Human Emotion Recognition using Hybrid approach

Darshita Shah, Swarndeep Saket
Student, Assistant Professor (CE)
ME (CE) LJIET
Ahmedabad, Gujarat

Abstract: *Speech processing one of the new areas for research in current generation. In spite of the fact that feeling location from discourse is a moderately new field of research and has numerous potential applications. In human computer or human-human connection frameworks, feeling acknowledgment frameworks could furnish clients with improved administrations by being versatile to their feelings. As of late takes a shot at impact location utilizing discourse and distinctive issues identified with influence discovery has been introduced. The essential difficulties of feeling acknowledgment are picking the feeling acknowledgment corpora (discourse database), recognizable proof of various highlights identified with discourse and a fitting decision of an arrangement model. Diverse sorts of strategies to gather passionate discourse information and issues identified with them are secured by this introduction alongside the past works audit. Here, we have used Mel Frequency Cepstral Coefficient (MFCC) And Hidden Markov Model (HMM) technique for proposed system to identify feature and then applied machine learning for classifying types of emotions from speech frame with improved accuracy.*

Keyword: Speech Emotion Recognition; MFCC (Mel Frequency Cepstral Coefficient); Support Vector Mechanism (SVM).

I. INTRODUCTION

Emotions play a compulsory vital role in human life. It might be known from the speech uttered by the person, it's a medium of expression of one's perspective or feelings to others. Speech recognition is a computer science term and is also known as automatic speech recognition [9]. Speech Recognition also known as computer speech recognition is a process in which speech signal is converted into a sequence of words, other linguistic units by making use of an algorithm which is implemented as a computer program [10]. In a speech recognition system we convert speech into text in which the text is the output of the speech recognition system which is equivalent to the recognized speech [10]. It is also a big advantage to people who may suffer from disabilities that affect their writing ability but can use their speech to create text on computers or other devices [8]. Machine Learning is a subset of artificial intelligence. It focuses mainly on the designing of system, thereby allowing them to learn and make prediction based on some experience which is data in case of machines. The primary purpose of this analysis is to spot the benefits and limitations of unimodal systems, and to point out that fusion approaches are a lot of appropriate for emotion recognition. Machine Learning enable computer to act and make data decision rather than be explicitly program to carry out a certain task this programs are designed to learn and improve our time when exposed to new data

II. Machine Learning

Machine learning is a paradigm that may refer to learning from past experience (which in this case is previous data) to improve future performance. The sole focus of this field is automatic learning methods. Learning refers to modification or improvement of algorithm based on past "experiences" automatically without any external assistance from human. While designing a machine (a software system), the programmer always has a specific purpose in mind. For instance, consider J. K. Rowling's Harry Potter Series and Robert Galbraith's Cormoran Strike Series. To confirm the claim that it was indeed Rowling who had written those books under the name Galbraith, two experts were engaged by The London Sunday Times and using Forensic Machine Learning they were able to prove that the claim was true. They develop a machine learning algorithm and "trained" it with Rowling's as well as other writers writing examples to seek and learn the underlying patterns and then "test" the books by Galbraith. The algorithm concluded that Rowling's and Galbraith's writing matched the most in several aspects. So instead of designing an algorithm to address the problem directly, using Machine Learning, a researcher seek an approach through which the machine, i.e., the algorithm will come up with its own solution based on the example or training data set provided to it initially.[6]
The signal level processing, artificial intelligence and machine learning technologies have boosted the machine intelligence, so that the machines gained the capability to understand human emotions. Incorporating the aspects of speech processing and pattern recognition algorithms an intelligent and emotions specific man-machine interaction can be achieved which can be harnessed to design a smart and secure automated home as well as commercial application. [7]

A. Speech emotion Description

Types of Speech Recognition System

I. Text-To-Speech.

Text-To-Speech (or TTS) will manipulate a string of text into an audio clip. It is useful for blind people to be able to use computers but can also be used to simply improve computer experience. There are several programs available that perform TTS, some of which are command-line based (ideal for scripting) and others which provide a handy GUI [8].

II. Simple Voice Control/Commands.

This is the most basic form of Speech-To-Text application. These are designed to recognize a small number of specific, typically one-word commands and then perform an action. This is often used as an alternative to an application launcher, allowing the user for instance to say the word *"Firefox"* and have his OS open a new browser window [8].

