Assessment of Patient Health Condition based on Speech Emotion Recognition (SER) using Deep Learning Algorithms

Dr. DNVSLS Indira[1], B. Lakshmi Hari Prasanna[2], Chunduri Pavani[3], Ganta Vandana[4]

[1]Associate Professor, Dept. of Information Technology, Gudlavalleru Engineering College, Gudlavalleru.
indiragamini@gmail.com[1]
[2,4]IV B.Tech. Student, Dept. of Computer Science and Engineering, Gudlavalleru Engineering College, Gudlavalleru. lakshmi.hariprasanna@gmail.com[2], vandana.ganta942@gmail.com[4]
[3] IV B.Tech. Student, Dept. of Computer Science and Engineering, PVPSIT, Vijayawada
pavanichunduri5028@gmail.com[3]

Abstract:
Human Emotion detection either through face or speech became a relatively nascent research area. Speech Emotion Acknowledgment concerns the undertaking of perceiving a speaker's feelings from their discourse chronicles. Perceiving feelings from discourse can go far in deciding an individual's physical and mental condition of prosperity. These emotions can be used for further assessment of patient's status for better diagnosis. This paper aims to categorize emotions in speech into four different categories which are happy, sad, angry and neutral. For this analysis, four different algorithms - the Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest (RF) and Convolutional Neural Network (CNN-1D) are developed. Detection of Emotion through speech of an individual might be a bit hectic, because of the dynamic changes in voice signal of the same person within a very subtle period of time. So, features like mfcc, chroma, tomez contrast and mel were extracted and given to the model in order to detect the emotions. Those features were given as input to the algorithms and the empirical results implicate that Convolutional Neural Network-1D performs well comparatively. RAVDESS database is chosen for the categorization. A good recognition rate of 89% was obtained from CNN-1D.

Keywords: Speech Emotion Recognition (SER) Random Forest, Multi-Layer Perceptron, CNN-1D, RAVDESS, Deep Learning, Diagnosis

1. Introduction

The investigation of feeling has progressed quickly in the course of the most recent decade, driven by minimal effort brilliant innovations and expansive enthusiasm from scientists in neuroscience, brain research, psychiatry, audiology, and software engineering. Feeling is a solid inclination that can be communicated by a person [2][3]. It tends to be communicated through voice, outward appearances and body signals. Typically, we people have normal capacity to identify the feelings of the person whom we are managing. In any case, machines or the frameworks can't distinguish the feelings and conditions of an individual and afterward act in a manner.

These machines can't have the option to get a handle on the condition of an individual and arrange the circumstance as a section, so there is a need to prepare machines to identify the feelings of an individual so as to accomplish a serious extent of human PC communication. Preparing a machine to take in consequently and improve for a fact without being expressly modified is called as Machine Learning (area in Artificial

Intelligence) Machine Learning is utilized anyplace from mechanizing ordinary errands to offering clever bits of knowledge, ventures in each area attempt to profit by it[13].

Discourse understanding frameworks are as yet restricted and not really viable. The fundamental issue right now is we are up 'til now unsuitable to have a system which is able to grasp a human's understanding level similarly as distinguishing energetic conditions. Accompanied by all endeavors within flow research, it would not be an amazement to notice in the approaching ages possessing their closest companions as PCs with whom they could talk and offer their feelings. What is feeling? In brain research and regular utilization, feeling is the language of an individual's inner condition, typically situated in or attached to their physical and social tangible inclination [2]. Love, disdain, fortitude, dread, bliss, and pity would all be able to be portrayed in both mental and physiological terms. The primary target of this exploration is to identify cheerful, tragic, and outrage feeling states from text sources, for example, messages and gatherings. We picked these three feelings in light of the fact that there are clear contrasts between them. The remainder of the feelings is firmly identified with each other, for example, love and euphoria, dread and pity, and so on. With an unmistakable differentiation we can recognize the feeling class with our present strategies. These emotions can further be used in recommender systems for patient analysis.

