A STUDY ON EFFICIENT AUTOMATIC SPEECH RECOGNITION SYSTEM TECHNIQUES AND ALGORITHMS

Yakubu A. Ibrahim [1], Tunji S. Ibiyemi [2]

[1] **Department of Computer Science, Bingham University, Karu, Nigeria**
[2] **Department of Electrical Engineering, University of Ilorin, Ilorin, Nigeria**

Corresponding Author: Yakubu A. Ibrahim, talktoibro80@gmail.com

ABSTRACT: Automatic speech recognition is a system by which computer recognizes and responds accordingly to a spoken words of a person on the basis of his or her voice signal waveform. ASR system is being adopted in different aspects of life, such as telephones and home computer control system. Despite their growing presence, a proper technique for efficient ASR system remains a major issue for researchers. This study explains an analysis of different techniques and algorithms that can be used in ASR system such as LPC, LPCC, PLP, MFCC for feature extraction and DTW, SVM, HMM, VQ, GMM, MLP, ANN, KNN for feature classification and pattern recognition.

KEYWORDS: ASR, DTW, HMM, Feature extraction, LPC, MFCC, SVM.

I. INTRODUCTION

Automatic Speech Recognition is the process by which a system recognizes a speech signal of acquiring the transcription (word sequence) of an utterance, given the speech waveform ([GRP17]). Speech signal processing systems have become available in all areas of life. However, the present natures of human computer interaction are directed towards living with the limitations of computer input and output devices such as keyboard rather than the convenience of human users. Speech is the basic way of human-to-human communication. The popular ways of input to computer system is through a keyboard or a mouse. Therefore, it would be good if speech waveform signals could be adopted to interact with the computer system. Human beings can vividly identify voice of a particular speaker while the person is speaking to another individual at any point in time. The direction of a person speaking may be different or the same location with a person spoken to. In fact, a blind individual can also correctly recognize the person speaking based solely on his or her acoustic speech sound. Animals such as mammals adopt these patterns to recognize their off springs as well as their familiar ones ([I+13]). An efficient speaker speech identification system has been a target of active research during last two decades because it has various numbers of useful applications in many areas that need correct user identification system such as shopping by telephone, voicemail, and accesses control services ([SRR10]).

II. CLASSIFICATION OF SPEECH

Speech recognition systems can be divided into many different parts by describing what types of utterances they have the ability to recognize. These are classified as the following:
A number of parameters define the capability of a speech recognition system.
a) Spontaneous speech: A System with spontaneous speech ability should be able to handle a variety of natural speech feature such as words being run together.
b) Continuous speech: It allows user to speak almost naturally, while the computer will check the content. There are special methods used to determine utterance boundaries and various difficulties occurred in it.
c) Connected word: The Connected word system is similar to isolated utterance but allow separate utterance to be run together with a minimum pause between them.
d) Isolated word: The Isolated utterance requires each word have sample windows and has a silence between both sides of the windows. It allows single utterance or single word at a given period of time.

III. AUTOMATIC SPEECH RECOGNITION SYSTEM (ASR)

Speech processing system involves the study of speech signal waveforms and the methods of processing these speech signals for proper recognition. The speech signal waveforms are normally processed in a digital format for good representation thereby speech processing can be referred to as the interaction of digital signal processing and language processing. Language processing, however, is a subarea of linguistics and artificial intelligence. It showcases the issues of automated natural language generation and understanding of human natural languages. Production of natural language systems converts information from computer databases into normal-speech of human language. The natural language understanding systems on the other hand converts samples of natural human language into more formal representations that are easier for computer programs to execute ([SRR10]). A typical ASR process comprise of many different components which is shown in figure 1.

IV. FEATURE EXTRACTION

In ASR system application, feature extraction is method of removing the vital and necessary information of the speech waveform signal while repeated and unwanted data information of a given speech signals are rejected. In other words, it is a way of analyzing speech waveform of a particular signal to generate useful as well as important speech signal while removing the unwanted information from the signal, in this process some useful information may be deleted. This method involves converting the acoustic speech waveform signal into a form which will be good and better for pattern classification by the system ([Ras14]). In ASR system development, some vital properties of the speech features which include high difference between sub-word classes, low speaker speech variability, invariance to degenerations in speech waveform signal due to noise and channel.

