Automatic Speech Recognition and Verification using LPC, MFCC and SVM

Aaron M. Oirere
Department of computer Science and Information Technology,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad- 431 004 (MS), India

Ganesh B. Janvale
MGM's Institute of Biosciences and Technology, MGM Campus, N-6, CIDCO,
Aurangabad 431 003 (MS) India

Ratnadeep R. Deshmukh
Department of computer Science and Information Technology,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad- 431 004 (MS), India

ABSTRACT
Speech has much capability as an interface between human and computer which comes under the Human Computer interaction (HCI). The major challenge has been the nature of voice is ever varying speech signal.

The paper presents the development of the speech recognition system using Swahili speech database which was collected in three sets: digits, isolated words and sentences from both native and non native speakers of Swahili language.

Different feature extraction techniques deployed in the system are: Linear Prediction Coding (LPC) and Mel-Frequency Coefficients (MFCC). We have used the 12 coefficient features from MFCC and 20 coefficients features from LPC. All these features extracted techniques are applied and tested for the own developed Swahili speech database.

Recognition and verification were done using confusion matrix and Support Vector Machine (SVM) as a classifier for the classification purpose. LDA was tested for the entire dataset for the dimension reduction. LDA gave a good clustering. The performance of the system was checked on basis of their accuracy; Confusion with MFCC 50.9%, confusion with LPC 50.1%, the higher recognition rate in each data set were as follows numeric data: MFCC: 75%, LCP:72% , isolated word data: MFCC: 65.2% LPC: 66.67%, sentence data MFCC: 63.8%, LPC: 59.6.

General Terms

Speech Database, Speech Recognition, Natural Language Processing, Human Computer Interaction, SVM, LDA

Keywords
Swahili, Swahili Text corpus, Phonetics, Text Corpus and Speech Corpus, Automatic Speech Recognition

1. INTRODUCTION
The most common and natural way of communication for human being is and has been speech. Technology world also have utilise voice extensively just to name a few are like radio, telephone, voice storage device (cassette, memory card), television etc. This has been there from the dark ages to the civilization ages and even the modern age it is still of great importance in the communication sector.

Digital signal processing has advance the speech technology and speech processing in various was for example: speech enhancement, speech synthesis, speech compression, speaker recognition and speech recognition and verification. Speech recognition can be categorized into two types that is speaker dependent and speaker independent. Speaker dependent speech recognitions is a system which can only recognise the speech based on particular individual(s) while speaker independent speech recognition system is a system that can recognise a speech regardless of the speaker, it doesn't depend on particular speakers it is open to any individual.

Speech recognition can be also called as computer speech recognition or text to speech (TTS) [1] it is a process of converting a speech signal to a word or sequence of words by means of computer algorithms. There have been many researchers worked in speech processing and communication, which has been quite motivating but still they haven't fully reached in the state-of-art due to the nature of speech is quasi-periodic/ non stationary. Speech signals are ever varying signals especially when examined over a short period of time (5-100 Msec.). it is assumed between 10-20 msec. the signals are fairly stationary. This can be reflected in different speech sounds being spoken.

Human voice commands can be followed and understand particular language via computer bit the main drawback is the limitation of speech vocabulary and computational duration. The system can work more effectively in small speech data and task oriented system unlike a dynamic system. There are a wider range of applications which require a human machine interaction such as query based information system, automatic call processing in telephone networks, stock price quotations, data entry, weather reports, voice dictation, and access to information: banking, travel, voice commands, automobiles portal, avionics, speech transcription, physical challenged people shopping, railways reservations etc [1]. The research in automatic speech recognition by machines has attracted a great deal of attention for last sixty years [2].

2. ABOUT SWAHILI LANGUAGE
The Swahili language mostly came from Arabic and Bantu languages, it serve as a first or second language in most countries in Eastern and few countries in western Africa. It is being spoken with approximately sixty to one hundred and fifty million Speakers [3]. The basic phone set of Swahili comprises of 5 vowels and 27 consonants [4]

Although Swahili language being one of most spoken language in Africa, yet very limited work is done in the language as compared to other language like English, hence this was the motivating factor to venture in it and utilize it more.

