ISSN:0975-9646

Speech Based Voice Recognition System for Natural Language Processing

Dr. Kavitha. R[1], Nachammai. N[2], Ranjani. R[2], Shifali. J[2],

[1]Assitant Professor-II,CSE, [2]BE..- IV year students
School of Computing,
SASTRA University, Thanjavur 613402, India

Abstract-A system that recognizes and authenticates the voice of a user by extracting the distinct features of their voice samples is usually termed as Voice recognition system. Voice identification is carried out by converting the human voice into digital data. The digitized audio samples then undergo feature excerption process to extract Mel Frequency Cepstral Coefficients features. These coefficients are subjected to feature matching through Dynamic Time Warping to match with the patterns existing in the database for limited Tamil words. This paper focuses on a secure system that deploys the voice recognition for a natural language (Tamil) by combining the digital and mathematical knowledge using MFCC and DTW to extract and match the features to improve the accuracy for better performance.

***Keywords:** Discrete Cosine Transform (DCT), Mel Frequency Cepstral Coefficients (MFCC), Feature Matching, Fast Fourier Transforms (FFT), Dynamic Time Warping (DTW).*

I. INTRODUCTION

Speech and Voice Recognition are the emerging scope of security and authentication for the future. Now-a-days text and image passwords are prone to attacks. In case of the most commonly used text passwords, users are required to handle different passwords for emails, internet banking, etc. Hence they tend to choose passwords such that they are easy to remember. But they are vulnerable in case of hackers. In case of image passwords, they are vulnerable to shoulder surfing and other hacking techniques. Advances in speech technology have created a large interest in the practical application of speech recognition. Therefore this system provides the users with the appropriate and efficient method of authentication system based on voice recognition.

Humans are always comfortable to communicate in their natural language. Communication is an integral part of human life, also a symbol of identity and authorization. However in case of computer, human interaction, language accent and dialects differ for different set of people.

The existing voice recognition systems are limited to the English language and its accuracy depends on whether the user is a native speaker or not and how much close the users' language commands are to the trained dataset. The proposed system for the natural language allows the users to select a password in the Tamil language which can be easy to remember for the native speakers. This system is mostly concerned on "who is speaking?" rather than "what they are speaking?" and hence it acts as a speech based authentication system.

II. METHODOLOGY

The analysis of a voice is performed by getting the input audio samples through a microphone from a user. Speech Recognition system for natural language can be achieved using two important phases such as Feature Excerption phase and Feature Matching phase. Feature excerption is the process of deriving the required data while discarding the noise and other disturbances in the sample. In the feature matching phase the MFCC features of voice signal are compared with the reference templates in the database that contains the user details with the help of Dynamic Time warping (DTW) [2] [6] .The Reference templates are created by continuously training the system with Tamil words and by the users at the time of sign up in to the system.

1. FEATURE EXCERPTION

Feature Excerption extracts features from voice samples that are unique to each individual which can be also used to differentiate various speakers. The voice samples of the speakers can vary with respect to their dialect, context, style of speaking and their emotional state [9]. Therefore excerption of the voice samples and representing them in an appropriate form is an important task as it can improve the recognition performance efficiently. Many algorithms such as Mel Frequency Cepstral Coefficients (MFCC), Human Factor Cepstral Coefficient (HFCC) and Linear Predictive Cepstral Coefficients (LPCC) can be used for the feature excerption [12]. The secure system proposed here uses Mel-frequency Cepstral Coefficients (MFCC) which has less complexity and provides high accuracy results. MFCC are the coefficients obtained by performing the processes as shown in Fig.1.1 on the digitized voice samples. The complete process of feature excerption consists of a set of computational steps having its own functionality and mathematical approach for easier processing.

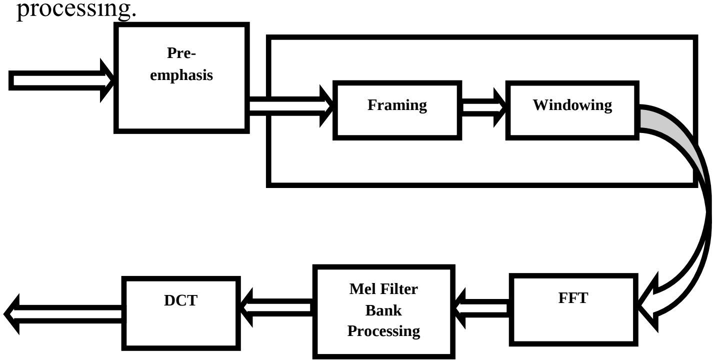


Figure1.1 Feature Excerption process

1.1 Pre-emphasis

The voice signal recorded by the user is taken as the input in this first phase feature excerption. Most of the recorded audio samples contain noise due to the disturbances in the recording environment. The noise and the silence in the audio samples should be removed before the features are extracted. Hence, the part of the signals with very high or very low amplitude is removed. A high order filter is used to pre-emphasize the voice sample [2].