This paper is isolated into few areas. To start with, the connected works in this field will be talked about in Section 2. In Section 3, the flow chart of proposed work, information preprocessing and the four AI algorithms for the emotional state location will be introduced. The outcomes and conversation part is there in the segment 4.The end and future work of this examination will be talked about in Section 5.

2. Related Work:

An ordinary Speech Emotion Recognition is separated amid dual sections:

(a) Attribute determination cycle for extrication of significant extent highlights over discourse information

(b) Classifiers determination to perceive feelings over speech efficiently.

2.1 The feature selection :

The strong element choice for SER is a difficult errand for scientists. There are a few analysts that utilized the high quality highlights utilized for SER. Henceforth, Dave et al. [9] assessed various highlights for discourse feelings and demonstrated the productivity for best Mel recurrence cepstral coefficient (MFCC) . The author Liu et al. [10] evaluated the GFCC i.e. gammatone recurrence cepstral coefficient things to see the speech emotion recognition for unlikely precision approximately 3.6% compared to MFCCs utilizing extra uttered highlights such as tremble and shine. Fahad [12] portrayed one strategy to choose highlights dependent at length velar along with MFCCs for preparation of Deep Neural Networks models for emotion detection..

2.2 The selection of classifiers

Wen et al. [13] utilized the arbitrary profound conviction grid for emotion detection through speech. Within the technique, first they remove short extent highlights over discourse signs utilizing LLD along with taking care for profound conviction organizations (DBNs) upon extricating significant extent discriminative highlights. The elevated level highlights are taken care of to the SVM classifier, which is associated with every conviction organization for foreseeing of orator's feelings along with afterward settles on choices dependent on greater part casting a ballot.

Lian et al. [14] used a confounded model, DBN utilized for highlights figuring out how to get concealed highlights from discourse and SVM classifier was used for feeling expectation to accomplish elevated level precision inside speech emotion detection utilizing Chinese dataset called CASIA.

Hajar and Hasan [16]presented one strategy upon emotion detection through speech, for parting discourse signs to outlines and discarded the Mel frequency cepstral coefficients included just changed over those to spectrograms upon choosing main image in general sound, which speak to the articulation of discourse. The point to point explanation of proposed work is clarified in the next segment and the effectiveness of output of initiated work is explained in the experimental area.

3. Proposed Work:

Detection of emotion over sound signs needs to highlight unheathing along with preparation of classifiers. Component trajectory comprises components for each sound sign that describes the orator's explicit highlights, for example, voice, timbre, vitality, for the preparation of a classifier model to perceive one specific feeling precisely. The North American English language open source dataset was separated amid preparing along with testing physically.Orator tone plot data, spoken to beside MFCC, removed over voice examples within preparing the dataset. Timbre, Short Term Energy(STE), and MFCC's consists of sound examples inside feelings of outrage, joy, along with bitterness are acquired. The extricated highlight trajectories shipped off the classifier model. Test dataset will go through the extraction methodology backing the classifier and settle on one choice with respect to the basic feeling in the test sound. Preparation along with test information bases utilized are in North American English along with characteristic discourse corpus.

The paper subtleties the four AI calculations appealed to highlight trajectories along with the impact of expanding the quantity of highlight vectors taken care of to the categorizer. This gives an investigation about exactness of characterization for Indian English discourse. The accomplished exactness for Indian English discourse was 80% of dark information.

