EMOTION RECOGNITION THROUGH SPEECH SIGNAL USING PYTHON

M. Aravind Rohan[1],K.Sonali Swaroop[2], ,B. Mounika[3], K. Renuka[4] , S. Nivas[5]
2.Assistant professor 1,3,4,5 UG Research scholars
sonali.k@sreyas.ac.in[2], maravindrohan@gmail.com[1]
Department Of ECE[1,2,3.4,5]
Sreyas Institute Of Engineering & Technology, Hyderabad[1,2,3.4,5]

Abstract-Emotion recognition helps in monitoring and understanding human emotionalstatewhich plays a key role in the current and future computational technologies. Speech emotion recognition is the demonstration to perceive human feeling and emotional states from speech. The emotions considered for the experiments include neutral, joy, and sadness. The emotions of an individual mainly depends on physical characteristics like muscular tension, skin elasticity, blood pressure, heart rate, breath, speech etc. Emotions of a person are unique in nature but their understanding, interpretation and reflections can be distinct. This analysis is done by using python libraries.

Keywords **- Emotion, speech and Python.**

I. INTRODUCTION

Emotions take a crucial role in day-to-day communication[3]. This is necessary as it helps to make wise and tactical decisions. It helps us to study the emotions of a speaker by intercommunicating it with our emotions and giving a report to others. Some researchers have proved that emotion plays an important role in human social interaction.

Human machine interaction is broadly used these days in lots of applications with speech as one of the medium. The fundamental challenges faced by Human-Machine communication are Emotion Recognition from Speech. When two people communicate with each other it is easy to identify their emotion through the nature of their speech[2]. The main aim of this project is to recognize emotion using their speech mechanisms.

It is very crucial to develop a mechanism that recognizes emotionaccurately. Emotion recognition can be achieved by collecting required features from the speech. To improve accuracy the Emotion recognition model must be trained using relatively large number of speech samples from database.
We use python libraries for speech Emotion detection and a large number of databases.

II. LITERATURE SURVEY

There has been research on speech recognition systems. Arti Rawat mentions the usage of a high pass filter before feature extraction which reduces noise.[11]And the features used is Mel Frequency cepstrum coefficient is proposed. Using high pass filter and neural network accuracy is high but execution time is more. Nithya Roopa proposed a method by using deep learning and image classifications to detect emotions with a recognition accuracy of 35.6%.[9] Prudhvi Raj Dachapally proposed two independent techniques[4]. The first one uses autoencoders to construct a unique representation of each emotion, while the secondone is an 8-layer CNN (Convolutional Neural Network). Though the first method did not give the best results, second method with 3-layer CNN gave the accurate results.

III. REVIEW OF DATABASES (SPEECH CORPORA)

To detect emotion, we need speech database. The speech database contains all the samples of required emotions [2][5]. There are three kinds of databases namely:
1)*Artificial speech database*: This kind of database(corpora) is collected by creating an artificial emotion situation to the speaker. This database is very much nearer to the natural database but the only disadvantage is that if the speaker gets to know that it's being recorded the emotions may not be natural.
2) *Actor based speech database*: This kind of speech database gathered from artists who are professionally trained. These databases are available in large number with different types of emotions and collecting them is easy but they are highly artificial in nature.
3) *Natural speech database:* This a completely natural database which is created based on real-world data. This type of database is useful for real-

978-1-7281-7213-2/20/$31.00 ©2020 IEEE

time emotion detection. The disadvantage of this data type is that it has a lot of background noise and all emotions may not be there for detection.

IV. FEATURES USED FOR ANALYSING SPEECH SIGNAL

The most critical part in the analysis is to study the properties or the characteristics of the speech signal. The speech signal is a complex signal to study as it changes continuously in a very shorter period of time. Therefore, it is necessary to segment the given input signal into various parts and analyze them. Practically, the duration of the segment needs to be 20ms as suggested by various researchers[1][8].

