IJST EXPLORE International Journal of Science and Technology CENTRE OF PROFESSIONAL RESEARCH PUBLICATIONS ISSN 2049-7318

International Journal of Science and Technology Volume 4 No. 1, January, 2015

Application of Mel Frequency Ceptrum Coefficients and Dynamic Time Warping For Developing an Isolated Speech Recognition System

Animasahun, I.O[*]. and Popoola, J.J.

Department of Electrical and Electronics Engineering, P.O.Box. 704, Federal University of Technology Akure, Ondo State, Nigeria

ABSTRACT

The recognition accuracy of speech recognition system has been a challenge due to insufficient combination of pre-processing techniques used for speech processing. In order to solve this problem, this system was adopted with well supported pre-processing techniques. Also, in order to enhance the recognition accuracy of the developed speech recognition for this study, it was developed using computational efficient end point detection algorithm that used probability function and linear classifier approach. The development of the speech recognition system was divided into four stages. The first stage involved speech recording from different speakers. Three isolated words; count, down and stop were taken from ten different speakers using microphone for speech template preparation. The second stage involved feature extraction from the recorded speeches using Mel Frequency Ceptrum Coefficients (MFCC). The Third stage was focussed on measuring the global dissimilarity between the stored speech templates and the test input speech samples for the isolated speech recognition using the dynamic time warping algorithm. In the fourth stage, the developed speech algorithm was tested to evaluate its recognition accuracy. The result obtained shows that the developed speech recognition system successfully recognised the isolated words; count, down and stop at different recognition rate of 100%, 60% and 70% respectively.

Keywords: *Speech recognition system, Classes of speech recognition system, End Point Detection Mel Frequency Cepstral Coefficient (MFCC), Dynamic Time Warping (DTW)*

1. INTRODUCTION

Speech recognition system is basically the process of converting a speech wave, which is usually human speech into acoustic features that are useful for further processing of the speech. It is a system that recognizes what is being said by the speaker. It recognises words as they are articulated and not the speaker as in the case of voice recognition [1]. It assumes that the speech signal is a realization of some messages encoded as a sequence of one or more symbols. Basically, speech recognition system can be divided into either speaker dependent or speaker independent recognition systems. The current areas of application of speech recognition system are security devices, asynchronous transfer module (ATM), cellular phones, home appliance, computers, global positioning system, military and artificial control, medical and legal transcription [2]. Whether speaker dependent or speaker independent, speech recognition can also be classified as isolated word, connected and complex word recognition system.

Isolated speech recognition are discrete words and it is good in cases where the user is required to give only one command [2]. Accurate endpoint detection is very important for isolated word recognition systems [3]. The recognition accuracy of speech recognition system has been limited due to inefficiency of the recognition techniques usually employed as well as lack of pre-processing techniques that can enhance better recognition accuracy. Endpoint detection, which aims to distinguish the speech and non-speech segments of a digital speech signal, is considered as one of the key pre-processing steps in automatic speech recognition (ASR) systems [3]. This is why; the paper presents isolated word recognition system using the newly developed and computationally efficient end point detection algorithm that uses probability function and linear classifier approach in order to get better recognition accuracy. The rest of the paper is organized as follows: in section 2, we present the activities involved in developing the isolated speech recognition for this study while the methodology employed in the study is provided in section 3. The result and discussion are presented in section 4 while concluding remarks are presented in section 5.

2. DEVELOPMENT OF THE ISOLATED SPEECH RECOGNITION ALGORITHM

Figure 1 shows the principle of operation of the developed isolated word recognition algorithm. For this study, the operational principles are divided into two stages or phases: training phase and the testing phase. During the training phase, speech samples from different speakers were recorded using Goldwave software as the digital recorder. Silence and unvoiced portion removal along with endpoint detection of the speaker utterance was performed using a linear pattern classifier.

The features of the digitized speech samples were extracted using Mel Frequency Ceptrum Coefficients (MFCC). The classification was done by template matching; therefore, initial reference templates were prepared beforehand from different speakers of the same utterance for the three words (count, stop and down) in order to make it speaker independent and the template features were stored in the database.

