Voice Recognition using Dynamic Time Warping and Mel-Frequency Cepstral Coefficients Algorithms

Abdelmajid H. Mansour
Department of IT
Faculty of Computers & IT
University of Jeddah
Khulais, Saudi Arabia

Gafar Zen Alabdeen Salh
Department of IT
Faculty of Computers & IT
University of Jeddah
Khulais, Saudi Arabia

Khalid A. Mohammed
Department of IT
Faculty of CS & IT
Alneelain University
Khartoum, Sudan

ABSTRACT
Voice recognition is an important and active research area of the recent years. This research aims to build a system for voice recognition using dynamic time wrapping algorithm, by comparing the voice signal of the speaker with pre-stored voice signals in the database, and extracting the main features of the speaker voice signal using Mel-frequency cepstral coefficients, which is one of the most important factors in achieving high recognition accuracy.

General Terms
Dynamic time wrapping "DTW" algorithm, Mel-frequency Cepstral Coefficients "MFCC" algorithm, vocal signal.

Keywords
Voice Recognition, Feature Extraction, Feature matching, voice signal.

1. INTRODUCTION
Voice recognition is one of the terms of biometric technology. It uses to provide any authentication to any system on the basis of acoustic features of voice instead of images. The behavioral aspect of human voice is used for identification by converting a spoken phrase from analog to digital format, and extracting unique vocal characteristics, such as pitch, frequency, tone and cadence to establish a speaker model or voice sample. In voice recognition, enrollment and verification processes are involved. Enrollment process describes the registration of speaker by training his voice features [1].

Voice is also physiological trait because every person has different pitch, but voice recognition is mainly based on the study of the way a person speaks, commonly classified as behavioral. Speaker verification focuses on the vocal characteristics that produce speech and not on the sound or the pronunciation of speech itself. The vocal characteristics depend on the dimensions of the vocal tract, mouth, nasal cavities and the other speech processing mechanism of the human body [2].

Voice-biometrics systems can be categorized as belonging in two industries: speech processing and biometric security. This dual parentage has strongly influenced how voice-biometrics tools operate in the real world. Speech processing. Like other speech-processing tools, voice biometrics extract information from the stream of speech to accomplish their work. They can be configured to operate on many of the same acoustic parameters as their closest speech-processing relative speech recognition [3].

Voice recognition has two categories text dependent and text independent. Text dependent voice recognition identifies the speaker against the phrase that was given to him at the time of enrollment. Text independent voice recognition identifies the speaker irrespective of what he is saying. This method is very often use in voice recognition as it require very little computations but need more cooperation of speakers. In this case the text in verification phase is different than in training or enrolment phase [1].

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 [4].

2. 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. 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. 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. [4].

3. MEL FREQUENCY CEPSTRAL COEFFICIENTS "MFCC"
MFCC is used to extract the unique features of human voice. It represents the short term power spectrum of human voice. It is used to calculate the coefficients that represent the frequency Cepstral these coefficients are based on the linear cosine transform of the log power spectrum on the nonlinear Mel scale of frequency. In Mel scale the frequency bands are equally spaced that approximates the human voice more accurate [1]. Equation (1) is used to convert the normal frequency to the Mel scale the formula is used as

$$m=2595 \log_{10}(1+f/700) \text{ -------- } (1)$$

Mel scale and normal frequency scale is referenced by defining the pitch of 1000 Mel to a 1000 Hz tones, 40 db above the listener's threshold. Mel frequency are equally spaced on the Mel scale and are applied to linear space filters below 1000 Hz to linearized the Mel scale values and logarithmically spaced filter above 1000 Hz to find the log power of Mel scaled signal. Mel frequency wrapping is the better representation of voice. Voice features are represented in MFCC by dividing the voice signal into frames and windowing them then taking the Fourier transform of a windowing signal. Mel scale frequencies are obtained by applying the Mel filter or triangular band pass filter to the transformed signal [1].