$Y (d) = X (d) – C * X (d-1)$

Where X (d) is the input signal, Y (d) is the output signal and C is a constant with value between 0.9-1.

1.2 Framing

Framing is a process of fragmenting the samples into small frames with lengths of about 20 to 40 milliseconds. The pre-emphasized signal is divided into number of frames with each having A samples so that each frame can be processed in quick succession of time instead of analysing the complete signal at once and hence decreasing the complexity [12]. Neighbouring frames can be separated using M (M<A) [5]. Since overlapping can be done on each individual frames, hamming window is preferred to avoid some information in the initial and final parts of the frame. Overlapping puts these data back into the derived features [12] [4]. The frame size is usually in the powers of 2, otherwise zero padding is done such that the signal is flexible for the further process.

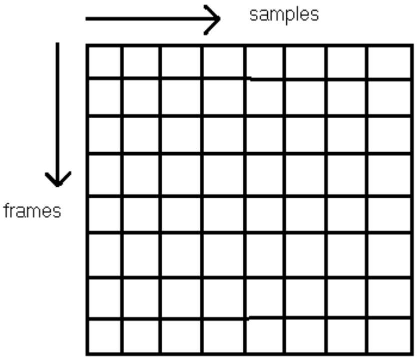


Figure 1.2 Framing of the voice sample

1.3 Windowing

Windowing is a process of integrating all the frequency lines that are close to each other in order to avoid any discontinuities and distortions at the initial and final parts of the each frame. When FFT is used on voice samples, it assumes that signal gets repeated and so the final part of one frame does not get connected with the initial of the next frame easily [5]. There occurs some disturbance on the frame at regular intervals. This can be overcome by windowing which makes the end of each frame smoother to get connected with other frames without any complications. The commonly used window in speech recognition is the Hamming Window defined as H (d)

$H (d) = 0.54 - 0.46 \cos (2\pi d/D-1), 0 \leq d \leq D-1$

The processed signal is obtained by the relation,

$Y (d) = X (d) * H (d)$

here D = number of samples in each frame, Y (d) = Output signal, X (d) = Input signal.

1.4 Feature Representation

In feature representation, the audio samples obtained from the speakers are converted into Mel-Frequency Cepstral coefficients. Feature representation is performed using Fast Fourier Transform (FFT), Mel-Filter Bank processing and Discrete Cosine Transform (DCT).

1.4.1 Fast Fourier Transform

FFT is employed to calculate DCT and its inverse. The FFT converts the signal from time function to frequency function. The fragmented frames are converted from time function into frequency function [4]. The transformation is performed by the following equation:

$Y (r) = F [H (i) * X (i)] = H (r) * X (r)$

Fast Fourier Transform computes the features of spectral-domain of the speech. Number of FFT elements is equal to the size of the time sample. The FFT is calculated separately for each frame as real and imaginary parts and the magnitude of the spectrum is calculated.

1.4.2 Mel Filter Bank Processing

Mel filter banks are used to map the powers of the cepstrum calculated to the Mel scale. The reference between the Mel frequency Cepstral scale and frequency scale (Hertz) is given by 1000 Mel to a 1000Hz tone [9].

$f (Mel) = [2595 * \log_{10} [1 + f/100]]$

1.4.3 Discrete Cosine Transform

DCT works by converting the signal from frequency function (Mel) to time function resulting in coefficients called as the Mel Frequency Cepstrum Coefficients. The mathematical representations for this conversion are given as follows:

$$\text{Log } Y = c (d) \cos (((2n+1) d \pi)/2n)$$

Where f (Mel) is Mel-frequency, Y is MFCC, D is the number of samples, n is the number of frames.