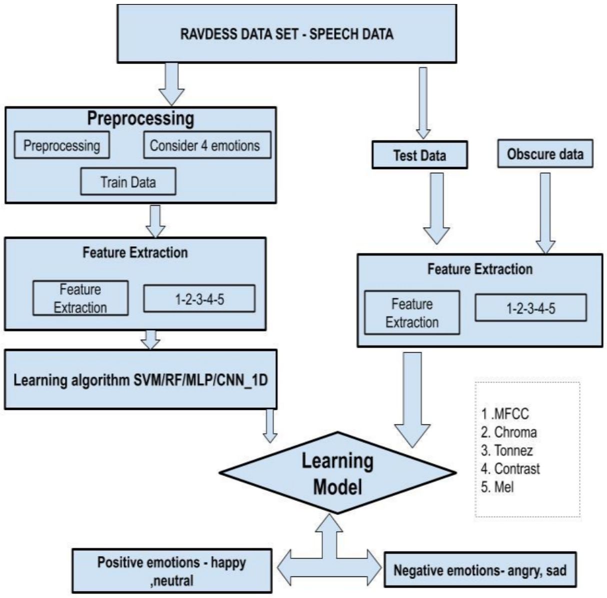

Figure 1: Architecture of Proposed Work

3.1 Data preparation and preprocessing

Preparing the Dataset: RAVDESS dataset is downloaded from kaggle.com after that it is converted so that it would be suited for extraction [11].

Loading the Datase: In this procedure, loading of dataset using python takes place that involves extracting or obtaining different features, such as power, timbre and tone tract design over speech signal, librosa is utilized to complete the task at hand.

Training the Model : A suitable sklearn model is detected and loading of the prepared dataset into that model is done in this step.

Testing the model: Analyzing the efficacy of our proposed model:

RAVDESS dataset is utilized which consists of twenty four actors voices composed of eight emotions neutral, angry, calm, happy, sad, fearful, disgust and surprise.

- Only four emotions were considered, those are happy, angry, sad and neutral.
- mfcc, chroma, tonnetz, mel, contrast are the features extracted in order to classify the emotions.
- The dataset is splitted using train_test_split with 75% train and 25% testing data.
- Number of training samples are 935, testing samples are 329 and total features extracted are 180.

3.2 Proposed model Features

Speech Recognition is such a model affirmation which allows a machine to change the talk signal into an automated eigenvector, a short time later setting up the model and the model planning with feeling features. The consistently used segment of talk feeling relies upon pitch, range, nature of voice and so forth A portion of the prosodic features are insinuates the pitch, length of sound, speed of discourse conveyed, reality thus. These don't impact our distinctive evidence of the word, yet they impact sounds of voice is normal or extraordinary. During the time burned through part extraction, picks the best worth, mean worth, reach, incline, and so forth of the beat feature. The prosodic credits have a particular powerful implementation in talk feeling affirmation [4].

Features dependent on the range shows the connection among the channel shape changes and the sound development. Regularly utilized unearthly attributes are direct ghastly highlights, for example, LPC, OSALPC, and cepstrum include, as LPCC, MFCC, [9][5] and so forth Highlights removed from sound quality: when individuals are talking, which is anything but difficult to influence sound by state of mind, now and then there will bebreathing, tremolo, stifling, and so on, these highlights are likewise an extraordinary assistance on feeling discourse acknowledgment. The auditory attributes utilized to gauge the nature of the sound are commonly utilized with formant recurrence and its data transmission, recurrence bother, adequacy encompass, glottal boundaries, etc.

The features utilized are mfcc, chroma, tomez difference and mel are extricated and it is important to be standardized before input is given, to guarantee that the significance of each measurement is nearly the equivalent.

3.3 Learning Algorithms

The for the most part used classifiers consolidate linear classifier, and nonlinear classifier, which can change the component from low dimensional space to high and from nonlinear space to direct space. Since the headway of AI, the normally used classifiers are support vector machine (SVM),K-nearest neighbor classifier (KNN), counterfeit neural association (ANN) HMM, SOFMNN, hence on[1] [7].

Since the solid nonlinear arrangement has the capacity of both SVM, ANN, the two sorts of classifiers has been broadly utilized in feeling acknowledgment, this paper utilizes SVM, Random Forest, Multi Level Perceptron and CNN 1D calculations to make discourse feeling acknowledgment, and gets higher acknowledgment rate after element learning with CNN 1D[16][17].