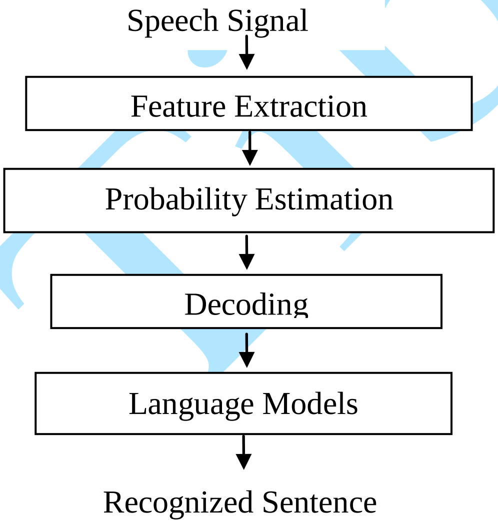

Figure 1: Speech Recognition Process ([GRP17])

The aim and target of feature extraction method in any ASR system is to produce certain set of properties of speech signal that have acoustic correlates in such speech signal, in the same vein, these properties makes parameters that can be estimated or computed through signal waveform processing. Such parameters are referred to as features. The parameterization of speech waveform signal is called as feature extraction ([S+12]). Feature extraction involves the process of the conversion of the speech waveform signal to a digital standard form, measuring some vital characters of the speech waveform signal such as frequency response, energy or appending these measurements with some few perceptually derived meaningful measurements and conditionally set statistically these numbers of features to form certain vectors called feature vectors. Feature extraction process is aimed at attaining the following set objectives: to break down the speech waveform signal into different acoustically discoverable components, to generate certain features with minimal rates of transformation so as to produce feasible signal computations. In a nut shell, feature extraction method can be classified into three basic parts which include spectral analysis, parametric transformation and statistical modeling. The components are shown in figure 2 below.

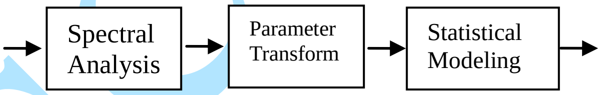

Figure 2: Feature Extraction Process ([Ras14])

IV.1.SPECTRAL ANALYSIS

The beginning point of the speech signal waveform processing is the digitized speech, which is emphasized to make-up for the properties of the glottis stop and the high pass filtering produced by human lips ([Wak73]). Because speech waveform signal is a non-stationary signal, the emphasized waveform is framed into short chunks, overlapping segments by multiplying the signal with a sliding window. In practical perspective, a sliding window that has a fixed length and shape is adopted to separate each segment from the speech waveform signal. The segments often have between 20 ms and 30 ms and they are 10ms overlapped ([GM00]). The resulting segments are almost stationary and can be processed further by means of a short term spectral analysis based on a DFT. To reduce speech signal distortion of the short term spectrum produced by windowing the signal, the window function is needed to meet-up with some spectral properties. When waveform of speech is generated with respect to time changing signal, the signal properties can be reproduced through parameterization of the spectral analysis activity. The main aspects of spectral analysis methods which can be adopted in ASR systems are classified into six components namely: Fourier Transform Derived Cepstral Coefficients, Fourier Transform Derived Filter Bank Amplitudes, Linear Prediction, Digital Filter Bank, Linear Prediction Derived Cepstral Coefficients, Linear Prediction Derived Filer Bank Amplitudes.

IV.2.PARAMETER TRANSFORMS

The acoustic system generates a stochastic description to get sequence of acoustic observation vectors in a given a word sequence. Owing to data sparsety, the model for individual words and model for entire sentences is produced by concatenating the acoustic models of basic sub-word units according to a pronunciation dictionary also called lexicon. In this aspect, speech signal parameters are produced from speech signal waveform computations using two basic operations which are differentiation and concatenation. The output of the two basic operations in this aspect is a parameter vector which contains the raw estimates of the speech signal.