During the testing phase, different set of speech samples were recorded from different speakers as test inputs in order to test the developed speech recognition system accuracy. The MFCC of the test inputs were matched with the template vocabulary using Digital Time Warping (DTW). The dynamic programming was used to find the best match for each spoken word by computing the optimum distance between the templates vectors and the test inputs vectors. The decision logic was based on the optimum distance. This is achieved by determining the word with the least optimum distance. Detailed information on the activities involved in each stage as shown in Figure 1 is presented in the following sub-section as follows:

The features of the digitized speech samples were extracted using Mel Frequency Ceptrum Coefficients (MFCC). The classification was done by template matching; therefore, initial reference templates were prepared beforehand from different speakers of the same utterance for different words in order to make it speaker independent and the template features were stored in the database.

During the testing phase, different set of speech samples were recorded from different speakers as test inputs in order to test their recognition accuracy. The MFCC of the test inputs were matched with the template vocabulary using Digital Time Warping (DTW). The dynamic programming was used to find the best match for each spoken word by computing of the optimum distance between the templates vectors and the test inputs vectors. The decision logic was based on the optimum distance and the word with the least optimum distance is the recognised word. Detailed information on the activities involved in each stage shown in Figure 1 is presented in the following sub-section as follows:

2.1 Silence Removal and End point Detection (EPD)

Incorrect endpoint detection of an utterance can produce two negative effects: recognition errors because of the incorrect boundaries and increased computations resulting in memory wasted on processing the non-speech segment of the speaker’s utterance [3]. In order to remove silence and the unvoiced parts in order to keep only the voiced segment of the speaker’s utterance needed for further processing of the speech, several end point detection algorithms can be adopted.

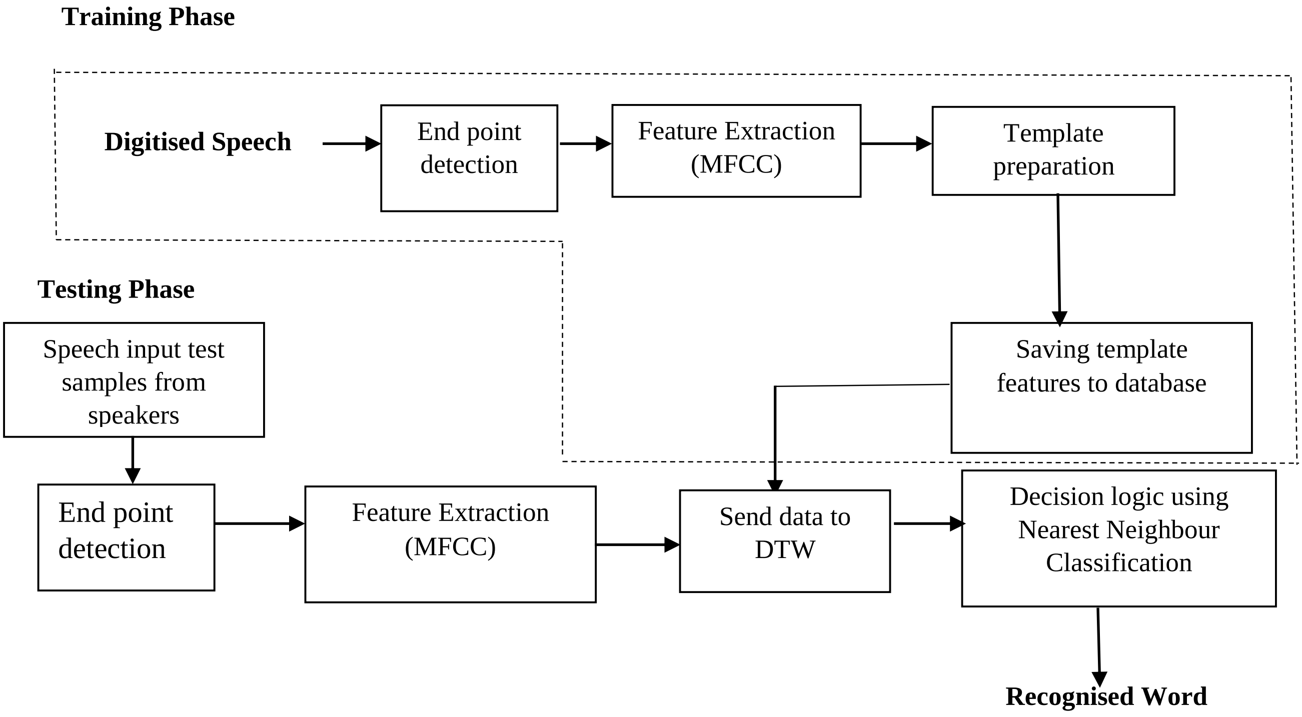
Training Phase
Digitised Speech
End point detection
Feature Extraction (MFCC)
Template preparation
Saving template features to database
Testing Phase
Speech input test samples from speakers
End point detection
Feature Extraction (MFCC)
Send data to DTW
Decision logic using Nearest Neighbour Classification
Recognised Word

