Effect of Pre-processing Stages on Recognition Accuracy of a Developed Isolated Word Recognition System

Animasahun, I, O*. and Popoola, J.J.
Department of Electrical and Electronics Engineering, Federal University of Technology Akure, Ondo State, Nigeria E-mail: ianimasahun8619@gmail.com; jidejulius2001@gmail.com
*Corresponding Author

Abstract— The major observed cause of inaccuracy of the earlier developed word recognition systems is the lack of sufficient combination of pre-processing stages adopted before extracting the speech features. In investigating this observed deficiency, two isolated word recognition systems were developed. The first was developed without pre-processing stages while the second was developed with pre-processing stages. The paper combined analog to digital conversion, end point detection and template preparation as paramount pre-processing techniques in developing the second word recognition system. The analog to digital conversion was used to obtain the digitized speech using a Gold wave software while linear pattern classifier was used as the end point detection technique to remove silence at the beginning and the ending of the digitized speech. Reliable master templates of the three speech samples words: count, stop and down were prepared using average template method. The performance evaluation carried out on the developed word recognition systems show a variation on the recognition accuracies of the three words employed with an average recognition of 23.3% and 76.7% without and with pre-processing stages respectively. The result of the study shows that pre-processing stages have significant effect on the accuracy of the word recognition system. Also, the results from the study show that differences in both ascent and pitch of the speakers have effects on the performance of the developed word recognition system.

Index Terms—Average Template Method, Dynamic Time Warping (DTW), Mel Frequency Cepstral Coefficient (MFCC), Nearest Neighbour Classification, Probability Density Function, Speech recognition, Nearest Neigbour Classiication (NNC).
.

1 Introduction

Speech recognition has been an area of research over decades. There are three types of speech recognition systems namely isolated, connected and complex word recognition systems. This paper focusses on the development of an isolated word recognition system. Isolated or discrete word recognition systems are systems that recognize digit words. A lot of works have been done which has to do with the development of isolated speech recognition system [1-3]. The targets of any developed recognition system are high recognition accuracy and run time. There are several reasons for the reduction in the recognition accuracies of most developed speech recognition systems [4]. The major cause of inaccuracy is the lack of sufficient combination of pre-processing stages adopted before extracting the speech features. Another reason for the deficiency in the recognition accuracy of some of the earlier developed speech recognition systems is inefficient pre-processing algorithms. As reported in [4] and [5], some of the crucial pre-processing stages needed to extract features necessary for speech recognition are analogue to digital conversion, silence removal/end point detection and the template preparation stages.

In [1], speech activated appliance was developed without carrying out end point detection and also vector quantization approach was used in template preparation. The recognition accuracies of the words used are low due to inefficient template preparation technique and the avoidance of end point detection technique.

This shows that before reasonable analysis can be carried out on speech, it must first be acquire in digital form as well as being recorded in a low noise environment [6]. Digital recorders such as Audacity, Adobe Audition, and Goldwave can be used. Audacity has editing properties but has fixed sampling frequency and the digitized speech cannot be read into Matlab file. Adobe Audition also has the limitation of fixed sampling frequency. Goldwave on the other hand, is preferable, since it allows the user to fix the sampling frequency and also has editing properties to enhance the quality of the digitized speech. One of such editing properties is the removal of background noise and the unvoiced segment of the digitized speech sample, which is a fundamental step in speech recognition systems [4].

In addition, there are different approaches that have been developed for detecting the speech end points in speaker's utterance. Two of them are energy based method which uses zero crossing rate and short time energy functions. The limitation in the energy based method is that there are still errors incurred in finding the exact end point of utterance [7]. Absolute Energy and Teager Energy (AETE) algorithm was proposed by Stephen [8], which is still far from perfect because of effect of background noise, which sometimes cause the algorithm to make the wrong start point sooner than the true starting point. A newly developed algorithm that adopted unidimensional Mahalanobis Distance and uses statistical properties of background noise as well as physiological aspect of speech production was formulated in [9].