4. RELATED WORKS

Voice Recognition is an emerging scope of security and authentication for the future, there are numerous studies and researches on this area. The concept of observing voice sample with MFCC for extracting acoustic features and then used to trained HMM parameters through forward backward algorithm which lies under HMM and finally the computed log likelihood from training is stored to database. It will recognize the speaker by comparing the log value from the database against the PIN code, were proposed by Shumaila Iqbal, Tahira Mahboob and Malik Sikandar Hayat Khiyal [1]. Presenting the viability of MFCC to extract features and DTW to compare the test patterns were proposed by Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi [5]. Sahil Verma, Tarun Gulati, Rohit Lamba were proposed the focuses on recognising voice corresponding to English alphabets using Mel-frequency cepstral coefficients (MFCCs) and Dynamic Time Warping (DTW) introduced by Sakoe Chiba [6]. Nidhi Desai, Prof.Kinnal Dhameliya, Prof. Vijayendra Desai were proposed an approach to recognize English words corresponding to control Robot in an isolated way by different male and female speakers. The aim is to focuses on recognizing voice using Melfrequency cepstral coefficients (MFCCs) and Dynamic Time Warping (DTW) [7]. Md. Akkas Ali, Manwar Hossain and Mohammad Nuruzzaman, were presented some technique for recognizing spoken words in Bangla. In this work we use MFCC, LPC, GMM and DTW [8]. MarutiLimkar, RamaRao & VidyaSagvekar were proposed an approach to recognize spoken English words corresponding to digits zero to nine in an isolated way by different male and female speakers [9]. The 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 were proposed by Dr. Kavitha. R, Nachammai. N, Ranjani. R, Shifali. J [11]. Ms. Savitha and S Upadhya were presented the two template matching techniques namely the Single and Average template matching techniques that were developed to recognize the English digits spoken in isolation. The algorithm was implemented both for speaker dependent and speaker independent type of isolated digit recognition and a comparison of recognition accuracies for the two template matching techniques was made [12]. An overview of major technological perspective and appreciation of the fundamental progress of speech recognition and gives overview technique developed in each stage of speech recognition and also summarize and compare different speech recognition systems and identify research topics and applications which are at the forefront of this exciting and challenging field were presented by Om Prakash Prabhakar, Navneet Kumar Sahu [13].

5. PROPOSED SCHEME

This paper build a system for voice recognition using dynamic time wrapping algorithm, and comparing the entered voice signal of the speaker with pre-stored voice signals in the database for the purpose of verifying. By using good statistical method for the process of comparing. Then extracting the main features of the speaker voice signal by using Mel-Frequency Cepstral Coefficients, which is one of the most important factors in achieving high recognition accuracy. In order to solve the problem of extracting the components and features of the voice signals that entered into the computers and performing the comparison to getting the best results for maintaining the confidentiality, security and integrity of the information. The extraction of the feature done by creating source for each digital voice from a set of vocabulary that forming the sound database. Which it is a voice signal for the voice called source signal. Where each signal is divided into blocks of equal length samples from beginning to end. Then each template converted to vector attributes that extract the signal features in that template. These include vector in groups are called Features Vectors. This processing repeated for each digital voice in the vocabulary set. The Features matching is called "recognition process", in this process the coming signal that to be recognized is transformed into a series of features vectors by using the conversion (begin – end), for processing of features extraction. These feature will be compared with all possible probability exist on the database by using pattern matching method. To give the recognition decision from matching quality by using Euclidean Distance between two series of features vectors, which one representing feature vectors of the source signal and the other feature vectors of the test signal. The proposed system moves through two phases:

5.1 Training Phase

At this stage the system will trained by creating training groups consists of different sounds samples, from which the system can create its own sound database by selecting samples with more accuracy and purity, as shown in the Fig 1.

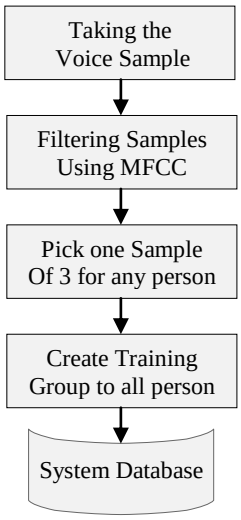

Fig 1: Training Phase

For the purpose of training, was taken 5 voice samples from every person, to be compared with all samples of the trainee person. Then taking the sample that have highest match with more purity, and store it on the system voice database. They selected models for 3 person from the training groups. The results of training process as shown in the following Tables.

