Sruthi Shivaramakrishnan

*Electronics Department, SPIT*

*Andheri (W), Mumbai, India*

sruthi.shivaramakrishnan@spit.ac.in

Pushkar Sutar

*Electronics Department, SPIT*

*Andheri (W), Mumbai, India*

pushkar.sutar@spit.ac.in

Darshan Tater

*Electronics Department,SPIT*

*Andheri(W), Mumbai, India*

darshan.tater@spit.ac.in

***Abstract ‒* This paper describes the development of a Python project model to recognize emotion from speech. Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. The employees of call center companies recognize customers’ emotions from speech, so they can improve their service and convert more people. In this way, they are using speech emotion recognition. This project includes an important role of Machine Learning by the testing and training operation of the algorithm.**

**This project presents a deep learning classifier able to predict the emotions of a human speaker encoded in an audio file. The classifier is trained using different datasets - RAVDESS and TESS which will be including multiple sample speech and song datasets rated on emotional validity, intensity and genuineness.**

***Keywords:-* SER, Machine Learning,Deep Learning, Python, RAVDESS dataset**

Ⅰ. Iɴᴛʀᴏᴅᴜᴄᴛɪᴏɴ

Human emotion is a very powerful tool in analysing and tapping various aspects of human feelings. Understanding human emotions can help in providing solutions to various human problems. This technology of recognising human emotions is used in various fields such as the automation industry, health care, video game testing,music industry, etc.

Various algorithms that are available for recognising human emotions, work on face detection, i.e detection of emotions through facial expressions. Most of the algorithms detect 7 basic emotions of anger, joy, fear, sadness, surprise , neutral and disgust.

The facial emotion detector results can be inaccurate at times. Facial expressions do not necessarily show the emotion or the mood of the person at the specific

time. Some emotion detectors use bimodular techniques that combine both speech as well as facial expressions to detect the emotions.

When the emotion lasts for a longer period of time it can be categorized as the mood of the person. Most of the

detectors used mostly identify the emotion or the feeling at a particular time.

Analysing speech signals can produce more accurate results as the duration for which the data is processed for analysing the mood is relatively shorter and accurate. This particular algorithm uses MLP Classifier. Unlike the Naive Bayes, MLPClassifier has an internal neural network for the purpose of classification. This is a feedforward ANN model. It is accurate up to 70%.

Ⅱ. Lɪᴛᴇʀᴀᴛᴜʀᴇ ʀᴇᴠɪᴇᴡ

This chapter gives a brief summary of all the researches done till now in the area of emotion recognition. It summarises the idea and the technology used to create the product. Consequently, it supplements the problem statement of the project which recognises the emotion through speech and it’s added advantages.

1. *HUMAN EMOTION DETECTION*

*THROUGH FACIAL EXPRESSIONS FOR COMMERCIAL ANALYSIS [2]*

This paper deals with the recognition of human emotion through facial expressions.The proposed work uses an Artificial Neural Network (ANN) which combines human expression and its corresponding emotion while viewing a commercial .

ANN also classifies the image set into the seven fundamental emotions. This algorithm was tested o Japanese Female Facial expression database.

# *EMOTION RECOGNITION THROUGH MULTIPLE MODALITIES: FACE, BODY GESTURE, SPEECH [3]*

This paper deals with the recognition of human emotions through facial expressions, body movement and gestures and speech.

This model was tested using a Bayesian classifier. It detects eight expressions and uses a multimodal approach. It was able to provide 10% better results at the decision level.

# *GENDER SPECIFIC EMOTION RECOGNITION THROUGH SPEECH SIGNALS [4]*

This paper aims at recognising the emotions of humans from their speech signals. It consists of Gender recognition and emotion recognition.

It can detect the 6 basic emotions with full accuracy. It uses the features of human speech like the pitch, frequency,etc to determine the results. It uses Naive Bayes’s approach for the algorithm.

Ⅲ. Mᴇᴛʜᴏᴅᴏʟᴏɢʏ

The three broad classes of recognizing emotions through speech audio signals are lexical features (vocabulary and words used), visual features (expressions and hand gestures) and acoustic features (pitch, tone, loudness etc.). However implementation using lexical features would require text transcripts which introduces a preceding requirement of speech to text conversion. Visual analysis would not offer real-time recognition of emotions. Hence, we choose to analyse acoustic features in this work.

Further, classification of emotions can be done in two ways,  
1) *Discrete classification* : classifying emotions with discrete labels like anger, sadness, joy, etc.  
2) *Dimensional representation* : classifying emotions on a scale of low to high like energy or dominance.

