Vol 7. No. 2 - June, 2014

African Journal of Computing & ICT

www.ajocict.net

An Enhanced Speech Recognition Algorithm Using Levinson-Durbin, DTW and Maximum Likelihood Classification

A.A. Adegun & E.O. Asani
Computer Science Programme
Landmark University
Omu-aran, Nigeria
adegun.adekanmi@lmu.edu.ng, asani.emmanuel@lmu.edu.ng
2348025717404

M. Yusuff
Computer Science and Information Systems
Achievers University
Owo, Nigeria

ABSTRACT

In this paper, we applied techniques such as Levinson-Durbin, DTW and maximum likelihood classification to achieve an enhanced speech recognition algorithm. Speech recognition has been adversely affected by noise and some other impairments factors making speech difficult to be recognized. Speech is distorted by a background noise and echoes, electrical characteristics. We used the combinatorial approach of Levinson-Durbin, DTW and maximum likelihood classification to develop a system for speech recognition. The system compares the speech with phonetics lattice and database of enrolled speeches from different speakers and output the enrolled speech with the recognition Id and name of the identified speech if it found a match. Otherwise error of unknown is returned to show speech mismatch. The system is able to distinguish two different speakers at any point in time using speech identity and phonetic pattern analysis.

Keywords: Levinson-Durbin, DTW and maximum likelihood.

African Journal of Computing & ICT Reference Format:
A.A. Adegun, E.O. Asani & M. Yusuff (2014). An Enhanced Speech Recognition Algorithm Using Levinson-Durbin, DTW and Maximum Likelihood Classification.. Afr J. of Comp & ICTs. Vol 7, No. 2. Pp 135-142.

1. INTRODUCTION

Accents, regional dialects, sex, age, speech impediments, emotional state and other factors cause people to pronounce the same word in different ways. Automatic Speech Recognition (ASR) is an important task in digital signal processing related applications. It is the process of automatically converting the spoken words into written text by the computer system [3]. Research on Speech recognition system has been on-going for over 80 years and quite a sizeable number of progresses have been attained. One of the major problems of all the system designed over the years is ability to still recognize a voice even if it has been affected by illness or some under uncontrollable circumstance. Speech recognition has been accomplished by combining various algorithms drawn from different disciplines such as statistical pattern recognition, signal processing and linguistics etc [1].

In this paper, we have combined three algorithms: Levinson-Durbin, DTW and maximum likelihood classification to develop an enhanced system for speech recognition. Each of these algorithms has outstanding performance in various aspect of speech recognition. We have successfully classified speech recognition processes into 3 categories and we have used an algorithm each of these aspects.

The three aspects of speech recognition identified in this work include:

- Speech feature extraction
- Pattern matching techniques
- Dynamic Programming-

Dynamic features are generally helpful in capturing temporal information. Features are extracted before template matching takes place. Firstly, the dynamic programming techniques have been proposed for spoken word recognition based on template matching approach.

We can use the Maximum Likelihood classification for matching a given utterance against a predefined vocabulary represented by DTW.

Linear prediction analysis of these features is done based on Levinson-Durbin Recursion. Because of the variability in a speech signal, it is better to perform feature extraction in short term interval that would reduce these variability. LP analysis is performed based on Levinson-Durbin recursion algorithm.

2. LITERATURE REVIEW

Data is the raw information which needs to be processed, they are entered into the system as input and processed, and the result of the processed data is referred to as the output. The data to be processed here is a sound signal recoded as a wave file generated from microphone connected to the system. The signals collected contain acoustic information which is process into a vector format and modeled appropriately for word recognition. Vimala and Radha [3] divided signal into smaller frames where the useful feature vectors were extracted. Then, these feature vectors are divided into training and testing features. Speech feature extraction is the signal processing frontend which converts the speech waveform into some useful parametric representation. These parameters are then used for further analysis in various speech related applications such as speech recognition, speaker recognition, speech synthesis and speech coding. It plays an important role to separate speech patterns from one another [2] [5][6].

2.1 Analysis Of Input

Early attempts to design systems for automatic speech recognition were mostly guided by the theory of acoustic-phonetics, which describes the *phonetic elements* of speech (the basic sounds of the language) and tries to explain how they are acoustically realized in a spoken utterance.