The second activity to be carried out in order to enhance accurate recognition system is template preparation. One of the main problems in speech recognition systems is the preparation of reliable reference templates for the set of words to be recognized [5]. Vector quantization (VQ) has been widely used as a solution to prepare reliable templates for the Dynamic Time Warping (DTW) based speech recognition systems but it requires many training examples to prepare a reliable codebook. In order to enhance the computational efficiency, a simple method is to use single reference template per word [5]. Another approach is by using average template method which according to [5] helps to mitigate bad samples from good samples by using a master template for each speech sample prepared from several speech samples. The paper focuses on the pre-processing algorithms and discussed the results of each algorithm. Also, pitch detection algorithm was also implemented in this paper in order to justify that the fundamental frequencies of each speaker varies and that it has a way of affecting the speech recognition accuracy. The speech features after these pre-processing stages were extracted and the speech samples classification was achieved using decision logic.

The rest of the paper is organized as follows: the preprocessing algorithms are presented in section 2 while the decision logic approach used during the testing phase are presented in section 3. The result and discussion are presented in section 4 while concluding remarks are presented in section 5.

2 Pre-Processing Stages for Development of Speech Recognition

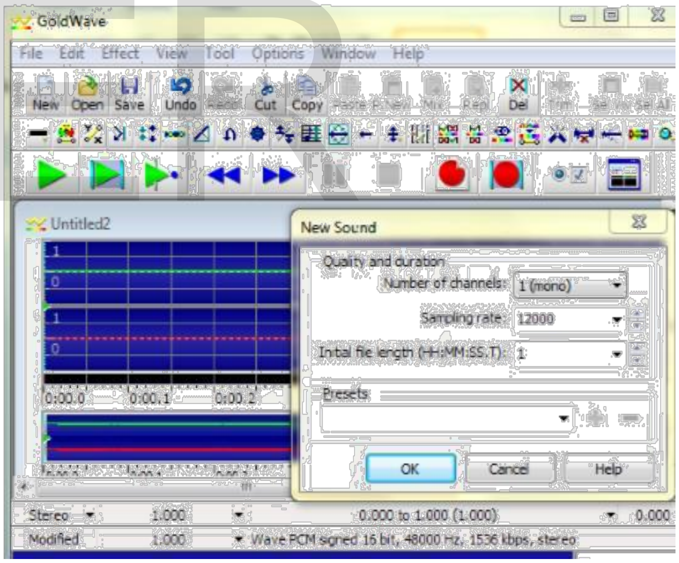
The pre-processing stages and the methods of implementation are presented in this section. The pre-processing stages discussed in the following sub-section are the analogue to digital conversion, end point detection, template preparation and finally pitch detection. All of these were implemented in the Matlab environment.

2.1 Analog to Digital Conversion

This is performed by sampling the analog speech signal using an analog-to-digital converter. Since speech is typically bandlimited to about 4,000Hz, the speech signal needs to be sampled fast enough such that aliasing is avoided [6]. The speech samples from the speakers were first recorded and converted into digitized speech using digital recorder. The recording requirements is shown in Table 1. The digital recorder used as the analog to digital converter is the Goldwave software, which is shown in Fig. 1.

Table 1: The Recording Requirements

Recording Requirement	**Description**
Number of channels	Mono(1)
Sampling rate	12000Hz
Duration	One(1) second
Isolated words	Count Stop Down
Environment	Relatively low noise environment



Number of channels: 1 (mono)
Sampling rate: 12000
Initial file length (HH:MM:SS.T): 1

Stereo 1.000 0.000 to 1.000 (1.000) 0.000
Modified 1.000 Wave PCM signed 16 bit, 48000 Hz, 1536 kbps, stereo

Fig. 1: The Gold wave's interface

2.2 End Point Detection

The end point detection method used is the linear pattern classifier approach presented in [4]. End point detection and silence removal was carried out on the digitized speech. The algorithm uses statistical properties of background noise to make a sample as voiced or silence/unvoiced. The algorithm also uses physiological aspects of speech production for smoothening and reduction of probabilistic errors in statistical marking of the voiced or silence/unvoiced.

The two basic parameters used for determining the probability density function used for the statistical mapping are the mean (μ) and variance (σ). The flowchart for the linear pattern classifier is shown in Fig. 2, where, x and K are random variable and threshold respectively.