Fig. 3.1 Developed Isolated Word Recognition System using MFCC and DTW Algorithm

In this paper, the linear classifier approach was adopted because it is computationally efficient for real time applications and it has better advantage over other conventional methods for speech samples collected from noisy as well as noise free environment [4].

The algorithm for the linear pattern classifier in this study was implemented using the flow chat adopted from [4] shown in Figure 2, where, x is a random variable, constants μ and σ and K are mean, the standard deviation and threshold value respectively. Figure 2 uses statistical properties of background noise and it also involves the smoothening of the physiological aspects of speech from the speech production process.

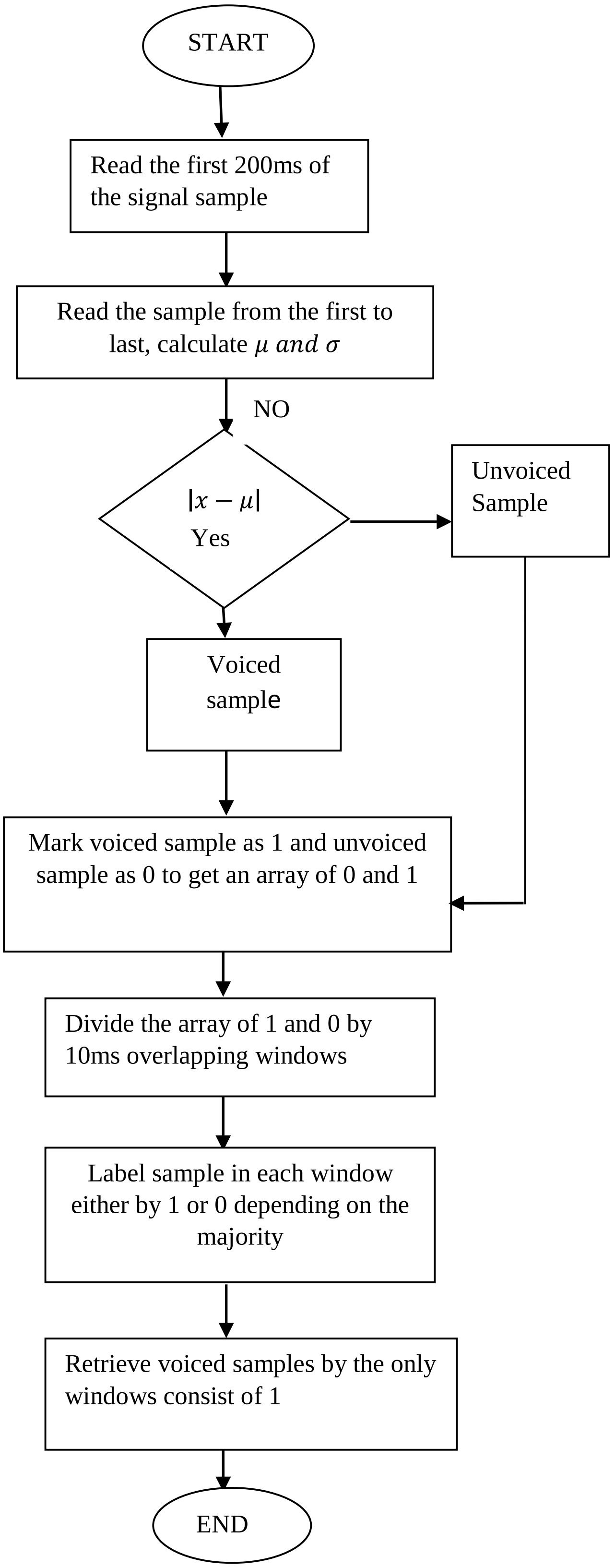
START
Read the first 200ms of the signal sample
Read the sample from the first to last, calculate μ and σ
NO
$|x-\mu|$
Yes
Unvoiced Sample
Voiced sample
Mark voiced sample as 1 and unvoiced sample as 0 to get an array of 0 and 1
Divide the array of 1 and 0 by 10ms overlapping windows
Label sample in each window either by 1 or 0 depending on the majority
Retrieve voiced samples by the only windows consist of 1
END