2. Feature Matching

The feature excerption process is followed by the feature matching phase. In feature matching, the comparison of MFCC coefficients of both speech and voice samples with the coefficients stored in the database obtained from feature excerption occurs. In this system for natural language, feature matching is performed under two different situations namely Login and Sign up which are very much necessary for ensuring the authenticity of the users. There are different feature matching techniques such as Hidden Markov Model (HMM), Neural Networks Vector, quantization and Dynamic Time Warping (DTW). The speech based system for natural language employs feature matching using the concept of DTW.

2.1 Dynamic Time Warping (DTW)

DTW is an algorithm that focuses on matching two sequences of feature vectors by repetitively shrinking or expanding the time axis till an exact match is obtained between the two sequences. It is generally used to calculate the distance between the two time series that vary in time [13].

A real time application of DTW in the voice recognition is that, it should be able to recognize the user’s voice even when spoken at different speeds. In order to check the similarity between two voice signals or the time series are warped non-linearly [5] .In other words we can say DTW is an optimal algorithm that looks for the similarity between two signals i.e., similar patterns. When the time series are wrapped, the time series or the signals are either "stretched" to match with the template available in the database when the speaker speaks fast or "shrunk" when the user speaks slowly since even with a small shift of the signal points leads to incorrect identification.

The similarity between the signals are identified by calculating distance or the difference between the each pair of the Mel-Frequency Cepstral Coefficient features extracted in the Feature Excerption phase of this system.

Consider two time series of feature vector calculated as,

G $\{g_1, g_2 \ldots g_n\}$ and K $\{k_1, k_2 \ldots k_n\}$

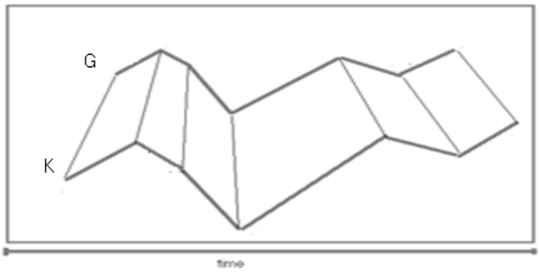


Figure 1.3 Warping of two time series

The figure 1.3 shows the two time series that is being wrapped non-linearly. In the time series G and K, the time series K seems to have taken a longer time, that is the user have spoken slowly. Each Vertical line of the same distance connects between the corresponding points in the time series G $\{g_1, g_2 \ldots g_n\}$ and K $\{k_1, k_2 \ldots k_n\}$. The line would have been straight vertical lines if the time series are identical ad also no warping is necessary [4]. This algorithm calculates the similarity by using the Euclidean's distance formula as follows,

$d\ (g,k) = |g-k| = [(g_1-k_1)^2+(g_2-k_2)^2+\ldots..+(g_n-k_n)^2]^{1/2}$

Mathematically, the two series are aligned by using an n-by-m matrix where the f^th and s^th element contains the distance d (g_f, k_s). [5] .The warping distance is calculated by,

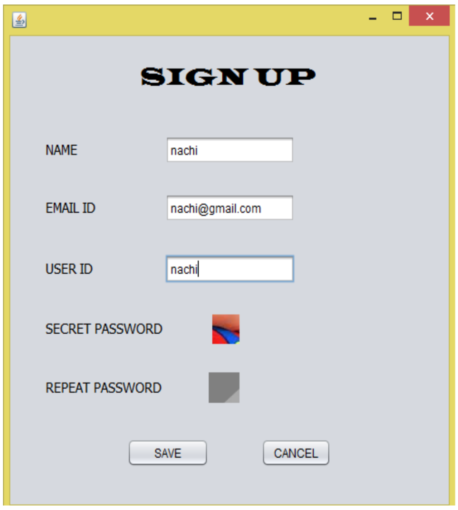
$D[f, s] = \min\ [D(f-1,s-1),D(f-1,s),D(f,s-1)]+d(f, s)$

It calculates the warping path which is the measure of the differences between the two voice samples. The two identical sequences after the warping will have the distance of zero when the voice signals are extracted from an environment, the distance will not be accurate but will be approximate.

In the proposed authentication system the feature matching phase is used in two ways such as when the users sign up into the account for the first time and when the users login to the account each time. When the users sign up the system requires the user to repeat the password twice and the system checks for the matching of these two samples to ensure that the password is correct and then it is again compared with the reference templates that is already stored in the dictionary for Tamil words to identify that word. After the speech recognition, the MFCC features calculated for this user is stored in the database. During each login process, the voice sample is recorded and is compared with the reference template created for the user at the time of sign up.

