

**ELECTION PREDICTION WITH AUTOMATED SPEECH EMOTION
RECOGNITION
at**

**Sathyabama Institute of Science and Technology
(DEEMED TO BE UNIVERSITY)**

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering Degree in Computer Science and Engineering

By

GangadasuAbhilash

Reddy, M Bharath

(Reg. No.

38110154,38110079)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING
SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600119, TAMILNADU**

APRIL 2020



SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade “A” by NAAC
(Established under Section 3 of UGC Act, 1956)
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI– 600119
www.sathyabama.ac.in



SCHOOL OF COMPUTING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of Gangadasu Abhilash Reddy, M Bharath (**Reg. No. 38110154, 3811007**), who carried out the project entitled “ELECTION PREDICTION WITH AUTOMATED SPEECH EMOTION **RECOGNITION**” under our supervision from November 2019 to April 2020.

Internal Guide

Ms. Brumancia E

Head of the Department

Dr. S. Vigneshwari, M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I Gangadasu Abhilash Reddy, M Bharath (**Reg. No. 3 8 1 1 0 1 5 4 , 3 8 1 1 0 0 7**) hereby declare that the Project Report entitled “**ELECTION PREDICTION WITH AUTOMATED SPEECH EMOTION RECOGNITION**” done by me under the guidance of **Ms.Brumancia E.**, is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala, M.E., Ph.D., Dean, School of Computing and Dr. S. Vigneshwari, M.E., Ph.D. and Dr. L. Lakshmanan M.E., Ph.D. Head of the Department, Department of Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Ms.Brumancia E.**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of **Computer Science and Engineering** who were helpful in many ways For the Completion of the project.

ABSTRACT

Even though artificial intelligence has made tremendous meticulous outcome, there is a gap and large thirst for speech emotion recognition in affective computing. Since machines don't understand the human emotions, natural human computer interaction is struggling a lot to strive and compete with today's market. Survey of past election results and prediction have become crucial and out-dated since the surveys could never clearly depicts the real mental state of an individual. Since the wealth and familiarity of the political people has big title role in many surveys, it is really not trustworthy. No surveys can really understand the truth in the heart of the common people. So, a novel approach is developed for day to day physical and psychological behavior of humans which is influenced by emotions. Other than behavioral expressions, emotion recognition from speech signal is a challenging modality. For the implementation, we take at least 100 audio signals of different area, village, town, district, community, religion and privileged people and we use a nice questionnaire section (20 questions) that could really record the real emotion of the people. MFCC (Mel-Frequency Cepstral Coefficient) feature is extracted and classified with Decision Tree, Random Forest and CNN (Convolutional Neural Network) to give the eight classes of emotions namely Neutral, Disgust, Calm, Angry, Happy, Sad, Fearful and Surprise. Any traditional algorithm can be compared to determine the performance of speech emotion recognition.

TABLE OF CONTENTS

SL.NO	TITLE	PAGE.NO
	ABSTRACT	v
	LIST OF FIGURES	vii
	LIST OF ABBREVIATIONS	viii
1	INTRODUCTION	1
	1.1 About Speech Emotion Recognition	1
	1.2 Classification of emotions	1
	1.3 Existing System	2
	1.4 Proposed System	2
	1.5 System Requirements	3
2	LITERATURE SURVEY	4
3	METHODOLOGY	8
	3.1 Aim	8
	3.2 Architecture	8
	3.3 Pre-processing	9
	3.3.1 Kalman Filter	10
	3.4 Feature Extraction	10
	3.5 Classification of Emotions	12
	3.5.1 Decision Tree	12
	3.5.2 Random Forest	14
	3.5.3 CNN	15
4	Results and Discussion	18
	4.1 Objective of the project	18
	4.2 Speech Emotion Recognition	18
5	Conclusion and Future work	23
	5.1 Conclusion	24
	5.2 Future work	24
	References	25
	Appendix	27
	A. Source Code	27
	B. Plagiarism report and paper	33

LIST OF FIGURES

FIG.NO	TITLE	PAGE.NO
3.1	Block Diagram of Speech Emotion Recognition	9
3.2	Enhanced speech	10
3.3.1	Plotting .wav form of audio signal	11
3.3.2	MFCC Feature	11
4.1	Classification of Emotions using Decision Tree	18
4.2	Classification of Emotions using Random Forest	19
4.3	Classification of Emotions using CNN	19
4.4	Confusion matrix	20
4.5	Model loss and Model Accuracy graph	20
4.6	Precision Scores	21
4.7	Recall Values	21
4.8	F1 scores	22
4.9	Comparison of accuracy	22
4.10	Pie chart of classified Emotions	23

LIST OF ABBREVIATIONS

ABBREVIATIONS	EXPANSION
MFCC	MEL-FREQUENCY CEPSTRAL COEFFICIENT
CNN	CONVOLUTIONAL NEURAL NETWORK
SKLEARN	SCIKIT LEARN
NUMPY	NUMERICAL PYTHON
SER	SPEECH EMOTION RECONITION
RF	RANDOM FOREST
D-TREE	DECISION TREE

CHAPTER 1

INTRODUCTION

1.1 ABOUT SPEECH EMOTION RECOGNITION:

Emotion plays a significant role in understanding and conveying the feelings in day to day life (Male or Female) that are expressed through multiple channels such as linguistic content, voice characteristics, facial expressions, and gestures. There are several ways to recognize human emotions. Some of them are recognition of emotions through speech, facial expressions and physiological signals. In recent years, there are many applications like entertainment, robots, audio surveillance, web-based E-learning for detecting human emotions. Many algorithms like CNN, SVM (Support Vector Machine), GMM (Gaussian Mixture Model), HMM (Hidden Markov Model) are used to classify different types of emotions like anger, fear, joy, sadness, surprise etc. by extracting acoustic features like pitch, MFCC (Mel-Frequency Cepstral Coefficient), frequency, wavelet packet cepstral coefficient.

Emotions can be either positive or negative based on the situation of the person. Researchers believe that speech is the fastest way of communication when compared to other natural methods. Different types of features like MFCC, WPCC are used for extracting spectrograms from speech signal, utterances from speech signals and waveforms from speech signals. By using Automated Speech Emotion Recognition, we can do several researches in the area of emotion recognition, general classification, pattern recognition etc which can also be used in detection of cyber-crimes. Several methods like MFCC extraction, Cost Matrix Method, LPCC (Linear Prediction Cepstral Coefficient) feature extraction methods are used to process speech signals.

1.2 CLASSIFICATION OF EMOTIONS:

Emotion recognition helps to understand the real intensity of the person by classifying the human emotions. In this proposed system, the audio files are taken as input and pre-processed audio files are used to identify acoustic MFCC (Mel-Frequency Cepstral Coefficient) feature. The extracted feature is classified with

Decision Tree, Random Forest and CNN (Convolutional Neural Network) to give the eight classes of emotions namely Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust and Surprise. Any traditional algorithm can be compared to determine the performance of speech emotion recognition.

1.3 EXISTING SYSTEM:

In existing system different types of emotions are recognized by using hybrid of cascaded Gaussian mixture model (GMM), deep neural network (DNN), CNN-BLSTM, SVM algorithms by taking different types of datasets like Emirati speech database, IEMOCAP, SAVEE, EMO-DB. They have extracted the most commonly used features like Mel-frequency cepstral coefficient, WPCC, pitch, energy from the speech signal to detect human emotions by taking several trials using trained models. In some model's utterances to classify emotions and to extract pattern of features and achieved an increase in accuracy for better system performance in the presence of spectrogram, noisy signals and wave models.

1.4 PROPOSED SYSTEM:

In this proposed system, the audio files are taken as input and pre-processed audio files are used to identify acoustic MFCC (Mel-Frequency Cepstral Coefficient) feature. The extracted feature is classified with Decision Tree, Random Forest and CNN (Convolutional Neural Network) to give the eight classes of emotions namely Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust and Surprise. Any traditional algorithm can be compared to determine the performance of speech emotion recognition.

1.5 SYSTEM REQUIREMENTS:

Hardware:

- Hard disk : 500GB and above
- Ram : 4 GB and above
- Processor : I5 and above

Software:

- Windows 8 and above
- Anaconda 3.5.4
- Jupyter notebook

CHAPTER 2

LITERATURE SURVEY

Ismail Shahin et.al (2019) proposed a novel methodology to identify six different emotions using a hybrid cascaded GMM and DNN by using Emirati Arabic - emphasized speech database. They have extracted Mel-frequency cepstral coefficient feature from the speech signal which gives pitch of human auditory system and logarithmic perception of onset to detect the emotions. For extracting MFCC feature speech signals are converted into discrete cosign transform to acquire spectral coefficients. They have achieved 83.97% by using GMM-DNN classifier while using SVM and MLP they achieved 80.33% and 69.78% to get better system performance in the presence of noisy signal they have provided dominant signal mask for hybrid classifier.

Yuan chao li et.al (2019) has classified emotion using spectrogram by extracting features like pitch, MFCC, energy from iemocap database to perform gender classification. They have used the combination of CNN-BLSTM in E2E manner for training the model and achieved an absolute increase in overall accuracy by 7.7% when compared to other traditional algorithms. Spectrogram with deltas and delta-deltas are extracted and given as input to neural network. They have implemented CNN-BLSTM model using keras and optimized by Adam method.

Santiago Pascual et.al (2019) has proposed an improved self-supervised method in which multiple workers like waveform, MFCC, LPS prosody solve different self-supervised tasks which is followed by a single neural coder to extract transferable robust problem agnostic features like speaker identity phonemes emotional cues from speech signal. They discovered robust representation direction from the raw speech waveform using PASE which can be used as pretrained network to avoid trained models from scratch.

John W. Kim et.al (2018) has proposed SVM and EMnet model which combines DNN to extract patterns of features with high level representation and use of common features which is useful for emotion recognition. In this model they have extracted multi dimension feature vectors and speech signal for every 10 msec using

30 msec hamming window. They have used EMO-DB database which contains 535 speech wave files and 10 short sentences labelled with one of the common emotions. They have achieved 88.9% accuracy from compare across 98 trails.

Jilt Sebastian et.al (2019) has investigated the use of long short-term memory (lstm) with pretrained text-based emotion recognition and CNN for emotion recognition with speech by combining fusion techniques. They have used IEMOCAP multi model dataset and adopted various fusion strategies to achieve overall score for each emotion. They have extracted features using CNN from utterance transcriptions by taking 300 dimensional pretrained word embeddings and for each utterance acoustic features are extracted in each stage for building join CNN model using early fusion to improve performance.

Zheng Lian et.al (2018) has proposed a novel method named pair wise discriminative task (PDT) to obtain high level audio vectors and compared with traditional classification task. Spectrogram based self- attentional CNN-BLSTM model is used to classify emotions in E2E manner. A total number of 5531 utterances are used to recognize the emotions. They have extracted acoustic and prosodic features like MFCC, pitch, energy using IEMOCAP database and achieved increase in overall accuracy by 7.7% when compared with other traditional methods.

Wei Jiang et.al (2019) used DNN architecture to extract the feature representations from Heterogeneous acoustic feature groups. They have extracted audio data from video data of IEMOCAP dataset to recognize performance of proposed DNN architecture. To learn discriminative acoustic feature representation to obtain informative features a fusion network is trained jointly. SVM is used to recognize task and achieved 64% when compared to existing approaches.

Zheng Lian et.al (2019) has combined unsupervised representation learning strategy - future observation prediction with transfer learning approaches to verify effectiveness of proposed system. They have conducted experiments to demonstrate the superiority of proposed system on advanced unsupervised learning strategies by using IEMOCAP database. They have tested each configuration 20 times. They have used Adam optimization scheme with a learning rate of 0.01 and

back size of 32 to outperform the main stream named semi-supervised and unsupervised learning strategies above 4% for speech recognition.