These elements include the phonemes and the corresponding place and manner of articulation used to produce the sound in various phonetic contexts. For example, in order to produce a steady vowel sound, the vocal cords need to vibrate (to excite the vocal tract), and the air that propagates through the vocal tract results in sound with natural modes of resonance similar to what occurs in an acoustic tube. These natural modes of resonance, called the *formants* or *formant frequencies,* are manifested as major regions of energy concentration in the speech power spectrum. In 1952, Davis, Biddulph, and Balashek of Bell Laboratories built a system for isolated digit recognition for a single speaker, using the formant frequencies measured (or estimated) during vowel regions of each digit. The Figure 1 below shows a block diagram of the digit recognizer developed by Daviset al., and Figure 2 shows plots of the formant trajectories along the dimensions of the first and the second formant frequencies for each of the ten digits, one-nine and oh, respectively. These trajectories served as the "reference pattern" for determining the identity of an unknown digit utterance as the best matching digit.[7]

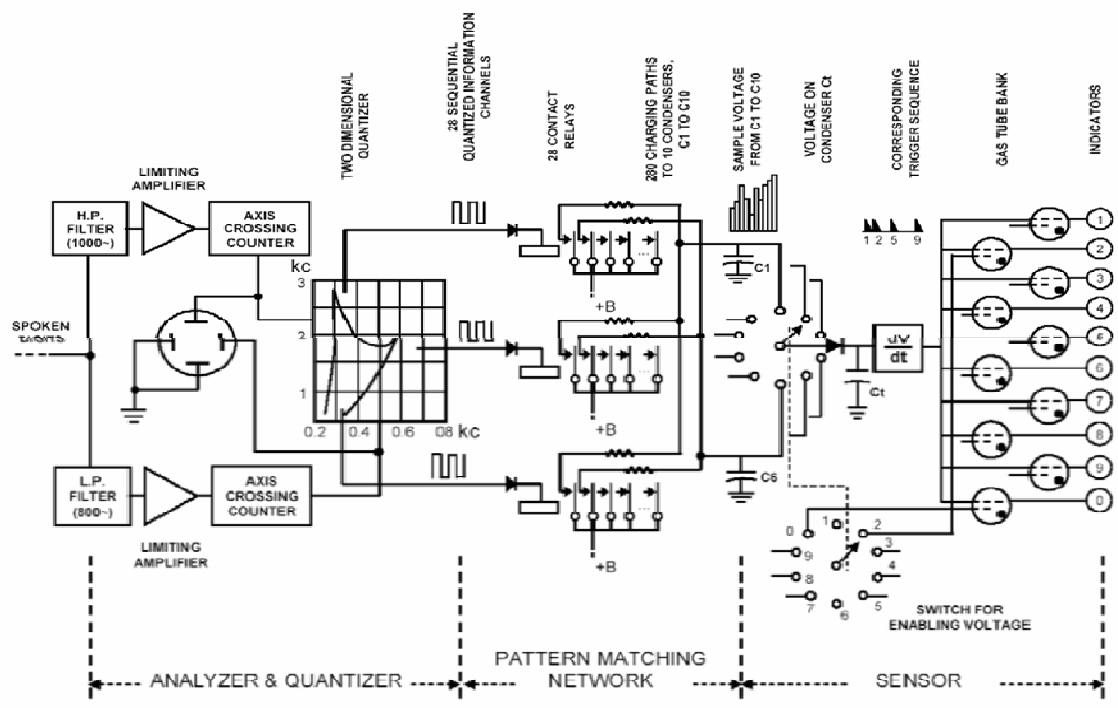

Figure 1: Block schematic of digit recognizer circuits. (Source [7])

