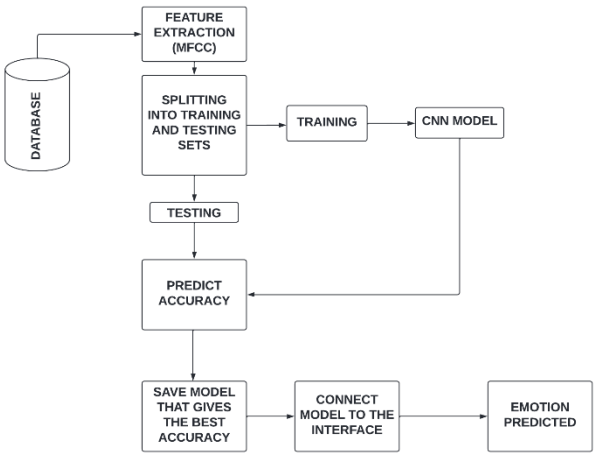

Fig 1: Block Diagram

a. Datasets

The system uses two large datasets from Kaggle namely RAVDESS and SAVEE.

i. RAVDESS

RAVDESS dataset is used in this project. It contains 1440 files that contain audio for the male and female voice, across a range of varied emotions. The speech

emotions encapsulated in these files are sad, happy, neutral, angry, disgust, fearful and surprise [16].

ii. SAVEE

SAVEE database is the other dataset used for effective speech emotion recognition. It consists of 480 audio recordings, made by 4 actors and each consisting of one of the seven emotions that is mentioned above. The resulting recordings are phonetically balanced for each emotion [17]. Increasing the datasets that our model is trained on would result in better accuracy of prediction as the training can be done over a wider range of audio files.

b. **Description of various libraries**

The ease of implementation of a software project is improved by the use of suitable libraries that can be imported. The various Python libraries used to help enhance our project are listed below:

i. Librosa

Librosa is a Python package that is used mainly for audio and music analysis [13]. It is used to load and decode audio as a time-series, which is represented as a one-dimensional NumPy floating point array. It can also be used to plot the waveform and spectrogram of the corresponding audio file, which enables us to visually notice high pitch and low pitch areas in the audio. The MFCC (Mel-frequency cepstral coefficients) features can also be extracted using Librosa. These extracted coefficients greatly help in classifying the emotion of an audio file, hence they're commonly used as features in speech recognition systems.

ii. Keras

Keras is a machine learning library developed by Google. It is an API that can be used on top of TensorFlow, Theano or CNTK (Microsoft Cognitive Toolkit) back-ends, and is used for implementing neural networks[14]. Keras has several in-built methods that can be used to compile a neural network model, fit the model to data, evaluate the network and make predictions.

ii. TensorFlow

TensorFlow is a popular open-source library for high performance numerical computation developed in Google [15]. Its name comes from its purpose of defining and running computations that involve tensors. Tensors are multi-dimensional arrays that have a uniform type and hold either character or data. It has great use in the domain of Machine Learning as it is used for the training and inference of deep neural networks.

c. Speech emotion recognition method used

i. Django

Django is a high-level python web framework. Django is a web application framework written in Python programming language. It is based on the MVT (Model-View-Template) design pattern. Django is very demanding due to its rapid development feature. It takes less time to build an application after collecting client requirements. This framework uses a famous tagline: The web framework for perfectionists with deadlines.

ii. CNN

The term neural network is derived from the human brain or human nervous system which contains a massive number of neurons. These neurons are interconnected which helps us achieve different tasks. The Convolutional Neural Network consists of three basic components namely the convolutional layer, the pooling layer which is optional and the output layer. Each layer is made up of a set of neurons. We give an input to the first convolutional layer, the output of which is obtained through an activation function. Pooling layers are then further added to reduce the number of parameters. Several convolutional and pooling layers are added before the prediction is made. It helps in extracting features. The output layer is a fully connected layer which presents the final output of network.

d. CNN Algorithm Used

1. Import
   required Librosa, Keras, TensorFlow, Scikit-learn and matplotlib libraries
2. Plot the audio file waveform and spectrogram using matplotlib
3. Set labels in an iterable for all the emotions (calm, happy, sad, angry, fearful) to be detected

4. Extract features (Mel Frequency Cepstral Coefficient) of audio files using Librosa
5. Divide the data into respective training and testing datasets in the ratio 80:20
6. Constructing CNN Model with appropriate Convolutional, Pooling, Fully-Connected and Activation layers
7. Fit the training data to the model, and plot model loss, and model accuracy graphs
8. Save the trained model as json in a Hierarchical Data(h5) file
9. Load the saved model
10. Evaluate the model's accuracy on the test data
11. Predict the emotions using the loaded model, and display the predicted and actual values

4. RESULTS

Our model detects emotions in speech with optimum accuracy. Model accuracy is defined as the number of classifications a model correctly predicts divided by the total number of predictions made. Various audio files were used to test the accuracy of the implemented model, the model gave accurate results in most of the cases. The emotions that were predicted were angry, happy, sad, fearful, disgust, calm, neutral and surprised. It involved the use of Django for taking the file input from the user and CNN model in the backend to detect the emotion.