3.3.1. Emotions Recognition from SVM

Support Vector Machine (SVM) was first proposed by Corinna Cortes and Vapnik during the 1990s, and it has been commonly used in the field of model affirmation. [8] SVM has various ideal conditions in handling the issue of little model, nonlinear model and high dimensional part plan affirmation. There are various central focuses with SVM, it has an outstandingly strong non-straight fitting limit, and it is routinely used in model backslide and course of action, it might be feasibly applied to feeling request.

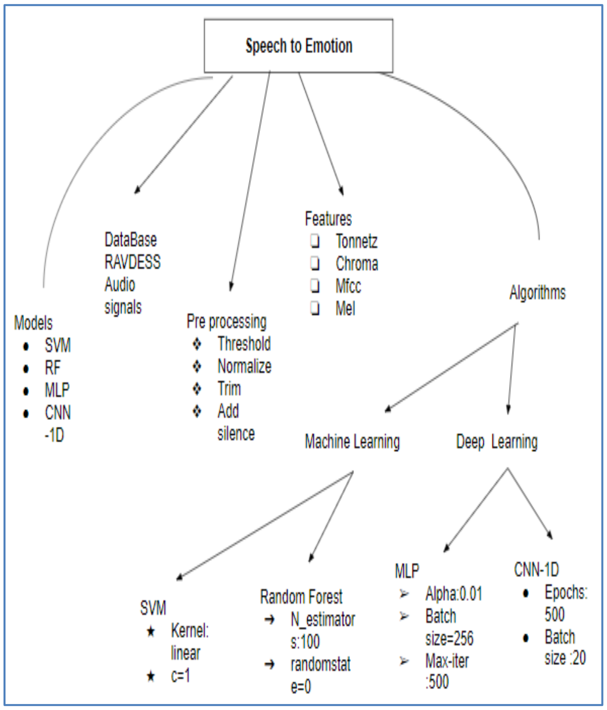

Figure 2: Tree Structure of Proposed Work

3.3.2. Random Forest

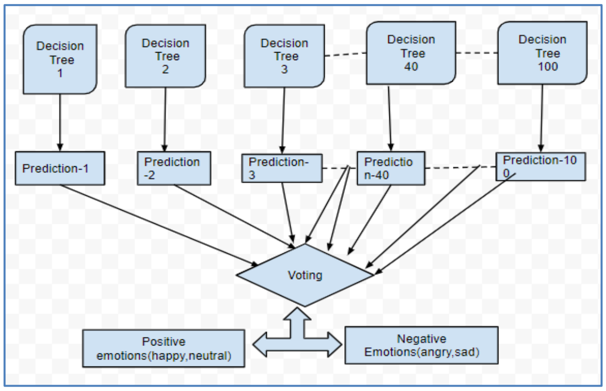

Figure 3: Working Model in Random Forest algorithm

3.3.3. Multi Layer Perceptron

Multi Layer Perceptron is the classifier model which used to group the feelings. It is a class of feed forward counterfeit neural networks[5][6]. MLP uses managed learning procedures got back to engendering for preparing. Its numerous layers and non-direct actuation recognize MLP from a straight perceptron. It can recognize information that isn't directly detachable. It has completely associated layers. It fundamentally comprises of three layers input, covered up, yield. The yield is determined utilizing equation for example duplicating loads with input and adding inclination to it.

- The training data is stacked into the MLP classifier. MLP classifier is a class of feed forward fake neural organizations.
- It comprises of three layers input, covered up, yield layers. But input hubs, all hubs utilize nonlinear enactment work, it utilizes regulated learning methods got back to engendering for preparing.
- It's various layers and nonlinear actuation work separates it from customary fake neural organizations.
    - The model parameters are :
        - ➢ alpha:0.01-regularization parameter which constraints the overfitting.
        - ➢ batch_size:256- to control the stability of the neural network.
        - ➢ epsilon:1e-0.8-value for numerical stability for adam.
        - ➢ hidden_layer_size : (300,1)-number of elements or neurons in the 1st hidden layer.
        - ➢ learning_rate: adaptive-for weight updates.
        - ➢ max_iter:500 - like number of epochs i.e., how many times each data point is used .
        - ➢ activation: relu-rectified linear unit function return f(x)= max(0,x).

