International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 1 Issue 11 (November 2014) www.ijirae.com

Comparative Analysis of MFCC, DTW&ANN for Arabic Speech Recognition

Bidoor Noori Ishaq
Research Student
Department of Computer Science & IT
Dr. B. A.M. University
Aurangabad, Maharashtra, India

Bharti W. Gawali
Professor
Department of Computer Science & IT
Dr. B.A.M. University
Aurangabad, Maharashtra, India

Abstract— This paper presents Arabic database and isolated Word recognition system based on Mel-frequency Cepstral coefficient (MFCC), Distance Time Warping (DTW) and Neural Network (NN) that investigates its performance in speech recognition. Arabic speech database has been designed by using the Computerized Speech laboratory (CSL) Lab. The database consists of the Arabic Character, digit, word and sentences. The size of database is 1590 .This paper presents the comparative recognition accuracy of DTW, MFCC and NN. The better recognition accuracy of about 90% was obtained with MFCC-based system

Keywords— MFCC, DTW, Neural Network, sampling rate, performance

I. INTRODUCTION

Speech recognition is the process of automatically recognizing the spoken words of person based on information in speech signal. Each spoken word is created using the phonetic combination of a set of vowel semivowel and consonant speech sound units. The most popular spectral based parameter used in recognition approach is the Mel Frequency Cepstral Coefficients called MFCC. MFCCs are coefficients, which represent audio, based on perception of human auditory systems [1]. Arabic is a Semitic language, and it is one of the oldest languages in the world. Currently it is the second language in terms of number of speakers [2]. Arabic is the first language in the Arab world, i.e., Saudi Arabia, Jordan, Oman, Yemen, Egypt, Syria, Lebanon, etc. Arabic alphabets are used in several languages, such as Persian and Urdu. Standard Arabic has basically 34 phonemes, of which six are vowels, and 28 are consonants [3]. A phoneme is the smallest element of speech units that indicates a difference in meaning, word, or sentence [4]. The Arabic language is considered nowadays as the fifth widely used language as there are more than 200 million people speak this language[5,6].

TABLE 1
THE PRONUNCIATION OF DIGIT IN ARABIC

Digit	Arabic writing	Pronunciation
1	واحد	Wahed
2	اثنين	Aathnay
3	ثلاثة	Thalathah
4	أربعة	Aarbaah
5	سة خ	Kaamsah
6	ستة	Settah
7	سبعة	Subaah
8	ثمانية	Thamaneyeh
9	تسعة	Tesah
0	صفر	Sefer

TABLE 2
THE COMPLETE PHONEME SET FOR DAILY LIFE

Write form	Pronunciation	Meaning
تلفاز	Telefaz	TV
كتب	Katab	he word
ذهب	Thahab	Gold
يتكلم	Yitkallim	he speaks
صيف	Saif	Summer
باب	Bab	Door
بنت	Bent	Girl
ولد	Wala	Boy

II. ARABIC SPEECH DATABASE

The database is volunteered by student from IRAQ of Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. For accuracy in the speech recognition; we need a collection of utterances. The generation of a corpus for Arabic database includes: characters, digits, words and sentences. The total number of speakers was 10 out of which 3 were Females and 7 were Males. The vocabulary size of the database consists of characters 840 samples, digit 300 samples, word 300 samples, sentences 150 samples.

A. *Acquisition setup*

To achieve a high audio quality, the recording took place in the room without noise sound and effect of echo. The Sampling frequency for all recordings was 11025 Hz in the Room temperature and normal humidity.

The speaker were seating in front of the direction of the microphone with the Distance of about 12-15 cm [7]. The speech data is collected with the help of Computerized speech laboratory (CSL) using the single channel. The CSL is most advanced analysis system for speech and voice. It is a complete hardware and software system with specifications and performance. It is an input/output recording device for a PC, which has special features for reliable for reliable acoustic measurements

III. FEATURE EXTRACTION TECHNIQUES

Transforming the input data into the set of features is called feature extraction. The features extracted are carefully chosen. It is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. Following are the most frequently used feature extraction techniques.