Table 1. Training of word "Khaled Ahmed"

Sample	Matching Rate (%)					Average Rate
	1	2	3	4	5	
1	100	94	91	95	93	94.6
2	94	100	96	93	95	95.6
3	91	96	100	97	94	95.6
4	95	95	97	100	93	96
5	93	95	94	93	100	95

Table 1 above the model number 4 achieved the highest rate of matching with 96%. So have been selected within the system database to represent the word "Khaled Ahmed".

Table 2. Training of the word "Abdel Rahim"

Sample	Matching Rate (%)					Average Rate
	1	2	3	4	5	
1	100	94	97	95	95	96.2
2	94	100	96	93	95	95.6
3	91	96	100	98	94	95.8
4	95	95	97	100	93	96
5	93	95	94	93	100	95

Table 2 above the model number 1 achieved the highest rate of matching with 96.2%. So have been selected within the system database to represent the word "Abdel Rahim".

Table 3. Training of word "Esraa"

Sample	Matching Rate (%)					Average Rate
	1	2	3	4	5	
1	100	94	91	95	93	94.6
2	94	100	96	93	92	95
3	91	93	100	97	94	95
4	95	95	97	100	93	96
5	96	95	98	97	100	97.2

Table 3 above the model number 5 achieved the highest rate of matching with 97.2%. So have been selected within the system database to represent the word "Esraa".

5.2 Testing Phase

At this stage the system will tested for recognizing the voice of the speaker after the matching process with samples taken from training stage. The system can take a decision whether the voice exist on the Database or not. As shown in the Fig 2.

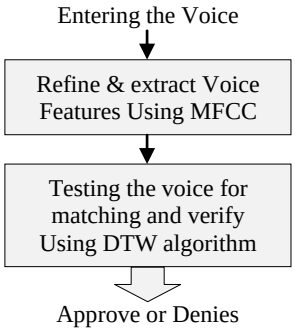

Fig 2: Testing Phase

For the purpose of testing recognition in the proposed scheme, they have been recorded 5 words sound for 20 different person and stored in 16 bit format with rate 11025 Hz and extension ".wav".

The model of recording the word "Khaled Ahmed" 5 times and selecting the best one to store it on the system voice database. This process repeated to all required samples to be stored, as shown in the following Figures.

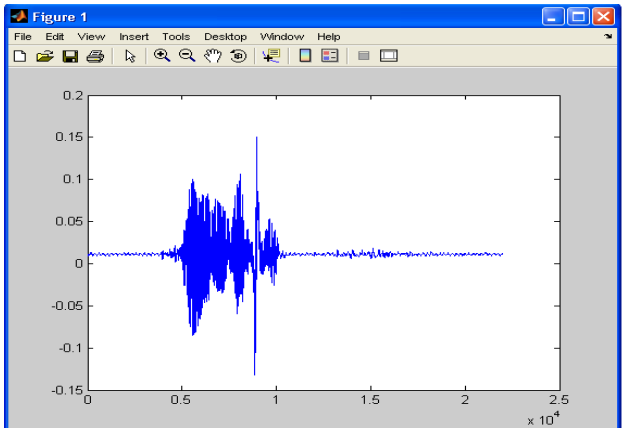

Fig 3: 1st Sample Model for Word "Khaled Ahmed"

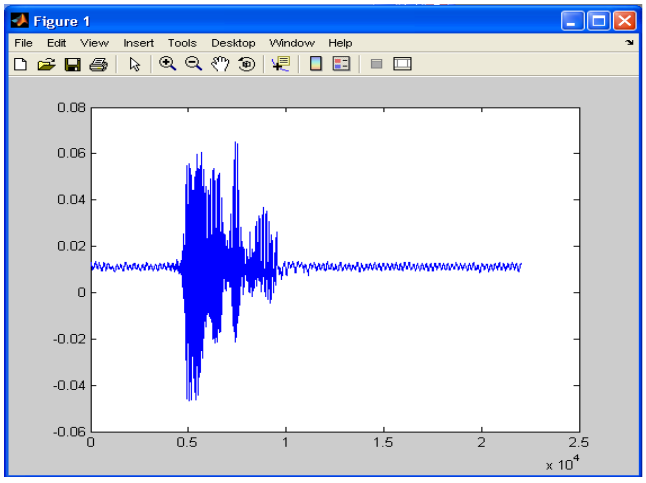

Fig 4: 2nd Sample Model for Word "Khaled Ahmed"