w=w + lr*(expect-predict)*x
y=f (wxT +b)
x=input vector or output vector of previous layer
b=bias vector

The weight is manipulated during back propagation so we use ADAM optimizer for the change in weight.
wt=(wt-1) - n(m ^t /(sqrt (v ^t)+E))
m ^t = mt/(1-B1t) v ^t = vt/(1-Bt2)------estimators
Moving averages of gradient and squared gradient
mt = B1mt-1+(1-B1)gt
vt=B2vt-1+(1-B2)(g^2)t

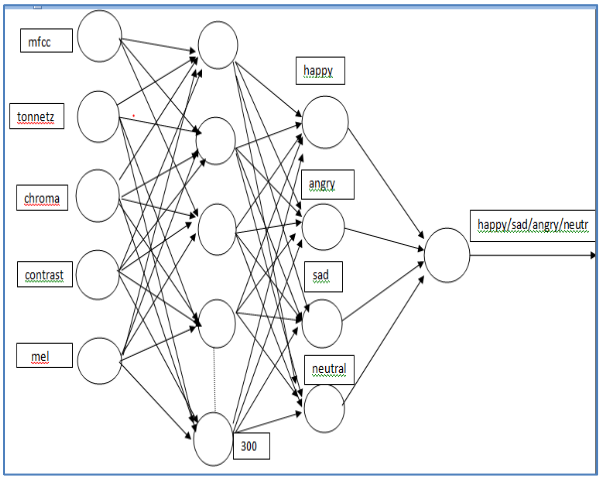

Figure 4: Flow diagram in MLP

3.3.4. CNN- 1D

Convolutional Neural Network's(CNN) are available status of workmanship models which are used to remove raised level features without any preparation level rough part information. CNN utilizes the amounts of bits to remove raised level features from pictures and those features are used to manufacture a CNN model for performing enormous gathering errands [6][17]. CNN designing is the mix of three sections; convolutional layers, which have a couple of amounts of channels to apply on the information. Each channel inspects the data using the spot thing just as convenience strategy to make the amounts of features to map into a single convolutional layer. Next resulting part is pooling layers, which are used to decrease or down-analyzing the components of features maps. Moreover there are a couple of plans used for decreasing measurements like max pooling, min pooling, mean pooling, ordinary pooling, etc. The last part is totally connected with Fully Connected(FC) layers of CNN[16], these are principally used for isolating the overall features which are dealt with to a SoftMax classifier is used to find the probability of every single class. CNN plans every one of these layers in reformist way, convolutional layers (CL), pooling layers (PL), and later Fully Connected layers(FC) trailed by the SoftMax classifier. The supposed structures are detailed in the following territory.

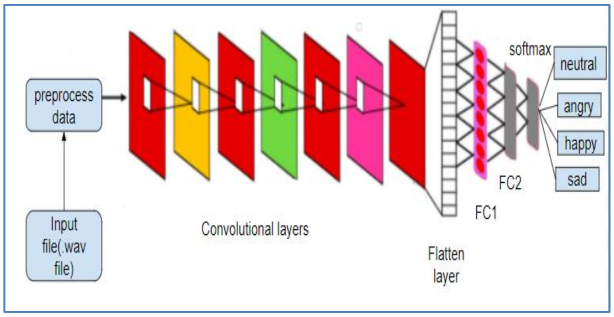

Figure 5: Working Model in CNN- 1D

The proposed CNN-1D model for SER is appeared in Figure 5. We investigate the plain organization from picture arrangement to discourse feeling acknowledgment and order.