3. DATABASE

3.1 Selection of Speaker

The speech data were collected from the native as well as non native speakers of Swahili language. The selected native speakers were resident of the countries where Swahili language/Kiswahili is one of the recognized national languages. The non native speakers were selected based on those who were from non Swahili speaking countries and are much comfortable in reading, writing and speaking the English language fluently. The speakers were selected to cover the complete diversity i.e. age, gender, ethnic group and language knowledge. For the non native speakers, the training was given to them about pronunciations.

3.2 Data Collection

The database was developed in three sets i.e. numeric speech data set (0-9) [5], agriculture speech data set (50 words), and general and commonly used sentence (30 sentences). Each data sets of the database contain 5 utterances summing up to a total of 3000 numeric speech data, 12,500 words and 15,000 sentences. The numeric speech database was collected from both native (30 person: 15 males, 15 females) and non native speakers (30 persons: 15 males, 15 females), second set was collected from native speakers (50persons: 25 males and 25 female) and third set was collected from 100 native speakers (50 male and 50 female).The age group of the speakers ranged from 15 to 60 years old randomly selected from Kenya in East Africa.

3.3 Recording Procedure

The database was recorded our database using two different high quality headsets i.e. (Sennheiser PC360 and PC350) and using the PRAAT Software. The data was recorded in noisy environment. The recording of the Speech samples in such environment was very crucial for the development of robust automatic speech recognition system. The speech samples were recorded in mono mode with a sampling frequency of 16000Hz. A microphone was at a distance of about 3 cm from the mouth. The PRAAT software is a freeware tool that is being widely used by the researchers who are working for the development of Speech technologies.

3.4 The steps followed while recording the speech

Step1

Selected speakers were asked if they had any problem with reading or speaking the Swahili language. They are trained to speak data samples.

Step 2

Speakers were demonstrated and thought about the headset used on when to speak the word.

Step 3

The sampling frequency set as 16 KHz with 16 bit in Mono sound type.

Step 4

The speaker was requested to read each word, sentence and the recorded sample was saved as ".wav file" simultaneously.

Step 5

Step 4 repeated for all utterances (digits, words and sentence) that were recorded from the speakers. All the steps were repeated for all the speakers.

3.5 Text corpus

The text corpus comprises both isolated words (digits and agriculture related words) and continuous sentence as shown in table 1 and 2.

Table 1 Samples of Isolated Words corpus

Numeric Speech Corpus		Agriculture Isolated Word Speech Recognition	
Swahili	**English**	**Swahili**	**English**
Sufuri	Zero	Pili pili	Chili pepper
Moja	One	Uyoga	Mushroom
Mbili	Two	Nyanya	Tomato
Tatu	Three	Sukuma wiki	Kales
Nne	Four	Kabichi	Cabbage
Tano	Five	Kunde	Cowpeas
Sita	Six	Kitunguu	Onions
Saba	Seven	Kiasi	Potato

Table 2 Sample of Continuous Speech Corpus

Swahili	English
Habari yako?	How are you?
Jina lako ni nani?	What is your name
Mzuri sana	Am very fine
Si mzuri sana	Am not very fine
Siku yako ilikuwaje	How was your day?
Habari ya asubuhi	Good morning
Umekula chakula	Had your food?
Umekula nini?	What have eaten?
Habari ya jioni	Good evening
Lala salama	Nice sleep

4. FEATURE EXTRACTION

The feature extraction techniques applied on the database collected is MFCC and LPC. MFCC we used 12 coefficients and LPC we used 20 coefficients

4.1 MFCC

MFCC is based on the human hearing perception auditory system which can't perceive frequency over 1KHz. Speech naturally is quasi periodic (non-stationary) hence the tone is measured in Hz and pitch is measured on a scale called the "Mel Scale" which contains two types of filters that are speech is linearly at low frequency below 1 KHz and logarithmic space above 1KHz [6]. The steps involved in the MFCC feature extraction are as follows:-

Speech signal

An exciting signal and impulse response of vocal tract is called speech signal. It is shown in figure 1.