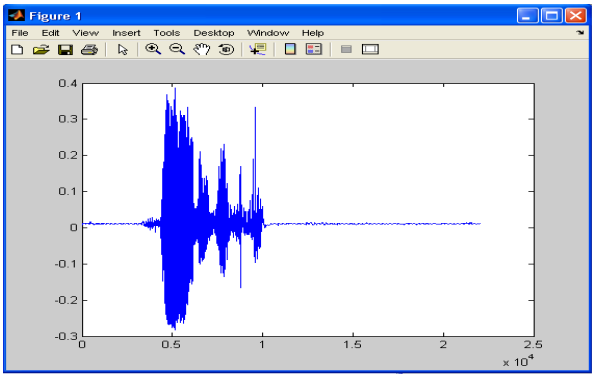

Fig 5: 3rd Sample Model for Word "Khaled Ahmed"

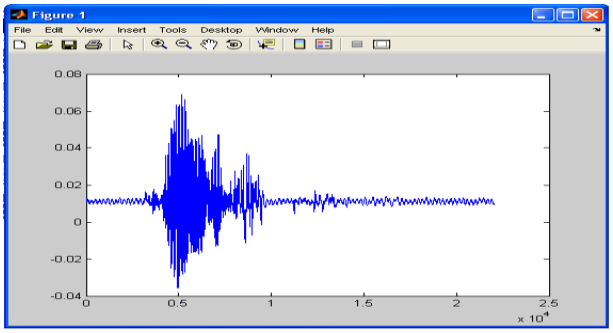

Fig 6: 4th Sample Model for Word "Khaled Ahmed"

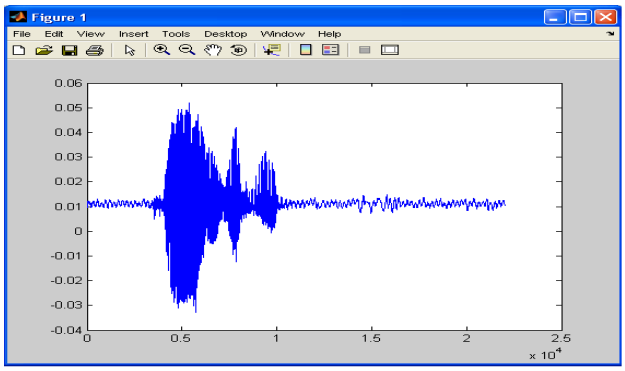

Fig 7: 5th Sample Model for Word "Khaled Ahmed"

The model of recorded voice for the word "Khaled Ahmed", at the process of performing feature extraction by using MFCC technology as shown in the Fig 8.

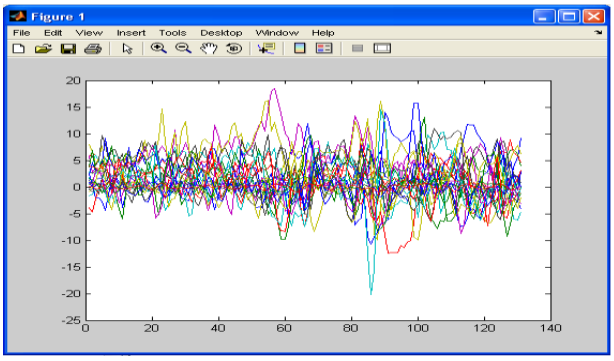

Fig 8: Processing Feature Extraction of word "Khaled Ahmed" using MFCC

The model of recorded voice for the word "Khaled Ahmed" after the process of performing feature extraction by using DTW technology, as shown in the Fig 9.