III. Full dictation/recognition.

Full dictation/recognition software allows the user to read full sentences or paragraphs and translates that data into text on the fly. This could be used, for instance, to dictate an entire letter into the window of an email client . In some cases, these types of applications need to be trained to your voice and can improve in accuracy the more they are used [8].

Types of speech

Speech Recognition System can be separated in different classes by describing what type of utterances they can recognize.

1. **Isolated Word**
2. **Connected Word**
3. **Continuous speech**
4. **Spontaneous speech**

Automatic Speech Recognition system classification:

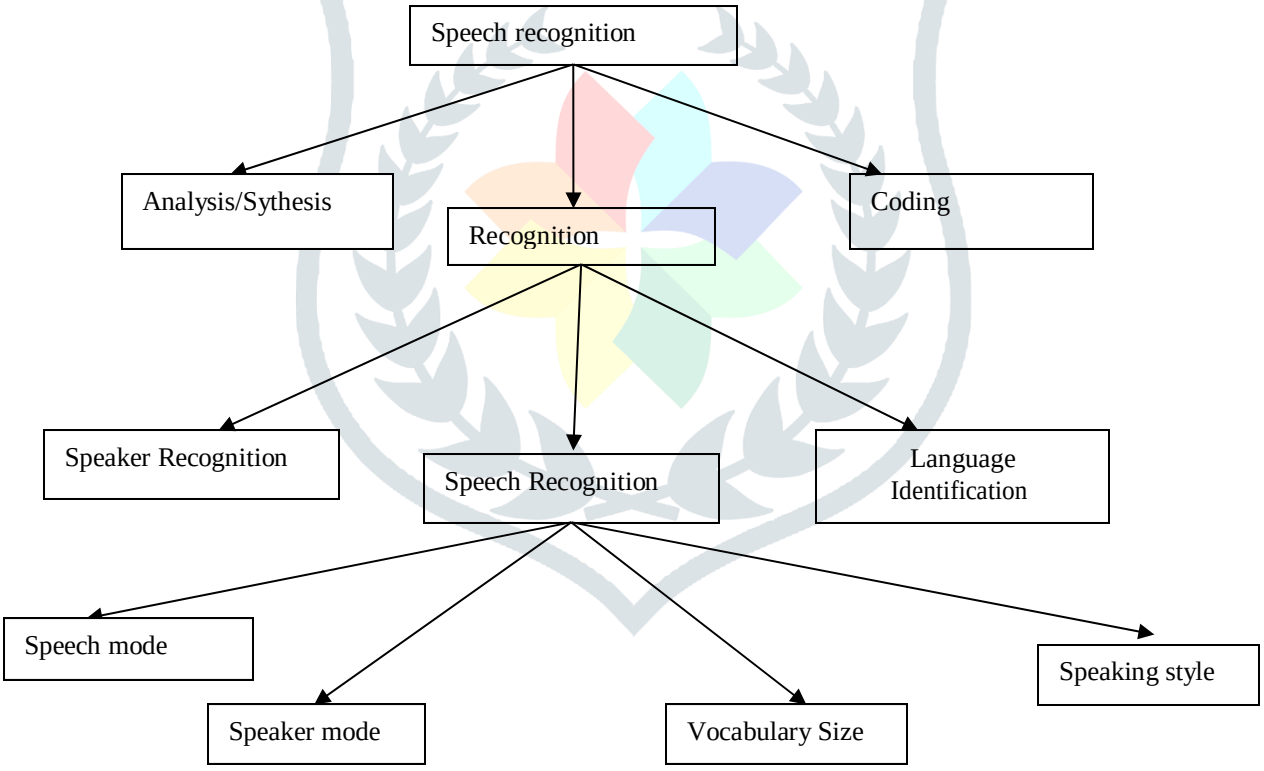


Fig.2 Speech Processing Classification [11]

The following tree structure emphasizes the speech processing applications. Depending on the chosen criterion, Automatic Speech Recognition systems can be classified as shown in figure 1.

a. Speech Mode (Type)

Discourse Recognition System can be isolated in various classes by depicting what sort of expressions they can perceive.

I. Isolated Word

Segregated word perceives achieve more often than not require every expression to have calm on both side of test windows. It acknowledges single words or single expressions at once .This is having "Tune in and Non Listen state". Disconnected articulation may be better name of this class [11].

II. Connected Word

Associated word framework are like disengaged words however enable separate articulation to be "run together least interruption between them [11].