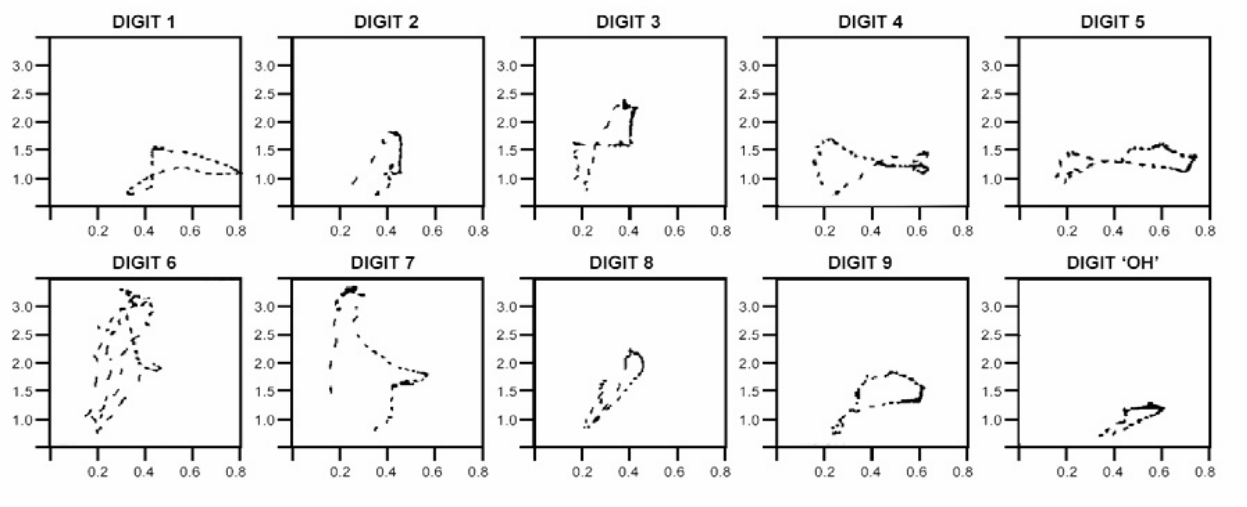

Figure 2: Photographs of formant 1 vs. formant 2 (Source: [7])

2.2 Maximum Likelihood classification

This paper work had been able to show the principles of the algorithms used for simple speech recognition tasks like the recognition of connected digits. It also shows how we can use the Maximum Likelihood classification for matching a given utterance against a predefined vocabulary represented by DTW. The algorithms used were quite straightforward and had a very regular structure. Saul and Rahim (2000) worked on maximum likelihood classification. According to them factor analysis uses a small number of parameters to model the covariance structure of high dimensional data. These parameters can be chosen in two ways: (1) to maximize the likelihood of observed speech signals, or (2) to minimize the number of classification errors. We derive an expectation-maximization (EM) algorithm for maximum likelihood estimation and a gradient descent algorithm for improved class discrimination. Speech recognizers are evaluated on two tasks, one small-sized vocabulary (connected alpha-digits) and one medium-sized vocabulary.

(New Jersey town names). We find that modeling feature correlations by factor analysis leads to significantly increased likelihoods and word accuracies. Saul and Rahim (2000). In their work, Saul and Rahim (2000) used factor analysis to model correlations between cepstra, delta-cepstra, and delta-delta-cepstra. It is worth emphasizing, however that the method applies to arbitrary features. Indeed, the ability to, model correlations efficiently should enable researchers to consider other features besides ceptra.

While cepstra have the advantage of being only weakly correlated, it may be that other features (e.g., narrow-band statistics) actually convey more information about the speech signal. Saul and Rahim (2000)

2.3 Dynamic Programming / Dynamic Time Warping

Dynamic time warping (DTW) is a well-known technique to find an optimal alignment between two given (time-dependent) sequences under certain restrictions. DTW technique is used for feature matching in speaker recognition. Intuitively, the sequences are warped in a non linear fashion to match each other. Originally, DTW has been used to compare different speech patterns in automatic speech recognition. In fields such as data mining and information retrieval, DTW has been successfully applied to automatically cope with time deformations and different speeds associated with time-dependent data. Muller,M (2007).

The time alignment of different utterances is the core problem for distance measurement in speech recognition. A small shift leads to incorrect identification. Dynamic Time Warping is an efficient method to solve the time alignment problem. DTW algorithm aims at aligning two sequences of feature vectors by warping the time axis repetitively until an optimal match between the two sequences is found. This algorithm performs a piece wise linear mapping of the time axis to align both the signals. Shivanker **,** Geeta and, Poonam (2013)

The procedure we defined to compare two sequences of vectors is also known as Dynamic Programming (DP) or as Dynamic Time Warping (DTW). In another words, Dynamic time warping is an algorithm for measuring similarity between two sequences that may vary in time or speed.