4. Results and Discussion:

Mfcc features of a signal are a miniature group of features which briefly express the comprehensive shape of a spectral envelope. Spectral contrast contemplates the spectral zenith, the spectral valley, and their difference in each frequency subband. Tonnetz determines the tonal centroid features of the audio signal and Chroma based features, which are also referred to as "pitch class profiles", are a powerful tool for analyzing music whose pitches can be meaningfully categorized and whose tuning approximates to the equal-tempered scale.

Firstly, we considered five features from the audio signals of RAVDESS dataset. Those are mfcc (mel frequency cepstral coefficient), Spectral contrast, MEL, tonnetz, chroma. The train and test data set are split in 75:25 ratio. We normalized the signals and trained two machine learning algorithms and two deep learning models. SVM classifier kernel is linear as it works rapidly and c=1 as it controls the tradeoff between decision boundary and training points. The accuracy we obtained for this model is approximately 75% and when coming to the emotions, 78% of angry, 70% of happy,72% of neutral and 77% of sad were predicted accurately. In the Random Forest algorithm, we used 100 esmators i.e, 100 decision trees and random state is set to Zero. The overall accuracy is not much different (76%) but individual emotion detection accuracy increased up to 6% in the angry (84%) category, 3 % in the happy (73%) category but neutral and sad emotions accuracy declined drastically.

Then we employed deep learning algorithms multi layer perceptron and CNN-1D. The accuracy expanded. In the MLP model, we took one hidden layer with 300 neurons, alpha value as 0.01 for regularization and learning is adaptive. The accuracy obtained is 80% and individual accuracies are 85%,80%,72%,86% for emotions angry, happy, neural and sad. The CNN-1D with rmsprop(root mean square propagation) optimizer for fast results ,500 epochs,2 hidden layers with relu activation function and softmax function for output layer as it consists of more than one category to detect. The accuracy obtained is 89% which is more than any other machine learning and deep learning models we analyzed and the individual emotions accuracies are 85.5 % angry,79.8 % for happy, 73 % for neutral and 86.2 % for sad. From the empirical data, we can infer that the CNN-1D(1 dimensional convolutional neural network) model outperforms the remaining models .

Emotion	angry	happy	neutral	sad
angry	78.8	14.4	4.4	2.2
happy	12.7	70.2	4.25	12.76
neutral	6.8	6.8	72.2	13.6
sad	4.95	8.9	8.9	77.2

Table 1: SVM Confusion Matrix

Emotion	angry	happy	neutral	sad
angry	84.3	11.1	0	2.2
happy	9.5	73	1	17
neutral	0	4.5	64	31.8
sad	2.97	16.8	2.9	77.2

Table 2: Random Forest model Confusion Matrix

Emotion	angry	happy	neutral	sad

angry	85.5	10	2.2	2.2
happy	5.3	79.7	5.3	9.6
neutral	0	4.5	72.8	22
sad	0	6.9	6.9	86.1

Table 3: MLP Confusion Matrix

Emotion	**angry**	**happy**	**neutral**	**sad**
angry	84	15.5	0	1.1
happy	5.3	81	3.1	10.6
neutral	0	11.3	66	25
sad	0	5.9	4.95	89.11

Table 4: CNN-1D Confusion Matrix

	Models	**Parameters**	**Dataset**	**No .of Training & Testing Samples**	**Report**			
						Precision	**recall**	**F1-score**
Machine Learning Algorithms	Support Vector Machine	Kernel : linear , C : 1	RAVDESS	985 & 325	accuracy	-	-	0.75
					macro avg	0.75	0.74	0.75
					weighted avg	0.75	0.75	0.75
	Random Forest	n_estimators :100, random_state =0		985 & 325	accuracy	-	-	0.77
					macro avg	0.79	0.75	0.77
					weighted avg	0.77	0.77	0.77
	Multi Layer	alpha : 0.01, batch_size:2		985	accuracy	-	-	0.82

Deep Learning Algorithms	Perceptron	56, h_l:300 epsilon:1e-0.8	RAVDESS	& 325	Macro avg	0.82	0.79	0.80
					weighted avg	0.82	0.82	0.82
	CNN-1D	b_s:20,epoch:500, optimizer:rmsprop,5e-6,d_o= .1		985 & 325	accuracy	-	-	0.89
					macro avg	0.82	0.82	0.82
					weighted avg	0.83	0.83	0.83

Table 5: Analysis of all four algorithms that are used in proposed work with parameters, training and test samples.