III. EXPERIMENTAL RESULTS

The results obtained by experimenting both MFCC and DTW in this proposed speech based authentication system for natural language are used for two functions namely sign up for the new users and login for the existing users. Sign up is generally a training session to train the system and create template for the new user.


SIGN UP

NAME: nachi

EMAIL ID: nachi@gmail.com

USER ID: nachi

Figure 1.4 Sign up page

In the sign up session, the user is made to create their password and to repeat it for confirmation. After the confirmation the average of the MFCC is taken and is compared with the features of Tamil words already stored in the dictionary to identify the corresponding word.

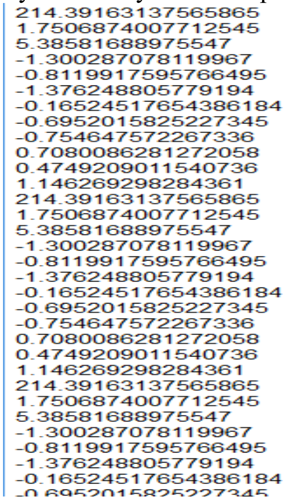
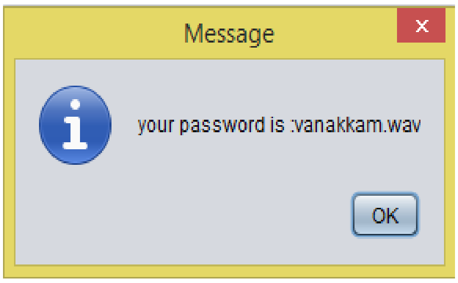
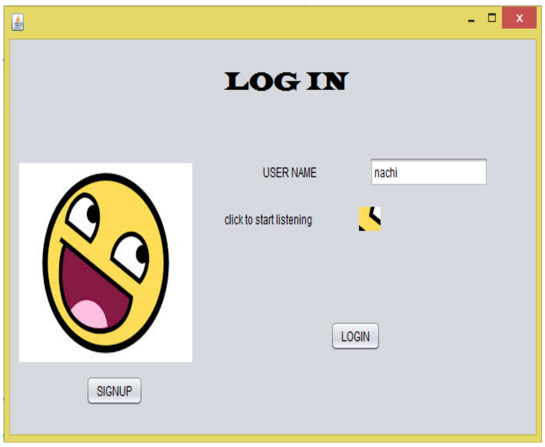

```
214.39163137565865
1.7506874007712545
5.38581688975547
-1.300287078119967
-0.8119917595766495
-1.376248805779194
-0.16524517654386184
-0.6952015825227345
-0.754647572267336
0.7080086281272058
0.4749209011540736
1.146269298284361
214.39163137565865
1.7506874007712545
5.38581688975547
-1.300287078119967
-0.8119917595766495
-1.376248805779194
-0.16524517654386184
-0.6952015825227345
-0.754647572267336
0.7080086281272058
0.4749209011540736
1.146269298284361
214.39163137565865
1.7506874007712545
5.38581688975547
-1.300287078119967
-0.8119917595766495
-1.376248805779194
-0.16524517654386184
-0.6952015825227345
```

Figure1.5 (a) MFCC of the word "vanakkam"


Message

your password is :vanakkam.wav

Figure1.5 (b) identified word is displayed in the signup session



LOG IN

USER NAME nachi

Figure 1.6 Login page

In the login session, when the users login with their voice password the feature excerption phase extracts the features which is then compared with corresponding users' MFCC features stored in the database during the sign up session. Login session is actually a testing phase. The DTW algorithm used in feature matching phase compares both the MFCC features and calculates the warping distance between them. The small warping distance ensures the authenticity of the users.