2.3 Template Preparation

The templates used for the development of an independent isolated word recognition system for this study were prepared using the average template method. For the independent isolated word recognition system developed for this study, each word was repeated 10 times by different speakers. After the end point detection, the features of each speech sample were extracted using Mel Frequency Ceptrum Coefficients (MFCC). 39 MFCC double delta coefficients were used per frame for each speech sample. Interested reader(s) can check the detailed procedures involved in our earliest article [10]. The block diagram for the average template method used in [5] and adopted from this study is shown in Fig. 3. Linear interpolation was used as the time normalization technique. This was achieved by using the expression;

$$V_q = \text{interp1}(X, V, X, \text{Method}) \tag{1}$$

where, interp1 is a one dimension interpolation, Vq is the interpolated speech sample X is the query points, X is the length of X , V is the speech sample to be interpolated and Method is the linear. The features of each speech sample after the end point detection were saved into the database.

2.4 Pitch Detection

The pitch detection was used to justify that the fundamental frequencies of each speaker varies and that it has a way of affecting the speech recognition accuracy. The Modified autocorrelation function (MACF) given in [11] was used as the pitch detection algorithm because it was more convenient for common usage compared to others. The block diagram for MACF employed is shown in Fig. 4. The autocorrelation function is expressed in [11] as;

$$R_x(i) = \frac{1}{N}\sum_{n=-N}^{N} x(n)x(n+i), 0 \le i \le T \tag{2}$$

Where, $x(n)$ is the speech samples, N is the length of analyzed frame, T is the number of autocorrelation points to be computed. The variable is called lag, or delay. Center-clipping was used to flatten the spectrum of the signal passed to the candidate generator. This is also expressed in [11] as;

$$y(n) = clc[x(n)] = \begin{cases} (x(n) - C_L), & x(n) \ge C_L \\ 0, & |x(n)| < C_L \\ (x(n) + C_L), & x(n) \le -C_L \end{cases} \tag{3}$$