IV.3.STATISTICAL MODELING

In this phase, speech signal waveform is assumed to be parameters generated from a process called multivariate random process. The basic idea behind this process is that a model is forced on the data, the model is then trained, and the vital quality aspect of the approximation is computed. The resultant outputs show the information about the process and the speech signal parameters that have been computed in the process. In this regard, the output of parameter vector from this process is called the speech signal observations. However, a statistical analysis of a given speech is to be computed on the feature vectors to ascertain whether they are part of a spoken speech word or whether they vectors are just a noise.

The major aim of speech recognition is to determine and act upon speaker speech utterances. Every speaker has a unique speech because every individual has his/her own characteristics. These properties of an individual are called speaker features which can be extracted from speech signal utterances. Different techniques can be adopted for feature extraction which include Fourier Transform, Digital filter bank, LPCC, MFCC, LPC, etc ([KSK15]).

A. LINEAR PREDICTIVE CODING (LPC) ANALYSIS

LPC method is a technique used for extracting features from a given speech signal and hence encoding the speech at low bit rate. However, LPC can be used for speech compression, synthesis and identification. As a matter of fact, LPC is spectral estimation method because it produces an estimate of the poles that belongs to the transfer function of the vocal tract. In analyzing the speech signal using window analysis, the speech is periodically stationary within the window and it becomes zero outside the window.

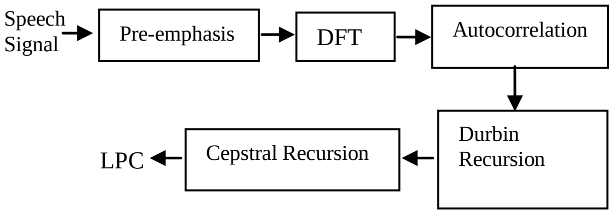

Figure 3: LPC Coefficients Extraction Process

LPC can also be defined as a digital technique for encoding an analog speech signal where a certain value is predicted by a linear function based on the past values of the speech signal ([KSK15]). LPCC method is just an extension to the LPC technique. If LP coefficient is represented in cestrum domain then the coefficients obtained are called linear predictive cepstral coefficients. However Cestrum is computed by calculating the inverse DFT of logarithm magnitude of the DFT of a given speech waveform signal ([KSK15]).

B. LINEAR PREDICTIVE CEPSTRAL COEFFICIENTS (LPCC)

This method is an extension to the LPC technique. When linear predictive coefficient is represented in cestrum domain the obtained coefficients are linear predictive cepstral coefficients. Cestrum is obtained by taking inverse DFT of logarithm of the magnitude of the DFT of the speech signal. They are more robust and reliable then LPC.

C. PERCEPTUAL LINEAR PREDICTION (PLP)

Perceptual linear prediction, similar to LPC analysis, is based on the short-term spectrum of speech. In contrast to pure linear predictive analysis of speech, perceptual linear prediction (PLP) modifies the short-term spectrum of the speech by several psychophysically based transformations. This technique uses three concepts from the psychophysics of hearing to derive an estimate of the auditory spectrum:

(1)The critical-band spectral resolution,
(2) The equal-loudness curve, and
(3) The intensity-loudness power law.

The auditory spectrum is then approximated by an autoregressive all-pole model. In comparison with conventional linear predictive (LP) analysis, PLP analysis method is happening more in a similar with human hearing.

D. MEL-FREQUENCY CEPSTRAL COEFFICIENT (MFCC)

The cepstrum coefficient is the output of a cosine change of the real logarithm of the short time energy spectrum defines on a Mel-frequency scale. This method of speech signal analysis is vaster and reliable feature set for speech recognition than the LPC coefficients method. The sensitive nature of the low order cepstrum coefficient to overall spectral slope, and the sensitive nature of the high-order cepstrum coefficient to noise, has proved it to be a standard method for speech analysis ([Jai14]). The major benefit of MFCC method is that it uses Mel frequency scaling which is very good to the human hearing system. The coefficients produced by this method are fine representation of speech signal spectra with great data compression ([JJ11]). The process of extracting MFCC's from continuous speech is illustrated in figure 4.