Both the approaches have certain limitations. Discrete classification is straightforward but it does not give a proper interpretation of speech signals and often gives a relatively lesser accuracy. Dimensional classification is more elaborate but it is much harder to implement and to calibrate the speech signals from low to high. We will use discrete classification in this implementation for lack of calibrated speech data in ordinary datasets.

1. *Libraries and Dependencies*

This work uses Python v3.5 in a miniconda environment along with a few open source libraries offered by conda. The libraries used are mentioned below.  
*1) Librosa* : It is an open source python package used for music creation and retrieval applications.  
*2) Numpy* : Used for high-level mathematical computations.

*3) Scikit-learn* : Predictive data analysis for machine learning in python.

1. *Dataset Description*

Using the right dataset is the most important step for building the learning algorithm. The data used in this project is based on the RAVDESS dataset.

RAVDESS dataset has 2452 audio files with 12 male speakers and 12 female speakers. The emotions used in this dataset are neutral, calm, happy, sad, angry, fearful, disgusted, surprised. However we will only focus on four emotions to observe viz. calm, happy, fearful, disgusted. The lexical features are kept constant as only 2 statements of equal length are spoken by 24 actors in above 8 emotions.

1. *Feature Extraction*

Features supported by librosa are Mel-frequency cepstral coefficients (MFCC), contrast and Mel-spectrogram

frequency(MEL). We read audio data from the dataset

using the default sample rate. For each feature we make a call to the corresponding librosa function and store it in a numpy array.  
 MFCCs are obtained by taking fourier transform of an excerpt of the audio input signal and representing it on the mel scale. The final coefficients are obtained by taking the amplitude of discrete cosine transform of the resulting spectrum.The output of the whole process is shown in Fig.1.