Figure 2: The Flowchart for the Linear Pattern Classifier

2.2. Acoustic Feature Extraction

This is the second stage in developing the speech recognition presented in this paper. As the name implies, the stage involves the extraction of some features from the speech samples. This is by transforming the speech signal to a set of feature vectors. The aim of this process is to obtain a new representation which is more compact, less redundant, and more suitable for statistical modeling. It was carried out because is a key to front-end process [5]. According to [5], the extraction technique to be adopted must be easy to extract, not easily influenced by the physical state of the speaker, not subject to mimicry and must not be affected by variation in speakers' utterances so as to be able to extract the important speech features that are suitable for speech recognition.

In this study, it was adopted because it fulfils the above features. The cepstral coefficients for MFCC were obtained through the following steps:

i. **Pre-emphasis**: This process increases the energy of signal at higher frequency [6]. The pre-emphasis stage is expressed mathematically in [7] as;

$$y\,(n) = \;x\,(n) – 0.95x(n-1) \qquad (1)$$

where, $y(n)$ is the pre-emphasised signal, $x\,(n)$ is the digitised speech sample. This means 95% of any one sample is presumed to originate from previous sample.

ii. **Frame Blocking**: This is the second step involved in obtaining the cepstral coefficients from the speech signal. The essence of frame blocking is to enhance the computation of the speech parameters in short time intervals to reflect the dynamic change of the speech signal. Typically, the spectral parameters of speech are estimated in time intervals of 10ms. The length of the segment is usually chosen from 16ms to 25ms, while the time window is shifted in time intervals of about 10ms to compute the next set of speech parameters.

iii. **Windowing**: This is to minimise the signal discontinuities at the beginning and end of each frame. It is expressed mathematically in [7] as;

$$Y(n) = y(n) \times w(n) \qquad (2)$$

The hamming window $w(n)$ is expressed mathematically in [7] as;

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \qquad (3)$$

iv. **Discrete Fourier Transform**: The Fourier transform was used to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain to frequency domain. The Discrete Fourier, $J(k)$, according to [8] is expressed as;

$$J(k) = \sum_{k=0}^{N-1} Y(n)e^{-\frac{j2\pi kn}{N}}; k = 0,1,\ldots.N-1 \qquad (4)$$

v. **Mel Filter Bank Processing:** The frequency range in the spectrum was very wide and voice signal does not follow the linear scale. Therefore, after the spectrum has been computed, the data was mapped on the Mel-scale using triangular, overlapped filters by following the four steps used in [9], which are:

a. Decision on the number of Mel bank filters m, where $m = 1, 2, \ldots, M$ is the number of bank filters.

b. Computation of the central frequencies: Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters [7]. The central frequency is expressed in [7] as;
$\phi_c = m.\Delta\phi_c$, where $\Delta\phi_c$ is given as;

$$\Delta\phi_c = \frac{(f_{max})_{(Mel)} - f_{min(Mel)}}{M+1}(Mel) \qquad (5)$$

where, $(f_{max})_{(Mel)}$ and $f_{min_{(Mel)}}$ are the correspondent of the Hertz scale in Mel scale. This was computed using the expression;

$$f(mel) = 2595\log_{10}\left(\frac{f(hertz)}{700} + 1\right) \qquad (6)$$

c. Computation of the Mel bank filters: This was used for filtering in the ceptrum main. It was achieved by using the converted central frequencies (Mel to Hz), which was expressed mathematically in [7] as;

$$R(k,m) = \begin{cases} 0 & , f(k) < f_c(m-1) \\ \frac{f(k)-f_c(m-1)}{f_c(m)-f_c(m-1)} & , f_c(m-1) \le f(k) < f_c(m+1) \\ \frac{f(k)-f_c(m+1)}{f_c(m)-f_c(m+1)} & , f_c(m-1) \le f(k) < f_c(m+1) \\ 0 & , f(k) \ge f_c(m+1) \end{cases} \qquad (7)$$

where the Hertz scale $f_c(m)$ correspondent of Mel scale ϕ_c was computed by using;

$$f_c(m) = 700\left(10^{\frac{f(mel)}{2595}} - 1\right) \qquad (8)$$

d. Mapping the frequencies from Hertz to Mel Scale: This is the last step in Mel bank processing. It was achieved by using;

$$J_1(m) = \sum_{k=0}^{N-1} |J(k)|\, R(k,m) \qquad (9)$$

where, $J_1(m)$ denotes the Mel spectrum.

