Analysis on Speech processing technique using Machine Learning

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

Abstract:*Speech processing one of the new areas for research with ANN(Artificial Neural Network) system in current generation. Although emotion detection from speech is a relatively new field of research, and has many potential applications. In human-computer or human-human interaction systems, emotion recognition systems could provide users with improved services by being adaptive to their emotions. Recent works on effect detection using speech and different issues related to affect detection has been presented. The primary challenges of emotion recognition are choosing the emotion recognition corpora (speech database),identification of different features related to speech and an appropriate choice of a classification model. Different types of methods to collect emotional speech data and issues related to them are covered by this presentation along with the previous works review. Here is use Mel Frequency Cepstral Coefficient (MFCC) with pre - emphasis technique for identify feature and then applied machine learning for classifying types of emotions from speech frame with high 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 assistancefrom 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 wereengaged by The London Sunday Times and using Forensic Machine Learning they were able to prove that the claimwas true. They develop a machine learning algorithm and “trained” it with Rowling’s as well as other writers writingexamples to seek and learn the underlying patterns and then “test” the books by Galbraith. The algorithm concludedthat 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:

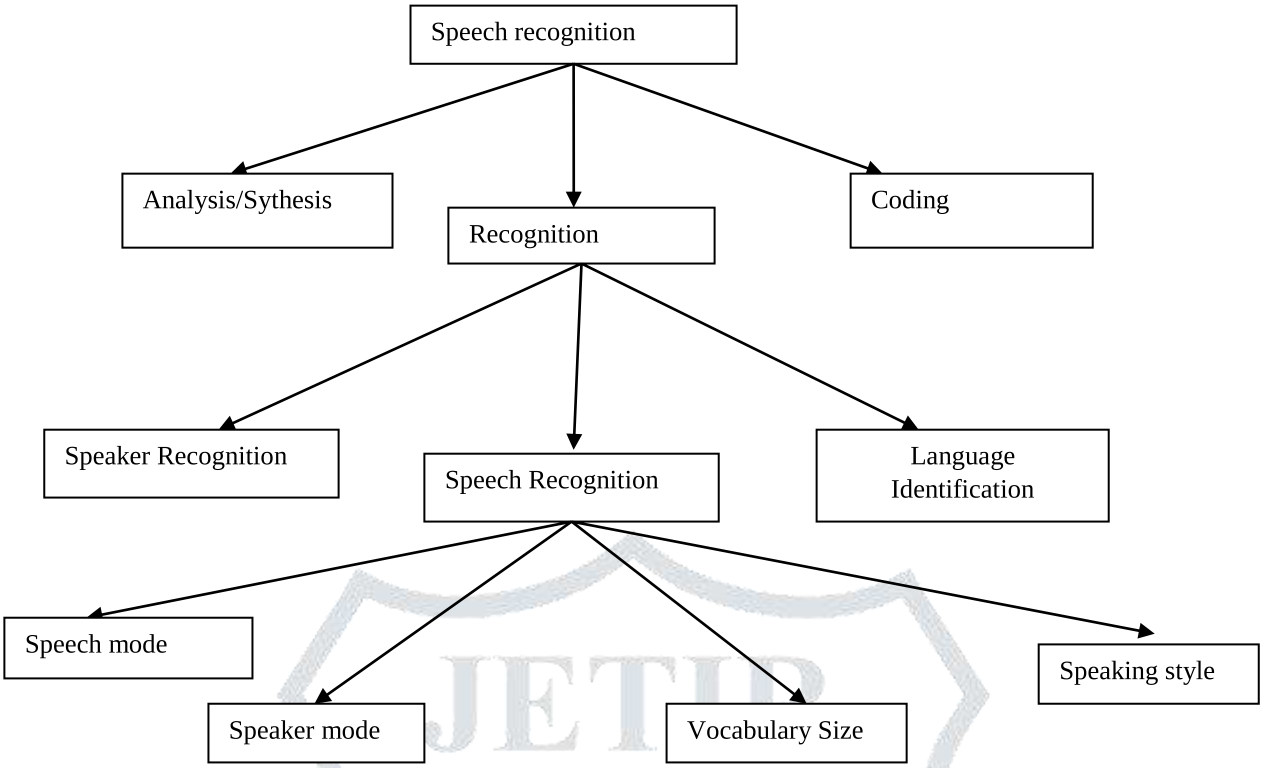


Fig.2 Speech Processing Classification[11]

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

Speech Mode (Type)

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

I. **Isolated Word**

Isolated word recognizes attain usually require each utterance to have quiet on both side of sample windows. It accepts single words or single utterances at a time .This is having “Listen and Non Listen state”. Isolated utterance might be better name of this class[9].

II. **Connected Word**

Connected word system are similar to isolated words but allow separate utterance to be “run together minimum pause between them[9].

III. **Continuous speech**

Continuous speech recognizers allows user to speak almost naturally, while the computer determine the content. Recognizer with continues speech capabilities are some of the most difficult to create because they utilize special method to determine utterance boundaries[9].