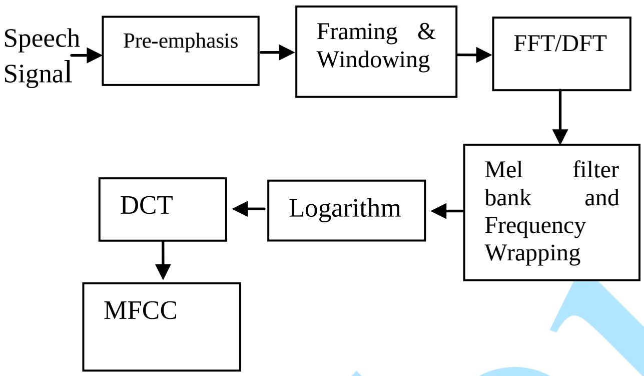

Figure 4: MFCC Coefficients Extraction Process ([KSK15])

V. FEATURE CLASSIFICATION METHODS AND ALGORITHMS

Feature classification method is used to represent the feature vectors of a speaker speech waveform signal. To generate these feature vectors for a given speech signal different classifier can be adopted. The classifier consists of the various speaker models and the decision logic ([BSD12]). The different classifiers discussed in the study are:

A. DYNAMIC TIME WARPING

DTW is a distance calculation method that has often been adopted solving speech recognition issues in aspect of ASR system. The method enables a non-linear mapping of one speech waveform signal to another speech waveform signal by reducing the distance between the two given signal. DTW can be said to be a pattern matching technique with a non-linear time normalization effect ([MBE10]). In other words, it is a method for calculating similarity between two sequences of speech which may vary in speed or time. Specifically, it is a method that enables a computer system to find an optimal match between two given sequences of speech signal with certain restrictions, which means that the sequences are warped non-linearly to match each other. ([MBE10]).

B. SUPPORT VECTOR MACHINES

SVM is a supervised technique under machine learning that demands the proper training of the tool before classification procedure begins in earnest. This algorithm is a good tool for binary classification of the data. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes form the input ([A+11]). The hyper plane is constructed defined by set of weights W, data points X and a bias or offset b, such that: W.X + b = 0 Where W.X denotes the dot product of the data and the normal vector to the hyper plane. The variable b specifies the hyper plane offset from the origin along the normal feature vector ([KJ15]). SVM is a very strong discriminative classifier that has been recently used in ASR system. It has been adopted in prosodic, spectral and high-level features. The major reason why SVM is popular is its good generalization performance to classify unseen data ([A+11]). SVMs are a set of related supervised learning methods that recognize patterns, analyze data used for classification and regression analysis.

C. HIDDEN MARKOV MODELS

HMM produces stochastic systems from known speech utterances and compares the probability that the unknown speech utterance was generated by each system. This uses theory from statistics in order to arrange our feature vectors into a Markov chains that stores probabilities of state transitions ([JJ11]). In other words, if each of the code words were to represent some state, the HMM would follow the sequence of state changes and build a model that includes the probabilities of each state progressing to another state. HMM is a popular algorithm in ASR because they are simple, can be trained automatically and is computationally feasible to use in speech signal. It breaks the feature vector of the signal into a number of states and finds the probability of a signal to transit from one state to another. HMM is a well known statistical tool for modeling a huge range of the time series data. It is a statistical algorithm for a Markov process with hidden parameters ([SM12]). In the aspect of natural language processing, HMM is normally applied with great success to problems such as noun-phrase chunking and part-of-speech tagging. Hidden Markov model is a stochastic finite-state automaton that produces a sequence of observable symbols. The sequence of states is a Markov chain, which means that the transition between states has an associated probability called transition probability. Each state has an associated probability function to generate an observable symbol. Only the sequence of observations is visible and the sequence of states is not observable and therefore hidden; hence the name hidden Markov model.