www.ajocict.net

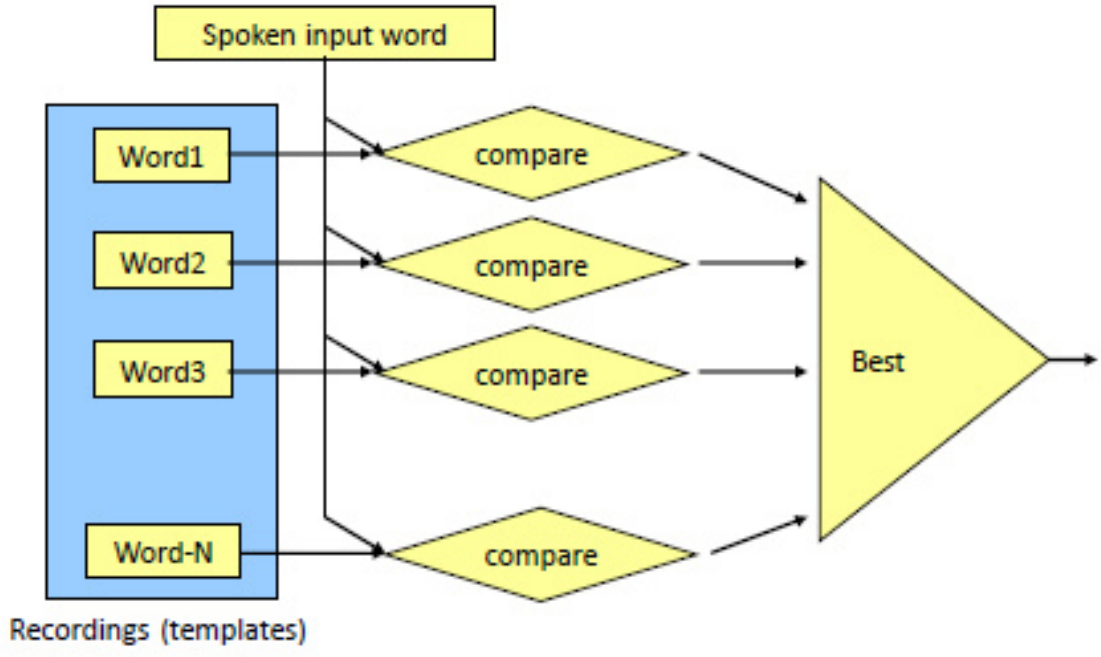

Fig 3 : Isolated word recognition scenario

2.4 Linear Prediction By Levinson-Durbin Algorithm

According to A G Constantinides 2012, the Durbin algorithm solves the following $\mathbf{R}_m \mathbf{w}_m = \mathbf{r}_m$

Where the right hand side is a column of R as in the normal equations.

Assume we have a solution for $\mathbf{R}_k \mathbf{w}_k = \mathbf{r}_k \quad 1 \le k \le m$

Where $\mathbf{r}_k = [r_1, r_2, r_3, ..., r_k]^T$

For the next iteration the normal equations can be written as

$$\begin{bmatrix} \mathbf{R}_k & \mathbf{J}_\mathbf{k}\mathbf{r}_\mathbf{k}^* \\ \mathbf{r}_\mathbf{k}^\mathbf{T}\mathbf{J}_\mathbf{k} & r_0 \end{bmatrix} \mathbf{w}_{k+1} = \mathbf{r}_{k+1}$$

Where $\mathbf{r}_{k+1} = \begin{bmatrix} \mathbf{r}_k \\ r_{k+1} \end{bmatrix}$

Set $\mathbf{w}_{k+1} = \begin{bmatrix} \mathbf{z}_k \\ \alpha_k \end{bmatrix}$

Multiply out to yield

$$\mathbf{z}_k = \mathbf{R}_k^{-1}(\mathbf{r}_k - \alpha_k \mathbf{J}_k \mathbf{r}_k^*) = \mathbf{w}_k - \alpha_k \mathbf{R}_k^{-1} \mathbf{J}_k \mathbf{r}_k^*$$

Note that

$$\mathbf{R}_k^{-1} \mathbf{J}_k = \mathbf{J}_k \mathbf{R}_k^{-1}$$

Hence $\mathbf{z}_k = \mathbf{w}_k - \alpha_k \mathbf{J}_k \mathbf{w}_k^*$

i.e the first k elements of $\mathbf{w}_{k+1}$ are adjusted versions of the previous solution