Rajeev Ranjan et.al (2019) has filtered audio signal by extracting features like pitch frequency and cepstral coefficients using feature extraction technique and perform comparative analysis using different types of MFCC feature extraction method. They have transformed audio waveform to frequency domain representation for advanced processing and analysis. In this proposed system they have analysed speech in each short frame window to obtain a spectrum by FFT and concluded delta-delta MFCC future extraction technique is better than other feature extraction technique.

Vladimir Chernykh et.al (2017) has proposed a novel approach named deep recurrent neural network with a sequence of trained acoustic features and compared with recent advances they have classified speech by emotion and time properties of emotions. They have used IEMOCAP database which consists 12 hours of audio-visual data. The audio signals are divided into small utterances and each utterance is evaluated from 3 to 4 assessors based on audio and video streams to recognize the emotions. They have compared BLSTM performance in one label approach and CTC approach and predicted the sequence of emotions for one utterance.

Sujay Angadi et.al (2019) has proposed a novel method named CNN-BLSTM to classify human emotions and measure parameters like precision, recall and accuracy. They have extracted spectral and temporal features to improve SER performance. IEMOCAP database is used to classify emotions by comparing CNN-BLSTM method with SVM and achieved an increase in 9.83% accuracy.

Hemantha Kumar Palo et.al (2018) has developed a feature extraction technique to reduce combination features by extracting features like wavelet coefficient, LPCC, MFCC for recognizing five different emotions. They have compared EMO-DB database with SAVEE database to validate accuracy and to recognize common arousal emotions by using different ratios of predefined training, testing and validation data. They have achieved 91.82% and 93.67% accuracy from EMO-DB and SAVEE for first 16 cepstral coefficients using RBFN classifier.

Noushin Hajarolasvadi et.al (2019) has proposed a novel method named k-means clustering to predict the emotion in THE speech signals. They have extracted 88-dimensional vector of audio features like MFCC, Pitch and Intensity of each frame which is splitted by analysis of speech signals. They have achieved 81.05%, 77% and 72.33% by extracting audio features using SAVEE, RML and ENTER05 database by applying SVM classifier.

Yea Hoon Kwon et.al (2018) has proposed a CNN model to classify emotion based on multi-model data and improved accuracy of human emotions. They have combined both Electro Encephalon Gram (EEG) and Galvanic Skin Response (GSR) signals which are pre-processed by Zero crossing rate. They have achieved 73.4% accuracy after noise removed and data processing.

Heekyung Yang et.al (2019) has proposed a novel method multi-column CNN to recognize emotions from Electro Encephalon Gram (EEG) signals. They have used DEAP dataset and circumplex model with balance and arousal to classify emotions and compare accuracies based on voting and weighted sum and also compared with different traditional algorithms.

CHAPTER 3

METHODOLOGY

3.1 AIM:

To predict election results from audio signal by classifying human emotions.

3.2: ARCHITECTURE:

In this project, we have collected dataset in audio format from 100 people with different perceptions, religion and region. Below figure.1 explains the block diagram of the proposed work. In the field visit, we have asked different set of questions to the people. A sample questionnaire is given below.

Sample Questionnaire for the individuals (Takes at least 10 minutes)

1. Do you plan to cast your vote in 2021 Tamil Nadu State election
2. How certain are you that you will vote in the upcoming elections?
3. Which do you identify the most with politically?
4. Do you think the present Government is well enough in decision making.(Justify your answer)
5. Do you think the present Government is well enough in law making. (Justify your answer)
6. Do you think the present Government is well enough in policy formation and implementation of policy. (Justify your answer)
7. Are you satisfied with the public facilities given by the government?
8. Do you think the present Government is well enough to support the youth
9. Are you satisfied with the women welfare schemes supported by the government?
10. What sort of changes you need to uplift the economical status of Tamil Nadu
11. Do you agree with the present facilities given by the government to your village/city/rural areas?
12. Are you satisfied with the Medical facilities given by the Primary Health Centres of your area?
13. Do you think the official leaders are frequently visiting the need and affected areas?
14. Do you think a great gap is made in agriculture area?
15. Do you think still Tamilnadu remains as federal parliamentary democratic
16. Was there a national unrest against any political party? Comment on this
17. Do you think casteism and corruption has given major roles in important sectors
18. Do you think the SC/ ST are happily moving in the state?
19. Do you think the present Government has done their job well enough to deserve re-elections, or do you think it is time to give a new person a chance?
20. Do your family members have the chance to opt for the same political party which you vote?

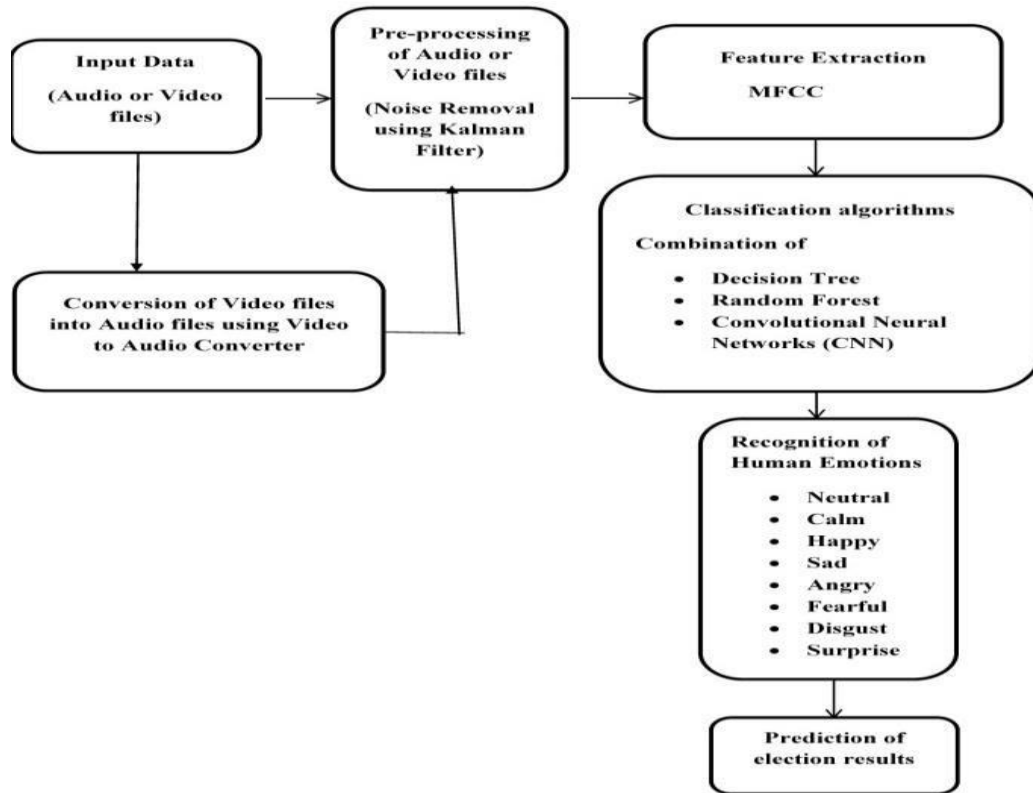


Fig.no. 3.1: Block Diagram of Speech Emotion Recognition

In this audio files have taken as input and pre-processed by using Kalman filter and extracted MFCC feature by using Decision Tree, Random Forest and Convolutional Neural Network(CNN) to classify 8 different classes of emotions like Neutral, Calm, Happy, Angry, Sad, Fearful, Disgust and Surprise to predict election results.

3.3: Pre-processing:

The .wav file is splitted into different chunks each of 3 seconds. Each chunk is fed to the Kalman Filter to remove the background noise. Since the background noise (from fan, vehicles) fluctuates over time, it is better to reduce or remove the noise from the wav file to enhance the speech quality. Figure 2 gives the pre-processing of the audio with Kalman Filter. The outcome of the filter is Enhanced audio signal without any loss of data.

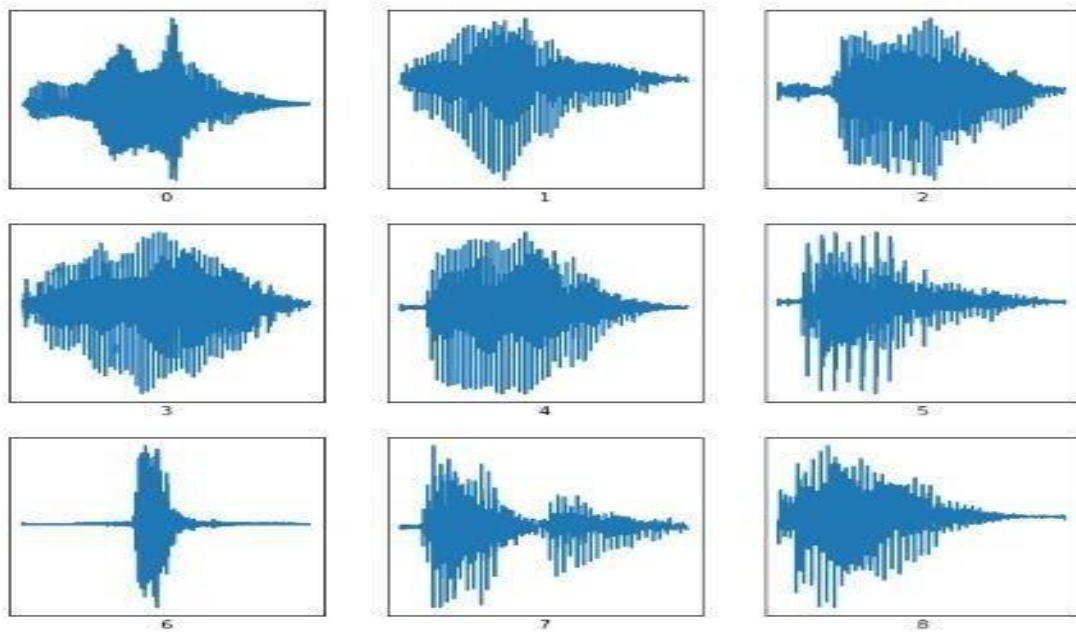


Fig no. 3.3.1: Plotting .wav form of audio signal

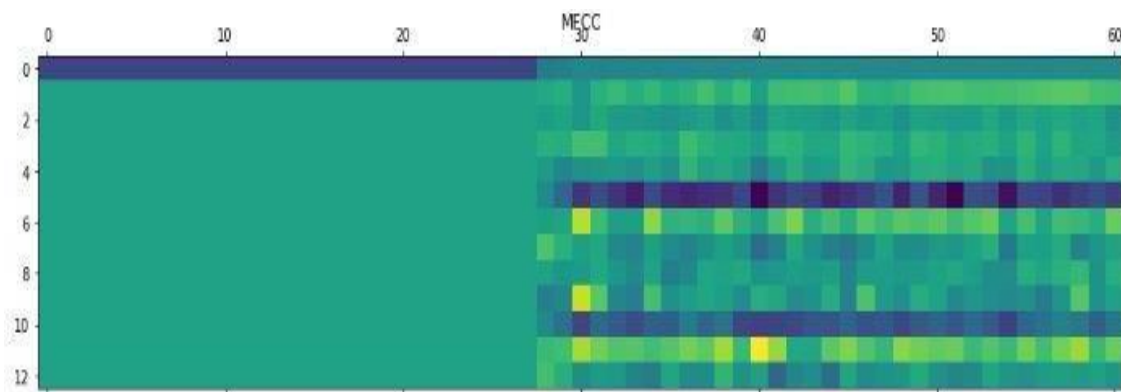


Fig no. 3.3.2 MFCC Feature

The coefficients that made up of Mel-Frequency cepstrum is called Mel-Frequency cepstral coefficient. MFCC are derived from a type of cepstral representation of an audio file equally spaced on Mel scale. In this Fourier transform of signal is taken and maps the power of spectrum that has been obtained on Mel scale along with these logs of power and discrete cosine transform is taken and calculated. The amplitude of the resulting spectrum is MFCC.