II. Continuous discourse

Constant discourse recognizers enables client to talk normally, while the PC decide the substance. Recognizer with proceeds with discourse capacities are the absolute most hard to make since they use unique strategy to decide articulation limits [11].

IV. Spontaneous discourse

At a Starting dimension, it tends to be thought of Voice that is Inartificial sounding and not practiced .Here is An ASR System with unconstrained Voice Caliber ought to have the capacity to deal with a Types of Inartificial Voice highlight, for example, words being run together[11].

b. Speaker Mode

Every speaker has exceptional voice, because of his one of a kind physical body and identity. Discourse acknowledgment framework is arranged into three principle classifications as pursues:

I. Speaker Dependent.

Speaker subordinate frameworks are produced for a specific sort of speaker. They are commonly progressively exact for the specific speaker, yet could be less precise for other sort of speakers. These frameworks are normally less expensive, simpler to create and increasingly precise. Be that as it may, these frameworks are not adaptable as speaker autonomous frameworks [10].

II. Speaker Indepented.

Speaker Independent framework can perceive an assortment of speakers with no earlier preparing. . A speaker free framework is created to work for a specific kind of speaker. It is utilized in Interactive Voice Response System (IVRS) that must acknowledge contribution from a substantial number of various clients. However, disadvantage is that it restricts the quantity of words in a vocabulary. Usage of Speaker Independent framework is the most troublesome. Likewise it is costly and its exactness is lower than speaker subordinate frameworks [10].

III. Speaker Adaptive.

Speaker versatile discourse acknowledgment framework utilizes the speaker subordinate information and adjusts to the most appropriate speaker to perceive the discourse and diminishes mistake rate by adaption. They adjust task as indicated by attributes of speakers [10].

c. Vocabulary Size

The extent of vocabulary of a discourse acknowledgment framework can influence the multifaceted nature, preparing and the rate of acknowledgment of ASR framework. So that ASR framework are arranged dependent on the vocabulary as following [10]:

I. Small

Little Vocabulary - 1 to 100 words or sentences

II. Medium

Medium Vocabulary - 101 to 1000 words or sentences

III. Large

Substantial Vocabulary-1001 to 10,000 words or sentences

2.5.4 Feature Extraction Techniques for Speech Recognition

Highlight extraction is the most significant grammatical form acknowledgment as it recognizes one discourse from other. The pron unciation can be grow from an Extensive position of highlight extraction strategies recommended and effectively used for Voice Admitted task, yet extend highlight should meet a few criteria while consulting of the Voice flag, for example, [19]:

- Easy to gauge extricated discourse highlight
- It ought not be open to mimicry
- It should indicate less variety starting with one talking condition then onto the next
- It ought to be adjusted after some time
- It ought to happen typically and normally in discourse

Distinctive strategies for highlight extraction are LPC, MFCC, AMFCC, PLP, PCA, cepstral investigation, RASTA Filtering and so forth [19].

I. LPC: It is one of the significant strategy for discourse examination since it can give a gauge of the shafts (subsequently the formant recurrence created by vocal tract) of the vocal tract exchange work. LPC (Linear Predictive Coding) investigations the discourse motion by evaluating the formants, expelling their belongings from the Voice flag, and assessing the recurrence of the rest of the automaton. The Activity of erasing the formants is called turn around sifting and the staying signal is known as the buildup. The essential thought behind LPC coding is that each example can be approximated as a straight mix of a couple past examples. The lineal forecast Process gives a hearty, dependable, action for assessing the parameters. The calculation associated with LPC preparing is significantly not exactly cepstrum investigation [14].

I. Cepstral Analysis: This examination is an extremely advantageous approach to demonstrate otherworldly vitality dissemination. Cepstral examination works in an area in which the glottal recurrence is isolated from the vocal tract resonances [20]. The short request co-variables of the cepstrum containing data about the vocal tract, while the higher request co-factors containing fundamentally data about the blackout. (All things considered, the higher request co-factors contain the two kinds of data, however the recurrence of periodicity commands). The word cepstrum determined by inversing the primary syllable in the word range. The cepstrum exists in an area alluded to as quefrency (inversion of the prime syllable in recurrence) which has units of time[14]. The cepstrum is characterized as the turn around Fourier change of the logarithm of the power range. The Cepstrum is the Forward Fourier Transform of a range. It is in this way the range of a range, and has certain properties that make it valuable in numerous sorts of flag examination. Cepstrum coefficients are determined in short casings after some time. Just the principal MF cepstrum coefficients are utilized as highlights (all coefficients model the exact range, coarse otherworldly shape is demonstrated by the main coefficients, exactness is chosen by the quantity of coefficients taken, and the primary coefficient (vitality) is generally disposed of). The cepstrum is determined in two different ways: LPC cepstrum and FFT cepstrum [14].