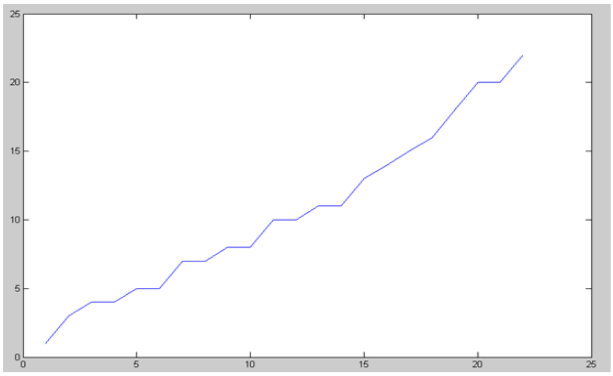

Fig 9: After Processing Feature Extraction of word "Khaled Ahmed" using DTW

6. VERFICATION USING DTW

This method uses distance measure to find the nonlinear matching between test signal and all source signal. Then selecting the source has less distance, and the voice that represented by that source, refers to the result of recognition. The system has been tested 3 times on 20 samples which represent the system database, and the verification result is in range from 0 to 1.

6.1 Result of the Test1

On this phase were selected the voice signal "AHMED" and compared it with prerecorded samples on the system, the result as shown in Table 4.

Table 4. Test1 of the word "AHMED"

Speaker	Rate
Khalid	0.62
Omer	0.70
Ahmed	0.94
Ali	0.66
Sara	0.70
Sahar	0.49
Ithar	0.55
Abdulrahim	0.72
Mohammed	0.66
Salah	0.71
Abdelmanie	0.64
Abdallah	0.76
Eltayeb	0.82
Soaad	0.52
Madina	0.46

Mahmoud	0.71
Haytham	0.64
Mohanad	0.77
Osman	0.61
Amna	0.61

The following Fig 10, show the result of Test 1.

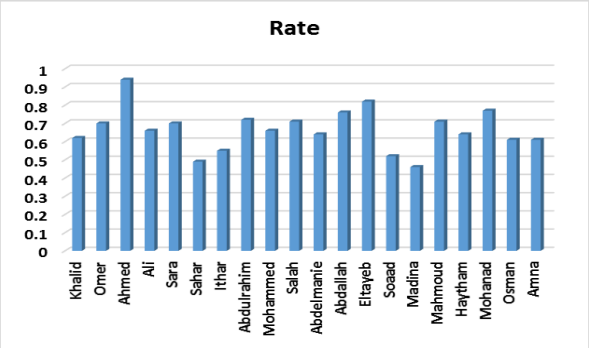

Fig 10: Test1 of the word "AHMED"

6.2 Result of the Test2

On this phase were selected the voice signal "SALAH" and compared it with prerecorded samples on the system, the result as shown in Table 5.

Table 5. Test2 of the word "SALAH"

Speaker	Rate
Khalid	0.48
Omer	0.53
Ahmed	0.57
Ali	0.36
Sara	0.80
Sahar	0.83
Ithar	0.37
Mohammed	0.34
Salah	0.93
Abdelmanie	0.30
Abdulrahim	0.26
Abdallah	0.39
Eltayeb	0.40
Soaad	0.89
Madina	0.31
Mahmoud	0.46
Haytham	0.50
Mohanad	0.44
Osman	0.76
Amna	0.30

The following Fig 11, show the result of Test 2.

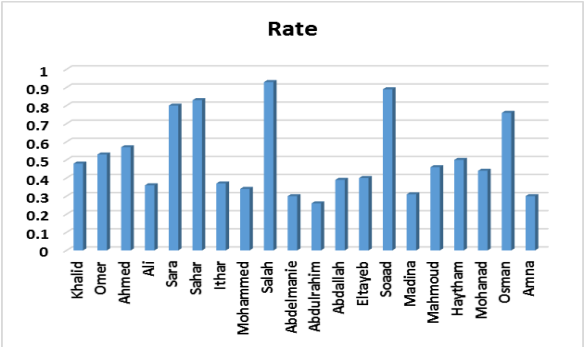

Fig 11: Test2 of the word "SALAH"

6.3 Result of the Test3:

On this phase were selected the voice signal "MOHAMMED" and compared it with prerecorded samples on the system, the result as shown in table 6.

Table 6. Test3 of the word "MOHAMMED"

Speaker	Rate
Khalid	0.30
Omer	0.41
Ahmed	0.81
Ali	0.41
Sara	0.36
Sahar	0.29
Ithar	0.45
Abdulrahim	0.52
Mohammed	0.96
Salah	0.33
Abdelmanie	0.29

Abdallah	0.51
Eltayeb	0.36
Soaad	0.54
Madina	0.31
Mahmoud	0.90
Haytham	0.50
Mohanad	0.88
Osman	0.42
Amna	0.31

The following Fig 12, show the result of Test 3.