The last element follows from the second equation of

$$\begin{bmatrix} \mathbf{R}_k & \mathbf{J}_\mathbf{k}\mathbf{r}_\mathbf{k}^* \\ \mathbf{r}_\mathbf{k}^\mathbf{T}\mathbf{J}_\mathbf{k} & r_0 \end{bmatrix} \begin{bmatrix} \mathbf{w}_k \\ \alpha_k \end{bmatrix} = \begin{bmatrix} \mathbf{r}_k \\ r_{k+1} \end{bmatrix}$$

i.e

$$\alpha_k = \frac{1}{r_0}(r_{k+1} - \mathbf{r}_k^\mathbf{T} \mathbf{J}_k \mathbf{z}_k)$$

The parameters α_k are known as the reflection coefficients. These are crucial from the signal processing point of view.

3. DESIGN METHODOLOGY

Data Flow Diagram (DFD) Of The System

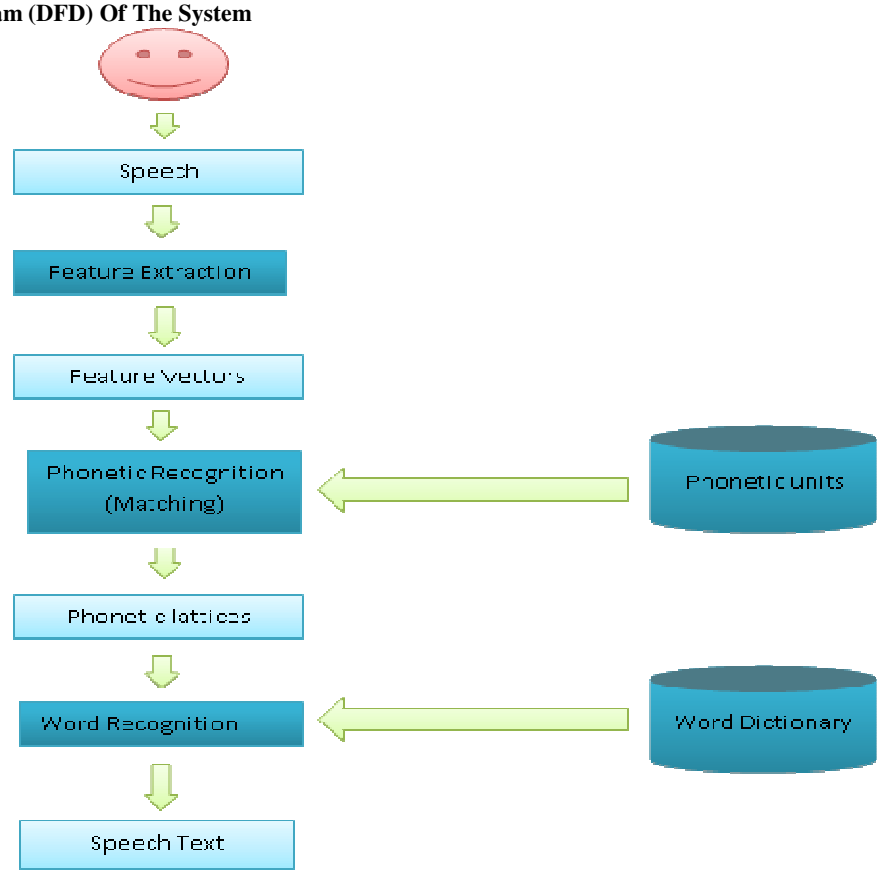

Figure 4: Data Flow Diagram

The algorithms are combined and put into use here. Each of the algorithm performs various tasks at each stage of the system. The system accepts input in form of speech. Features extraction is performed on the speech input and this is further broken into vectors that can be worked on. Phonetic recognition also known as matching is done. This is done through the template matching. Phonetic lattice are extracted and combined and word recognition is eventually done.