II. MFCC: This strategy is considered as one of the standard technique for highlight extraction and is acknowledged as the benchmark. MFCCs depend on the known deviation of the human ear's basic transmission capacities with recurrence; channels dispersed at short frequencies and logarithmically at high frequencies have been utilized to take phonetically significant qualities of voice[14]. This is communicated in the Mel-recurrence scale (the Mel scale was utilized by Mermelstein and Davis to extricate highlights from the discourse motion for improving the acknowledgment execution). MFCC are the consequences of the momentary vitality communicated of Mel-recurrence scale The MFCCs are

demonstrated more productive preferable enemy of commotion capacity over other vocal tract parameters, for example, LPC [15].

2.1.1 Types of ML

1.1 Supervised learning: In layman language regulated learning can be characterized as "Preparing information incorporates wanted yields". Managed learning is the Data mining undertaking of inducing a capacity from marked preparing information. The training information comprise of a lot of preparing models. In managed adapting, every precedent is a couple comprising of an information object (ordinarily a vector) and ideal yield esteem (additionally called the supervisory flag). [15]

Directed AI frameworks give the learning calculations realized amounts to help future decisions. Identifying Diseases (Medical Treatment), Chat-bots, self-driving autos, facial acknowledgment programs, master frameworks and robots are among the frameworks that may utilize either managed or unsupervised learning. Managed learning frameworks are for the most part connected with recovery based AI however they may likewise be fit for utilizing a generative learning model. [15]

2.1 Unsupervised learning

In Unsupervised Learning the "Preparation information does exclude wanted yields". Unsupervised learning is the preparation of a man-made brainpower (AI) calculation utilizing data that is neither characterized nor named and enabling the calculation to follow up on that data without direction. [15]

In unsupervised learning, an AI framework may amass unsorted data as indicated by likenesses and contrasts despite the fact that there are no classifications given. Computer based intelligence frameworks equipped for unsupervised learning are frequently connected with generative learning models, in spite of the fact that they may likewise utilize a recovery based methodology (which is regularly connected with managed learning). [15]

3.1 Semi-supervised learning

Semi-regulated learning is a class of AI undertakings and systems that likewise utilize unlabelled information for preparing – normally a modest part of named information with a major parcel of unlabelled information. Semi-directed learning falls between unsupervised learning (with no marked preparing information) and managed learning (with totally named preparing information). Many AI scientists have discovered that unlabelled information, when utilized related to a little measure of named information, can deliver impressive improvement in learning exactness over unsupervised realizing (where no information is named), however without the time and costs required for regulated realizing (where all information is marked).[15]

4.1 Reinforcement learning

Fortification Learning enables the machine or programming specialist to gain proficiency with its conduct dependent on input from the earth. This conduct can be adapted unequivocally, or continue adjusting as time passes by. On the off chance that the issue is displayed with consideration, some Reinforcement Learning calculations can consolidate to the worldwide ideal; this is the perfect conduct that boosts the reward. [16]

B. APPLICATIONS OF MACHINE LEARNING [6]

1. SPEECH RECOGNITION
2. COMPUTER VISION.
3. BIO-SURVEILLANCE
4. ROBOT OR AUTOMATION CONTROL
5. EMPIRICAL SCIENCE EXPERIMENTS