D. VECTOR QUANTIZATION

VQ is a classical quantization technique for preprocessing speech waveform signal which enables the designing of probability density functions by dividing the prototype featurw vectors. The technique starts by breaking a large set of points into group of clusters having definitely the exact number of points that is nearest to them whereas the Centroid point stands for each group ([M+12]). The pattern matching property density of this quantization method is efficiently powerful, purposely in the large density and data with high dimensional identification. Data with points are displayed by their closest indexing Centroid; data that occur frequently have low error while the rare data high error. Therefore, this method is not only effective but also appropriate for lossy data compression process. VQ is a technique for mapping feature vectors from a large feature vector space to a finite number of regions in that space. Every region is referred to as Cluster and can be replaced by its center also known as a centroid. The collection of all code words is called a codebook.

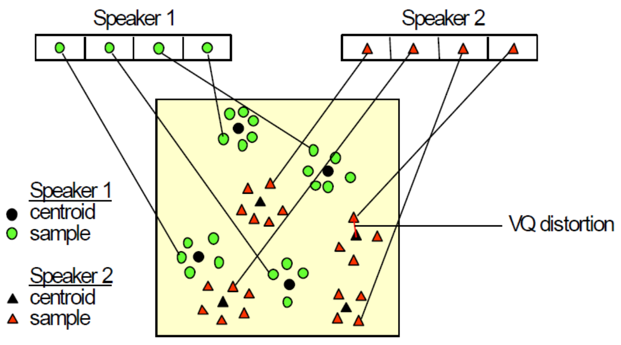

Figure 5: Formation of VQ Codebook ([KR11])

A given speech of a speaker can be distinctly differentiated from speech of another speaker by the location of centroid figure five shows a conceptual diagram to illustrate the recognition process. However, the figure consists of speech of two different speakers and two acoustic feature vectors dimensions. The circles in the diagram refer to the acoustic feature vectors from the speech of speaker one while the triangles symbols are the acoustic feature vectors of speech signal are from the speaker two. During the training section, a speech speaker specific VQ codebook is produce for evry known speech of the speaker by clustering his/her training acoustic feature vectors. The resultant code words as shown in figure five means that the black circles represents speech of speaker one and black triangles represents the speech of speaker two. The space distance from a feature vector to the closest codeword of a codebook is also referred to as a VQ distortion. In aspect of the speech recognition section an input speech utterance of an unknown speech utterance is vector quantized through each trained codebook and thereafter the total VQ codebook with lowest total distortion is identified as the correct speech.

E. GAUSSIAN MIXTURE MODEL

GMM is a stochastic technique which has become the de facto reference method in ASR system. The GMM technique is an expansion of the VQ technique, in GMM the group clusters are overlapping. GMM technique is made-up of a finite mixture of multivariate Gaussian components ([A+13]). The GMM technique is a function density estimator and distribution of the feature vector using a certain mixture of Gaussians. In recognition stage, a sequence of features vector is removed from the input speech signal the space distance of the given sequence from the model is obtained by calculating the log likelihood of the given sequence. The technical model which gives the highest likelihood score is examined as the identity speech utterance of the speaker.

F. MULTI-LAYER PERCEPTRONS

MLP is a neural network based classifiers. They are used mainly for the powerful structure in classifying complex, nonlinear instants and in regression. Critical parameters like the size of hidden layer, learning rate, transfer functions in both hidden and output layers can be well optimized to get best results for the specific purpose ([D+12]).

G. ARTIFICIAL NEURAL NETWORKS

ANN was inspired by the sophisticated functionality of the human brain where neurons process information in parallel. ANN consists of a layer of input nodes, then one hidden layer of nodes and finally a layer of output nodes ([KK13]). ANN came as a better acoustic modeling approach which has been adopted in many areas of speech recognition for isolated word recognition, speaker adaptation and phoneme classification. In contrary to HMM methods, ANNs make no estimations about statistical feature properties and have several qualities making them good recognition models for ASR system. To calculate the probabilities of a speech feature segment, ANNs allow discriminative training in an efficient and natural manner ([DS16]).