4. IMPLEMENTATION, RESULTS AND DISCUSSION

The Speech Recognition algorithm is implemented here using Visual Basic 2008 programming language. The feature extraction process here is applied using the Levinson and Durbins LPC algorithm and the Dynamic Time Warping for the matching. This software is expected to prompt the user for input which is the user utterance, it capture it search the inbuilt vocabularies using (Dynamic Time Wrap) DTW and compare the captured sound with the inbuilt sound using levinson - durbin algorithm. It give out the sound captured inform of text as the output after all the processing. Thus, identifying the enrolled speech through the speech Id that identifies distinct speakers. The testing here was done with varieties of utterances from people with different speed of talking and accent. The other testing done here was carried out where we have some sound distortion like background and a desired output was generated.

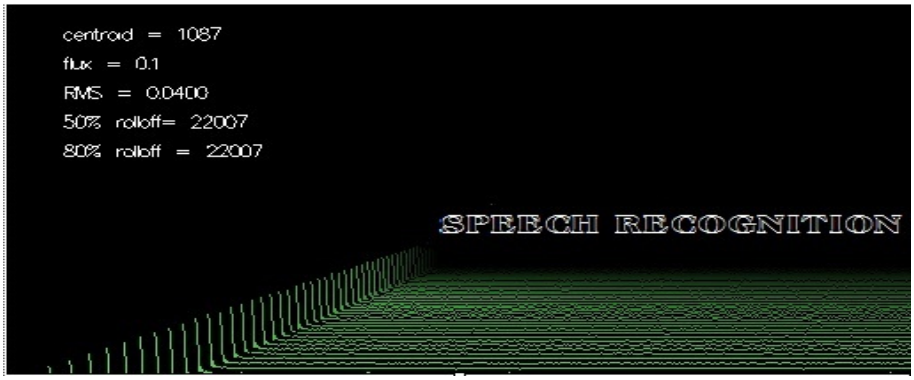

Figure 5 Speech Recognition Welcome page Snapshot

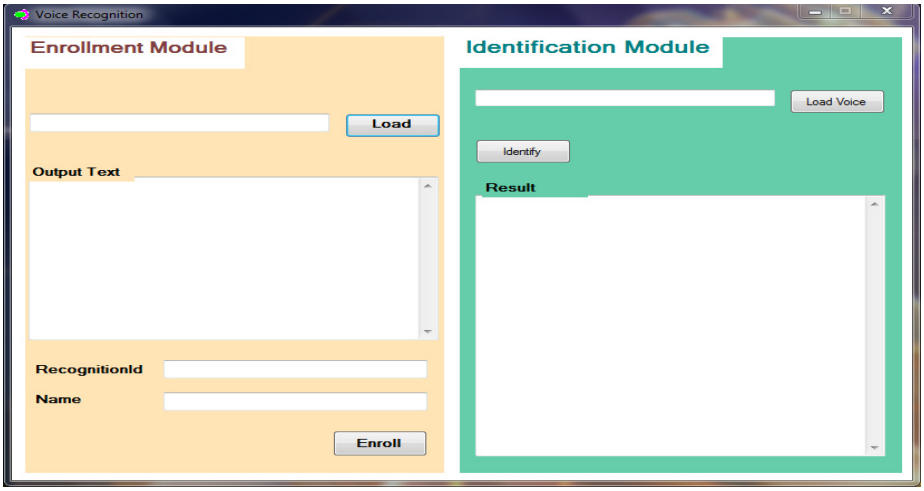

Figure 6. Input: Speech enrollment and Identification Interface Snapshot

Input to the program is obtained by grabbing the sound in ***.wav*** form from windows based sound recorder or other sound grabber or voice recorder software such as Atube catcher, Audacity, Wave engine, wavepad editor etc. the recorded sound is then loaded from enrollment interface.

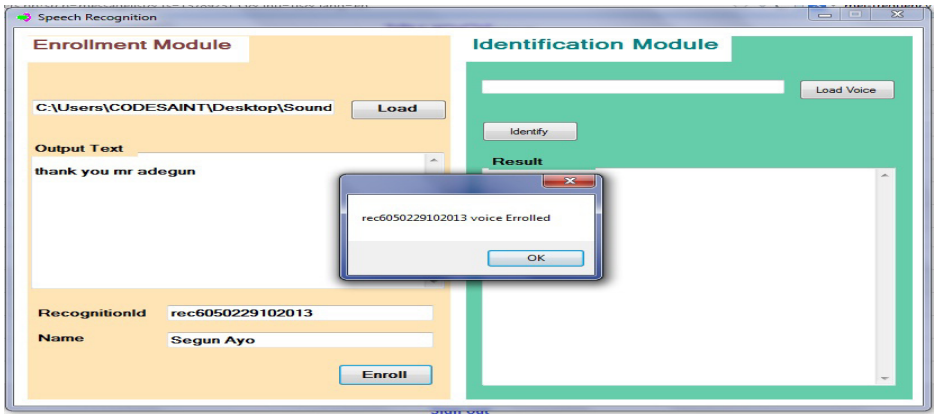

Figure 7: Input (Enrollment) Snapshot.