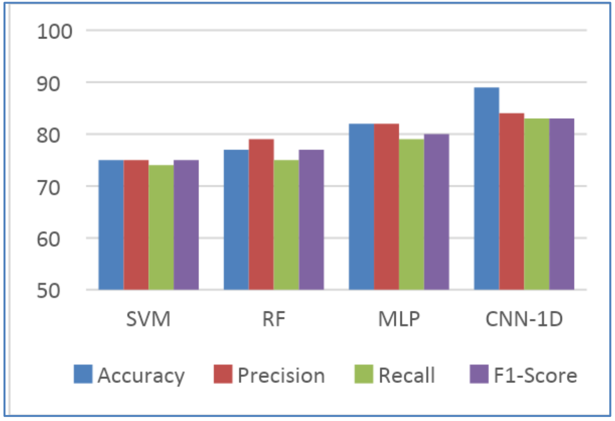

Figure 6: Analysis of different algorithms used in proposed model.

	SVM	RF	MLP	CNN-1D
Angry	79	87	86	91
Happy	73	76	83	94
Sad	77	77	86	92
Neutral	73	64	73	80
	75.54	75.77	81.94	89.30

Table 6: Total Number of test samples are 325. Among them 90 are angry, 90 are happy, 101 are sad and 44 are neutral. The above table shows correctly classified values by four algorithms in four emotions. Last row gives the accuracy of each algorithm in %.

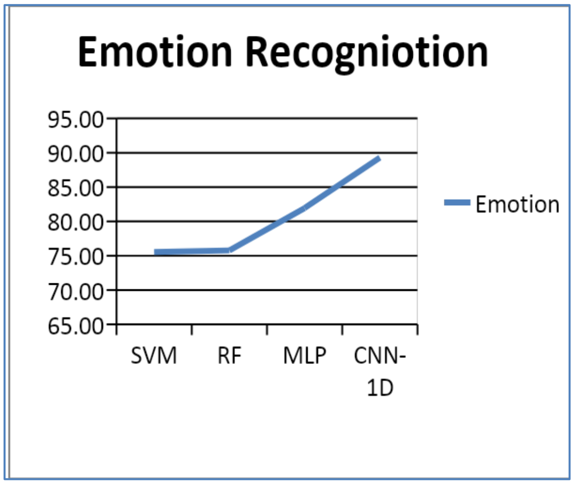

Figure 7: Shows that CNN-1D recognizes emotions accurately with nearly 90%

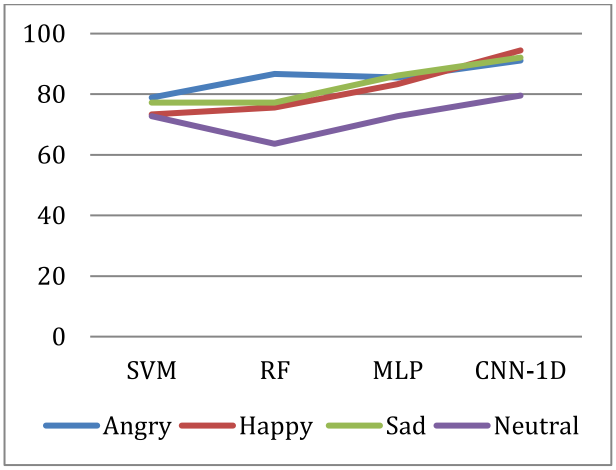

Figure 8: % of accuracy correctly classified by 4 algorithms for 4 emotions

5. Conclusion and Future scope:

The proposed system is used to detect one's emotion through speech. Considering both Positive and Negative Emotional modulations. When an audio file which is been trained under Four classifier algorithms SVM, RF, MLP and CNN-1D is given, our resultant model will be able to differentiate a positive or a negative Emotion. These emotions can be used many fields. One of the best suited applications of SER is in the area of health applications. The purpose of SER is used to identify the patient satisfaction on a diagnosis.