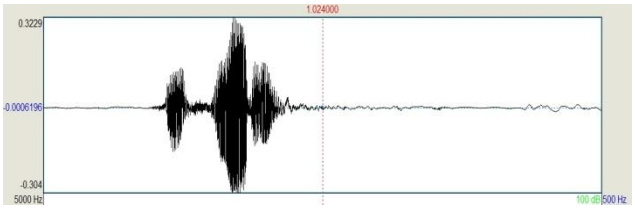


Fig 1: Speech signal of digit "sufuri" zero

Pre-emphasizing:
The speech signal at time contain noise, hence in order to remove noises pre-emphasis is performed to enhance the speech signal.

Framing and Windowing:
Speech signal is non stationary but it remain stationary at 0-20ms, hence we did the segmentation in the framing 20ms. Windowing is performed to avoid unnatural discontinuities in the speech segment and distortion in the spectrum. Calculation of number of frame is done by multiplying the signal, consisting of N samples, with a rectangular window function of a finite length.

Fast Fourier Transform:
The purpose of performing Fast Fourier Transform is to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain from frequency domain. Spectral analyses signify that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore, FFT is executed to obtain the magnitude frequency response of each frame and to prepare the signal for the next stage i.e. Mel Frequency Warping.

Mel-Frequency Filter Bank:
A 24 triangular shaped band-pass filter banks are created by calculating a number of peaks uniform spaced in the mel – scale and then transforming them back to the normal frequency scale. Human hearing of speech sound signal perception's frequency does not follow a linear scale. Individual tone's frequency f is measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The mel frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000Hz. To compute the mels for a given frequency f in Hz, a the following approximate formula is used [7]. The transmission from linear frequency to mel-frequency is shown in equation 1.

$$mel = 2595 \cdot \log(1 + \frac{frequency}{700}) \quad (1)$$

The mel frequency spectrum furthermore reduces the amount of data without losing vital information in a speech signal. The resolution as a function of the frequency is logarithmic.

Discrete Cosine Transform:
The DCT transform the frequency domain into time domain. It is widely used in the area of speech processing and is often used when working with cepstrum coefficients. In a frame, there are 24 mel cepstral coefficients obtained, out of 24 only 13 coefficient has been selected for the recognition system. For the word recognition system, we have selected 13 Mel-Cepstral Coefficients frame wise [8]. Figure 2 shows the 13 Mel Cepstral coefficients of five utterances of isolated word digit"zero".

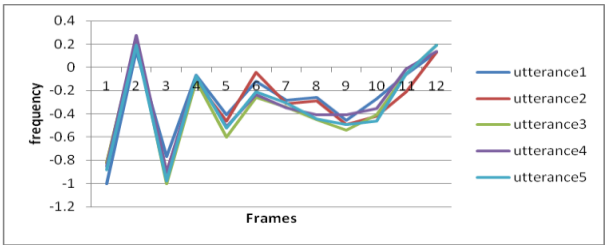


Fig.2 The graph of MFCC features of "zero" (numeric isolated word sample)

4.2 LPC

Linear Predictive coding (LPC) analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue. In LPC system, each sample of the signal is expressed as a linear combination of the previous samples. This equation is called a linear predictor and hence it is called as linear predictive coding .The coefficients of the difference equation (the prediction coefficients) characterize the formants. The figure 4 shows the steps involved in LPC feature extraction.