vi. **Discrete Cosine Transform**: It was used to compute the Mel ceptrum coefficients. This was achieved by first computing the logarithm of the Mel ceptrum and then transformed into quenfrency domain using the discrete cosine transform.

vii. **Delta energy and delta spectrum**: The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, changes in features were added. These are the delta (velocity) features and double-delta (acceleration) features.

2.3 Template Preparation

During the training stage, there is a need to prepare reference template for each of the word to be recognised. The reference template can either be prepared for the representation of a single segment of speech with a known transcription, or for some sort of average of a number of different segments of speech. Therefore, if the template created is corrupt, it must be connected or changed because a suitable template for the tested word [9]. The accuracy of the speech recognition systems greatly relies on the quality of the prepared reference templates. This was achieved in this study by averaging the speech features of different speakers using the cross word template technique also known as average template method as outlined in [10]. Table 1 gives detailed information of the training session. A master template was formed from 10 samples of each word and stored in database.

Table 1: Training Requirement

Process/Method	Description
1) Speaker	Male
2) Environment	Relatively low noise environment
3) Utterance	Repeat 10 utterances from different speakers for recognition for the following words: i. Count ii. Stop iii. Down
4) Template preparation method.	Average template method
5) Time normalisation technique adopted	Linear interpolation
6) Sampling Frequency	12000Hz
7) Feature computation	39 double delta MFCC

2.4 Testing Phase

This is the second stage in developing the speech recognition system reported in this paper. This stage like the training stage also involves different activities. Some of the activities permed in the training stage were involved in this stage. However, two peculiar activities in this stage are feature matching and system testing. The two activities are presented in the following sub-sections.

2.4.1 Feature Matching

This is used for speech classification. The MFFC features of each reference word template are matched with that of the test input speech features. This process is called feature matching. In this paper, the dynamic time warping (DTW) was used as the feature matching technique. It is a non-linear technique that involves stretching or shrinking (mapping) of one signal to another by minimizing the distance between the reference templates and the test input template. It is widely used for isolated word recognition systems [11]. In this algorithm, the time-axis fluctuation is approximately modeled with a non-linear warping function of some carefully specified properties. Timing differences between two speech patterns are eliminated by warping the time axis of one so that the maximum coincidence is attained with the other. Then, the time-normalized distance is calculated as the minimized residual distance between them [11].

The local distance measure used to measure the time difference between the reference template and input speech is given in [8] as;

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})' * (\vec{x} - \vec{y})} \qquad (10)$$

where, x and y are the components of the reference sample and the input test sample respectively and $d(\vec{x}, \vec{y})$ is the Euclidean distance. Similarly, the optimum distance was estimated using the dynamic programming algorithm given in [11] as;

$$H(i,j) = min \begin{bmatrix} H(i,j-1) + d(i,j) \\ H(i-1,j-1) + 2d(i,j) \\ H(i-1,j) + d(i,j) \end{bmatrix} \qquad (11)$$

where, $H(i,j)$is global distance, $d\ (i,j)$ is the local distance and the initial condition is expressed mathematically as;

$$H_{\ 1}(1,1) = 2d(1,1) \qquad (12)$$

The expected or standard result of the dynamic time warping reported in [7] is shown in Figure 3.

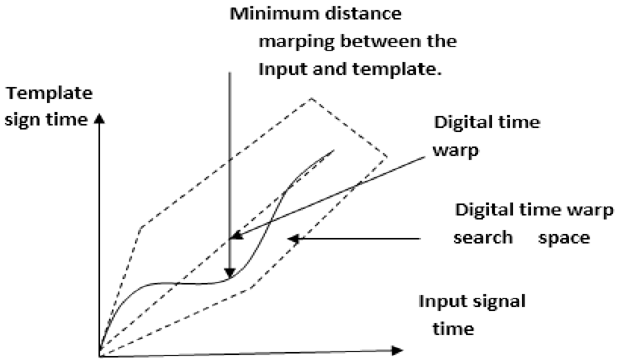

Figure 3: The Expected Result of the DTW

2.4.2 Testing

In order to evaluate the performance of the developed speech recognition system, ten extra speech input samples were collected for each word. The developed algorithm was tested with 20 speech samples each. DTW was used to estimate the optimum distance between the templates and the input speech sample. The word with the least optimum distance is the recognised word.