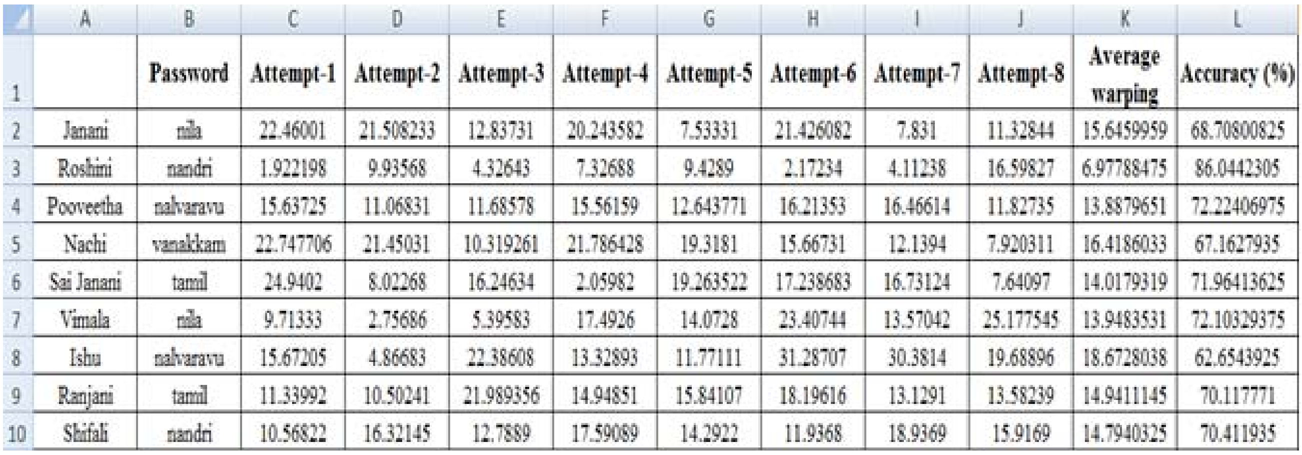
When the user's voice is matched with the reference template the user can successfully login to their account.

IV. ANALYSIS

This paper describes about the speech-based voice authentication system that identifies the Tamil words and the voice of the users. It employs secure system for natural language Tamil. An analysis when the Tamil words 'Vanakkam', 'Nandri', 'Tamil', 'Nila' and 'Nalvaravu' are spoken by different users.

Comparative Analysis

A comparative study on the Tamil words spoken by different users is represented using a bar chart to depict the warping distances for each word in the following figures. The password match is found to be exact when the warping distance is small.



	A	B	C	D	E	F	G	H	I	J	K	L
1		Password	Attempt-1	Attempt-2	Attempt-3	Attempt-4	Attempt-5	Attempt-6	Attempt-7	Attempt-8	Average warping	Accuracy (%)
2	Janani	nila	22.46001	21.508233	12.83731	20.243582	7.53331	21.426082	7.831	11.32844	15.6459959	68.70800825
3	Roshini	nandri	1.922198	9.93568	4.32643	7.32688	9.4289	2.17234	4.11238	16.59827	6.97788475	86.0442305
4	Pooveetha	nalvaravu	15.63725	11.06831	11.68578	15.56159	12.643771	16.21353	16.46614	11.82735	13.8879651	72.22406975
5	Nachi	vanakkam	22.747706	21.45031	10.319261	21.786428	19.3181	15.66731	12.1394	7.920311	16.4186033	67.1627935
6	Sai Janani	tamil	24.9402	8.02268	16.24634	2.05982	19.263522	17.238683	16.73124	7.64097	14.0179319	71.96413625
7	Vimala	nila	9.71333	2.75686	5.39583	17.4926	14.0728	23.40744	13.57042	25.177545	13.9483531	72.10329375
8	Ishu	nalvaravu	15.67205	4.86683	22.38608	13.32893	11.77111	31.28707	30.3814	19.68896	18.6728038	62.6543925
9	Ranjani	tamil	11.33992	10.50241	21.989356	14.94851	15.84107	18.19616	13.1291	13.58239	14.9411145	70.117771
10	Shifali	nandri	10.56822	16.32145	12.7889	17.59089	14.2922	11.9368	18.9369	15.9169	14.7940325	70.411935