4.3.2 Output (Identification) Module

The system compares the speech with phonetics lattice and database of enrolled speeches from different speakers and output the enrolled speech with the recognition Id and name of the identified speech if it found a match. Otherwise error of unknown is returned to show speech mismatch. However, the system is able to distinguish two different speakers at any point in time using speech identity and phonetic pattern analysis.

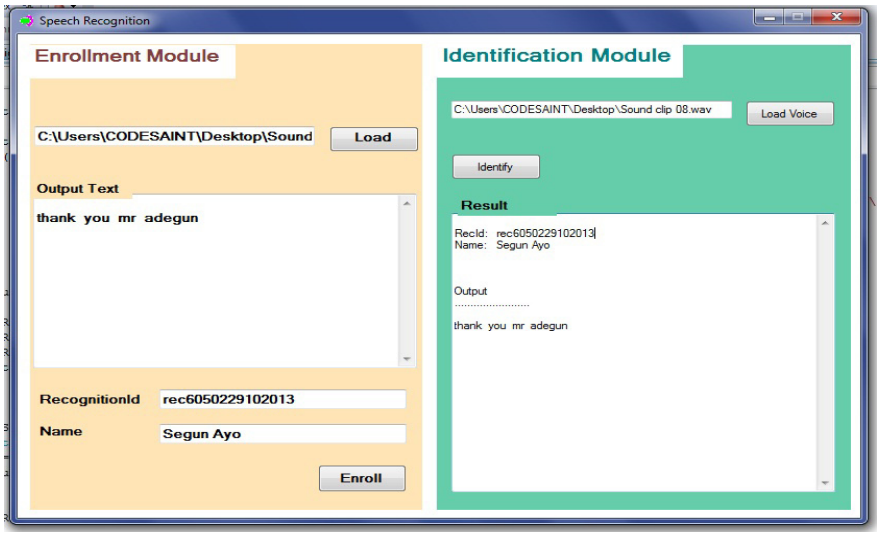

Figure 8:Output(identification)Snapshot.

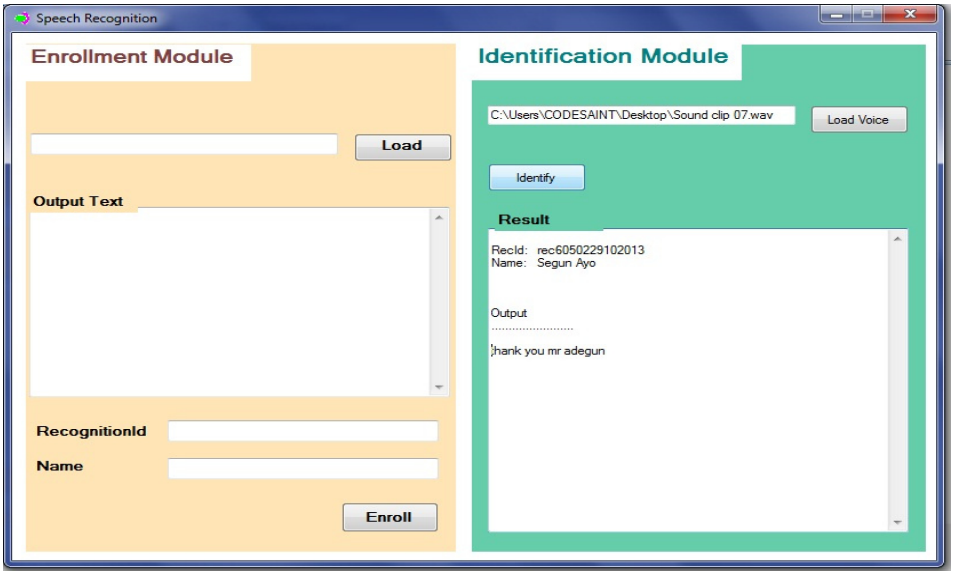

Figure 9: Output: Querying to match an existing speech Snapshot.