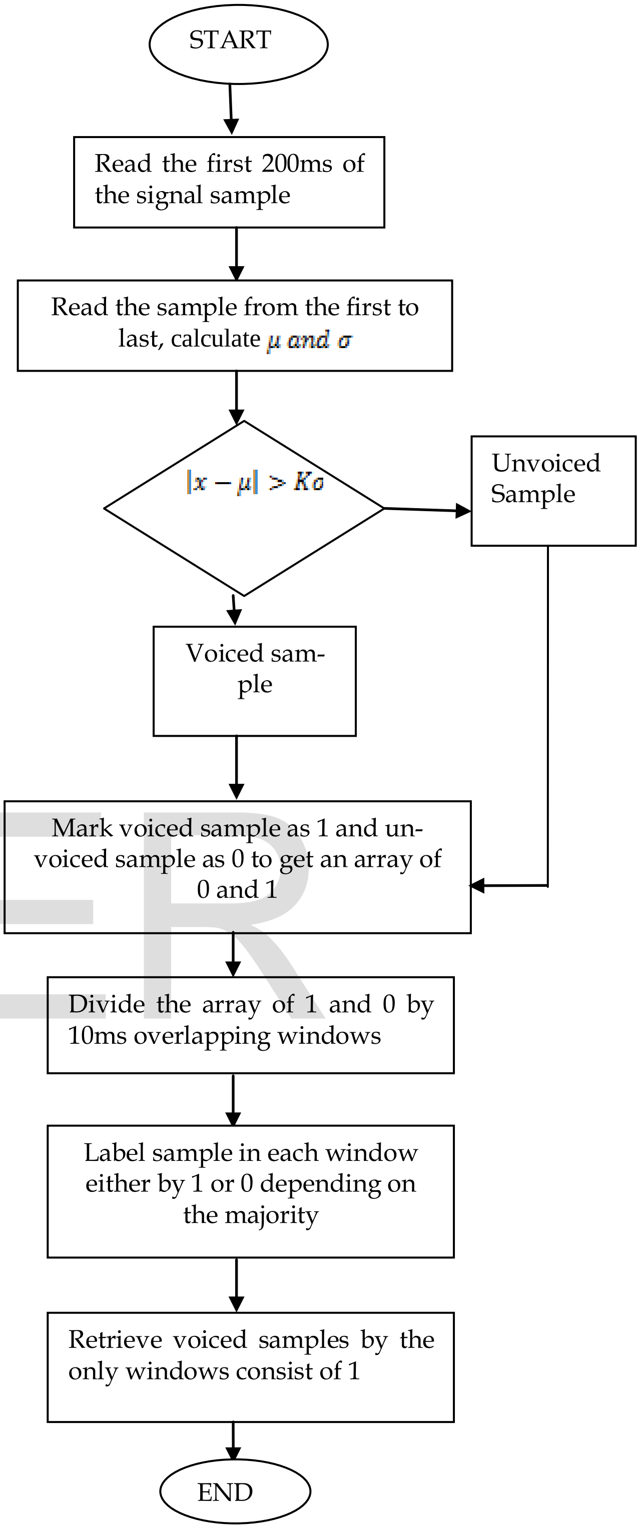

START
Read the first 200ms of the signal sample
Read the sample from the first to last, calculate μ and σ
$|x - \mu| > K\sigma$
Unvoiced Sample
Voiced sample
Mark voiced sample as 1 and unvoiced sample as 0 to get an array of 0 and 1
Divide the array of 1 and 0 by 10ms overlapping windows
Label sample in each window either by 1 or 0 depending on the majority
Retrieve voiced samples by the only windows consist of 1
END

Fig. 2: The flowchart for the Linear Pattern Classifier [4]

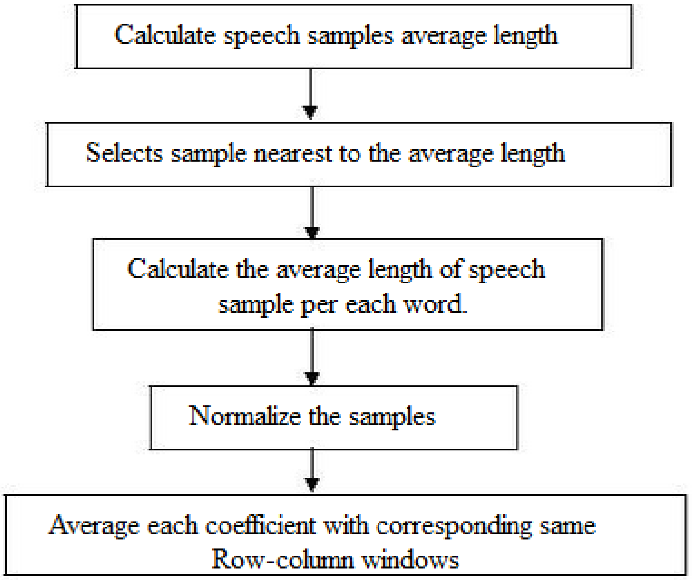

Calculate speech samples average length
Selects sample nearest to the average length
Calculate the average length of speech sample per each word.
Normalize the samples
Average each coefficient with corresponding same Row-column windows

Fig. 3: Block Diagram of the Average Template Method [5]

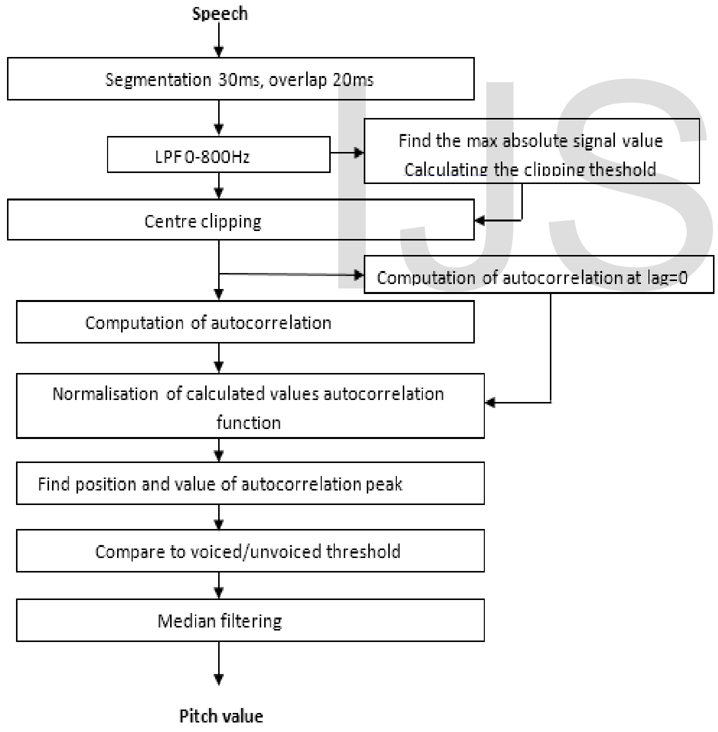

Speech
Segmentation 30ms, overlap 20ms
LPF 0-800Hz
Find the max absolute signal value
Calculating the clipping theshold
Centre clipping
Computation of autocorrelation at lag=0
Computation of autocorrelation
Normalisation of calculated values autocorrelation function
Find position and value of autocorrelation peak
Compare to voiced/unvoiced threshold
Median filtering
Pitch value