Traditionally this creates a better understanding and prioritizes the importance of feelings and emotions. But technically it tends in sloving problems like slowing down a car when a person is in fear or in danger considering these voice modulations, or simply automating self driving cars is one the example among the huge. This Developed system contributes a lot in Robotics, Artificial Sciences and related stuff. Apart from that it can be a diagnosis tool which can create a pleasant environment to deal with patients. This technically can detect the emotion of the patient and help the consultant to react accordingly. Provides a great zeal in Voice Assistance (like Alexa, siri).

6. References:

[1] Valery A. Petrushin,"Emotion Recognition in Speech Signal: Experimental Study, Development, And Application", 6th International Conference on Spoken Language Processing, ICSLP,Oct-2000.

[2] Huan Su Ling, Ranaivo Bali, Rosalina Abdul Salam, “Emotion Detection using Keywords Spotting and Semantic Network” , IEEE ICOCI 2006

[3] Won-Joong Yoon, Youn-Ho Cho, and Kyu-Sik Park, “A Study of Speech Emotion Recognition and Its Application to Mobile Services”, UIC 2007, LNCS 4611, pp. 758–766, 2007

[4] Lugger M, Yang B. The Relevance of Voice Quality Features in Speaker Independent Emotion Recognition[J]. 2007, 4.

[5] Guobao Zhang, Qinghua Song, ShuminFei,”Speech Emotion Recognition System Based on BP Neural Network in Matlab Environment”, International Symposium on Neural Networks ISNN 2008, pp 801-808

[6] Yongming Huang, Guobao Zhang, Xiaoli Xu, ” Speech Emotion Recognition Research Based on Wavelet Neural Network for Robot Pet”, International Conference on Intelligent Computing, ICIC 2009, pp 993-1000

[7] Xu Huahu, Yuan Jian, “Application of Speech Emotion Recognition in Intelligent Household Robot”, International Conference on Artificial Intelligence and Computational Intelligence, 2010

[8] Hu Y, Wu L, Gao L, et al. Research on speech emotion recognition based on SVM[J]. Electronic Test, 2011.

[9] Dave, N. “Feature extraction methods LPC, PLP and MFCC in speech recognition”, International Journal of Advanced Research Engineering Technology. 1, 1–4, 2013

[10] Gabrielle K. Liu, “Evaluating Gammatone Frequency CepstralCoe_cients with Neural Networks for Emotion Recognition from Speech”,Computer Science and Engineering, arXiv: 1806.09010, Published 2018,

[11] Livingstone, S.R. Russo, F.A. “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)”: A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 2018.

[12] Fahad, M.; Yadav, J.; Pradhan, G.; Deepak, A. “DNN-HMM based Speaker Adaptive Emotion Recognition using Proposed Epoch and MFCC Features”, arXiv:1806.00984, 2018.

[13] Wen, G.; Li, H.; Huang, J.; Li, D.; Xun, E. “Random deep belief networks for recognizing emotions from speech signals”. Comput. Intell. Neurosci., 2017.

[14] Zhu, L.; Chen, L.; Zhao, D.; Zhou, J.; Zhang, W. Emotion recognition from Chinese speech for smart a_ective services using a combination of SVM and DBN. Sensors 2017, 17, 1694.

[15] Anjali Bhavan, Pankaj Chauhan, Hitkul, Rajiv Ratn Shah, “Bagged support vector machines for emotion recognition from speech”, Knowledge-Based Systems , June 2019.

[16] Hajarolasvadi, N.; Demirel, H. 3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms. Entropy, 21, 479, 2019.

[17] Mustaqeem and Soonil Kwon, “A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition”, Journal Sensors, Dec 2019.