The LPC model characterizes the vocal tract of a person. Linear prediction coefficients are a highly effective representation of the speech signal. In this analysis, each speech sample is represented by a weighted sum of *p* past speech samples plus an appropriate excitation. The corresponding formula for the LPC model is shown by equation 2.

$$S_n = \sum_{k-1}^{p} a_k s_{n-k} + Gu_n \quad (2)$$

Where *p* is the order of the LPC filter, s_n is n^{th} speech sample and a_k is the k^{th} coefficients of the LPC vector. The LPC are found by Durbin algorithm which minimizes the mean square prediction error of the model [9, 10, 11].

The fig 3 shows the LPC features extraction of 20 coefficients of the five utterances of the isolated word: digit "zero".

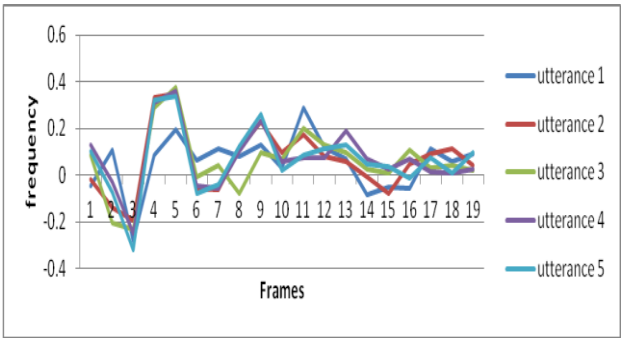


Fig. 3 The graph of LPC features of zero (numeric isolated word sample)

5. CLUSTERING

5.1 Linear Discriminant Analysis (LDA)

LDA is dimensional reduction algorithm, it was used for clustering purpose; LDA has a potential of creating different classes of dataset and reduce dimensions.

Linear Discriminant Analysis (LDA) with MFCC

In this paper, we utilized the classification of ten classes of numeric numbers (0-9). Each class had a group 12 MFCC cepstral coefficients. The linear discriminant (fisher Algorithm) has been implemented for class-within class matrix dataset. Our total samples were 100: 50 for male speakers and 50 for female. MFCCs features of numerical data (from 0 to 9 digits) are classified by LDA. Likewise the same was done for isolated words and sentence speech data. Figure 4 (a, and b) separates the groups with less confusion.

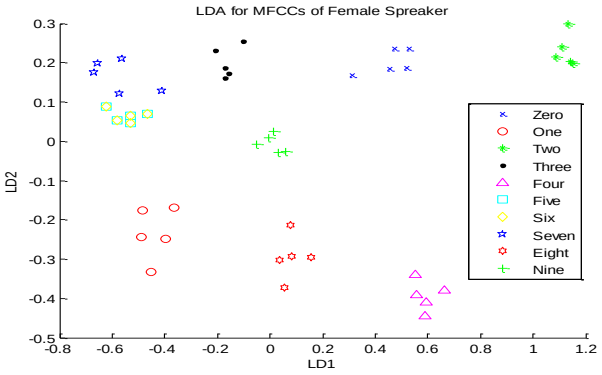


Fig: 4a Cluster of LPC by LDA Female Speaker

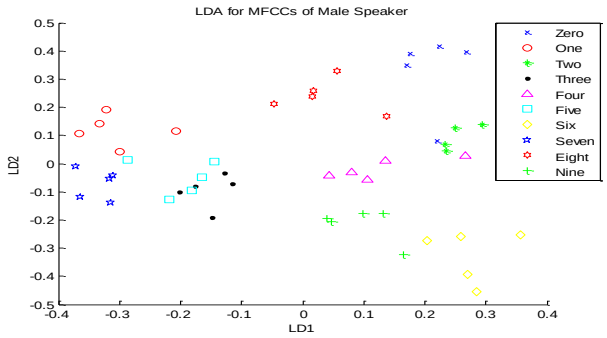


Fig: 4b Cluster of LPC by LDA Male Speaker

5.1.1 Linear Discriminant Analysis (LDA) with LPC

LPC is based on the source-filter model of speech signal. This was used as a feature extraction and we focused on the 20 coefficients, which were stored in different feature vector and by help of LDA classification was done accordingly. This was done for all the data sets. Below are the figs of the clustering of 0 to 9 numbers. From the figure 7 a, it is observed that all ten groups are separated except seven and three. But in figure 7 b group of eight and nine are mixed each others.