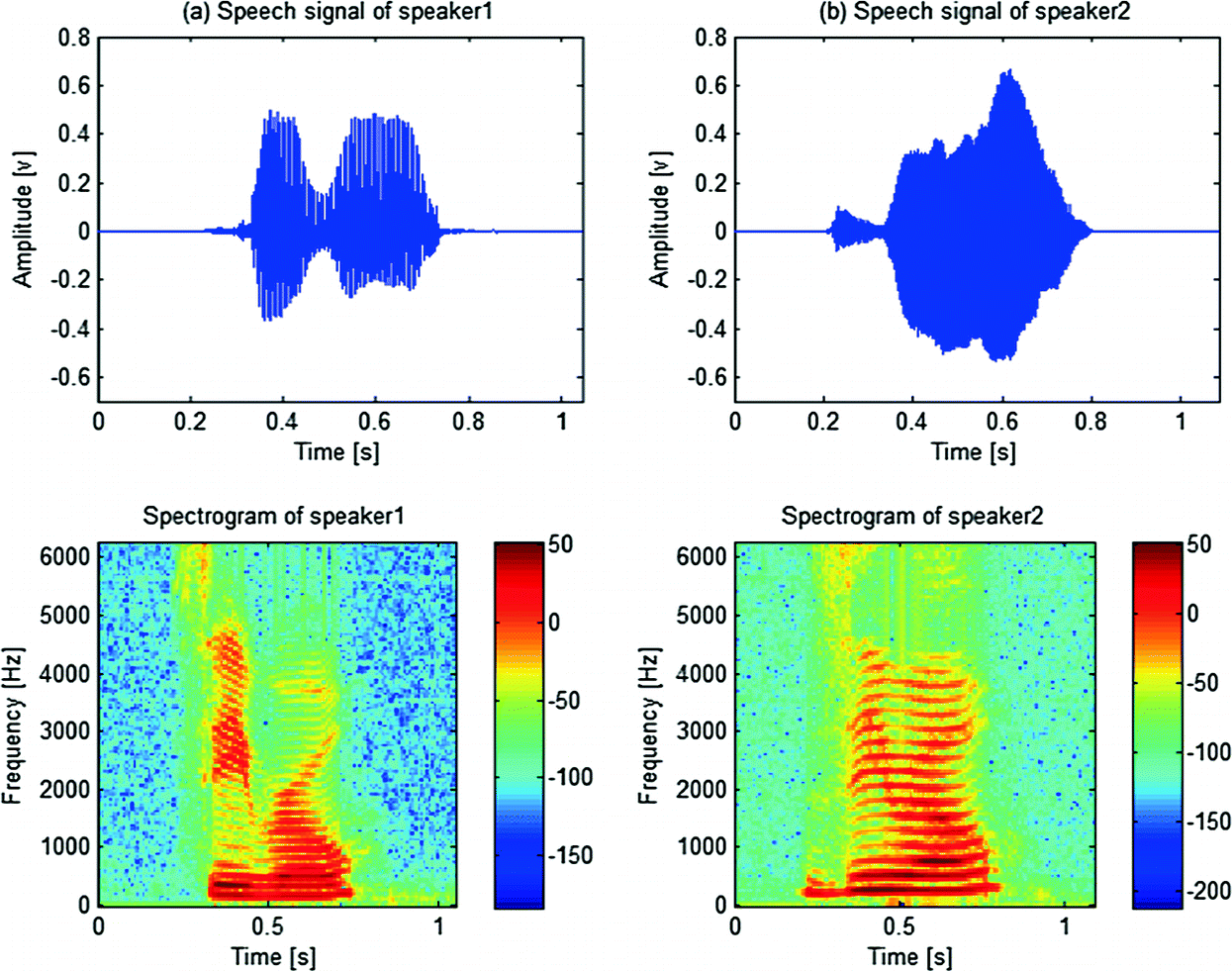


Fig. 1 MFCC feature extraction spectrogram output

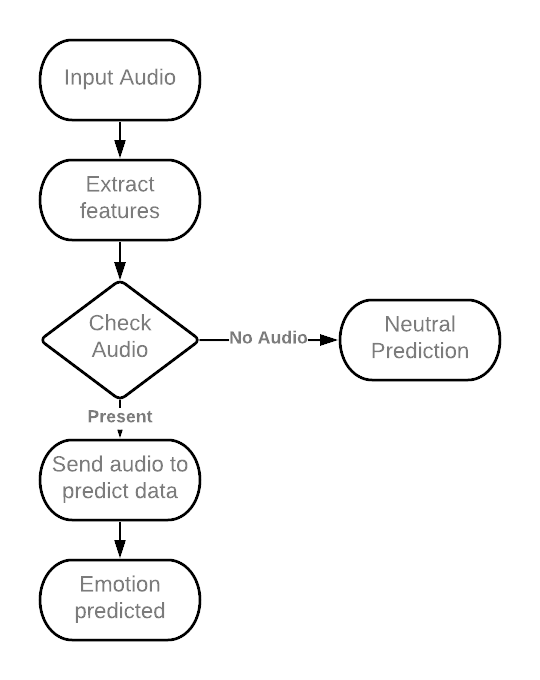
1. *Implementation of the Model*

We implement the model using a Multi-layer Perceptron (MLP) classifier to connect to the neural network. Unlike other classification algorithms , MLP has a built-in neural network for classification. We use this to optimize the extracted feature using stochastic gradient descent. This classifier is particularly useful as it is efficient for time series based data, in our case the audio dataset used for emotion prediction.

We fit and train the dataset by feeding the input to the classifier. Particularly the three extracted features are passed together to MLP as a single feature is not enough to make a prediction. Passing the features separately or together gives different results.

The RAVDESS dataset is split into the ratio of 0.75:0.25 to train the model. 75% of the dataset is used as a training set and the rest is used for a testing set. We first have to complete the training process, then we use the remaining 25% in order to predict the emotions correctly.  
 The accuracy of the model is then determined using the functions imported from the Scikit-learn library. The output can then directly be printed as the data obtained from MLP classifier and Scikit library.

The Fig. 2 shows the basic workflow of the training model.

  
Fig 2. Training Model Workflow

The remaining 25% of the data is passed to the testing set for prediction as shown in Fig. 3.