3. RESULTS AND DISCUSSION

The recognition results during the developmental stage were discussed followed by the experimental results when tested with test input speech samples. The Figure 4 shows the speech waveform for a digitised speech segment and Figure 5 shows its speech waveform after silence removal and end point detection set at 0.5 threshold.

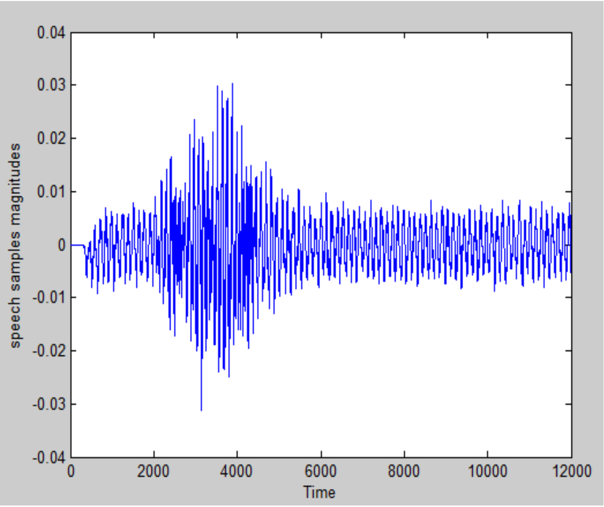

Figure 4: Digitised Speech Frame

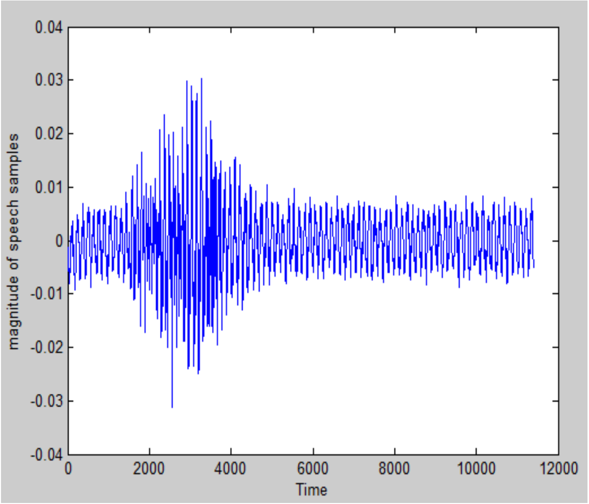

Figure 5: Speech waveform after EPD

The MFCC was carried out on the digitised speech after silence was removed. The power spectrum after performing discrete Fourier transform is shown in Figure 6 and the MFCC of the speech after applying the Mel filter bank is shown in figure 7. It was observed from Figure 6 that the power spectrum is periodic and by the application of symmetric property of the discrete Fourier transform, only a total number of $(N/2 + 1)$ samples were used or sufficient for further processing of the speech. Figure 7 shows the extracted features of the speech after reassembling the whole speech segment.