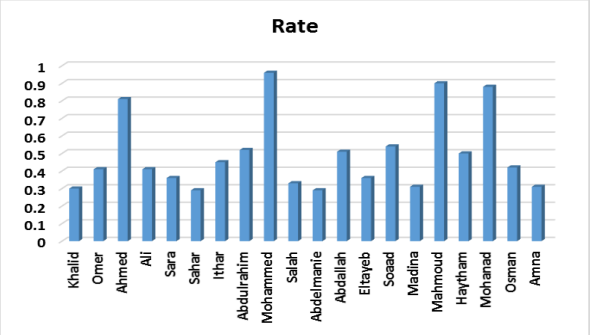

Fig 12: Test3 of the word "MOHAMMED"

6.4 Result of the Test4:

On this phase were selected the voice signal "OSMAN" and compared it with prerecorded samples on the system, the result as shown in table 7.

Table 7. Test4 of the word "OSMAN"

Speaker	Rate
Khalid	0.26
Omer	0.89
Ahmed	0.33
Ali	0.57
Sara	0.47
Sahar	0.44
Ithar	0.53
Abdulrahim	0.37
Mohammed	0.61
Salah	0.41
Abdelmanie	0.22
Abdallah	0.35
Eltayeb	0.58
Soaad	0.39
Madina	0.51
Mahmoud	0.61
Haytham	0.49
Mohanad	0.43
Osman	0.91
Amna	0.66

The following Fig 13, show the result of Test 4.

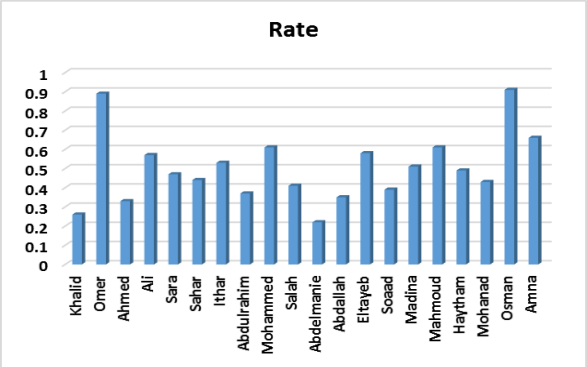

Fig 13: Test4 of the word "OSMAN"

6.5 Result of the Test5:

On this phase were selected the voice signal "HAYTHAM" and compared it with prerecorded samples on the system, the result as shown in table 8.

Table 8. Test5 of the word "HAYTHAM"

Speaker	Rate
Khalid	0.65
Omer	0.42
Ahmed	0.39
Ali	0.41
Sara	0.25
Sahar	0.29
Ithar	0.39

Abdulrahim	0.76
Mohammed	0.71
Salah	0.44
Abdelmanie	0.35
Abdallah	0.74
Eltayeb	0.66
Soaad	0.23
Madina	0.79
Mahmoud	0.32
Haytham	0.98
Mohanad	0.56
Osman	0.78
Amna	0.59

The following Fig 14, show the result of Test 5.

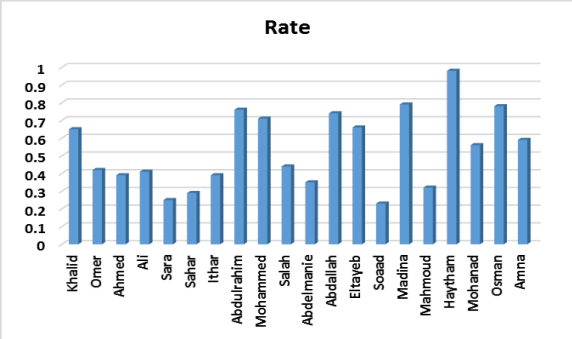

Fig 14: Test5 of the word "HAYTHAM"

7. CONCLUSION

The performance of recognition systems can be determined by voice sources including the speaker voice specification, the rate of voice issuance, delivery source, and recording media. The accuracy of recognition come from the behavior of speaker in issuing the voice. From the tests and training above we found that the DTW is flexible mathematical method, they gives high accuracy results. The performance can be improved by selecting the sources carefully, which have significant role in influencing the accuracy of recognition.