III. LITERATURE SURVEY

Sr No.	Paper Title	Method / Classification	Advantages	Disadvantages
1	Emotion Detection using MFCC and Cepstrum Features	Mal Frequency Cepstral Coefficients (MFCC), Artificial Neural Network (ANN).	One method of SER has been presented in this paper and an accuracy of 85.7% is obtained in detecting 7 emotions.	In this paper seven emotions are considered but the emotion sad could not be recognized. Sad and happy are two extreme emotions having a very narrow feature set and this is leading to a misclassification.
2	Performance Analysis	Mal Frequency Cepstral Coefficients (MFCC),Short ti Network (ANN), Support Vector Mechanism (SVM)	We can achieve 88.4% accuracy using ANN classifier and 78.2 % accuracy using SVM classifier.	SVM had a problem with classification if dataset is small. We can achieve 88.4% or 78.2% accuracy but only four different emotion recognized into the 365 types.
3	An Experimental Study of Speech Emotion Recognition Based on Deep Convolutional Neural Networks	Principle component analysis – Deep convolution Neural Networks- Speech Emotion Recognition(PCA-DCNNS-SER), Support Vector Mechanism (SVM)	We have extracted 85 dimensional hand- crafted features from the speech files for each 25ms frame.	SVM had a problem with classification if dataset is small. We can achieve 88.4% or 78.2% accuracy but only four different emotion recognized into the 365 types.
4	Emotion Recognition Using Machine Learning	Mal Frequency Cepstral Coefficients (MFCC), Fast Fourier Transform (FFT)	The different classification strategies the maximum accuracy of 81.05 % is obtained for the database by using Random Decision Forest classifier.	This system was valid only for 3 types of emotions. Accuracy go down if more types of emotion detect.
5	Speech Based Human Emotion Recognition Using MFCC	Mal Frequency Cepstral Coefficients (MFCC),	Efficiency was found to be about 80%. This efficiency performance continued even in noisy environment.	The designed system was validated only for Happy, Sad, and Anger emotions.

Table 1: Comparative Study

A. RELAVANCE SURVEY ON EMOTION SPEECH RECOGNITION

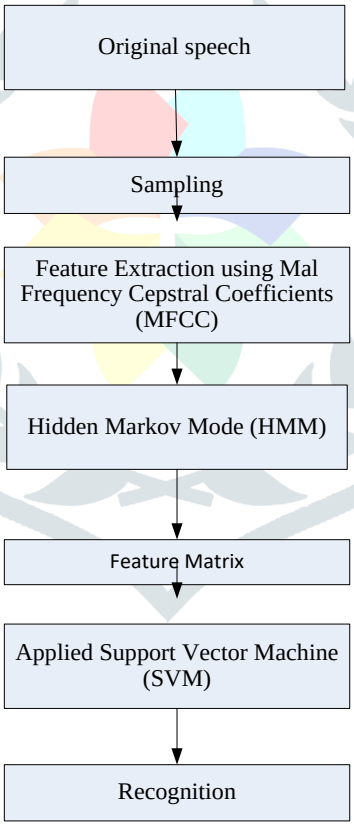
In this paper [1] One method of SER has been presented in this paper and an accuracy of 85.7% is obtained in detecting 7 emotions. These results are achieved using cepstral based features compact feature vector, and a simple Neural Network Classifier. Since a supervised testing is done, it is better compared to six. In this work, seven emotions are considered but the emotion Sad could not be recognized. Sad and Happy are two extreme emotions having a very narrow feature set and this is leading to a misclassification.

Rajisha T. M.[a], Sunija A. P.[b], Riyas K. S[c] [2] carried out automatic recognition of four different emotions anger, happy, sad and neutral by using features Mel frequency cepstral coefficients (MFCCs), Pitch and Short Time Energy (STE). The experiments on dataset shows that speech emotion recognition with ANN classifier has better recognition accuracy of 88.4 % as compared to SVM, 78.2 %.

This paper [3] Accordingly, a speech emotion recognition algorithm termed as PCA-DCNNs-SER is proposed. Preliminary experiments have been conducted to evaluate the performance of PCA-DCNNs-SER on the IEMOCAP database. Results show that our proposed PCADCNNs-SER (containing 2 convolution and 2 pooling layers) is able to obtain about 40% classification accuracy, which outperforms the SVM based SER using hand-crafted features.

In this paper [4] This method of speech emotion recognition has proven to be 80% efficient. This efficiency in performance continued even in noisy environment. Hence this system can serve as noise robust emotion recognition system. Such efficiency in noisy environment extends the scope of the work wherein emotion recognition systems can be utilized in military.

This paper [5] Hence, we concluded that inclusion of energy as a feature along with other 13 MFCC features led to better assessment of the emotion attached with the speech. It can be seen that integrating frames into overlapping segments led to a greater continuity in samples and also resulted in each data point having many more features. PROPOSED WORK



PROPOSED SYSYTEM

Steps of Flow chart

Step 1: Load original speech using laptop audio recorder.