Fig. 4: The block diagram for Modified Autocorrelation Function [11]

The clipping threshold (C_{T}) was set to be the maximum of the absolute value signal value. In order to enhance the pitch detection accuracy, median filtering was employed. The median filtering was carried out by using a Matlab function given as;

$$Y = \mathrm{medfilt1}(X,N) \qquad (4)$$

where, medfilt1 is one dimensional median filter, X is the speech sequence, N is the order of the output, Y is the filtered speech sequence.

3 DECISION LOGIC

After developing the speech recognition for this study, it was tested. The speech recognition was tested using the dynamic time warping to determine the optimum or the time-normalized distance. The input speech test was compared to the acoustic vectors of the three templates (count, stop and down). Templates already formed in the database for ten different speakers were compared with the input speech test for the independent isolated word. Nearest Neighbour Classification (NNC) was used as the decision logic.

The Nearest Neighbour Classification is given as;

$$wr = \arg\min Dtw(IST,Rr) \qquad (4)$$

where, wr is the recognized word, Dtw is the dynamic time warping (DTW), IST is the input speech test. The Nearest Neighbour Classification is explained by the block diagram given in Fig. 5.

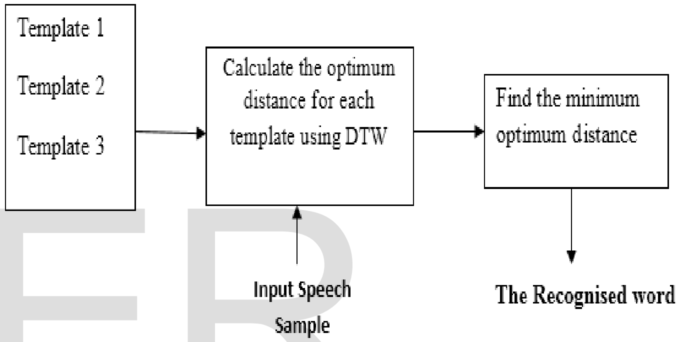

Template 1
Template 2
Template 3
Calculate the optimum distance for each template using DTW
Find the minimum optimum distance
Input Speech Sample
The Recognised word

Fig. 5: The block diagram for the Nearest Neighbour Classification

4 RESULT AND DISCUSSION

The results of the pre-processing stages are presented and discussed in this section. The digitized speech samples of count, stop and down are given in Fig. 6.