H. K-NEAREST NEIGBOUR

KNN classification technique is a very concise and it is yet a strong classification method often use in speech recognition system. In KNN model the basic idea about this classification method is that similar observations are said to belong to similar classes. Hence, seeking for the class designators of a number of the nearest neighbors and adding up their class numbers to set a class number to the unknown. In aspect of practical, for instance, KNN gets the number of closest neighbors to the unlabeled data from the training distance based on the chosen distance measure ([Rey95]).

VI. EXPERIMENTAL RESULTS

The results are presented in figures 6 to 11.

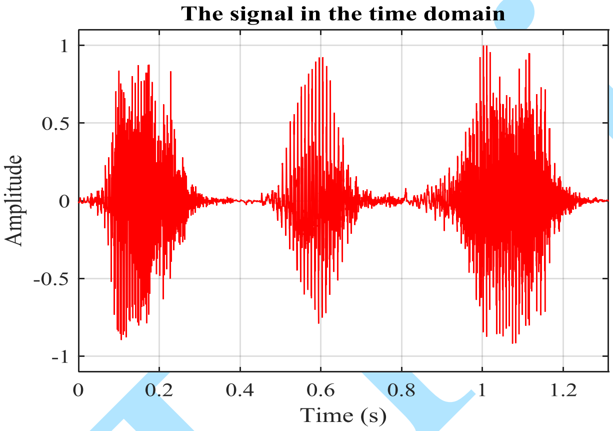

Figure 6: The speech waveform of the word GODIYA

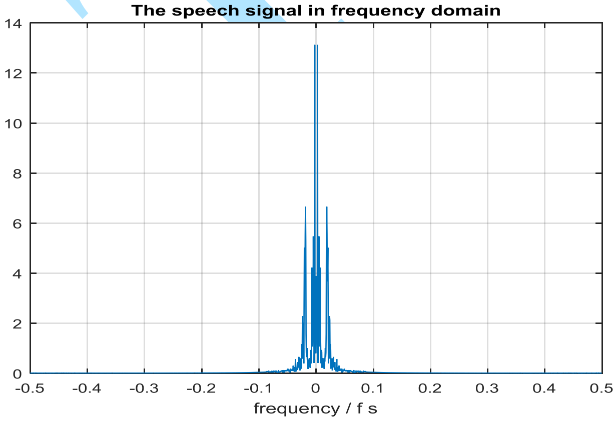

Figure 7: The FFT of the speech word GODIYA

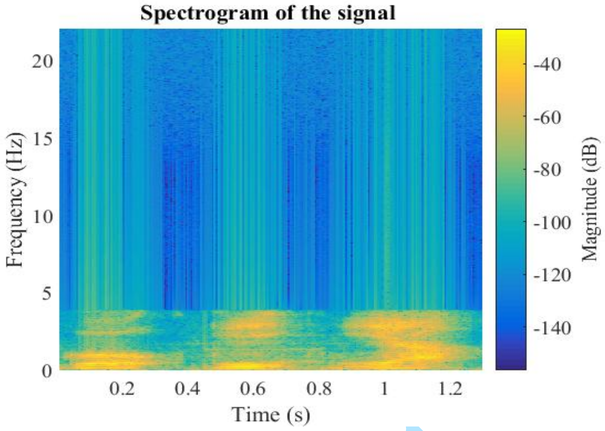

Figure 8: Spectrogram of the word GODIYA

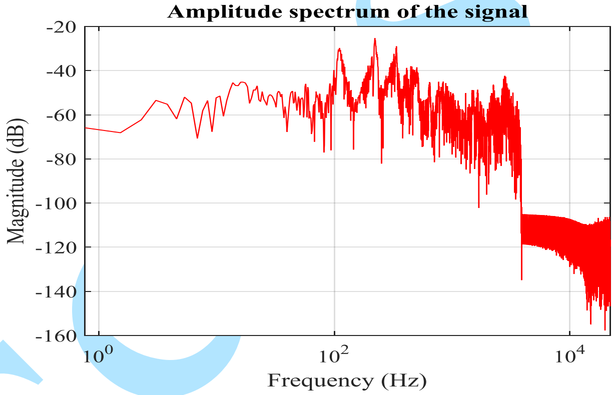

Figure 9: Magnitude spectrum of the word GODIYA

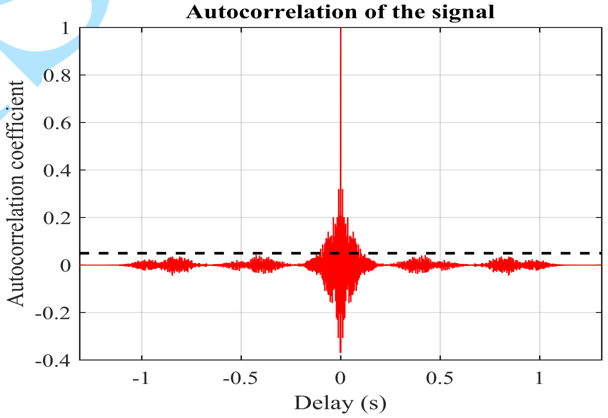

Figure 10: The autocorrelation of the word GODIYA

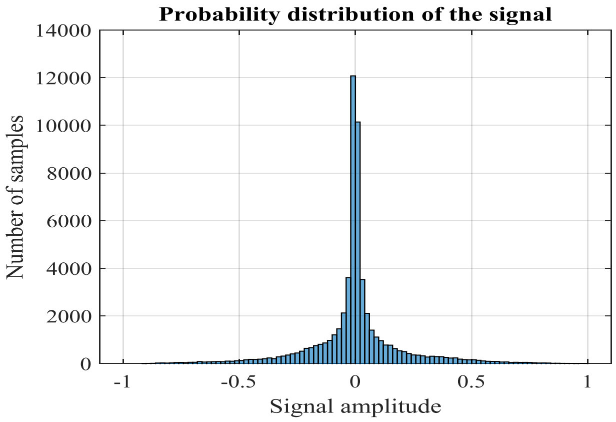

Figure 11: The probability distribution of the word GODIYA

VII. CONCLUSION