3.5 CLASSIFICATION OF EMOTIONS:

Emotion recognition helps to understand the real intensity of the person by classifying the human emotions. In this proposed system, the audio files are taken as input and pre-processed audio files are used to identify acoustic MFCC (Mel-Frequency Cepstral Coefficient) feature. The extracted feature is classified with Decision Tree, Random Forest and CNN (Convolutional Neural Network) to give the eight classes of emotions namely Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust and Surprise. Any traditional algorithm can be compared to determine the performance of speech emotion recognition.

3.5.1 DECISION TREE:

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of three types of nodes:

Decision nodes – typically represented by squares

Chance nodes – typically represented by circles

End nodes – typically represented by triangles

Decision trees are commonly used in operations research and operations management. If, in practice, decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision tree is as a descriptive means for calculating conditional probabilities. Decision trees, influence diagrams, utility functions, and other decision analysis tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or management science methods.

ADVANTAGES:

Among decision support tools, decision trees (and influence diagrams) have several advantages. Decision trees:

1. Are simple to understand and interpret. People are able to understand decision tree models after a brief explanation.
2. Have value even with little hard data. Important insights can be generated based on experts describing a situation (its alternatives, probabilities, and costs) and their preferences for outcomes.
3. Help determine worst, best and expected values for different scenarios.
4. Use a white box model. If a given result is provided by a model.
5. Can be combined with other decision techniques.

DISADVANTAGES:

They are unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree.

1. They are often relatively inaccurate. Many other predictors perform better with similar data. This can be remedied by replacing a single decision tree with a random forest of decision trees, but a random forest is not as easy to interpret as a single decision tree.
2. For data including categorical variables with different number of levels, information gain in decision trees is biased in favor of those attributes with more levels.

3. Calculations can get very complex, particularly if many values are uncertain and/or if many outcomes are linked.

3.5.2 RANDOM FOREST:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method,[2] which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg. An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark. The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho and later independently by Amit and Geman in order to construct a collection of decision trees with controlled variance.

Instead of decision trees, linear models have been proposed and evaluated as base estimators in random forests, in particular multinomial logistic regression and naive Bayes classifiers.

As part of their construction, random forest predictors naturally lead to a dissimilarity measure among the observations. One can also define a random forest dissimilarity measure between unlabelled data: the idea is to construct a random forest predictor that distinguishes the "observed" data from suitably generated synthetic data. The observed data are the original unlabelled data and the synthetic data are drawn from a reference distribution. A random forest dissimilarity can be attractive because it handles mixed variable types very well, is invariant to monotonic transformations of the input variables, and is robust to outlying observations. The random forest dissimilarity easily deals with a large number of semi-continuous variables due to its intrinsic variable selection; for example, the "Addcl 1" random forest dissimilarity weighs the contribution of each

variable according to how dependent it is on other variables. The random forest dissimilarity has been used in a variety of applications, e.g. to find clusters of patients based on tissue marker data.

Advantages of using Random Forest:

As with any algorithm, there are advantages and disadvantages to using it. In the next two sections we'll take a look at the pros and cons of using random forest for classification and regression.

1. The random forest algorithm is not biased, since, there are multiple trees and each tree is trained on a subset of data. Basically, the random forest algorithm relies on the power of "the crowd"; therefore, the overall biasedness of the algorithm is reduced.
2. This algorithm is very stable. Even if a new data point is introduced in the dataset the overall algorithm is not affected much since new data may impact one tree, but it is very hard for it to impact all the trees.
3. The random forest algorithm works well when you have both categorical and numerical features.
4. The random forest algorithm also works well when data has missing values or it has not been scaled well (although we have performed feature scaling in this article just for the purpose of demonstration).

Disadvantages of using Random Forest:

1. A major disadvantage of random forests lies in their complexity. They required much more computational resources, owing to the large number of decision trees joined together.
2. Due to their complexity, they require much more time to train than other comparable algorithms.

3.5.3 CONVOLUTIONAL NEURAL NETWORK(CNN):

In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analysing visual imagery.

They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics. They have applications in image and video recognition, recommender systems, image classification, medical image analysis, natural language processing, and financial time series.

The name “convolutional neural network” indicates that the network employs a mathematical operation called convolution. Convolution is a specialized kind of linear operation. Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers.

A convolutional neural network consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of a series of convolutional layers that convolve with a multiplication or other dot product. The activation function is commonly a RELU layer, and is subsequently followed by additional convolutions such as pooling layers, fully connected layers and normalization layers, referred to as hidden layers because their inputs and outputs are masked by the activation function and final convolution.

Though the layers are colloquially referred to as convolutions, this is only by convention. Mathematically, it is technically a sliding dot product or cross-correlation. This has significance for the indices in the matrix, in that it affects how weight is determined at a specific index point.

When programming a CNN, the input is a tensor with shape (number of images) x (image width) x (image height) x (image depth). Then after passing through a convolutional layer, the image becomes abstracted to a feature map, with shape (number of images) x (feature map width) x (feature map height) x (feature map channels). A convolutional layer within a neural network should have the following attributes:

Convolutional kernels defined by a width and height (hyper-parameters).

The number of input channels and output channels (hyper-parameter).

The depth of the Convolution filter (the input channels) must be equal to the number channels (depth) of the input feature map.

Convolutional layers convolve the input and pass its result to the next layer. This is similar to the response of a neuron in the visual cortex to a specific stimulus. Each convolutional neuron processes data only for its receptive field. Although fully connected feedforward neural networks can be used to learn features as well as classify data, it is not practical to apply this architecture to images. A very high number of neurons would be necessary, even in a shallow (opposite of deep) architecture, due to the very large input sizes associated with images, where each pixel is a relevant variable. For instance, a fully connected layer for a (small) image of size 100 x 100 has 10,000 weights for each neuron in the second layer. The convolution operation brings a solution to this problem as it reduces the number of free parameters, allowing the network to be deeper with fewer parameters. For instance, regardless of image size, tiling regions of size 5 x 5, each with the same shared weights, requires only 25 learnable parameters. In this way, it resolves the vanishing or exploding gradients problem in training traditional multi-layer neural networks with many layers by using backpropagation.

ADVANTAGES OF CNN:

- The usage of CNNs are motivated by the fact that they can capture are able to learn relevant features from an image video at different levels similar to a human brain. This is feature learning ! Conventional neural networks cannot do this.
- In terms of performance, CNNs outperform NNs on conventional image recognition tasks and many other tasks. Look at the Inception model, Resnet50 and many others for instance.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 OBJECTIVE OF THE PROJECT:

To provide a flawless, trustworthy model for prediction of Election results from the speech signal to capture the audio signal in order to measure emotion levels of at least 100 individuals. To extract features like pitch, MFCC, speech rate, formant, frequency, Wavelet packet cepstral coefficient from the audio signal. To classify and determine the classes of emotion(happy, angry, neutral).

4.2 SPEECH EMOTION RECOGNITION:

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.

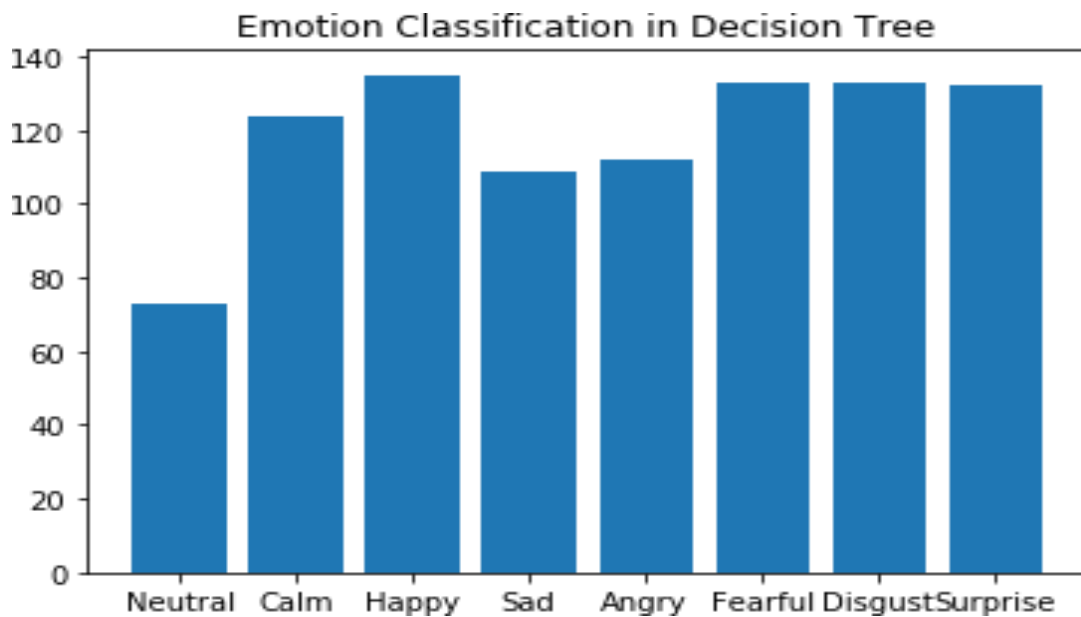


Fig.4.1: Classification of Emotions using Decision Tree

Emotions are classified by extracting the values from Decision tree.

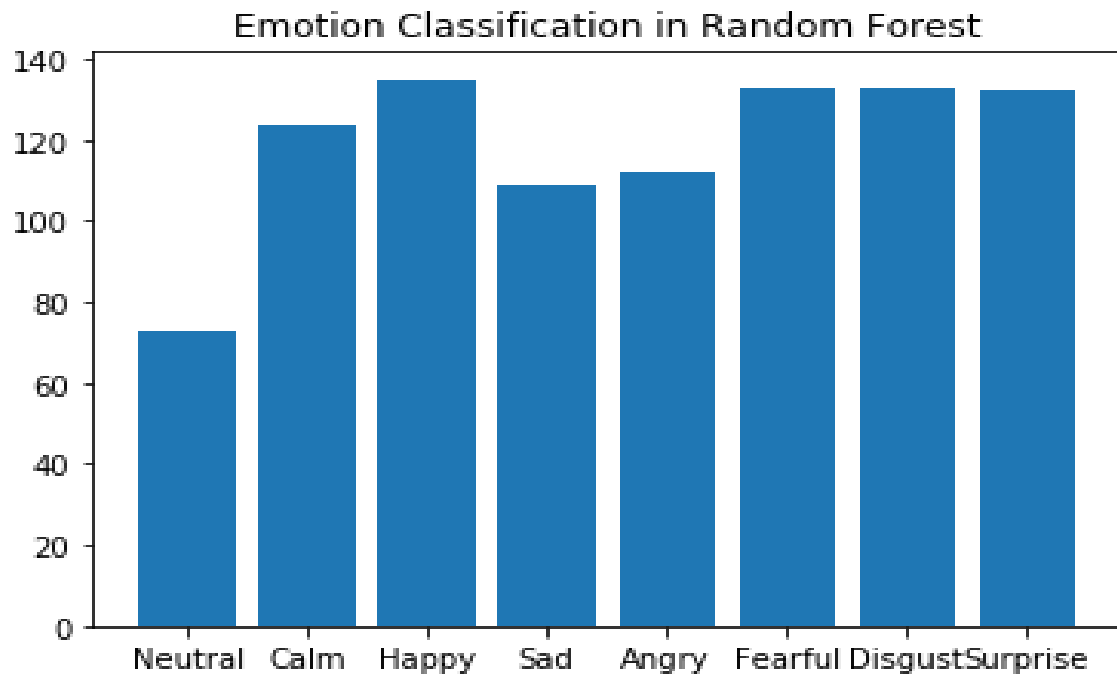


Fig.4.2: Classification of Emotions using Random Forest

Emotions are classified by extracting the values from Random Forest.