5.2 Confusion Matrx

5.2.1 Confusion matrix applied on MFCC

The research randomly checked the accuracy of MFCC extracted features with confusion matrix and the following were our results. The high accuracy for female speaker: 73.2% least: 20%, average: 50.7% male speaker: accuracy- high 80%, least 36.6%, average 55.2% all speakers in general: accuracy high75.3%, least 38.6% and average 50.9%.

Table 3 Average distance matrix for all speakers' MFCC with confusion matrix

D	2	7	0	6	4	3	1	8	5	9	N of S	Per%
2	**49**	0	7	0	1	4	4	2	2	0	69	71.01
7	0	**58**	1	0	0	4	8	0	6	0	77	75.32
0	2	0	**52**	6	1	4	3	1	0	5	74	70.27
6	0	3	7	**34**	2	4	2	9	5	17	83	40.96
4	9	0	2	3	**33**	6	1	21	0	2	77	42.85
3	0	4	3	3	0	**38**	10	2	16	10	86	44.18
1	0	9	4	3	0	8	**29**	1	14	7	75	38.66
8	0	0	2	4	14	8	3	**34**	0	6	71	47.88
5	0	6	0	1	0	7	11	9	**28**	9	71	39.43
9	0	0	8	20	1	6	4	2	0	**26**	67	38.80
												50.94

D=Digits, N of S =Number of tested Sample, Per= performance in percentage

5.2.2 LPC with confusion matrix

The LPC coefficients features of our speech data were randomly tested using the confusion matrix. The performance of the systems was rated as shown table 4:

Table 4 Performance of the system

LPC	Highest accuracy	Least accuracy	Average accuracy
Female speakers Male speakers Both (female + male)	84.2% 85.3% 50.1%	32.5% 28.9% 25.7%	49.99% 51.6% 50.1%

Table 5: Average distance matrix for all speakers' LPC with confusion matrix

D	4	1	7	3	6	0	5	9	8	2	N of S	Per%
4	**39**	3	0	1	20	2	1	4	4	12	86	45.34
1	0	**41**	15	2	0	4	3	3	4	0	72	56.94
7	0	6	**36**	9	0	0	12	6	2	0	71	50.70
3	0	12	11	**22**	4	2	23	4	4	0	82	26.82
6	4	1	0	2	**27**	9	1	27	4	2	77	35.06
0	0	3	0	0	1	**63**	0	6	0	0	73	86.30
5	0	9	9	12	0	0	**18**	0	22	0	70	25.71
9	1	3	1	3	11	10	0	**37**	1	3	70	52.85
8	0	8	1	10	4	3	10	1	**36**	0	73	49.31
2	10	1	0	0	2	3	0	3	2	**55**	76	72.36
											750	**50.14**

D=Digits , N of S =Number of tested Sample, Per= performance in percentage

5.3 Support Vector Machine (SVM)

One of the tools for pattern recognition that uses a discriminative approach is a SVM [12 13]. SVMs use linear and nonlinear separating hyper-planes for data classification. However, since SVMs can only classify fixed length data vectors, this method cannot be readily applied to task involving variable length data classification. The support vector classifier uses the function is given as equation 3

$$f(x) = \langle \propto \cdot K_s(x) \rangle + b \quad (3)$$

The $K_s(x) = [k(x, s_1), \dots, k(x, s_d]^T$ is the vector of evaluation of kernel functions centred at the support vectors $S = \{s_1, \dots, s_d\}, s_i \epsilon R^n$ which are usually subset of the training data. The classification rule is defined as in equation 4