IV. **Spontaneous speech**

At a basic level, it can be thought of as speech that is natural sounding and not rehearsed .an ASR System with spontaneous speech ability should be able to handle a variety of natural speech feature such as words being run together[9].

B. **Feature Extraction Techniques for Speech Recognition**

Feature extraction is the most important part of speech recognition as it distinguishes one speech from other. The utterance can be extracted from a vast range of feature extraction techniques suggested and successfully utilized for speech recognition task, but extracted feature should meet some criteria while negotiating with the speech signal such as [12]:

- Easy to measure extracted speech feature
- It should not be receptive to mimicry
- It should show less variation from one speaking environment to another
- It should be balanced over time
- It should occur normally and naturally in speech

Different techniques for feature extraction are LPC, MFCC, AMFCC, PLP, PCA, cepstral analysis, RASTA Filtering etc. [12].

I. **LPC:** It is one of the important method for speech analysis because it can provide an estimate of the poles (hence the formant frequency- produced by vocal tract) of the vocal tract transfer function. LPC (Linear Predictive Coding) analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering and the remaining signal is called the residue.The basic idea behind LPC coding is that each sample can be approximated as a linear combination of a few past samples. The linear prediction method provides a robust, reliable, and accurate method for estimating the parameters. The computation involved in LPC processing is considerably less than cepstrum analysis [13].

II. **Cepstral Analysis:** This analysis is a very convenient way to model spectral energy distribution. Cepstral analysis operates in a domain in which the glottal frequency is separated from the vocal tract resonances [13]. The low order coefficients of the cepstrum contain information about the vocal tract, while the higher order coefficients contain primarily information about the excitation. (Actually, the higher order coefficients contain both types of information, but the frequency of periodicity dominates). The word cepstrum was derived by reversing the first syllable in the word spectrum. The cepstrum exists in a domain referred to as quefrency (reversal of the first syllable in frequency) which has units of time[13]. The cepstrum is defined as the inverse Fourier transform of the logarithm of the power spectrum. The Cepstrum is the Forward Fourier Transform of a spectrum. It is thus the spectrum of a spectrum, and has certain properties that make it useful in many types of signal analysis [13].

III. **MFCC:** This technique is considered as one of the standard method for feature extraction and is accepted as the baseline. MFCCs are based on the known variation of the human ear's critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech[13]. This is expressed in the Mel-frequency scale (the Mel scale was used by Mermelstein and Davis to extract features from the speech signal for improving the recognition performance). MFCC are the results of the short-term energy spectrum expressed on a Mel-frequency scale The MFCCs are proved more efficient better anti-noise ability than other vocal tract parameters, such as LPC [13].

C. Types of ML

1. Supervised learning

In layman language supervised learning can be defined as "Training data includes desired outputs". Supervised learning is the Data mining task of inferring a function from labelledtraining data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). [14]

Supervised machine learning systems provide the learning algorithms with known quantities to support future judgments. Detecting Diseases (Medical Treatment), Chat-bots, self-driving cars, facial recognition programs, expert systems and robots are among the systems that may use either supervised or unsupervised learning. Supervised learning systems are mostly associated with retrieval-based AI but they may also be capable of using a generative learning model. [14]

Training data for supervised learning includes a set of examples with paired input subjects and desired output (which is also referred to as the supervisory signal). In supervised learning for image processing, for example, an AI system might be provided with labelled pictures of vehicles in categories such as cars and trucks. After a sufficient amount of observation, the system should be able to distinguish between and categorize unlabelled images, at which time training can be said to be complete. [14]

2. Unsupervised learning

In Unsupervised Learning the "Training data does not include desired outputs". Unsupervised learning is the training of an artificial intelligence (AI) algorithm using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance.[14]

In unsupervised learning, an AI system may group unsorted information according to similarities and differences even though there are no categories provided. AI systems capable of unsupervised learning are often associated with generative learning models, although they may also use a retrieval-based approach (which is most often associated with supervisedlearning). [14]

Unsupervised learning algorithms can perform more complex processing tasks than supervised learning systems. However, unsupervised learning can be more unpredictable than the alternate model. [14]

3. Semi-supervised learning

Semi-supervised learning is a class of machine learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy over unsupervised learning (where no data is labeled), but without the time and costs needed for supervised learning (where all data is labeled).[15]

4. Reinforcement learning

Reinforcement Learning allows the machine or software agent to learn its behaviour based on feedback from the environment. This behaviour can be learnt once and for all, or keep on adapting as time goes by. If the problem is modelled with care, some Reinforcement Learning algorithms can converge to the global optimum; this is the ideal behaviour that maximises the reward.[16]

This automated learning scheme implies that there is little need for a human expert who knows about the domain of application. Much less time will be spent designing a solution, since there is no need for hand-crafting complex sets of rules as with Expert Systems, and all that is required is someone familiar with Reinforcement Learning.[16]

D. 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

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.