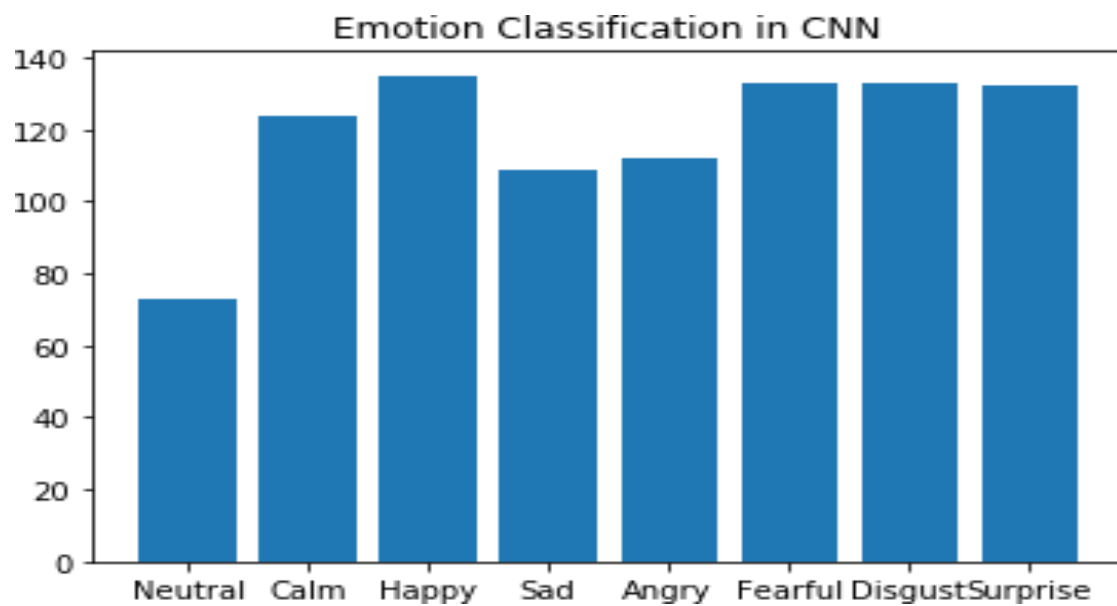


Fig.4.3: Classification of Emotions using CNN

Emotions are classified by extracting the values from CNN.

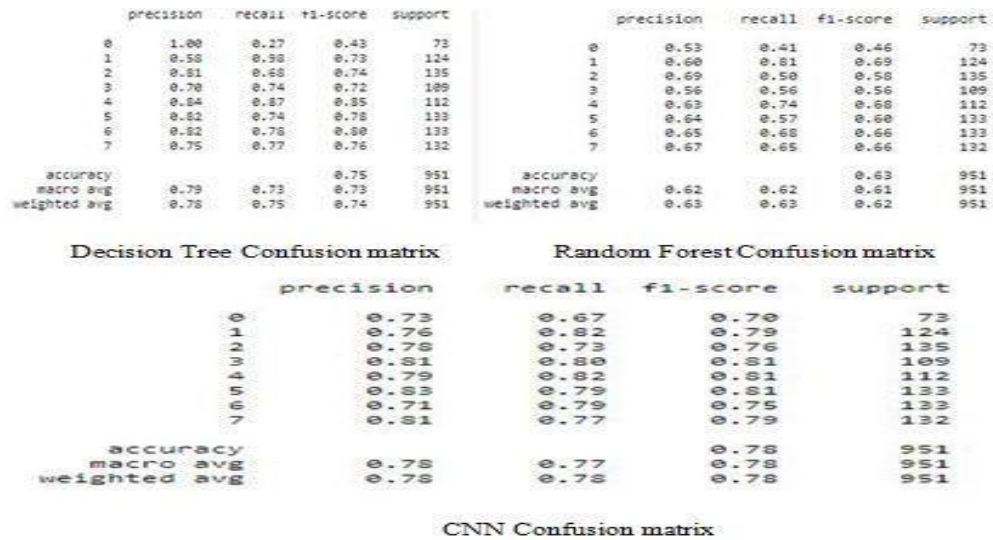


Fig.4.4: Confusion matrix

Confusion matrix has been drawn from the values extracted from input audio signal for Decision tree, Random Forest and CNN. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions.

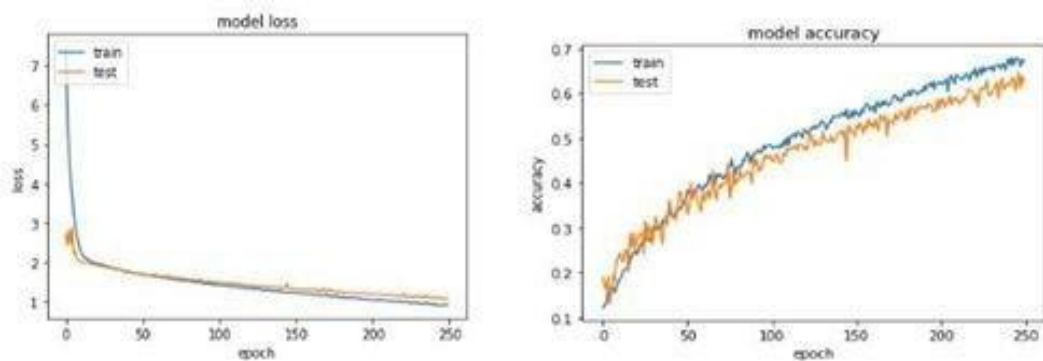


Fig.4.5: Model loss and Model Accuracy graph

Model loss and Model accuracy has been calculated by extracting loss and accuracy values from audio signal.

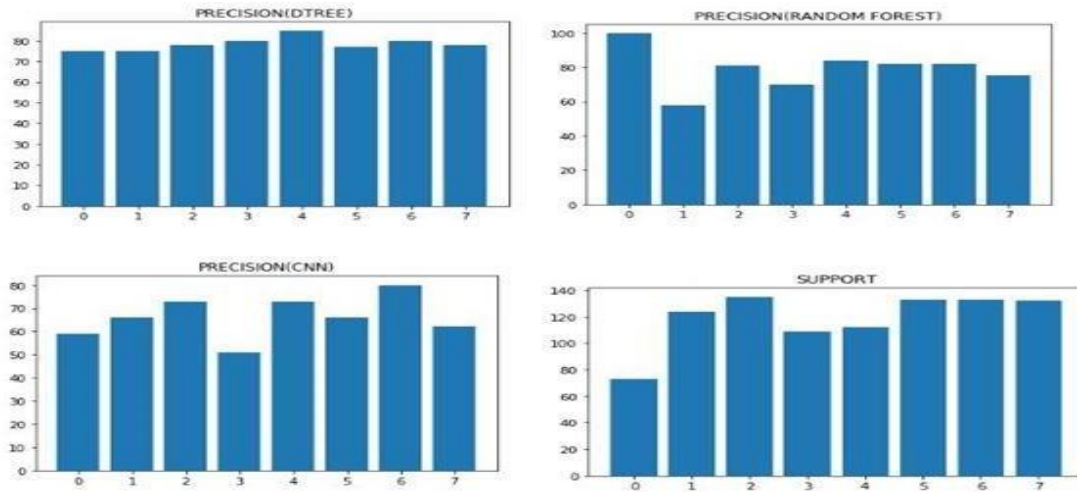


Fig.4.6: Precision Scores

Precision is defined as the fraction of relevant instances among retrieved instances in pattern recognition, informational retrieval and classification of machine learning algorithms. Precision scores of Decision tree, Random forest and CNN were calculated and plotted on the graph.

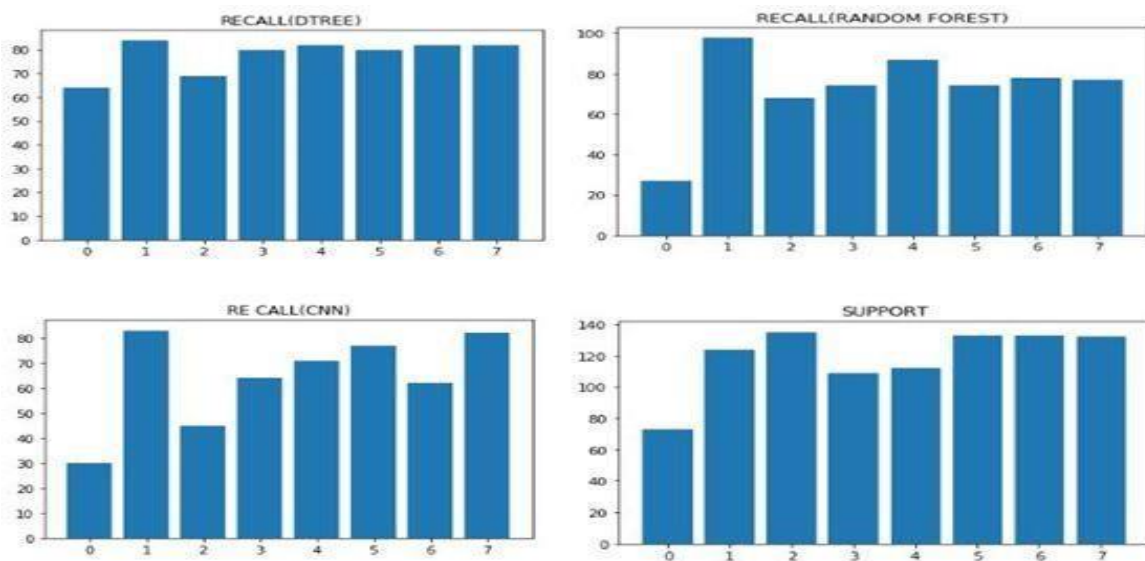


Fig.4.7: Recall values

Recall is defined as the fraction of the total amount of relevant instances that were actually retrieved in pattern recognition, informational retrieval and classification of machine learning algorithms. Recall values of Decision tree, Random forest and CNN were calculated and plotted on the graph.

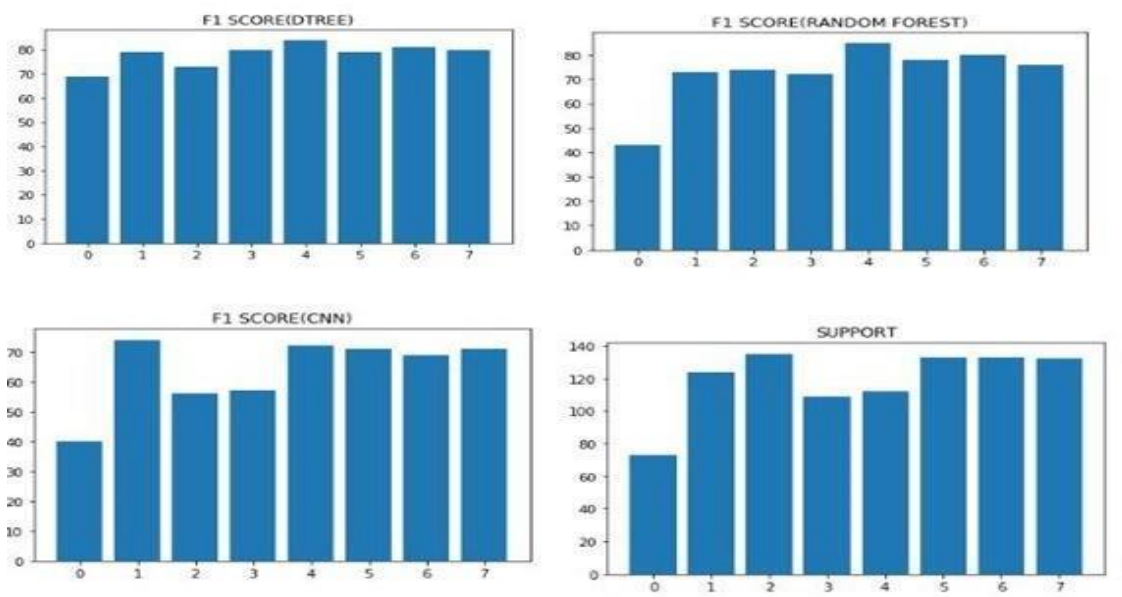


Fig.4.8: F1 Scores

F1 score values of Decision tree, Random forest and CNN were calculated and plotted on the graph. F1 score might be a better measure to use if we need to seek a balance between precision and recall values.