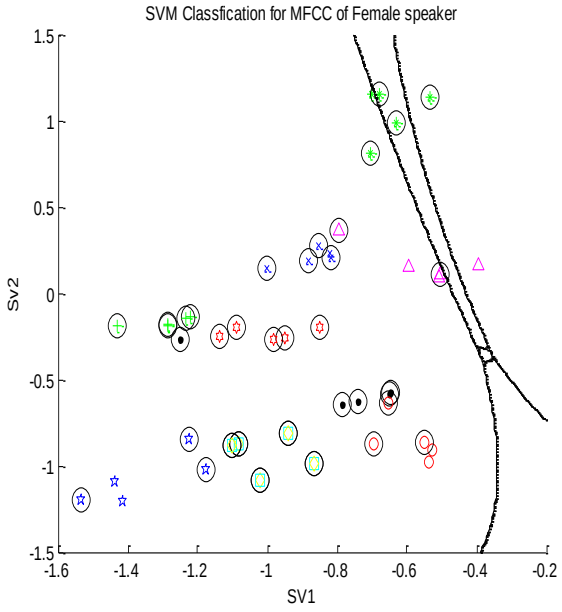
$$q(x) = \begin{cases} 1 \; for \; f(x) \geq 0, \\ 2 \; for \; f(x) < 0. \end{cases} \quad (4)$$

And multiclass classification function and rule is defined in equation 5 and 6 respectively.

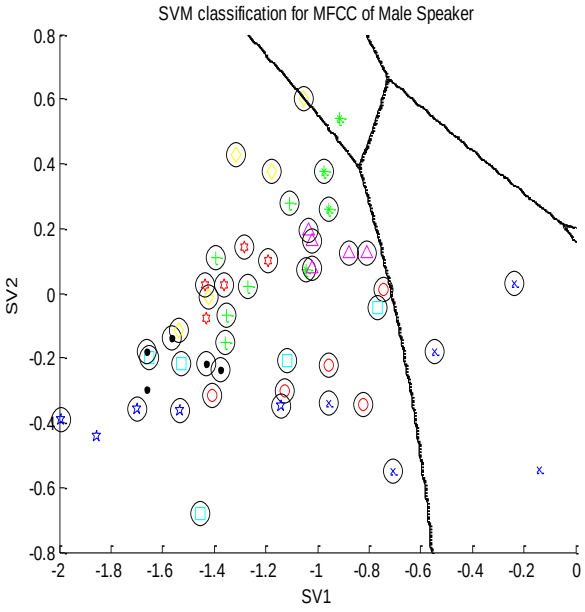
$$f_y(x) = \langle \propto_y \cdot k_s(x) \rangle + b_y, \qquad y \in Y \quad (5)$$

$$q(x) = argmax \; f_y(x), y \in Y \quad (6)$$

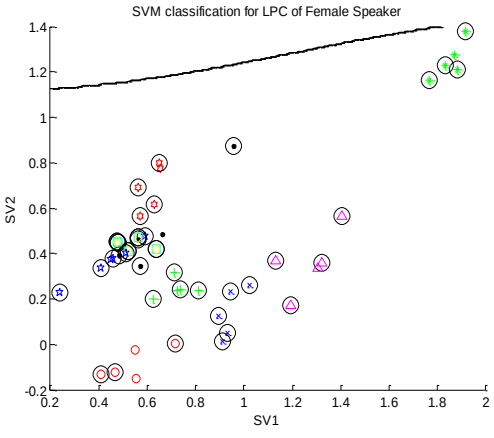
We have tested all numerical data by SVM classifiers. Clustering is given in figure 5



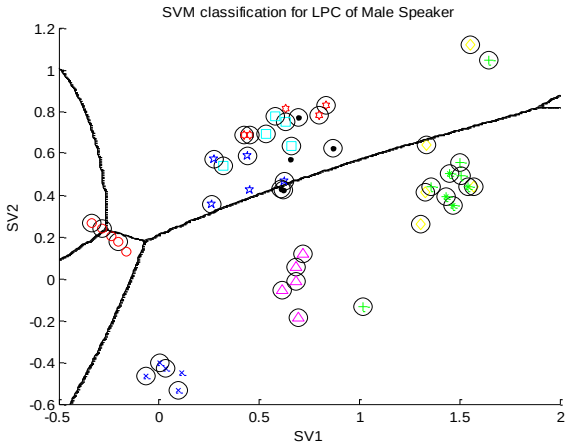
a.



b.