Generally, features of speech are of 2 types
1.Qualitative (vocal : harsh, tense, breathy).
2. Spectral (LPC, MFCC, LFPC).
Spectral features like MFCC, pitch and carry emotional information because of their smallest p-values.
Mel Frequency Cepstrum Coefficient (MFCC) is widely used for speech and speech emotion recognition models for their high accuracy [7]. Mel frequency cepstrum is an illustration of short-term power spectrum of sound.In MFCC high frequency resolution with minimum noise can be achieved in low frequency region compared to high frequency region[7]. Linear prediction cepstrum coefficient (LPCC) is most effective representation of speech signal provides the information about the properties of particular signal of any individual person and due to accordance of different emotions this channel characteristic will get change, so emotions in speech can be extracted by using these features.

MFCC are a set of coefficients which represents cosine transformation and logarithmic frequency on Mel scale.

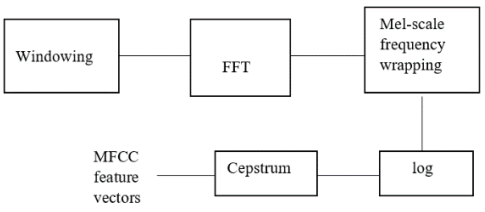

Fig 1: MFCC feature extraction flow graph[4]

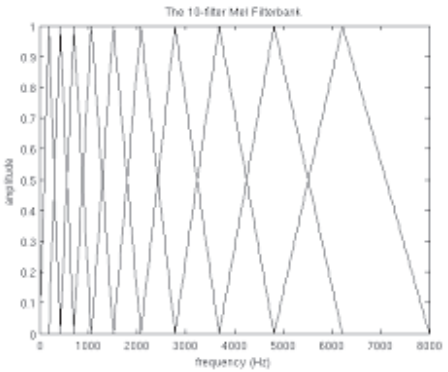

Fig 2: Mel scale[4]

The different types of classifiers can be used to detect the emotions from human speech once the features are extracted. The different classifiers are described below.

SVM: The Support Vector Machines (SVM) is an ML (Machine Learning) algorithm works on group problems. It can be used for both Linear and non-linear type of classification. In linear type of classification, two support vectors are taken and a hyper plane is drawn which is equidistant to the support vectors[12]. This hyper plane is also called as decision boundary. Any thing that falls above the hyperplane is regarded as one group and any thing that falls below hyper plane is considered as another group. For non-linear classification of data, we take another plane namely z-plane and z plane represents the equation of a circle. Through z-plane which is parallel to x-plane we create a hyper plane which separates one group from another.

RF: The Random Forest classifier (RF)is a group tree-based learning algorithm. The Random Forest Classifier is a group of decision trees from arbitrarily chosen subset of preparing set. It totals the votes from various choice trees to choose the last class of the test object.

It is one of the most efficient learning calculations accessible. For some, databases, it delivers a profoundly exact result. It runs effectively on huge databases. It can deal with a huge number of information without variable attenuation. It gives assessments of what factors that are significant in the characterization. It produces an inside fair gauge of the generalization error as the forest building advances.

MLP: Multiple Layer Perception (MLP) has three layers namely input layer, hidden layer and output

layer. The input layer is used for giving inputs. After allocating inputs they are matched to the neurons in hidden layer. The hidden layer is used to increase the accuracy of prediction[6]. The more number of hidden layers the more accurate the model will be and then the hidden layer is mapped to the output which gives the prediction.

KERAS: Keras is a high level neural network API which runs on tensorflow. It has two major models namely sequential model and functional model. Sequential model works like a linear stack of layers[10]. It is used for building simple classification network and encoder decoder models. In this model we treat each layer as a object which leads to next layer. Functional model is a multi-input and multi-output model. Each complex node in this model gets divided into two or more branches. This is a model with shared layers and gives accurate results to 95% of the cases.