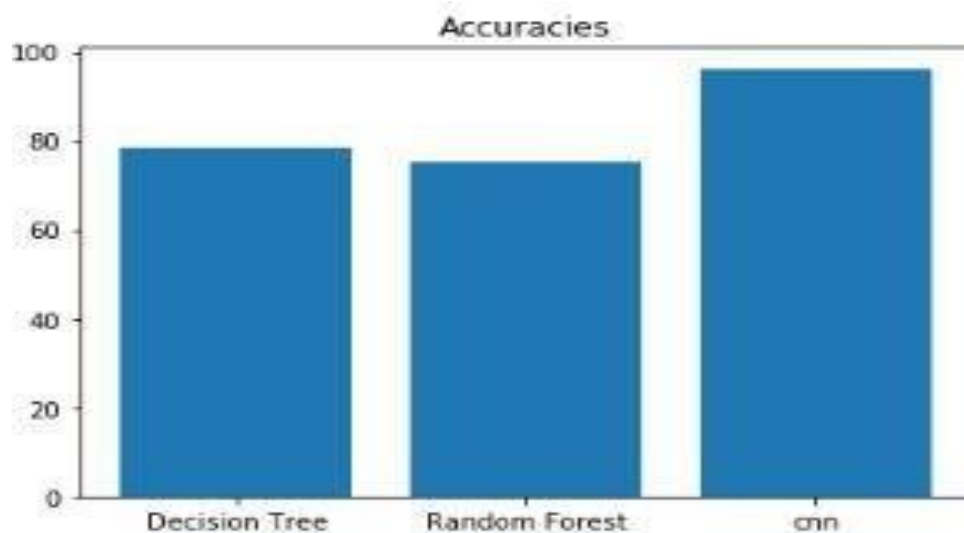


Fig.4.9: Comparison of accuracy

The accuracies calculated from the audio signal using three algorithms has been compared and plotted as graph.

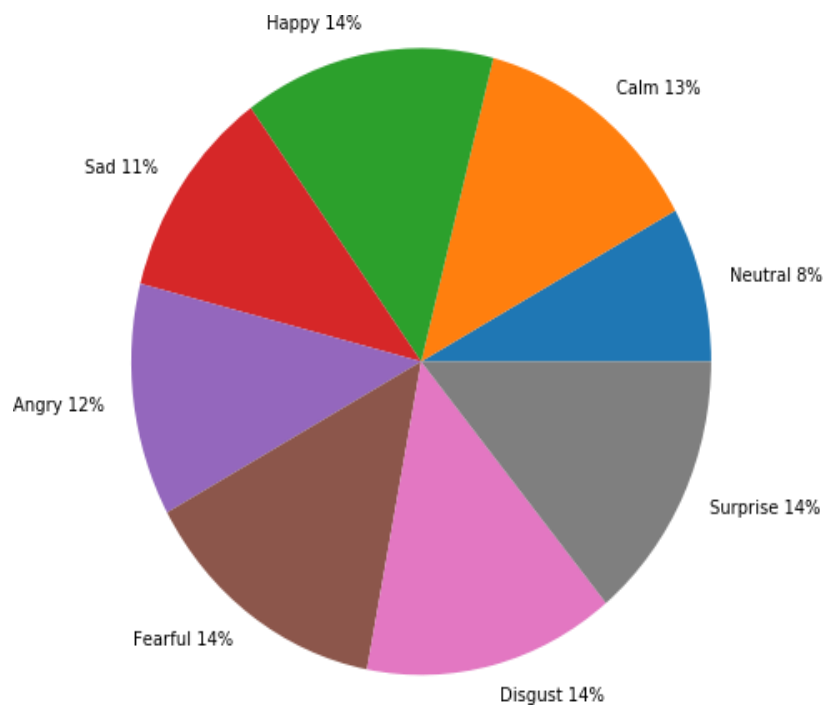


Fig.4.10: Pie chart of classified Emotions

From the above figure we are inferred that 14% of the people are happy, 12% of the people are calm, 8% of the people are neutral, 14% of the people are surprise, 14% of the people are disgust, 14% of the people are fearful, 12% of the people are angry and 11% of the people are sad.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 CONCLUSION:

Our aim is to provide a flawless model for speech emotion recognition by extracting acoustic MFCC feature from the audio signals and classify the classes of emotion like Neutral, Calm, Disgust, Happy, Sad, Angry, Fearful and Surprise using Decision tree, Random forest and CNN (Convolutional Neural Network) to achieve an overall increase in accuracy when compared to all traditional algorithms.

After we acquire the accuracy values of all algorithms, we can use the best algorithm to obtain the emotions from audio signal, then we have calculated eight type of emotions happy, sad, angry, neutral, surprise, calm, fearful, disgust. By this we can predict the Election results.

5.2 FUTURE WORK:

In future we can add more audio and video files and extract more features from the enhanced speech to classify emotions. We can extract a greater number of features like WPCC, LPCC, Filter bank etc. By using different algorithms, we can improve accuracy.

REFERENCES:

1. ISMAIL SHAHIN, ALI BOU NASSIF, SHIBANI HAMSA, Emotion Recognition Using. Hybrid Gaussian Mixture Model and Deep NeuralNetwork,http://www.ieee.org/publications_standards/publications/rights/index.html.
2. Jilt Sebastian, Piero Pierucci, Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts, <http://dx.doi.org/10.21437/Interspeech.2019-3201>.
3. John W. Kim1, Rif A. Saurous, Emotion Recognition from Human Speech Using Temporal Information and Deep Learning, 10.21437/Interspeech.2018-1132.
4. Rajeev Ranjan, Abhishek Thakur, Analysis of Feature Extraction Techniques for Speech Recognition System, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-7C2, May 2019.
5. Sujay Angadi1* Venkata Siva Reddy1, Hybrid Deep Network Scheme for Emotion Recognition in Speech, International Journal of Intelligent Engineering and Systems, Vol.12, No.3, 2019.
6. Vladimir Chernykh, Grigoriy Sterling, Pavel Prihodko, Emotion Recognition from Speech with Recurrent Neural Networks, arXiv:1701.08071v1 [cs.CL] 27 Jan 2017.
7. Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Unsupervised Representation Learning with Future Observation Prediction for Speech Emotion Recognition, arXiv:1910.13806v1 [eess.AS] 24 Oct 2019.
8. Santiago Pascual, Mirco Ravanelli, Joan Serra, Antonio Bonafonte, Yoshua Bengio, Learning Problem-agnostic Speech Representations from Multiple Self-supervised Tasks, arXiv:1904.03416v1 [cs. LG] 6 Apr 2019.
9. Yuanchao Li, Tianyu Zhao, Tatsuya Kawahara, Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning, <http://dx.doi.org/10.21437/Interspeech.2019-2594>.

10. Yea-Hoon Kwon, Sae-Byuk Shin and Shin-Dug Kim, Electroencephalography Based Fusion Two-Dimensional (2D)-Convolution Neural Networks (CNN) Model for Emotion Recognition System, www.mdpi.com/journal/sensors.
11. Heekyung Yang, Jongdae Han, Kyungha Min, A Multi-Column CNN Model for Emotion Recognition from EEG Signals, www.mdpi.com/journal/sensors.
12. Noushin Hajarolasvadi, Hasan Demirel, 3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms, www.mdpi.com/journal/entropy.
13. Wei Jiang, Zheng Wang, Jesse S. Jin, Xianfeng Han, Chunguang Li, Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network, www.mdpi.com/journal/sensors.
14. Zheng Lian, Ya Li, Jianhua Tao, Jian Huang, A PAIRWISE DISCRIMINATIVE TASK FOR SPEECH EMOTION RECOGNITION, www.researchgate.net/publication/322305986_A.
15. Hemanta Kumar Palo, Mihir Narayana Mohant, Wavelet based feature combination for recognition of emotions, <https://www.researchgate.net/publication/330263647>.

APPENDICES

A. SOURCE CODE:

```
! pip install librosa

import librosa

from librosa import display

data, sampling_rate = librosa.load('C:/Users/hp/Desktop/Emotion-Classification-
Ravdess-master/Emotion-Classification-Ravdess-master/data/user 1/chunk0.wav')

%pylab inline

import os

import pandas as pd

import glob

plt.figure(figsize=(12, 4))

librosa.display.waveplot(data, sr=sampling_rate)

import time

path = 'C:/Users/hp/Desktop/Emotion-Classification-Ravdess-master/Emotion-
Classification-Ravdess-master/data'

lst = []

start_time = time.time()

for subdir, dirs, files in os.walk(path):

    for file in files:

        try:
```

```
#Load librosa array, obtain mfccs, store the file and the mfcc information in a new array
```

```
X, sample_rate = librosa.load(os.path.join(subdir,file), res_type='kaiser_fast')  
  
mfccs      =      np.mean(librosa.feature.mfcc(y=X,      sr=sample_rate,  
n_mfcc=40).T,axis=0)
```

```
# The instruction below converts the labels (from 1 to 8) to a series from 0 to 7  
  
# This is because our predictor needs to start from 0 otherwise it will try to  
predict also 0.
```

```
file = int(file[7:8]) - 1  
  
arr = mfccs, file  
  
lst.append(arr)  
  
# If the file is not valid, skip it  
  
except ValueError:  
  
    continue
```

```
print("--- Data loaded. Loading time: %s seconds ---" % (time.time() - start_time))
```

```
# Creating X and y: zip makes a list of all the first elements, and a list of all the  
second elements.
```

```
X, y = zip(*lst)
```

```
import numpy as np
```

```
X = np.asarray(X)
```

```
y = np.asarray(y)
```

```
X.shape, y.shape
```

```
# Saving joblib files to not load them again with the loop above
```

```
import joblib
```

```
X_name = 'X.joblib'
```

```

y_name = 'y.joblib'

save_dir = 'C:/Users/hp/Desktop/Emotion-Classification-Ravdess-master/Emotion-
Classification-Ravdess-master/features'

savedX = joblib.dump(X, os.path.join(save_dir, X_name))

savedy = joblib.dump(y, os.path.join(save_dir, y_name))

# Loading saved models

X = joblib.load('C:/Users/hp/Desktop/Emotion-Classification-Ravdess-
master/Emotion-Classification-Ravdess-master/features/X.joblib')

y = joblib.load('C:/Users/hp/Desktop/Emotion-Classification-Ravdess-
master/Emotion-Classification-Ravdess-master/features/y.joblib')

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33,
random_state=42)

from sklearn.tree import DecisionTreeClassifier

dtree = DecisionTreeClassifier()

DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                        max_depth=None, max_features=None, max_leaf_nodes=None,
                        min_impurity_decrease=0.0, min_impurity_split=None,
                        min_samples_leaf=1, min_samples_split=2,
                        min_weight_fraction_leaf=0.0, presort='deprecated',
                        random_state=None, splitter='best')

dtree.fit(X_train, y_train)

predictions = dtree.predict(X_test)

from sklearn.metrics import classification_report, confusion_matrix

print(classification_report(y_test, predictions))