C



D

Fig.5. SVM Classification of MFCCs and LPCs (a & c female speaker) (b & d male speaker)

From the figure 5 (a, b, c, and d), only LPC features are better classified compare MFCCs from both male and female and LPC from female.

6. CONCLUSION

The speech database was designed and developed for Swahili language in three sets: digits, agriculture isolated words and sentences spoken by native and non native speakers. The features are extracted by using MFCCs and LPCs. LDA was used for clustering speech features. LDA gives better clustering results. Classifications of the features between the training and test dataset were done with confusion matrix and SVM. The performance of the system was tested on the basis of the accuracy. These were as follows: Confusion with MFCC 50.9%, confusion with LPC 50.1%, the higher recognition rate in each data set were as follows numeric data: MFCC: 75%, LCP:72% , isolated word data: MFCC: 65.2% LPC: 66.67%, sentence data MFCC: 63.8%, LPC: 59.6

7. FUTURE SCOPE

Swahili database and recognition systems will be used for robust ASR. The complete recognition system will be useful in garniture field. The excellent performance of our system may be proposed for many other under resource languages in Africa as well.

8. ACKNOWLEDGMENTS

We would like to thank the University Authorities to provide the basic facilities for carrying out the research work. This work is supported by University Grants Commission.

9. REFERENCES

[1] M.A.Anusuya and S.K.Katti , (2009) "Speech Recognition by Machine: A Review" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3.

[2] Dat Tat Tran, Fuzzy "Approaches to Speech and Speaker Recognition", A thesis submitted for the degree of Doctor of Philosophy of the university of Canberra.

[3] Irele, Abiola and Biodun Jeyifo, (2010) The Oxford encyclopedia of African thought, Volume 1. Oxford University Press US. New York City. 2010. ISBN 0-19-533473-6

[4] Gakuru, Mucemi Iraki, Frederick K. Tucker, Roger Shalonova, Ksenia Ngugi, Kamanda, (2005) "Development of a Kiswahili text to speech system", In INTERSPEECH, 1481-1484.

[5] Aaron M. Oirere, Ratnadeep R. Deshmukh and Pukhraj P. Shirshrimal, (2013) "Development of Isolated Numeric Speech Corpus for Swahili Language for Development of Automatic Speech Recognition System" International Journal of Computer Applications (0975 – 8887) Volume 74– No.11, July 2013

[6] Kashyap Patel, R.K. Prasad, (2003) "Speech Recognition and Verification using MFCC & VQ" international journal of Emerging Science and Engineering (IJESE) volume 1 issue 7, 33-37.

[7] Shivanker Dev Dhingra, Geeta Nijhawan and Poonam Pandit , (2013) " Isolated Speech Recognition Using MFCC and DTW" International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, p 4085- 4092.

[8] Daniel Jurafsky & James H. Martin, (2007)"Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition.

[9] Makhoul J. Linear Prediction: (1975) A Tutorial Review. *Proceedings of the IEEE.* Vol 63, 561-579.

[10] Campell J.P. and Jr. (1997) Speaker recognition: A tutorial. *Proceeding of the IEEE.* Vol 85, 1437-1462.

[11] Ganesh B. Janvale , Vishal Waghmare, Vijay Kale, Ajit Ghodke, "Recognition of Marathi Isolated Spoken Words Using Interpolation and DTW Techniques", ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India- Vol I

[12] Volume 248 of the series Advances in Intelligent Systems and Computing pp 21-29

[13] R.K.Moore, (1994) "Twenty things we still don't know about Speech", Proc.CRIM/ FORWISS Workshop on Progress and Prospects of speech Research and Technology.