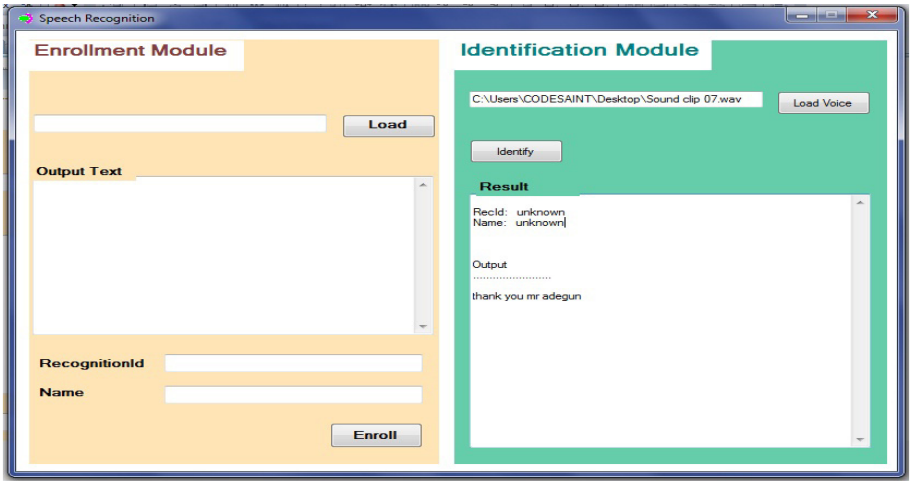

Fig 10 Output: Speech mismatches (identification) Snapshot.

5. CONCLUSION AND FUTURE WORKS

This project covered the basics of speech recognition up to the principles of connected word recognition and matching. We also pointed out the disadvantages of these primitive algorithms. However, the algorithm was able to handle isolated word recognition (IWR) and continuous speech recognition (CSR) process it by looking (searching) for the best match from the inbuilt vocabularies.

6. RECOMMENDATIONS

There are challenges facing these algorithms. Like other supervised learning algorithms. Speech recognition (by a machine) is a very complex problem. Vocalizations vary in terms of accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed. Speech is distorted by a background noise and echoes, electrical characteristics. Accuracy of speech recognition varies with the following:

- Vocabulary size and confusability
- Speaker dependence vs. independence
- Isolated, discontinuous, or continuous speech
- Task and language constraints
- Read vs. spontaneous speech
- Adverse conditions

So there is a need for improvement or more efficient algorithm.

REFERENCES

[1] Saul, L.K. & Rahim, M.G (2000). Maximum likelihood and minimum classification error factor analysis for automatic speech recognition Speech and Audio Processing, IEEE Transactions on (Volume:8 , Issue: 2); AT&T Bell Labs., Florham Park, NJ, USA.

[2] Tomyslav Sledevi˘c, Art¯uras Serackis, Gintautas Tamulevi˘cius, Dalius Navakauskas (2013). Evaluation of Features Extraction Algorithms for a Real-Time Isolated Word Recognition System World Academy of Science, Engineering and Technology International Journal of Electrical, Electronic Science and Engineering Vol:7 No:12, 2013

[3] Vimala.C & Vimala.C, Radha.V (2014). Suitable Feature Extraction and Speech Recognition Technique for Isolated Tamil Spoken Words Vimala.C, Radha.V *Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women,Coimbatore, Tamil Nadu, India* / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 378-383

[4] Shivanker Dev Dhingra, Geeta Nijhawan & Poonam Pandit (2013). Isolated Speech Recognitionusing MFCC and DTW International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering *(An ISO 3297: 2007 Certified Organization)* Vol. 2, Issue 8, August 2013

[5] Muller,M. (2012). Information Retrieval for music and motion 2007XXVI,318 p..136 illus.39 in color., Hardcover. Springer http://www.springer.com/978-3-540-74047-6 ISBN: 978-3-540-74047-6

[6] Plannerer, B. (2005). An Introduction to Speech Recognition March 28, 2005 plannerer@ieee.org

[7] B. Pellom (2004) http://www.cis.hut.fi/Opinnot/T-61.184/. September , 2004