```

```

from sklearn.ensemble import RandomForestClassifier

rforest = RandomForestClassifier(criterion="gini", max_depth=10,
max_features="log2",

                                max_leaf_nodes = 100, min_samples_leaf = 3,
min_samples_split = 20,

                                n_estimators= 22000, random_state= 5)

RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                        criterion='gini', max_depth=10, max_features='log2',
                        max_leaf_nodes=100, max_samples=None,
                        min_impurity_decrease=0.0, min_impurity_split=None,
                        min_samples_leaf=3, min_samples_split=20,
                        min_weight_fraction_leaf=0.0, n_estimators=22000,
                        n_jobs=None, oob_score=False, random_state=5, verbose=0,
                        warm_start=False)

```

```

rforest.fit(X_train, y_train)

```

```

predictions = dtree.predict(X_test)

```

```

from sklearn.metrics import classification_report, confusion_matrix

```

```

print(classification_report(y_test, predictions))

```

```

from sklearn.ensemble import RandomForestClassifier

```

```

rforest = RandomForestClassifier(criterion="gini", max_depth=10,
max_features="log2",

                                max_leaf_nodes = 100, min_samples_leaf = 3,
min_samples_split = 20,

                                n_estimators= 22000, random_state= 5)

```

```

RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                        criterion='gini', max_depth=10, max_features='log2',

```

```

        max_leaf_nodes=100, max_samples=None,
        min_impurity_decrease=0.0, min_impurity_split=None,
        min_samples_leaf=3, min_samples_split=20,
        min_weight_fraction_leaf=0.0, n_estimators=22000,
        n_jobs=None, oob_score=False, random_state=5, verbose=0,
        warm_start=False)
rforest.fit(X_train, y_train)

predictions = rforest.predict(X_test)

print(classification_report(y_test, predictions))

x_traincnn = np.expand_dims(X_train, axis=2)
x_testcnn = np.expand_dims(X_test, axis=2)

x_traincnn.shape, x_testcnn.shape

import keras

import numpy as np

import matplotlib.pyplot as plt

import tensorflow as tf

from keras.preprocessing import sequence

from keras.models import Sequential

from keras.layers import Dense, Embedding

from keras.utils import to_categorical

from keras.layers import Input, Flatten, Dropout, Activation

from keras.layers import Conv1D, MaxPooling1D

from keras.models import Model

from keras.callbacks import ModelCheckpoint

model.add(Dropout(0.1))

```

```

model.add(Flatten())

model.add(Dense(8))

#model.add(Activation('softmax'))

model.add(Dense(10))

model.add(Activation('softmax'))

opt = keras.optimizers.rmsprop(lr=0.00006, decay=1e-6)

#opt = keras.optimizers.rmsprop(lr=0.00005, rho=0.9, epsilon=None, decay=0.0)

model.summary()

model.compile(loss='sparse_categorical_crossentropy',
              optimizer=opt,
              metrics=['accuracy'])

plt.plot(cnnhistory.history['loss'])

plt.plot(cnnhistory.history['val_loss'])

plt.title('model loss')

plt.ylabel('loss')

plt.xlabel('epoch')

plt.legend(['train', 'test'], loc='upper left')

plt.show()

plt.plot(cnnhistory.history['accuracy'])

plt.plot(cnnhistory.history['val_accuracy'])

plt.title('model accuracy')

plt.ylabel('accuracy')

plt.legend(['train', 'test'], loc='upper left')

```

```
plt.show()
```

```
loss, acc = loaded_model.evaluate(x_testcnn, y_test)
```

```
print ("Restored model, accuracy: {:.2f}%".format(100*acc))
```

B. PUBLICATION WITH PLAGIARISM REPORT:

ELECTION PREDICTION WITH AUTOMATED SPEECH EMOTION RECOGNITION

K.R.N.V.V.D.ARAVIND

Computer science and
engineering
Sathyabama institute of science
and technology
Chennai, India
aravind26051999@gmail.com

K.KAJA KARTHEEK

Computer science and
engineering
Sathyabama institute of science
and technology
Chennai, India
kajakartheek1@gmail.com

S.PRAYLA SHYRY

Computer science and
engineering
Sathyabama institute of science
and technology
Chennai, India
praylashyry@gmail.com

Abstract- Even though artificial intelligence has made tremendous meticulous outcome, there is a gap and large thirst for speech emotion recognition in affective computing. Since machines don't understand the human emotions, natural human computer interaction is struggling a lot to strive and compete with today's market. Survey of past election results and prediction have become crucial and out-dated since the surveys could never clearly depicts the real mental state of an individual. Since the wealth and familiarity of the political people has big title role in many surveys, it is really not trustworthy. No surveys can really understand the truth in the heart of the common people. So a novel approach is developed for day to day physical and psychological behaviour of humans which is influenced by emotions. Other than behavioural expressions, emotion recognition from speech signal is a challenging modality. For the implementation, we take at least 100 audio signals of different area, village, town, district, community, religion and privileged people and we use a nice questionnaire section (20 questions) that could really record the real emotion of the people. MFCC (Mel-Frequency Cepstral Coefficient) feature is extracted and classified with Decision Tree, Random Forest and CNN (Convolutional Neural Network) to give the eight classes of emotions namely Neutral, Disgust, Calm, Angry, Happy, Sad, Fearful and Surprise. Any traditional algorithm can be compared to determine the performance of speech emotion recognition.

Keywords: Speech Emotion Recognition, MFCC, Decision Tree, Random Forest, CNN.

I. INTRODUCTION

Emotion plays a significant role in understanding and conveying the feelings in day to day life (Male or Female) that are expressed through multiple channels such as linguistic content, voice characteristics, facial expressions, and gestures. There are several ways to recognize human emotions. Some of them are recognition of emotions through speech, facial expressions and physiological signals. In recent years, there are many applications like entertainment, robots, audio surveillance, web-based E-learning for detecting human emotions. Many algorithms like CNN, SVM (Support Vector Machine), Gaussian Mixture Model(GMM), Hidden Markov Model(HMM) are used

to classify different types of emotions like anger, fear, joy, sadness, surprise etc. by extracting acoustic features like pitch, MFCC, frequency, WPCC(wavelet packet cepstral coefficient).

Emotions can be either positive or negative based on the situation of the person. Researchers believe that speech is the fastest way of communication when compared to other natural methods. Different types of features like MFCC, WPCC are used for extracting spectrograms from speech signal, utterances from speech signals and waveforms from speech signals. By using Automated Speech Emotion Recognition, we can do several researches in the area of emotion recognition, general classification, pattern recognition etc which can also be used in detection of cyber-crimes. Several methods like MFCC extraction, Cost Matrix Method, LPCC (Linear Prediction Cepstral Coefficient) feature extraction methods are used to process speech signals.

II. STATE OF ART

Ismail Shahin et.al (2019) proposed a novel methodology to identify six different emotions using a hybrid cascaded GMM and DNN by using Emirati Arabic -emphasized speech database. They have extracted MFCC feature from the speech signal which gives pitch of human auditory system and logarithmic perception of onset to detect the emotions. For extracting MFCC feature speech signals are converted into discrete cosign transform to acquire spectral coefficients. They have achieved 83.97% by using GMM-DNN classifier while using SVM and MLP they achieved 80.33% and 69.78% to get better system performance in the presence of noisy signal they have provided dominant signal mask for hybrid classifier.

Yuan chao li et.al (2019) has classified emotion using spectrogram by extracting features like pitch, MFCC, energy from iemocap database to perform gender classification. They have used the combination of CNN-BLSTM in E2E manner for training the model and achieved an absolute increase in overall accuracy by 7.7%when compared to other traditional algorithms. Spectrogram with deltas and delta-deltas are extracted and given as input to neural network. They have implemented CNN-BLSTM model using keras and optimized by Adam method.

Santiago Pascual et al (2019) has proposed an improved self-supervised method in which multiple workers like waveform, MFCC, LPS prosody solve different self-supervised tasks which is followed by a single neural coder to extract transferable robust problem agnostic features like speaker identity phonemes emotional cues from speech signal. They discovered robust representation direction from the raw speech waveform using PASE which can be used as pretrained network to avoid trained models from scratch.

John W. Kim et al (2018) has proposed SVM and EMnet model which combines DNN to extract patterns of features with high level representation and use of common features which is useful for emotion recognition. In this model they have extracted multi dimension feature vectors and speech signal for every 10 msec using 30 msec hamming window. They have used EMO-DB database which contains 535 speech wave files and 10 short sentences labelled with one of the common emotions. They have achieved 88.9% accuracy from compare across 98 trails.

Jilt Sebastian et al (2019) has investigated the use of LSTM (long short-term memory) with pretrained text-based emotion recognition and CNN for emotion recognition with speech by combining fusion techniques. They have used IEMOCAP multi model dataset and adopted various fusion strategies to achieve overall score for each emotion. They have extracted features using CNN from utterance transcriptions by taking 300 dimensional pretrained word embeddings and for each utterance acoustic features are extracted in each stage for building join CNN model using early fusion to improve performance.

Zheng Lian et al (2018) has proposed a novel method named pair wise discriminative task (PDT) to obtain high level audio vectors and compared with traditional classification task. Spectrogram based self-attentional CNN-BLSTM model is used to classify emotions in E2E manner. A total number of 5531 utterances are used to recognize the emotions. They have extracted acoustic and prosodic features like MFCC, pitch, energy using IEMOCAP database and achieved increase in overall accuracy by 7.7% when compared with other traditional methods.

Wei Jiang et al (2019) used DNN algorithm to extract the feature representations from Heterogeneous acoustic feature groups. They have extracted audio data from video data of IEMOCAP dataset to recognize performance of proposed DNN architecture. To learn discriminative acoustic feature representation to obtain informative features a fusion network is trained jointly. SVM is used to recognize task and achieved 64% when compared to existing approaches.

Zheng Lian et al (2019) has combined future observation prediction-unsupervised representation learning strategy with transfer learning approaches to verify effectiveness of proposed system. They have conducted experiments to demonstrate the superiority of proposed system on advanced unsupervised learning strategies by using IEMOCAP database. They have

tested each configuration 20 times. They have used Adam optimization scheme with learning rate of 0.01 and back size of 32 to outperform the main stream named semi-supervised and unsupervised learning strategies above 4% for speech recognition.

Rajeev Ranjan et al (2019) has filtered audio signal by extracting features like pitch frequency and cepstral coefficients using feature extraction technique and perform comparative analysis using different types of MFCC feature extraction method. They have transformed audio waveform to frequency domain representation for advanced processing and analysis. In this proposed system they have analysed speech in each short frame window to obtain a spectrum by FFT and concluded delta-delta MFCC feature extraction technique is better than other feature extraction technique.

Vladimir Chernykh et al (2017) has proposed a novel approach named DRNN (deep recurrent neural network) with a sequence of trained acoustic features and compared with recent advances they have classified speech by emotion and time properties of emotions. They have used IEMOCAP database which consists 12 hours of audio-visual data. The audio signals are divided into small utterances and each utterance is evaluated from 3 to 4 assessors based on audio and video streams to recognize the emotions. They have compared BLSTM performance in one label approach and CTC approach and predicted the sequence of emotions for one utterance.