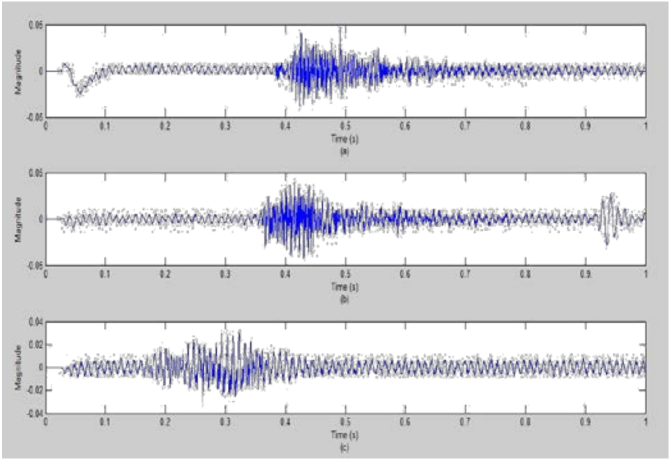


Fig. 6: Speech Waveforms: (a) Count, (b) Stop and (c) Down

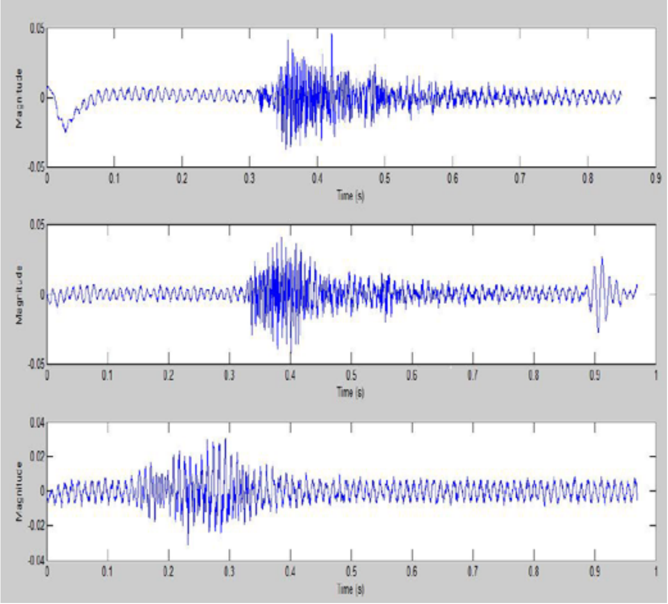


Fig. 7: Speech Waveforms after for End Point Detection for (a) Count, (b) Stop, (c) Down

Table 2 shows the threshold values (K) at which the voiced segment of the speaker utterance were extracted. It also shows the accuracy of the voiced segment obtained after end point detection. The Table 2 also shows whether the extracted voice segments of the utterances are correct or not when played back. The desired threshold value differs from one speech sample to the other even when a sample was repeated more than once. This is due to the fact that a speaker cannot utter the same word, the same way at different times. The results obtained show that the higher the threshold, the lower the accuracy of the voiced segment of the speaker's utterance. In order to prevent distortion of the voice segment, the threshold value was set at 0.3 for all the speech samples. The 0.3 threshold was used because at 0.3, the extracted voice segment of the utterances were all correctly obtained. Each end pointed utterance was obtained at a lesser number of samples. Figure 7 shows the speech waveform of count, stop and down after silence has been removed. The beginning and the ending of the speaker's utterance were detected when set at a threshold value of 0.3. Comparison between Fig. 6 and Fig. 7 shows that the silence in each speech sample was removed and the beginning and the ending of the speaker's utterance were detected. The speech samples were played back so as to confirm whether voiced segment were not eliminated with the unvoiced and silence portion of the speakers' utterances.

After the template for the study had been prepared, it robustness and accuracy was tested by comparing the reference template or prepared template with one of the ten speakers for the speaker independent isolated word recognition system to be developed. The result obtained is presented in Fig. 8. The obtained results show that the count reference template and count 1 sample follows the same pattern with reduced local distance. The effect of preparing a master template is evident in the minimization of the local distance compared to when a single sample is used as the reference template.

Table 2: Voiced Segment after EPD at different Thresholds

Threshold values	Count1	Stop1	Down1	Count2	Stop2	Down2
0.1	Correct	Correct	Correct	Correct	Correct	Correct
0.2	Correct	Correct	Correct	Correct	Correct	Correct
0.3	Correct	Correct	Correct	Correct	Correct	Correct
0.5	Incorrect	Correct	Correct	Incorrect	Correct	Correct
0.8	Incorrect	Correct	Incorrect	Incorrect	Correct	Incorrect
1.1	Incorrect	Incorrect	Incorrect	Incorrect	Correct	Incorrect
1.4	Incorrect	Incorrect	Incorrect	Incorrect	Correct	Incorrect