GNB: Gaussian Navie Byes classifier is a powerful algorithm for predictive learning. It runs on Bayes theorem. GNB is used for classification and assumes feature for a normal distribution[12].

KNN ALGORITHM:k-Nearest Neighbor is a simple algorithm which is used for prediction purpose. Here K=N, if k=n it chooses the n nearest neighbors and votes them[9]. The neighbor that has the maximum number of votes is taken as the predicted value. It is an effective classifier if the training data is large and it is also robust to nosier databases. The only drawback of this classifier is that for different values of k we get different predictions.

V. EXPERIMENTATION& RESULTS

In this paper, a comprehensive study of all the algorithms mentioned above is performed. The database used is actor based and the emotions considered are happy, sad and neutral. A total of 140 samples were used for training the system. The features extracted from the speech samples are Mel Frequency Cepstrum Coefficients (MFCC).

BLOCK DIAGRAM

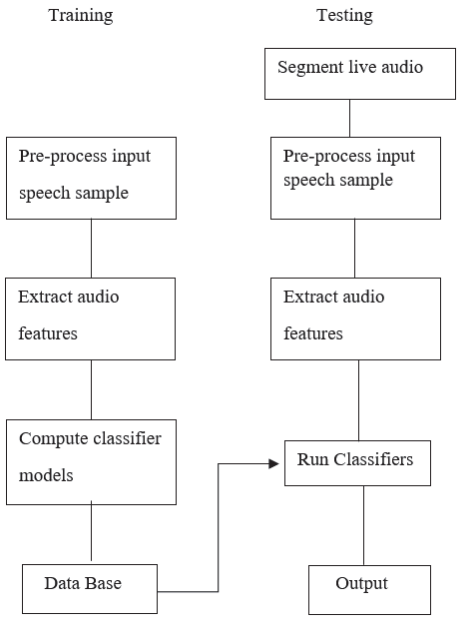

Fig 3: Block diagram of Emotion detection model

The emotion recognition system consists of two phases which are training and testing. The training phase is nothing but the creation of database, Database is a collection of features of the input speech signal. the procedure is as follows:

Initially speech signal is given to the system. As the speech signal is not a stationary signal it must be divided into sub parts called as framing. The typical size of the frame is 20 milliseconds. Once the frames are calculated the features of the signal are extracted. The features which are extracted are classified into modules using the above-mentioned classifiers. This process has to be repeated for large number of speech signals of different emotions. The classified models will form the database

The testing phase consists of preprocessing, feature extraction ,classification of the features and finally the comparison of the testing signal with the database.

The figures of the speech samples with respect to the emotions are shown below.

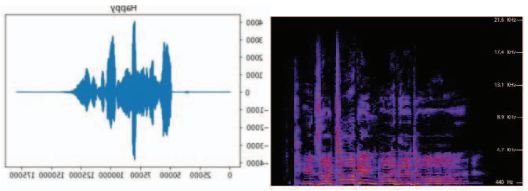

Fig 4: Spectral Samples for the Emotion Happy

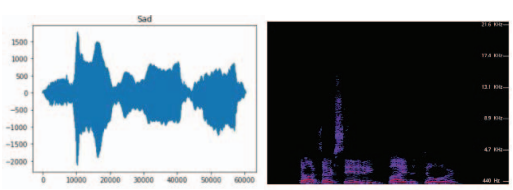

Fig 5: Spectral Samples for the Emotion Sad

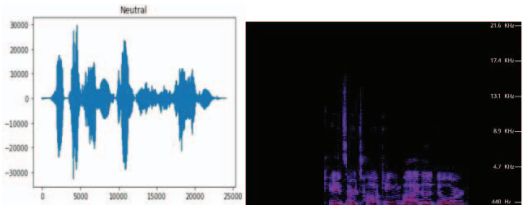

Fig6: Spectral Samples for the Emotion Neutral.