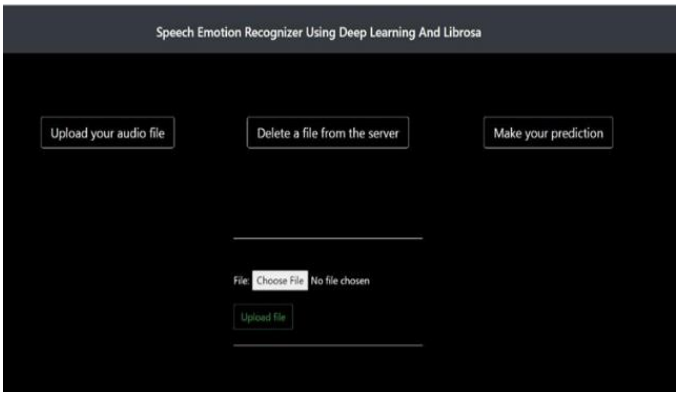

Fig 2 : Main Page

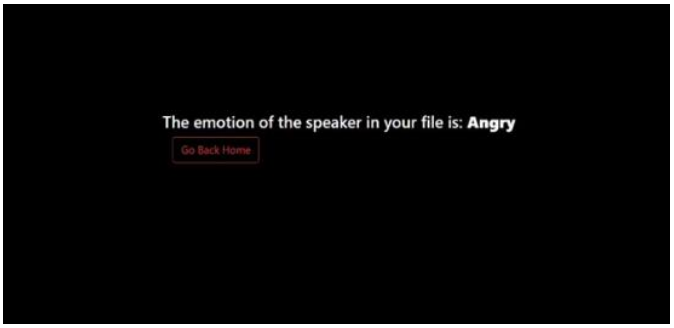

Fig 3 : Output

5. CONCLUSION AND FUTURE SCOPE

Thus, we have developed a speech emotion recognition system using Django rest framework, and Python's machine learning libraries - Librosa, Keras and TensorFlow. It enables users to upload an audio from their device, which can then be used to detect the emotion of the person in the audio file. This has a vast number of applications and can prove to be beneficial in areas like music and movie recommendation systems, virtual assistants and the likes.

If better and more efficient systems for facial emotion detection and speech emotion detection are devised, the overall accuracy of this project will rise. These two systems can be merged for more better accuracy. We can collaborate this model with a robot and/or chat bot on e-commerce websites to help improve customer experience.

ACKNOWLEDGEMENT

The progress and outcome of this project required a lot of guidance and assistance from many people and we are extremely privileged to have got this all along the completion of our project. All that we have done is only due to such supervision and assistance and we would not forget to thank them.

We respect and thank our project guide **Prof. S. S. Bhong,** for his valuable inputs.

We extend our gratitude to **Prof. Dr. R. H. Borhade**, Head of Department, Computer Engineering, for encouragement, and moreover for their timely help and counsel till the completion of our project work. We heartily thank **Prof. Dr. A. V. Deshpande, Principal**,

Smt. Kashibai Navale College of Engineering, Pune for his assistance and encouragement.

REFERENCES

[1] Senthil Kumar, N. K., and N. Malarvizhi. "Bi-directional LSTM–CNN combined method for sentiment analysis in part of speech tagging (PoS)." *International Journal of Speech Technology* 23, no. 2 (2020): 373-380.

[2] Jermsittiparsert, Kittisak, Abdurrahman Abdurrahman, Parinya Siriattakul, Ludmila A. Sundeeva, Wahidah Hashim, Robbi Rahim, and Andino Maseleno. "Pattern recognition and features selection for speech emotion recognition model using deep learning." *International Journal of Speech Technology* 23, no. 4 (2020): 799-806.

[3] Sajjad, Muhammad, and Soonil Kwon. "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM." *IEEE Access* 8 (2020): 79861-79875.

[4] Bhavan, Anjali, Pankaj Chauhan, and Rajiv Ratn Shah. "Bagged support vector machines for emotion recognition from speech." *Knowledge-Based Systems* 184 (2019): 104886.

[5] Issa, Dias, M. Fatih Demirci, and Adnan Yazici. "Speech emotion recognition with deep convolutional neural networks." *Biomedical Signal Processing and Control* 59 (2020): 101894.

[6] Kwon, Soonil. "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach." *Expert Systems with Applications* 167 (2021): 114177.

[7] Nagoev, Zalimhan, Larisa Lyutikova, and Irina Gurtueva. "Model for Automatic speech recognition using multi-agent recursive cognitive architecture." *Procedia computer science* 145 (2018): 386-392.

[8] Koduru, Anusha, Hima Bindu Valiveti, and Anil Kumar Budati. "Feature extraction algorithms to improve the speech emotion recognition rate." *International Journal of Speech Technology* 23, no. 1 (2020): 45-55.

[9] Khalil, Ruhul Amin, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. "Speech emotion recognition using deep learning techniques: A review." *IEEE Access* 7 (2019): 117327-117345.

[10] Roopa, S. Nithya, M. Prabhakaran, and P. Betty. "Speech emotion recognition using deep learning." *Int. J. Recent Technol. Eng* (2018).

[11] Zhang, Yong, Cheng Cheng, and Yidie Zhang. "Multimodal emotion recognition using a hierarchical fusion convolutional neural network." *IEEE Access* 9 (2021): 7943-7951.

[12] Lotfian, Reza, and Carlos Busso. "Curriculum learning for speech emotion recognition from crowdsourced labels." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, no. 4 (2019): 815-826.