The signal analysis by using MFCC provide spectrum factors which represents the exact vocal system for stored words. MFCC provide a high level of perception of the human voice, where they work to remove all unimportant information, then give a better representation of the signal, which leads to a higher resolution in the performance of recognition.

The conclusion of this paper is the spectrum factors that have high-specification show their importance depending on the speaker himself and the method of producing the voice and vocal pronunciation style which can be used in many applications such as security systems, where the voices of the people are different as fingerprints differ.

8. REFERENCES

[1] Shumaila Iqbal, Tahira Mahboob and Malik Sikandar Hayat Khiyal, "Voice Recognition using HMM with MFCC for Secure ATM", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011, ISSN (Online): 1694-0814, pp. 297-303.

[2] Debnath Bhattacharyya, Rahul Ranjan, Farkhod Alisherov A. and Minkyu Choi, "Biometric Authentication: A Review", International Journal of u- and e- Service, Science and Technology Vol. 2, No. 3, September, 2009.

[3] Judith A. Markowitz, "Voice Biometrics", September 2000/Vol. 43, No. 9 Communications of the ACM.

[4] Dr. Kavitha. R, Nachammai. N, Ranjani. R, Shifali. J," Speech Based Voice Recognition System for Natural Language Processing", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014, ISSN: 0975-9646, pp. 5301-5305.

[5] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi," Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, MARCH 2010, ISSN 2151-9617, pp. 138 -143.

[6] Sahil Verma, Tarun Gulati, Rohit Lamba," RECOGNIZING VOICE FOR NUMERICS USING MFCC AND DTW" International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 5, May 2013 ISSN 2319 – 4847, pp. 127 -130.

[7] Nidhi Desai 1, Prof.Kinnal Dhameliya2, Prof. Vijayendra Desai, "Recognizing voice commands for robot using MFCC and DTW", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 5, May 2014, ISSN (Online), 2278-1021, pp. 6456 6459.

[8] Md. Akkas Ali, Manwar Hossain and Mohammad Nuruzzaman Bhuiyan," Automatic Speech Recognition Technique for Bangla Words", International Journal of Advanced Science and Technology Vol. 50, January, 2013, pp. 51-60.

[9] MarutiLimkar, RamaRao & VidyaSagvekar, "Isolated Digit Recognition Using MFCC AND DTW", International Journal on Advanced Electrical and Electronics Engineering, (IJAEEE), ISSN (Print): 2278-8948, Volume-1, Issue-1, 2012, pp. 59-64.

[10] R. Bellman and S. Dreyfus, "Applied Dynamic Programming", Princeton, NJ, Princeton University Press, 1962.

[11] M.A.Anusuya, S.K.Katti, Classification Techniques used in Speech Recognition Applications: A Review -– Int. J. Comp. Tech. Appl., July-August 2011 -Vol 2 (4), 910-954.

[12] Ms. Savitha and S Upadhya," Digit Recognizer Using Single and Average Template Matching Techniques", International Journal of Emerging Technologies in

Computational and Applied Sciences, 3(3), Dec.12-Feb.13, ISSN (Online): 2279-0055, pp. 357-362

[13] Om Prakash Prabhakar, Navneet Kumar Sahu," A Survey On: Voice Command Recognition Technique", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013 ISSN: 2277 128X, pp. 576-585.

9. AUTHOR'S PROFILE

Abdelmajid Hassan Mansour Emam, Assistant Professor, Department of Computers and Information Technology, University of Jeddah, Faculty of Computers and Information Technology, Khulais, Jeddah, Saudi Arabia.

Permanent Address: Department of Information Technology, Faculty of computer Science and Information Technology, Alneelain University, Khartoum, Sudan.

Gafar Zen Alabdeen Salh Hassan, Assistant Professor, Department of Computers and Information Technology, University of Jeddah, Faculty of Computers and Information Technology, Khulais, Jeddah, Saudi Arabia..

Permanent Address: Department of Information Technology, Faculty of computer Science and Information Technology, Alneelain University, Khartoum, Sudan.

Khalid Ahmed Mohammed Ismaiel, Lecturer, Department of Information Technology, Faculty of computer Science and Information Technology, Alneelain University, Khartoum, Sudan.