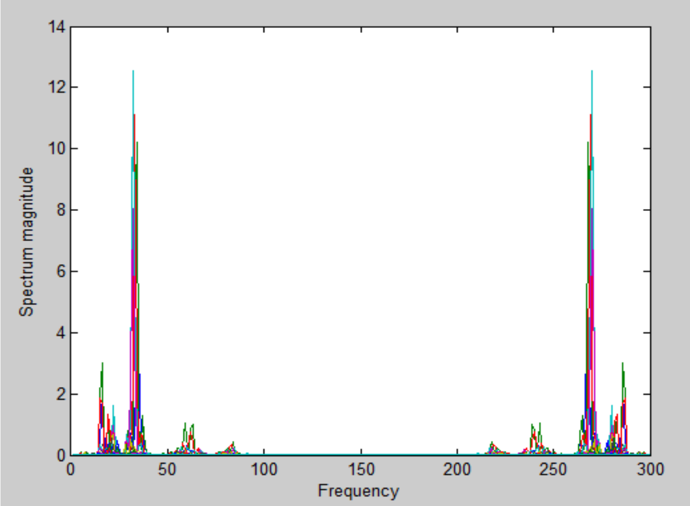

Figure 6: Power Spectrum of Speech Segment

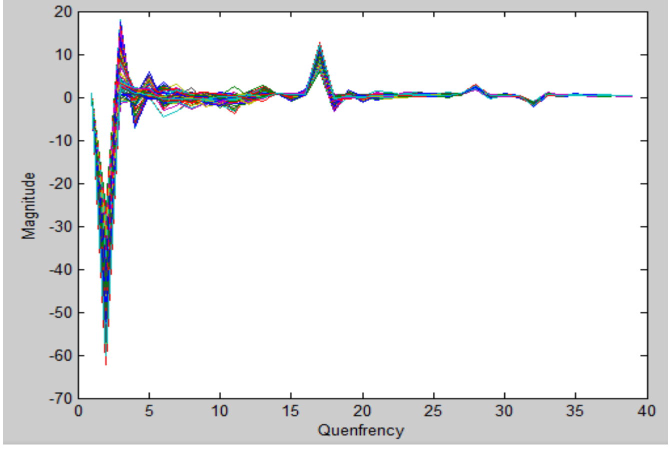

Figure 7: MFCC after Frame Reassembling

Figure 8 shows the warping path between a down sample and the down template while Figure 9 shows the warping path between a count sample and the down template. It was observed from Figure 8 that when the input test sample ‘down’ was compared with its template, it has reduced global dissimilarity and when compared with a different speech sample ‘count’, it has high global dissimilarity. By using DTW, the minimum optimum distance for Figure 8 was 8.5118 and that of figure 9 was 13.5003. As reported in [7], the expected or standard warping path between test speech sample and reference speech template using DTW is illustrated in Figure 3. Hence, comparison between Figure 8 and Figure 3 shows that the result obtained in this study, is similar to warping path of a standard or expected result of DTW in Figure 3. However, when Figure 9 was compared with Figure 3, there is significant difference as a result of the differences in the test input speech and the reference speech template used.

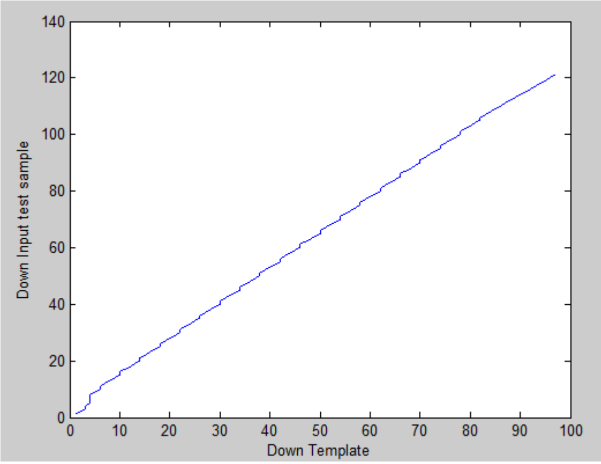

Figure 8: Warping Path between the Down Sample

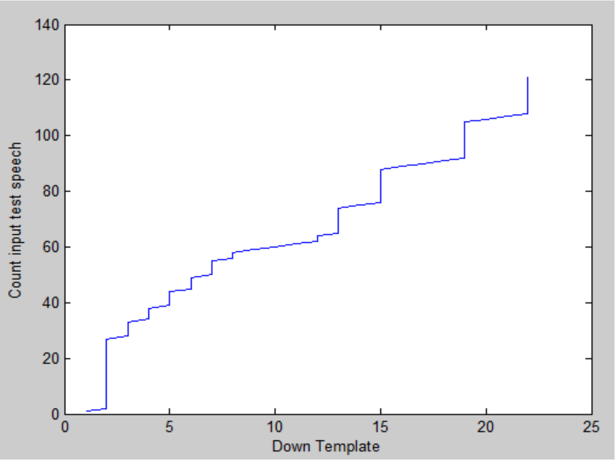

Figure 9: Warping Path between the Down Template and the Down Test Template and the Down Test Sample

In order to further estimate the performance of the developed speech recognition system, experimental result of the developed algorithm was carried out using DTW. This helps to compare the reference template of each speech sample with an unknown input speech and the one with the least optimum distance was displayed as the recognised word. The obtained experimental recognition accuracy was carried out using equation (13).

$$\text{Recognition Accuracy} = \frac{\text{The number of correctly recognised words}}{\text{The number of test inputs}} \times 100\% \qquad (13)$$

The recognition accuracy for the three speech samples in this study is presented in Table 2.