Step 2: Apply pre-processing and Fast Fourier's transformation an input speech.

Step 3: Apply Mal Frequency Cepstral Coefficients (MFCC) & Hidden Markov Mode (HMM) model for extent meaningful feature audio.

Step 4: Divide data & speech in training and testing samples.

Step 5: Use machine learning for recognition of speech & classify it with high accuracy.

IV. RESULT

Number	Wave File	Accuracy	Emotions
1.Wave	F1 Happy	59.2544	Happy
2.Wave	F2 Happy	94.2660	Happy
3.Wave	F3 Happy	84.6360	Happy
4.Wave	F1 Sad	85.2888	Sad
5.Wave	F1 Sad	60.9569	Sad
6.Wave	F1 Normal	90.5771	Normal
7.Wave	F2 Normal	77.67	Normal
8.Wave	F3 Normal	60.69	Normal
9.Wave	F4 Normal	57.06	Normal

Table: Emotion Accuracy

V. CONCLUSION

Speech based emotion recognition find numerous application in automated speech voice recognition system. In this dissertation work on Speech emotion identification that emotion like sad, happy, Normal and so on. It has good implication in analysis application like live detection, diagnosis of mental depression etc. We have utilized proposed system work on MFCC and HMM based feature matrix calculation for identification meaningful features from the voice signal. We have improved accuracy and time complexity.

VI. REFERENCES

[1] Mohan Ghai, Shamit Lal, Shivam Dugga l and Shrey Manik."Emotion Recognition On Speech Signals Using Machine Learning" 978-1-5090-6399-4/17/$31.00c 2017 IEEE.

[2] M.S. Likitha[1] ,Sri Raksha R. Gupta[2] ,K. Hasitha[3] and A. Upendra Raju[4] " Speech Based Human Emotion Recognition Using MFCC" 978-1-5090-4442-9/17/$31.00c 2017 IEEE.

[3] Rajisha T. M.[a], Sunija A. P.[b], Riyas K. S[c] " Performance Analysis of Malayalam Language Speech Emotion Recognition System using ANN/SVM" doi: 10.1016/j.protcy.2016.05.242 pg No. 1098 – 1104.

[4] W. Q. Zheng, J. S. Yu, Y. X. Zou " An Experimental Study of Speech Emotion Recognition Based on Deep Convolutional Neural" 978-1-4799-9953-8/15/$31.00 ©2015 IEEE pg No. 827-831.

[5] S Lalitha [a], D Geyasruti [a], R Narayanan[a], Shravani M[a] " Emotion Detection using MFCC and Cepstrum Features" 1877-0509 © 2015 pg No. *29 – 35*

[6] Kajaree Das[1], Rabi Narayan Behera[2] " A Survey on Machine Learning: Concept, Algorithms and Applications" IJIRCCE Vol. 5, Issue 2, February 2017 P.P 1301 – 1309

[7] Chul Min Lee "Toward Detecting Emotions in Spoken Dialog" IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 13, NO. 2, MARCH 2005.

[8] Anchal Katyal, Amanpreet Kaur, Jasmeen Gill, "Automatic Speech Recognition: A Review", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-3, Issue-3, February 2014.

[9] Santosh K.Gaikwad,Bharati W.Gawali,Pravin Yannawar, "A Review On Speech Recognition Technique", International Journal of Computer Applications(0975-8887) Volume 10-No3,November 2010.

[10] Shreya Narang1, Ms. Divya Gupta, "Speech Feature Extraction Techniques: A Review", IJCSMC, Vol. 4, Issue. 3, March 2015.

[11] M.A.Anusuya, S.K.Katti,“ Speech Recognition by Machine: A Review”, International Journal of Computer Science and Information Security,Vol. 6, No. 3, 2009

[12] Namrata Dave, “Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition”, INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY, Volume 1, Issue VI, July 2013.

[13] Varsha Singh1, Vinay Kumar Jain2, Dr. Neeta Tripathi, “A Comparative Study on Feature Extraction Techniques For Language Identification”, International Journal of Engineering Research and General Science Volume 2, Issue 3, April-May 2014.

[14] https://dataaspirant.wordpress.com/2014/09/19/supervised-and-unsupervised-learning/

[15] https://en.wikipedia.org/wiki/Semi-supervised_learning/

[16] http://reinforcementlearning.ai-depot.com/