Sujay Angadi et al (2019) has proposed a novel method named CNN-BLSTM to classify human emotions and measure parameters like precision, recall and accuracy. They have extracted spectral and temporal features to improve SER performance. IEMOCAP database is used to classify emotions by comparing CNN-BLSTM method with SVM and achieved an increase in 9.83% accuracy.

Hemantha Kumar Palo et al (2018) has developed a feature extraction technique to reduce combination features by extracting features like wavelet coefficient, LPCC, MFCC for recognizing five different emotions. They have compared EMO-DB database with SAVEE database to validate accuracy and to recognize common arousal emotions by using different ratios of predefined training, testing and validation data. They have achieved 91.82% and 93.67% accuracy from EMO-DB and SAVEE for first 16 cepstral coefficients using RBFN classifier.

Noushin Hajarolasvadi et al (2019) has proposed a novel method named k-means clustering to predict the emotion in the speech signals. They have extracted 88 types of different dimensional vectors of audio features like MFCC, Pitch and Intensity of each frame which is splitted by analysis of speech signals. They have achieved 81.05%, 77% and 72.33% by extracting audio features using SAVEE, RML and ENTER05 database by applying SVM classifier.

Yea Hoon Kwon et al (2018) has proposed a CNN model to classify emotion based on multi-model data and improved the accuracy of human emotions.

They have combined both EEG (Electro Encephalon Gram) and GSR (Galvanic Skin Response) signals which are pre-processed by Zero crossing rate. They have achieved 73.4% accuracy after noise removed and data processing.

Heekyung Yang et.al (2019) has proposed a novel method multi-column CNN to recognize emotions from Electro Encephalon Gram (EEG) signals. They have used DEAP dataset and circumplex model with balance and arousal to classify emotions and compare accuracies based on voting and weighted sum and also compared with different traditional algorithms.

III. SUMMARY OF EXISTING WORK

In existing system different types of emotions are recognized by using hybrid of cascaded GMM, DNN, CNN-BLSTM, and SVM algorithms by taking different types of datasets like Emirati speech database, IEMOCAP, SAVEE, EMO-DB. They have extracted the most commonly used features like Mel-frequency cepstral coefficient, WPCC, pitch, energy from the speech signal to detect human emotions by taking several trials using trained models. In some model's utterances to classify emotions and to extract pattern of features and achieved an increase in accuracy for better system performance in the presence of spectrogram, noisy signals and wave models.

IV. METHODOLOGY

In this project, we have collected dataset in audio format from 100 people with different perceptions, religion and region. Below figure.1 explains the block diagram of the proposed work. In the field visit, we have asked different set of questions to the people. A sample questionnaire is given in table 1.

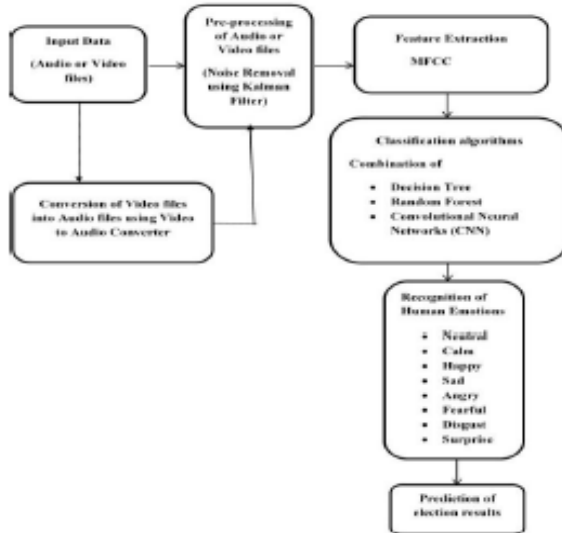


Figure 1. Block Diagram of Speech Emotion Recognition.

Initially the audio files are pre-processed to remove noise in the audio using kalman filter and the MFCC feature is extracted. The extracted feature is classified with 3 different classification algorithms and the people with different emotions are evaluated. Also the accuracy of emotion recognition is calculated to reveal the best classification algorithm.

Table.1:

Sample Questionnaire for the individuals (Takes at least 10 minutes)

1. Do you plan to cast your vote in 2021 Tamil Nadu State election
2. How certain are you that you will vote in the upcoming elections?
3. Which do you identify the most with politically?
4. Do you think the present Government is well enough in decision making.(Justify your answer)
5. Do you think the present Government is well enough in law making. (Justify your answer)

1. Pre-processing:

The .wav file is split into different chunks each of 3 seconds. Each chunk is fed to the Kalman Filter to remove the background noise. Since the background noise (from fan, vehicles) fluctuates over time, it is better to reduce or remove the noise from the wav file to enhance the speech quality. Figure 2 gives the pre-processing of the audio with Kalman Filter. The outcome of the filter is Enhanced audio signal without any loss of data.

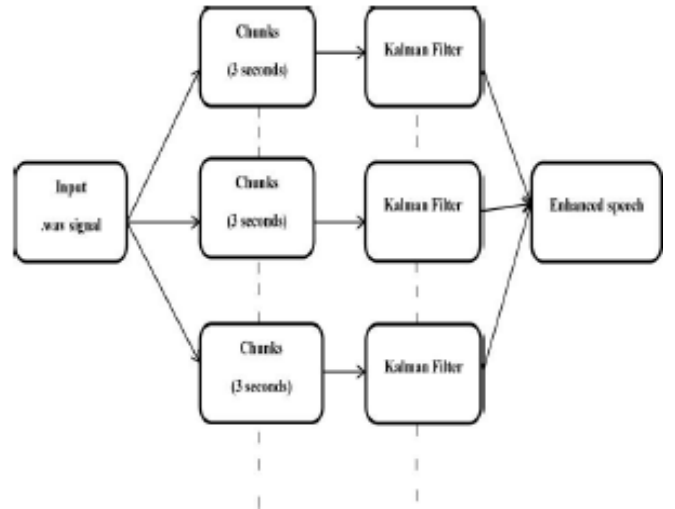


Figure 2. Enhanced Speech

Kalman Filter:

Kalman filter is used to predict unknown values more accurately by taking data inputs from multiple sources. It estimates unknown variables from high level of signal noise to get enhanced speech from the wav signal. It is mostly used in navigation and control technology. An Estimation equation $X_k^{\wedge}$ is calculated by adding

measured value Z_k and Kalman gain K_k to previous estimation equation $\hat{X}_{k-1}$. Extracting audio features plays a vital role in classification of emotions.

$$\hat{X}_k = K_k \cdot Z_k + (1 - K_k) \cdot \hat{X}_{k-1}$$

Kalman Gain

2. Feature Extraction:

In the proposed system we have extracted MFCC feature from audio signal to identify the human emotion.

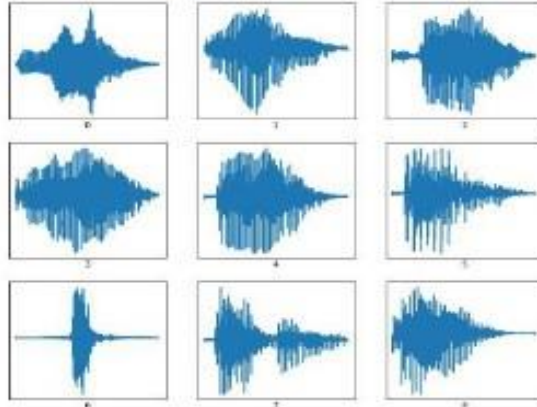


Figure 3. Plotting .wav form of Audio Signal

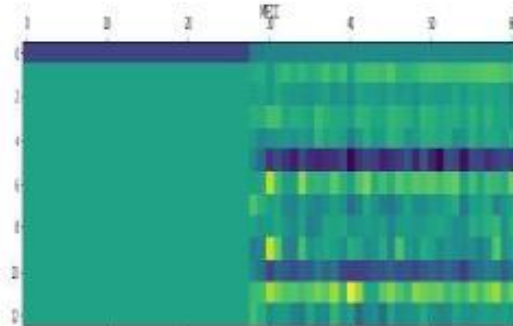


Figure 4. MFCC feature

The coefficients that made up of Mel-Frequency cepstrum is called Mel-Frequency cepstral coefficient. MFCC are derived from a type of cepstral representation of an audio file equally spaced on Mel scale. In this Fourier transform of signal is taken and maps the power of spectrum that has been obtained on Mel scale along with this logs of power and discrete cosine transform is taken and calculated. The amplitude of the resulting spectrum is MFCC.

3. Classification of Emotions:

Emotion recognition helps to understand the real intensity of the person by classifying the human emotions. In this proposed system, the audio files are taken as input and pre-processed audio files are used to identify acoustic MFCC (Mel-Frequency Cepstral

Coefficient) feature. The extracted feature is classified with Decision Tree, Random Forest and CNN (Convolutional Neural Network) to give the eight classes of emotions namely Happy, Sad, Neutral, Calm, Disgust, Angry, Fearful and Surprise. Any traditional algorithm can be compared to determine the performance of SER.

3.1 Decision Tree:

Decision tree is easy to understand and used for solving classification and regression problems. The aim of the decision tree is to predict class or value of the target variable by simple decision terms taken from training data from the created training model.

3.2 Random Forest:

A large number of individual decision trees that operates as an ensemble is called as random forest. Class prediction is done in each individual tree. In random forest the class with more predictions is taken as prediction model.

3.3 CNN (Convolutional Neural Network):

CNN is a class of deep neural network mostly used in image processing, speech recognition. CNN consists of Multi-Layer perceptron's (fully-connected networks), each neuron in one layer is connected to all neurons in next layer. In CNN we can take Hierarchical patterns in data and assemble complex patterns using simple and smaller patterns.

V. RESULTS AND DISCUSSION

Precision:

Precision is defined as the fraction of relevant instances among retrieved instances in pattern recognition, informational retrieval and classification of machine learning algorithms.

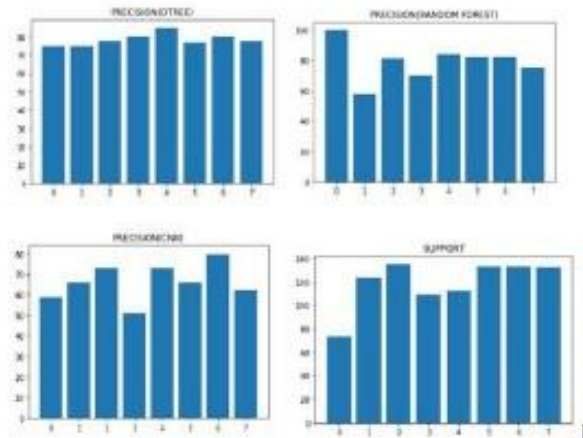


Figure 5. Precision scores

Recall:

Recall is defined as the fraction of the total amount of relevant instances that were actually retrieved in pattern recognition, informational retrieval and classification of machine learning algorithms.

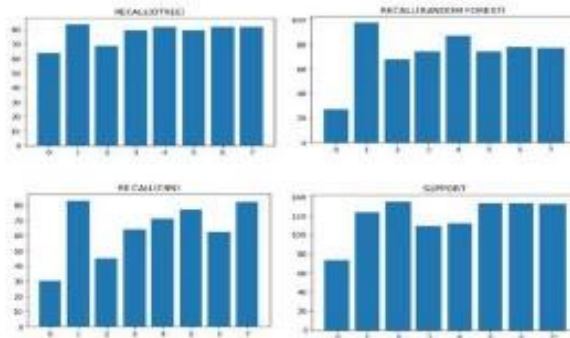


Figure 6. Recall values

F1-Score:

F1 score is defined as harmonic mean of the precision and recall in statistical analysis of binary classification.