From the figures, the variations of the speech signal spectrum of different emotions is clearly visible. The results of the emotion recognition system using the various classifiers are described in table below

Table 1: Accuracy of the speech emotions as categorized by each classifier.

Emotion Classifier	Neutral	Happy	Sad
RF	92.5	87.5	67.5
SVM	67.5	57.5	87.5
KERAS	75	57.5	87.5
MLP	90	80	42.5
GNB	82.5	82.5	55
KNN	87.5	77.5	47.5

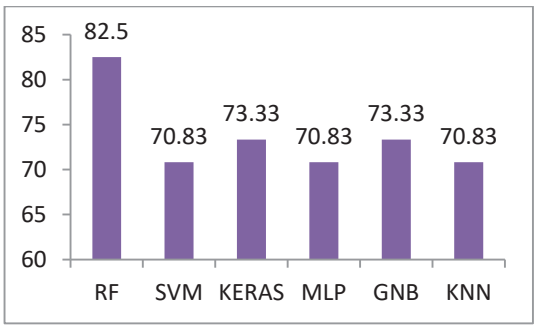

Fig 6: Comparing accuracy levels of different classifiers.

```
print("prediction:", RF.predict("data/validation/Actor_08/03-02-01-01-02-02-08_neutral.wav"))
prediction: neutral
print("prediction:", RF.predict("data/validation/Actor_20/20_02_01_01_kids-talking_happy.wav"))
prediction: happy
print("prediction:", RF.predict("data/validation/Actor_20/20_02_01_01_dogs-sitting_sad.wav"))
prediction: sad
```

Fig 7: Predicting Emotions using RF classifier.

VI.CONCLUSION

This paper represented a comprehensive review on emotion recognition through speech using various python libraries and comparison of various classifiers is also presented. An accuracy of 82% is obtained for RF algorithm. This accuracy can be improved further by an implementation of hybrid classifiers for emotion recognition through speech.

REFERENCES

[1] Anasuya, M.A. and Katti, S.K., 2009. Speech recognition by machine: A review. *International Journal of Computer Science and Information Security*, *6*(3), pp.181-205.

[2] Basu, S., Chakraborty, J., Bag, A. and Aftabuddin, M., 2017, March. A review on emotion recognition using speech. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 109-114). IEEE.

[3] Cen, L., Yu, H.L.Z.L., Dong, M. and Chan, P., 2010. *Machine learning methods in the application of speech emotion recognition*. INTECH Open Access Publisher.

[4] Dachapally, P.R., 2017. Facial emotion detection using convolutional neural networks and representational autoencoder units. *arXiv preprint arXiv:1706.01509*..

[5] El Ayadi, M., Kamel, M.S. and Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), pp.572-587.

[6] Hossain, M.S. and Muhammad, G., 2019. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, *49*, pp.69-78.

[7] Ingale, A.B. and Chaudhari, D.S., 2012. Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE)*, *2*(1), pp.235-238.

[8] Jason, C.A. and Kumar, S., An Appraisal on Speech and Emotion Recognition Technologies based on Machine Learning. *language*, *67*, p.68.

[9] Nithya Roopa, S., Prabhakaran, M. and Betty, P., 2018. Speech emotion recognition using deep learning. *Int J Recent Technol Eng (IJRTE)*, *7*(4S), pp.247-250

[10] Raju, K.S., Govardhan, A., Rani, B.P., Sridevi, R. and Murty, M.R. eds., 2020. *Proceedings of the Third International Conference on Computational Intelligence and Informatics: ICCII 2018* (Vol. 1090). Springer Nature.

[11] Rawat, A. and Mishra, P.K., 2015. Emotion recognition through speech using neural network. *Int. J*, *5*, pp.422-428.

[12] Shah, R.D., Anil, D. and Suthar, C., 2016. Speech Emotion Recognition Based on SVM Using MATLAB. *International Journal of Innovative Research in Computer and Communication Engineering*..