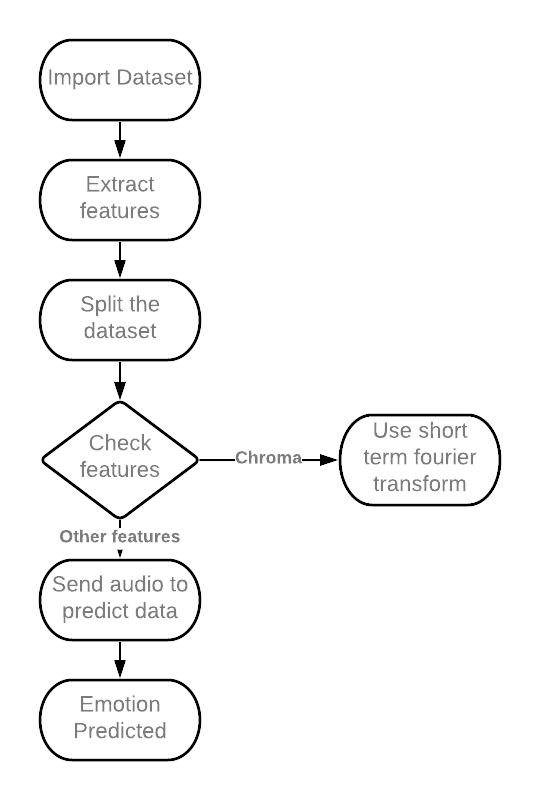


Fig 3. Testing Model Workflow

IV. Rᴇsᴜʟᴛ

The result of the model is based on the accuracy metrics of model evaluation in machine learning. We compare the results obtained with the actual values from the RAVDESS dataset. In MLP classifiers the actual result returns the subset accuracy. If the entire set of predicted emotions exactly match the true values in the dataset the accuracy is 100%. The model is trained on a training dataset using the same sampling rate and lexical features. We find that a total of 180 features are identified by the model and accuracy is 70.24%. The Fig. 4 shows the output of the training model.

However, it is important to note that using raw audio to predict emotions would often produce undesirable output and hence pre-processing the audio becomes necessary.

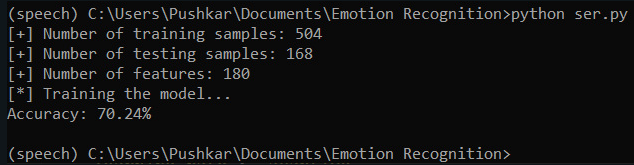


Fig. 4 Output of training model

V. Cᴏɴᴄʟᴜsɪᴏɴ

Through this project, we showed how we can leverage Machine learning to obtain the underlying emotion from speech audio data and some insights on the human expression of emotion through voice. This system can be employed in a variety of setups like Call Centre for complaints or marketing, in voice-based virtual assistants or chatbots, in linguistic research, etc.

VI. Fᴜᴛᴜʀᴇ ᴡᴏʀᴋ

A few possible steps that can be implemented to make the models more robust and accurate are the following**:**

* An accurate implementation of the pace of the speaking can be explored to check if it can resolve some of the deficiencies of the model.
* Figuring out a way to clear random silence from the audio clip
* Exploring other acoustic features of sound data to check their applicability in the domain of speech emotion recognition.
* Following lexical features based approach towards SER and using an ensemble of the lexical and acoustic models. This will improve the accuracy of the system because in some cases the expression of emotion is contextual rather than vocal.
* Adding more data volume either by other augmentation techniques like time-shifting or speeding up/slowing down the audio or simply finding more annotated audio clips.

VII.Rᴇғᴇʀᴇɴᴄᴇsᴇʀᴇɴᴄᴇs

[1] “*Data flair training: python speech emotion recognition*”, DataFlair, 2021 [Online], Available:<https://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/>

[2] L. Z. Ruiz, R. P. V. Alomia, A. D. Q. Dantis, M. J. S. San Diego, C. F. Tindugan and K. K. D. Serrano, "Human emotion detection through facial expressions for commercial analysis," 2017IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), Manila, Philippines, 2017, pp. 1-6, doi: 10.1109/HNICEM.2017.8269512.

[3] Castellano G., Kessous L., Caridakis G. (2008) Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech. In: Peter C., Beale R. (eds) Affect and Emotion in Human-Computer Interaction. Lecture Notes in Computer Science, vol 4868. Springer, Berlin, Heidelberg. <https://doi.org/10.1007/978-3-540-85099-1_8>

[4] Vinay, S. Gupta and A. Mehra, "Gender specific emotion recognition through speech signals," 2014 International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2014, pp. 727-733, doi: 10.1109/SPIN.2014.6777050.

[5] Giannakopoulos T (2015) pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. PLOS ONE 10(12): e0144610. <https://doi.org/10.1371/journal.pone.0144610>

[6] McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. “librosa: Audio and music signal analysis in python.” In Proceedings of the 14th python in science conference, pp. 18-25. 2015.

[7] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. *Array programming with NumPy*. Nature 585, 357–362 (2020). DOI: [0.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2). ([Publisher link](https://www.nature.com/articles/s41586-020-2649-2)).

[8] [Scikit-learn: Machine Learning in Python](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.