Table 2: Recognition Accuracy

Speech sample	**Number of speech input test**	**Number of recognised words**	**Recognition accuracy (%)**
Count	20	20	100
Stop	20	14	70
Down	20	12	60

Results presented in Table 2 shows that 100% recognition accuracy was achieved for the ‘‘count’ ’sample. The percentage recognition accuracy for the speech sample ‘‘stop’’ and ‘‘down’’ are 60% and 70% respectively. The differences in the recognition accuracy according to [12] might due to difference in pitch and ascent of the speakers.

4. CONCLUSION

In this study, speaker independent isolated word recognition using MFCC and DTW was developed. The processes involved as well as the result obtained when the developed isolated word recognition was evaluated has been presented in detailed. . The result obtained shows that the developed speech recognition system successfully recognised the isolated words; count, down and stop at different recognition rate of 100%, 60% and 70% respectively. The study has provided an efficient and non-complex computational algorithms for improving the recognition accuracy for speaker independent isolated word recognition system. This results obtained from the developed speech recognition system for this study also show better recognition accuracy compared to earlier isolated recognition developed systems which their results are low compared to the results obtained in the study. The study is gender dependent and require vocal tract length normalisation to make it gender independent in future research.

REFERENCES

[1] Blackburn, J., Miles, C., and Wing, B. (2006) ‘‘Speaker Recognition ’’*Journal of National Science and Technology Council*, available at www. Nistgov/speech/tests/spk/index.htm, pp 1-9.

[2] Baumann, J (2010), "Voice Recognition", Human Interface Technology Laboratory, *IEEE proceedings* Vol.1, pp. 1-4.

[3] Stephen, A. Z (2002) ‘‘A New Robust Algorithm for Isolated Word Endpoint Detection”, Lingyun GU Old Dominion University, *International Conference on Acoustics, Speech, and Signal Processing, Orlando,* pp 23, 25.

[4] Saha, S C., Suman, S., (2011) “A New Silence Removal, Endpoint detection and algorithm for Speech and Speaker Recognition Applications”, *Proceedings of the 13th international conference on Rough sets, fuzzy sets, data mining and granular computing India,* pp. 1-5.

[5] Reynolds, D., and Heck, L.P., (2000) “Automatic Speaker Recognition”, *Advancement of Science (AAAS) Symposium.*

[6] Anjali B.**,** Abhijeet, K., and Nidhika, B (2010), ‘‘Voice Command Recognition System Based On MFCC and DTW’’, *International Journal of Engineering Science and Technology*, Vol. 2 No. 12, pp. 7335-7342.

[7] Lindasalwa, M., and Elamvazuthi, I. (2010), ‘‘Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques’’ *Journal of Computing*, Vol. 2, No. 3, pp.1-6.

[8] Plannerer B. (2005), ‘‘An Introduction to Speech Recognition’’, Bernd Plannerer, Munich, Germany available at plannerer@ieee.org , *International* Journal *of Computer Applications, Vol. 12 No. 2, pp.1–7.*

[9] Rabiner, L.R and Sambur, M.R (1974), ‘‘An Algorithm for determining the endpoints detection of Isolated utterances’’, *American Telephone and Telegraph Company*, Vol.54, N0.2, pp.297.

[10] Mutcha, S.R. (2012), ‘‘Pattern Normalization/Template Optimization in order to Optimize Speech Recognition Process’’, *Journal of Research and Reviews*, available at www.ijsrr.org, ISSN: International 2279-0543, pp. 69-74., available at www.ijsrr.org, pp. 1.

[11] Hiroaki, S. and Seibi, C. (1978), ‘‘Dynamic Programming Algorithm Optimization for Spoken Word Recognition ’’, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 26, No. 1, pp. 44-49.

[12] [12] Patil, S.A., and Hassen, H.L (2007), ‘‘ Speech under stress: Analysis , Modelling and Recognition ’’, centre for robust speech systems university of Texas at Dallas, Richardson, USA, *Journal of the Acoustic Society of America*, Vol. 96, No. 6, pp. 108-137.