Figure 1.7 Warping Distance obtained for each user

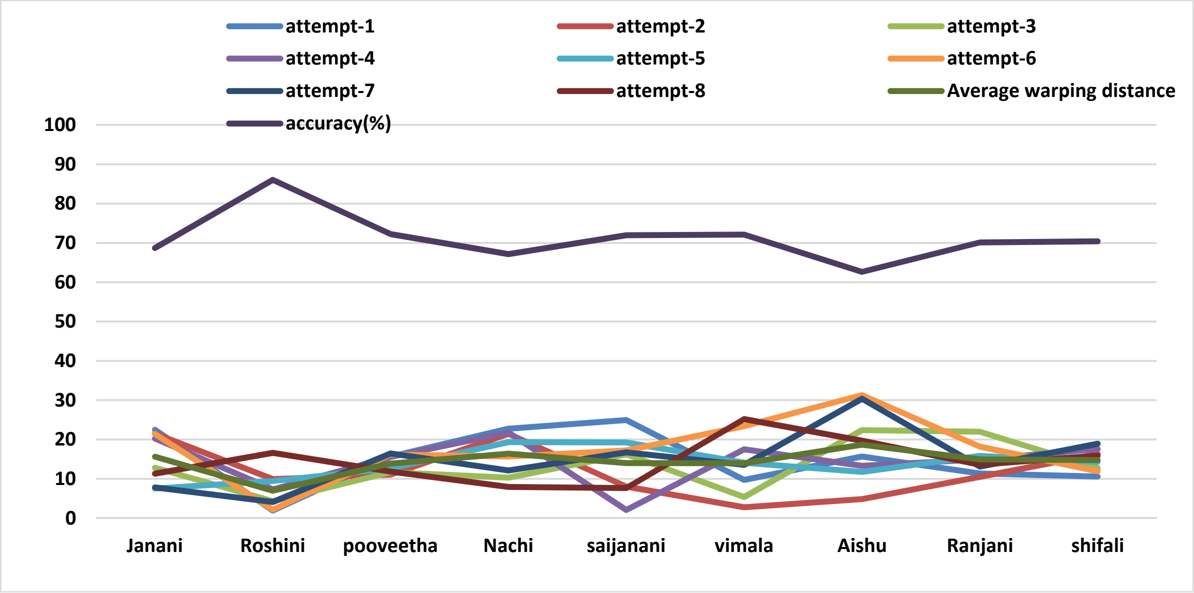


Figure 1.8 Analysis of the system accuracy

V. CONCLUSION

This paper explains about the speech based voice authentication system for Tamil language that has two major phases, feature excerption phase and feature matching phase. In the feature excerption phase, the system extracts the MFCC features from the voice input sample. In the feature matching module it identifies the Tamil words and the user using the Dynamic Time Warping algorithm that computes the warping distance between two time sequences. The two time series is similar when the warping distance between them is very small. Thus the system creates the voice password using Tamil words.

REFERENCES

[1] Stephen J. Wright, Dimitri Kanevsky, LiDeng, Xiaodong He, Georg Heigold and Haizhou Li (2013); "*Optimization Algorithms and Applications for Speech and Language Processing*", IEEE Transactions on Audio, Speech and Language Processing, 1558-7916.

[2] Shivanker Dev Dhingra**,** Geeta Nijhawan, Poonam Pandit (2013); "*Isolated Speech Recognition using MFCC and DTW*", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2278-8875.

[3] Dharun V S, Karnan M, "*Voice and Speech Recognition for Tamil words and Numerals*", International Journal of Modern Engineering Research (IJMER) Vol.2, Issue.5, 3406-3414, ISSN: 2249-6645.

[4] Lindasalwa Muda, Mumtaj Begam and Elamvazuthi I (2010); "*Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques*", Journal of Computing, Volume 2, Issue 3, ISSN 2151-9617.

[5] Anjali Bala, Abhijeet Kumar, Nidhika Birla (2010); "*Voice Command Recognition System Based on MFCC and DTW*", International Journal of Engineering Science and Technology Vol. 2 (12), 7335-7342.

[6] Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk (2012); "*Speech Recognition using MFCC*", International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) Pattaya (Thailand)

[7] Gold B and Morgan N (2000); Speech and Audio Signal Processing, John Wiley and Sons, New York, NY.

[8] Becchetti C and Lucio Prina Ricotti (1999); Speech Recognition, John Wiley and Sons, England.

[9] Karpov E (2003); "Real Time Speaker Identification", Master`s thesis, Department of Computer Science, University of Joensuu.

[10] Vibha Tiwari (2010); "*MFCC and its applications in speaker recognition*", Gyan Ganga Institute of Technology and Management, Bhopal, (MP) INDIA.

[11] Shumaila Iqbal, Tahira Mahboob and Malik Sikandar Hayat Khiyal (2011); "*Voice Recognition using HMM with MFCC for Secure ATM*", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, 2011 ISSN (Online): 1694-0814, ISSN (Online): 1694-0814.

[12] Sreejith C , Reghuraj P C, " *Isolated Spoken Word Identification in Malayalam using Mel-frequency Cepstral Coefficients and K-means clustering*", International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064 .

[13] Hao Feng, "*A Cryptosystem with Private Key Generation from Dynamic Properties of Human Hand Signature*", Nanyang Technological University.