This study has shown the different methods used in area of automatic speaker recognition. The different techniques adopted for feature vectors extraction process and speech pattern recognition process have been discussed. Moreover, the MFCC method of feature vector extraction is better when a clean speech is used but it is very sensitive to noise when noisy speech is involved in the recognition process. Some techniques often used for ASR systems includes over others such as MFCC for feature vector extraction has better performance rather than LPC, PLP or LPCC, because MFCC is very similar to human hearing due to Mel scale representation and the techniques like DTW, VQ, ANN, HMM, SVM for pattern matching.

REFERENCES

[AMA11] **M. Alsulaiman, G. Muhammad, Z. Ali** - *Comparison of Voice Features for Arabic Speech Recognition*, IEEE, pp.90-95, 2011.

[A+11] **W. Astuti, A. M. Salma, A. M. Aibinu, R. Akmeliawati, Momoh Jimoh, E. Salami** - *Automatic Arabic Recognition System based on Support Vector Machines (SVMs)*, IEEE, 2011.

[A+13] **J. Amini, A. S. Shahrebabaki, N. Shokouhi, H. Sheikhzadeh** - *Speech Analysis/Synthesis by Gaussian Mixture Approximation of the Speech Spectrum for Voice Conversion*, IEEE Transaction on Audio Speech processing, and Language, pp.000428-000433, 2013.

[BSD12] **N. S. Bansod, K. Seema, S. B. Dabhade** - *Review of different techniques for speaker recognition system*, Vol. 4, Issue 1, pp. 57-60, 2012.