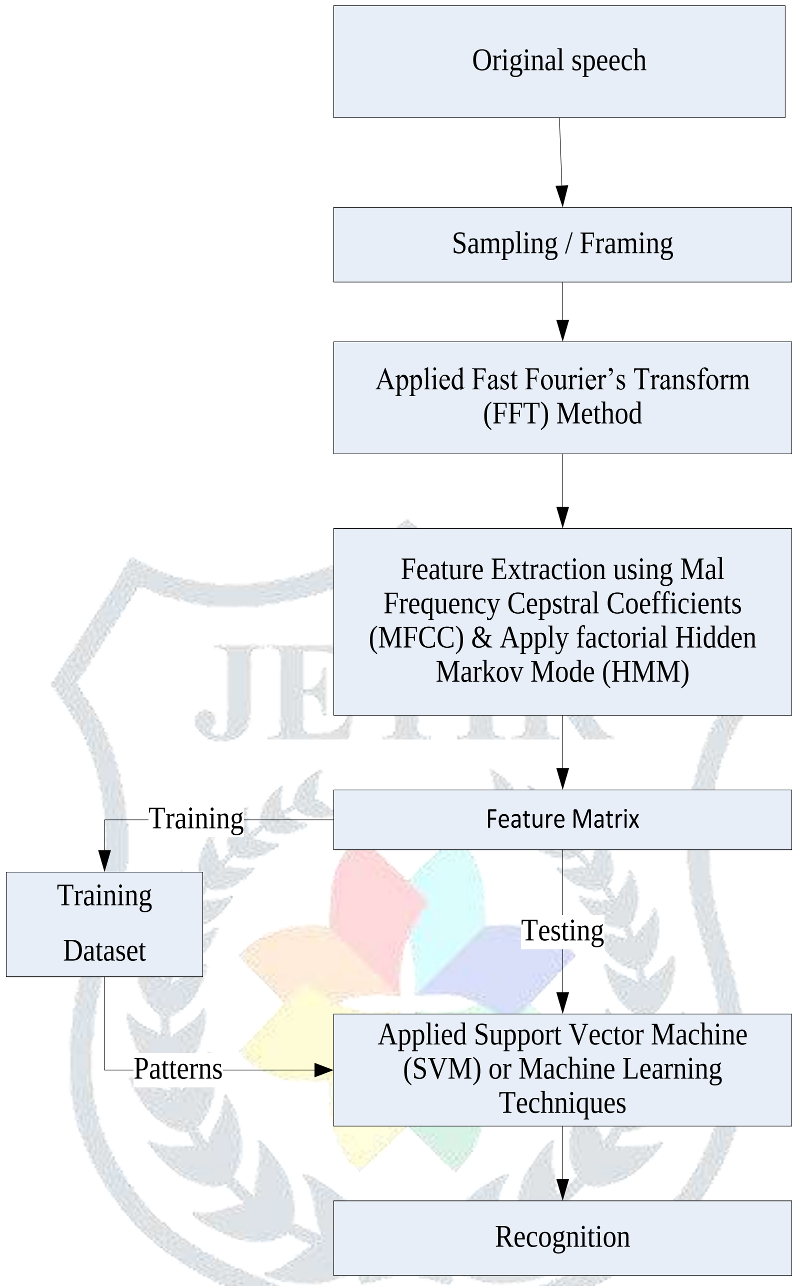
B. Comparison table

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 of Malayalam Language Speech Emotion Recognition System using ANN/SVM.	Mal Frequency Cepstral Coefficients (MFCC),Short time energy (STE), Pitch, Artificial Neural Network (ANN), Support Vector Mechanism (SVM)	We can achieve 88.4% accuracy using ANN classifier and 88.4 % 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	Principle component analysis – Deep	We have extracted 85 dimensional hand-	SVM had a problem with classification if dataset is

	Emotion Recognition Based on Deep Convolutional Neural Networks	convolution Neural Networks- Speech Emotion Recognition(PCA-DCNNS-SER), Support Vector Mechanism (SVM)	crafted features from the speech files for each 25ms frame.	small. We can achieve 88.4% or 78.2% accuracy but only four different emotion recognized into the 365 types.
4	Emotion Recognition On Speech Signals 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 in performance continued even in noisy environment.	The designed system was validated only for Happy, Sad, and Anger emotions.

Table 1: Comparative Study

IV. PROPOSED WORK


Original speech
Sampling / Framing
Applied Fast Fourier's Transform (FFT) Method
Feature Extraction using Mal Frequency Cepstral Coefficients (MFCC) & Apply factorial Hidden Markov Mode (HMM)
Feature Matrix
Training
Training Dataset
Testing
Patterns
Applied Support Vector Machine (SVM) or Machine Learning Techniques
Recognition

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.

V. CONCLUSION

In proposed work increase Accuracy and time complicity which is improving using MFCC & HMM Feature extraction method and used other classification algorithm. According to competitive analysis certain method's and Algorithm's available for Emotion recognition base on speech. In future design and implement proposed model on Matlab 2018a.

VI. REFERENCES

[1] Mohan Ghai, ShamitLal, ShivamDugga l and ShreyManik."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] AnchalKatyal, AmanpreetKaur, 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] SantoshK.Gaikwad,BharatiW.Gawali,PravinYannawar, "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/