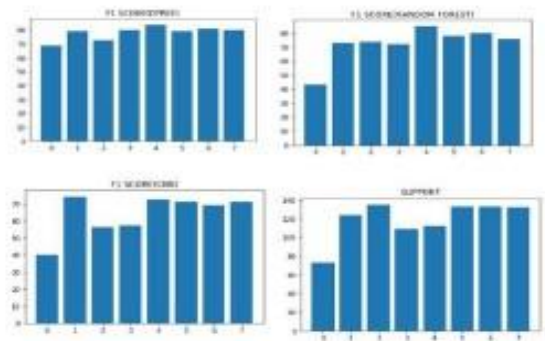


Figure 7. F1 Scores

	precision	recall	f1-score	support
0	0.80	0.27	0.41	75
1	0.82	0.38	0.53	124
2	0.81	0.48	0.59	128
3	0.78	0.78	0.78	148
4	0.64	0.67	0.65	131
5	0.82	0.74	0.78	131
6	0.82	0.78	0.80	131
7	0.78	0.77	0.78	131
macro avg	0.79	0.59	0.69	951
weighted avg	0.79	0.78	0.79	951

	precision	recall	f1-score	support
0	0.51	0.41	0.44	75
1	0.48	0.31	0.39	124
2	0.48	0.39	0.43	128
3	0.70	0.85	0.76	148
4	0.40	0.74	0.52	131
5	0.64	0.57	0.61	131
6	0.40	0.53	0.46	131
7	0.67	0.65	0.66	131
macro avg	0.51	0.52	0.51	951
weighted avg	0.49	0.62	0.52	951

	precision	recall	f1-score	support
0	0.78	0.67	0.72	75
1	0.75	0.82	0.79	124
2	0.78	0.78	0.78	128
3	0.85	0.80	0.83	148
4	0.79	0.83	0.81	131
5	0.83	0.79	0.81	131
6	0.78	0.79	0.79	131
7	0.81	0.77	0.79	131
accuracy			0.78	951
macro avg	0.78	0.77	0.78	951
weighted avg	0.78	0.78	0.78	951

CNN Confusion matrix:

Figure 8. Confusion matrices

In the above figure the confusion matrix of Decision Tree, Random Forest and CNN are drawn by extracting f1 score, precision and recall values

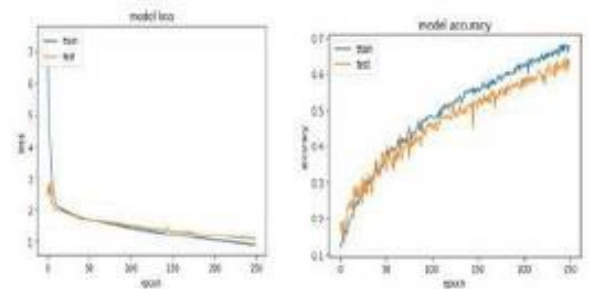


Figure 9. Model loss and Model Accuracy graph

In the above figure Model loss and Model accuracy is plotted by extracting the loss and accuracy values from the audio signal.

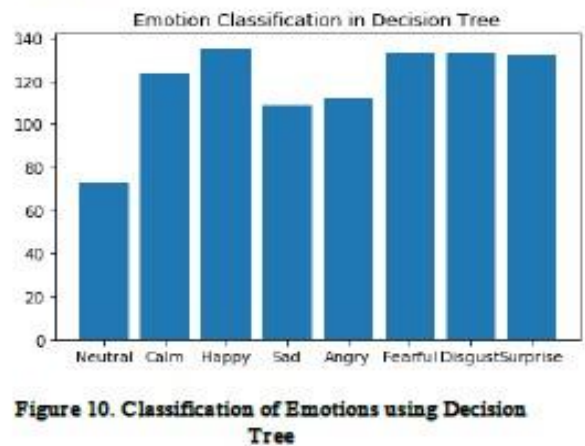


Figure 10. Classification of Emotions using Decision Tree

In the above figure we have classified and plotted the emotions that have been extracted from the audio signals using Decision Tree.

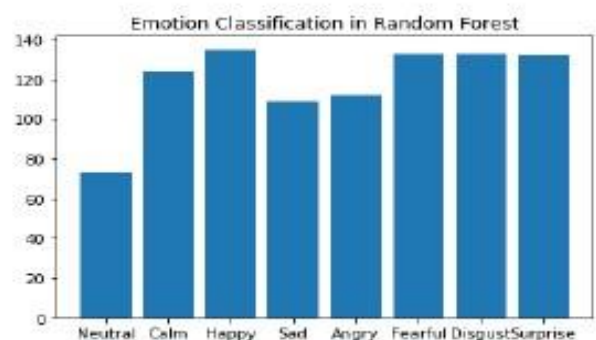


Figure 11. Classification of Emotion using Random Forest

In the above figure we have classified and plotted the emotions that have been extracted from the audio signals using Random Forest.

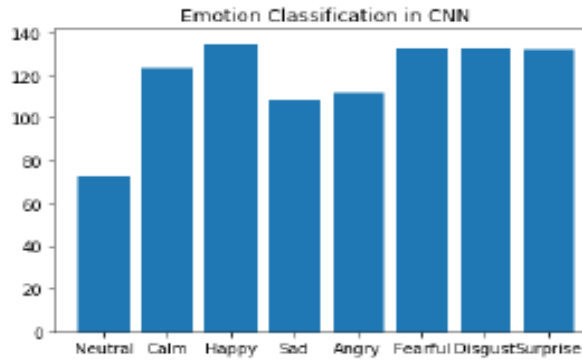


Figure 12. Classification of Emotions using CNN

In the above figure we have classified and plotted the emotions that have been extracted from the audio signals using CNN.

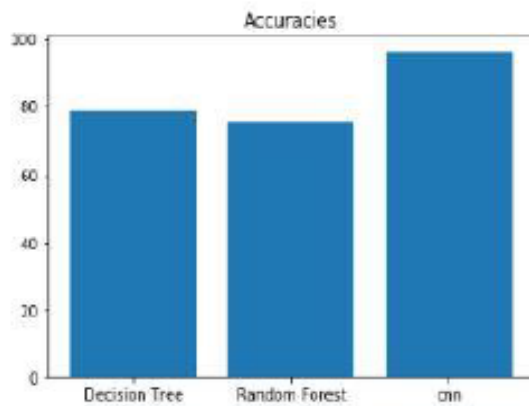


Figure 13. Comparison of accuracy

In the above figure we have compared the accuracies of Decision Tree (79%), Random Forest (75%) and CNN (96%).

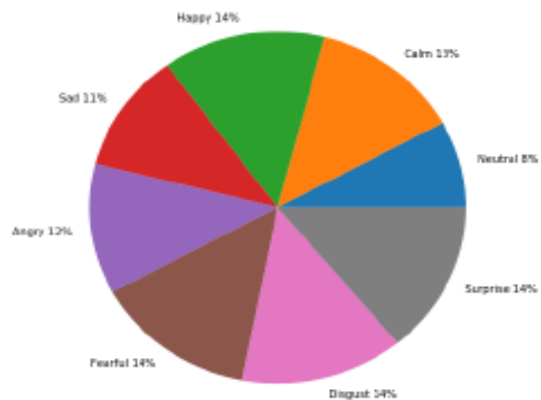


Figure 14. Pie chart of classified Emotions

From the above figure we are inferred that 14% of the people are happy, 12% of the people are calm, 8% of the people are neutral, 14% of the people are surprise, 14% of the people

are disgust, 14% of the people are fearful, 12% of the people are angry and 11% of the people are sad.

VI. CONCLUSION

Our aim is to provide a flawless model for speech emotion recognition by extracting acoustic MFCC feature from the audio signals and classify the classes of emotion like Neutral, Calm, Disgust, Happy, Sad, Angry, Fearful and Surprise using Decisiontree, Randomforest and CNN (Convolutional Neural Network) to achieve an overall increase in accuracy when compared to all traditional algorithms.

REFERENCES:

1. ISMAIL SHAHIN, ALI BOU NASSIF, SHIBANI HANSA, Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep NeuralNetwork, http://www.ieee.org/publications_standards/publications/rights/index.html.
2. Jilt Sebastian, Piero Pierucci, Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts, <http://dx.doi.org/10.21437/Interspeech.2019-3201>.
3. John W. Kiml, Rif A. Saurous, Emotion Recognition from Human Speech Using Temporal Information and Deep Learning, [10.21437/Interspeech.2018-1132](http://dx.doi.org/10.21437/Interspeech.2018-1132).
4. Rajeev Ranjan, Abhishek Thakur, Analysis of Feature Extraction Techniques for Speech Recognition System, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-7C2, May 2019.
5. Sujay Angadi* Venkata Siva Reddy1, Hybrid Deep Network Scheme for Emotion Recognition in Speech, International Journal of Intelligent Engineering and Systems, Vol.12, No.3, 2019.
6. Vladimir Chernykh, Grigoriy Sterling, Pavel Prihodko, Emotion Recognition from Speech with Recurrent Neural Networks, [arXiv:1701.08071v1 \[cs.CL\]](https://arxiv.org/abs/1701.08071v1) 27 Jan 2017.
7. Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Unsupervised Representation Learning with Future Observation Prediction for Speech Emotion Recognition, [arXiv:1910.13806v1 \[eess.AS\]](https://arxiv.org/abs/1910.13806v1) 24 Oct 2019.
8. Santiago Pascual, Mirco Ravaselli, Joan Serra, Antonio Bonafonte, Yoshua Bengio, Learning Problem-agnostic Speech Representations from Multiple Self-supervised Tasks, [arXiv:1904.03416v1 \[cs.LG\]](https://arxiv.org/abs/1904.03416v1) 6 Apr 2019.
9. Yuanchao Li, Tianyu Zhao, Tatsuya Kawahara, Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning, <http://dx.doi.org/10.21437/Interspeech.2019-2594>.
10. Yea-Hoon Kwon, Sae-Byuk Shin and Shin-Dug Kim, Electroencephalography Based Fusion Two-Dimensional (2D)-Convolution Neural Networks (CNN) Model for Emotion Recognition System, www.mdpi.com/journal/sensors.
11. Heekyung Yang, Jongdae Han, Kyungha Min, A Multi-Column CNN Model for Emotion Recognition from EEG Signals, www.mdpi.com/journal/sensors.
12. Noushin Hajarolasvadi, Hasan Demirel, 3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms, www.mdpi.com/journal/entropy.
13. Wei Jiang, Zheng Wang, Jesse S. Jin, Xianfeng Han, Chunguang Li, Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network, www.mdpi.com/journal/sensors.
14. Zheng Lian, Ya Li, Jianhua Tao, Jian Huang, A PAIRWISE DISCRIMINATIVE TASK FOR SPEECH EMOTION RECOGNITION, www.researchgate.net/publication/322301986_A.
15. Hemanta Kumar Palo, Mihir Narayana Mohant, Wavelet based feature combination for recognition of emotions, <https://www.researchgate.net/publication/330263647>.

12%

SIMILARITY INDEX

6%

INTERNET SOURCES

7%

PUBLICATIONS

8%

STUDENT PAPERS

PRIMARY SOURCES

1

www.mdpi.com

Internet Source

1%

2

Hui Wang, Chengdong Zhou, Leixiao Li. "Design and Application of a Text Clustering Algorithm Based on Parallelized K-Means Clustering", Revue d'Intelligence Artificielle, 2019

Publication

1%

3

export.arxiv.org

Internet Source

1%

4

Submitted to University of Southern California

Student Paper

1%

5

Submitted to Vels University

Student Paper

1%

6

arxiv.org

Internet Source

1%

7

Submitted to The Robert Gordon University

Student Paper

1%

8

Submitted to University of Birmingham

Student Paper

1%