[DS16] **V. A. Devi, V. Suganya** - *An Analysis on Types of Speech Recognition and Algorithms* International Journal of Computer Science Trends and Technology (I JCS T) – Volume 4 Issue 2, ISSN: 2347-8578, www.ijcstjournal.org, pp 350-355, Mar - Apr 2016.

[D+12] **R. Djemili, R. Bourouba, M. Cherif, A. Korba** - *A Speech Signal Based Gender Identification System Using Four Classifiers*, IEEE, 2012.

[GM00] **B. Gold, N. Morgan** - *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley, New York (2000).

[GRP17] **A. Gupta, P. Raibagkar, A. Palsokar** - *Speech Recognition Using Correlation Technique*, International Journal of Current Trends in Engineering & Research (IJCTER) e-ISSN 2455–1392 Volume 3 Issue 6, pp. 82–89, June 2017.

[I+13] **M. M. Islam, F. H. Khan, A. Ahsan, M. M. Haque** - *A Novel Approach for Text-Independent Speaker Identification Using Artificial Neural Network*, International Journal of Innovative Research in Computer and Communication Engineering, vol. 1, 2013.

[Jai14] **A. Jain** - *Evaluation of MFCC for Speaker Verification on Various Windows*, IEEE International Conference on Recent Advances and Innovations in Engineering, pp.1-6, 2014.

[JJ11] **W. Junqin, Y. Junjun** - *An Improved Arithmetic of MFCC in Speech Recognition System*, IEEE Transaction on Audio Speech processing and Language, pp.719-722, 2011.

[KJ15] **K. Kaur, N. Jain** - *Feature Extraction and Classification for Automatic Speaker Recognition System: A Review*, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, 2015.

[KK13] **G. Kaur, H. Kaur** - *Multi Lingual Speaker Identification on Foreign Languages Using Artificial Neural Network with Clustering*, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, 2013.

[KR11] **C. S. Kumar, P. M. Rao** - *Design of an automatic speaker recognition system using MFCC, vector quantization and LBG algorithm*

International Journal on Computer Science and Engineering (IJCSE) vol. 3 No. 8 August 2011.

[KSK15] **G. Kaur, D. Singh, G. Kaur** - *A Survey on Speech Recognition Algorithms* International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume 4, Issue 5). May 2015.

[MBE10] **L. Muda, M. Begam, I. Elamvazuthi** - *Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques*, Journal of Computing, pp 138-143, Vol. 2, Issue 3, March 2010.

[M+12] **J. Martinez, H. Perez, E. Escamilla, M. Mabo** - *Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector Quantization (VQ) Techniques*, IEEE, pp 248-251, 2012.

[Ras14] **C. R. Rashmi** - *Review of Algorithms and Applications in Speech Recognition System* International Journal of Computer Science and Information Technologies, Vol. 5 (4), pp. 5258-5262, 2014.

[Rey95] **D. A. Reynolds** - *Speaker Identification and Verification Using Gaussian Mixture Speaker Models*. Speech Communication, vol. 17, pp. 91-108, 1995.

[SM12] **G. I. Sapijaszko, W. B. Mikhael** - *An Overview of Recent Window Based Feature Extraction Algorithms for Speaker Recognition*, IEEE, pp 880-883, 2012.

[SRR10] **V. Sailaja, K. S. Rao, K. V. S. Reddy** - *Text Independent Speaker Identification with Finite Multivariate Generalized Gaussian Mixture Model and Hierarchical Clustering Algorithm*, International Journal of Computer Applications (0975-8887), vol. 11, 2010.

[S+12] **B. Singh, R. Kaur, N. Devgun, R. Kaur** - *The process of Feature Extraction in Automatic Speech Recognition System for Computer Machine Interaction with Humans: A Review*, IJARCSSE, Volume 2, Issue 2, February 2012.

[TJ11**]** **J. Tao, X. Jiang** - *A Domestic Speech Recognition Based on Hidden Markov Model*, IEEE CCIS, pp.606-609, 2011.

[Wak73] **H. Wakita** - *Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveform*. IEEE Transactions on Audio Electroacoustics, Vol. 21, pp. 417–427, 1973.