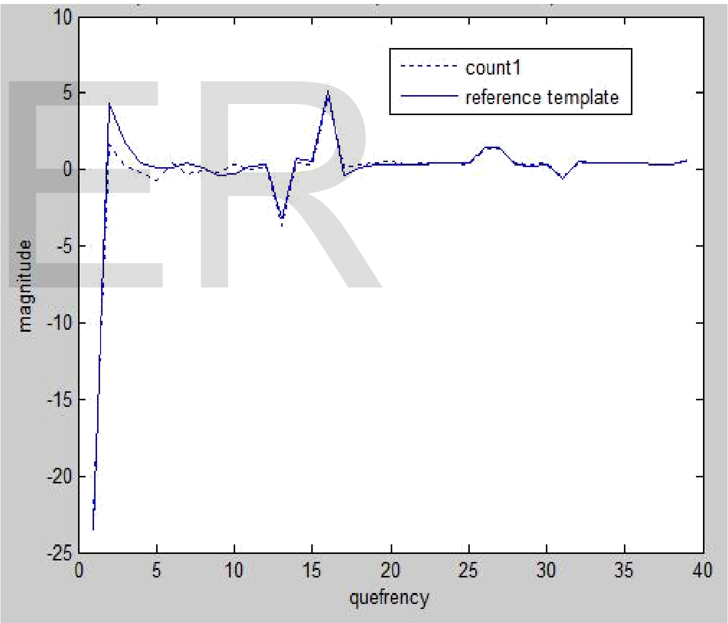


Fig. 8: Comparison between the Count Reference Template and Count 1 Sample

The result of the pitch detection is presented in Table 3, which shows that the fundamental frequencies of ten speakers for ''count'', ''stop'' and ''down'' using MACF. The results show significant variation in the fundamental frequencies of ten speakers. However, the variation is not an indication of error in the template preparation but due to the fact that individual speaker has different pitch values. The result obtained, though varied for the 10 speakers fall within 91-209Hz with average fundamental frequency of 120.1Hz. The higher values are due to the age difference of the speakers and also the ambiguity of peak detection in the MACF used for estimation of the pitch.

The recognition of the isolated words were achieved using the Nearest Neighbour Classification (NNC) as the decision logic. The decision logic used the DTW as the measure of global dissimilarity. The developed speech recognition system was experimentally evaluated by collecting twenty (20) samples each for each word and were tested against the master templates already in the database using the 39 double delta acoustic vectors.

The overall average values of 23.3% and 76.7% were obtained as recognition accuracies before and after pre-processing stages respectively. The Fig. 9 and Fig. 10 shows the number of recognized speech samples for each word before and after the pre-processing stages. The comparative results obtained when pre-processing activities were employed and when they were not employed are presented in Fig. 9. The results show that the system recognition accuracy is better when pre-preprocessing stages were employed. This show that adequate pre-processing stage has significant effect on recognition accuracy of the word recognition system.

The variation observed in the recognition accuracies of the three speech samples buttresses the results presented in [13] that the ascent and the pitch differences have effects on the speakers' utterances.

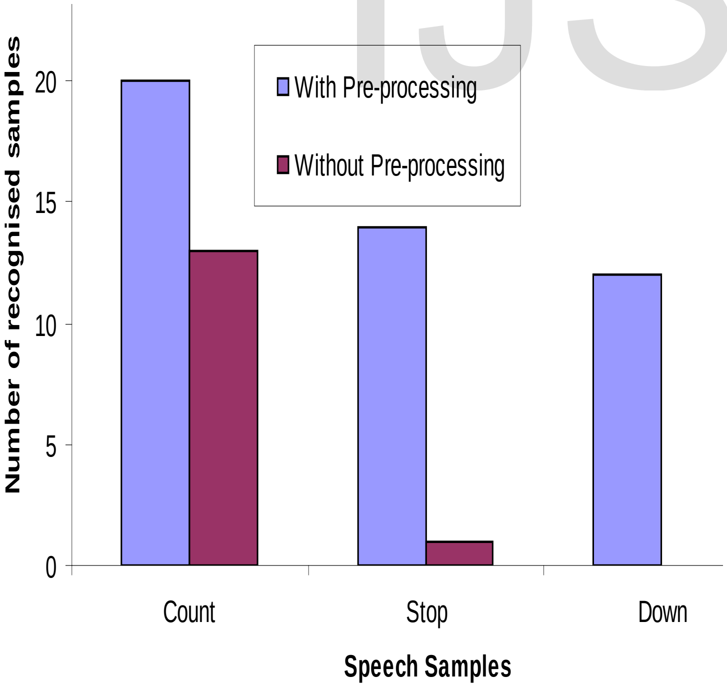


Fig. 9: Comparative Recognition Accuracy with and without pre-processing activities

S

Table 3: Fundamental Frequencies of Speech Samples Using MACF

Speaker	**Fundamental Frequency/Pitch (Hz)**		
	Count	Stop	Down
1	99	91	134
2	178	136	124
3	169	209	121
4	162	209	125
5	181	127	119
6	112	122	93
7	92	92	93
8	92	92	91
9	92	92	94
10	95	91	92

5 Conclusion

In the study, the effect of the pre-processing algorithms on accuracy of word recognition have been examined. The detailed information on the development of an isolated word recognition algorithms for the study was presented. The recognition was done using 39 MFCC delta coefficients and NNC as decision logic. The waveforms of speech samples after silence has been removed and the beginning and the ending of the speaker's utterance were detected at a threshold value of 0.3. The average recognition accuracy for sixty test input speech samples is 23.3% before pre-processing stages and 76.7% after the pre-processing stages. The effect of the preparing a master template is evident in the minimization of the local distance compared to when a single sample is used as the reference template. This aids easy warping of speech features and hence; improved recognition accuracy. The result of the study shows that both ascent and pitch has effect on word recognition accuracy.

REFERENCES

[1] D. Beckstrom, A. Harun, S. Nicole and B. Zorawar, "Speech Activated Appliances," Uiniversity of Victoria, 2007. (Thesis or Dissertation)

[2] A. B, K. Abhijeet and B. Nidhika, "Voice Command Recognition System Based On MFCC and DTW," *International Journal of Engineering Science and Technology,* vol. 2, no. 12, pp. 7335-7342, 2010. (Journal)

[3] L. Antanas, Joana L.E. and Laimutis T., "Development of Isolated Word Speech Recognition System," *Institute of Mathematics and Informatics,* vol. 113, no. 1, pp. 39-41, 2002. (Journal)

[4] S. C. Saha and S. Suman, "A Silence Removal, End Point Detection Algorithm for Speech and Speaker's Recognition Applications," Indian Institute of Technology, khragpur, 2011. (Journal)

[5] S. Mutcha, "Pattern Normalization/Template Optimization in order to Optimize Speech Recognition Process," *Journal of Research and Reviews,* vol. 1, no. 2, pp. 69-74, 2012. (Journal)

[6] A. Panaithep, "Implementation of a Connected Digit Recognizer Using Continuous," Bradley Department of Electrical and Computer Engineering Polytechnic Institute and State University, Virginia, 1998. (Thesis or Dissertation)

[7] L. Rabiner and M. Sambur, "An Algorithm for determining the endpoints detection of Isolated utterances," *American Telephone and Telegraph Company,* vol. 54, no. 2, pp. 297-315, 1974. (Journal)

[8] A. Z. Stephen, A New Robust Algorithm for Isolated Word Endpoint Detection, Lingyun: Lingyun GU Old Dominion University, 2002, pp. 22,25. (Thesis or Dissertation)

[9] O. Richard, P. Duda and David G.S., Pattern Classification, second ed., Wiley, Ed., Califonia: A Wiley- interscience publication, John Wiley & Sons, 2001, p. 1. (Journal)

[10] I. O. Animasahun and J. J. Popoola, "Application of Mel Frequency Ceptrum Coefficients and Dynamic Time Warping in developing an isolated Speech Recognition System," *International Journal of Science and Technology, UK.,* vol. 4, no. 1, pp. 1-8, 2015. (Journal)

[11] D. Vydáno, "Performance Evaluation of Pitch Detection Algorithms," *Journal of the Acoustical Society of America,* vol. 41, no. 2, pp. 293-309, 2009. (Journal)

[12] C. Jan and H. Valentina, "Speech Recognition: Introduction and Dynamic Time Warping," *Journal of Statistical Software,* vol. 31, no. 7, pp. 1-24, 2007. (Journal)

[13] S. Patil and H. H.L., "Speech under stress: Analysis , Modelling and Recognition," *Journal of the Acoustic Society of America,* vol. 96, no. 6, pp. 108-137, 2007. (Journal)

- *Animasahun I.O. has B.ENG and M.ENG in Electrical and Electronics engineecing, communication option in Federal university of Technology Akure, Ondo State. E-mail: ianimasahun8619@gmail.com*

- *Popoola J.J has B.S.C and M.ENG in the Federal University of Technology Akure, Ondo State and Ph.D in the University of Witwatersrand, South Africa. E-mail: jidejulius2001@gmail.com*