A. Mel Frequency Cepstral Coefficient(MFCC)

The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency [8-9]. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech.[10].The overall process of the MFCC is shown in Figure 1*:*

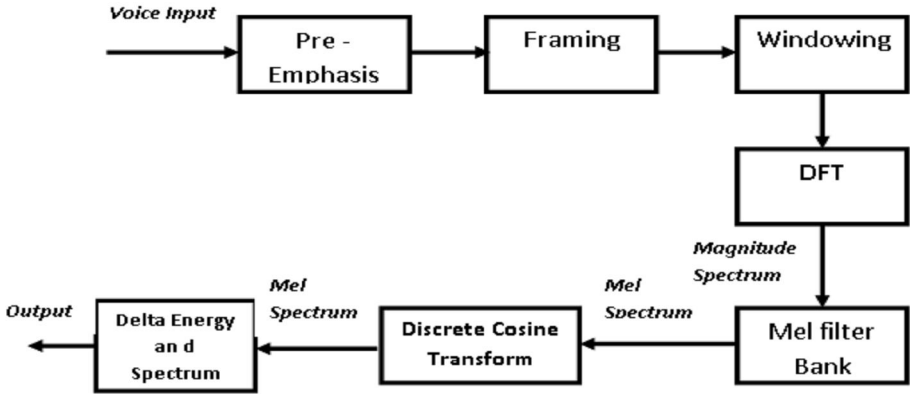

Fig. 1 MFCC Block Diagram

As shown in Figure 1, MFCC consists of seven computational steps. Each step has its function and mathematical. Figure 2 represents Extraction of MFCC Feature for a Frame.

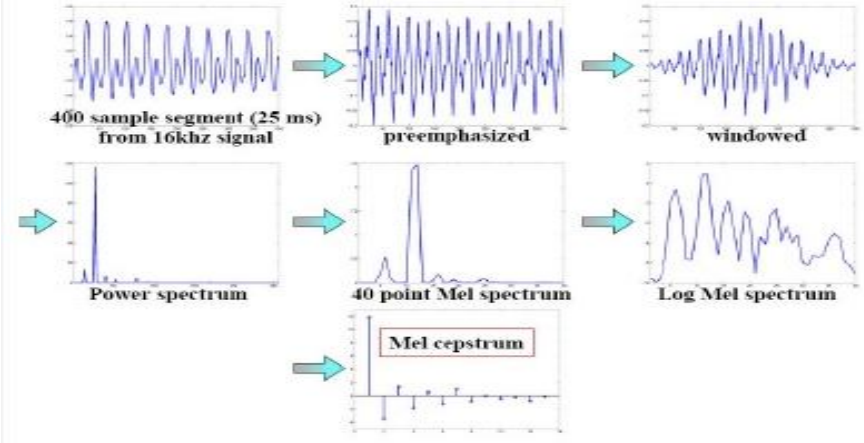

Figure 2. Extraction of MFCC Feature for a Frame

TABLE 3
DISTANCE MATRIX FOR MFCC COEFFICIENT FOR DIGITS

	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9
D0	0.0227	0.0482	0.0365	0.0239	0.0474	0.1346	0.2855	0.1206	0.0853	0.1397
D1	0.0482	0.0117	0.0355	0.0243	0.0956	0.1828	0.3337	0.1687	0.0371	0.1879
D2	0.0365	0.0646	0.0117	0.0127	0.084	0.1712	0.322	0.1571	0.0487	0.1762
D3	0.0239	0.0243	0.1866	0.0127	0.0713	0.1585	0.3094	0.1444	0.0614	0.1635

D4	0.178	0.0956	0.084	0.0713	0.0474	0.0872	0.2381	0.0731	0.1327	0.0922
D5	0.1346	0.1828	0.1712	0.1585	0.0872	0.011	0.1508	0.0141	0.2199	0.005
D6	0.2855	0.3337	0.322	0.3094	0.2381	0.1508	0.0532	0.1649	0.3707	0.1458
D7	0.1206	0.1687	0.1571	0.1444	0.0731	0.0172	0.1649	0.0141	0.2058	0.0191
D8	0.0853	0.1034	0.0487	0.0614	0.1327	0.2199	0.3707	0.2058	0.0371	0.2249
D9	0.1397	0.1879	0.1762	0.1635	0.0922	0.0678	0.1458	0.0191	0.2249	0.005

TABLE 4
DISTANCE MATRIX FOR MFCC COEFFICIENT FOR WORDS

	W1	**W2**	**W3**	**W4**	**W5**	**W6**	**W7**	**W8**	**W9**	**W10**
W1	0.0317	0.0415	0.3767	0.1681	0.2552	0.275	0.2023	0.2406	0.2036	0.2542
W2	0.0415	0.0098	0.3352	0.1266	0.2137	0.2335	0.1608	0.1991	0.0197	0.2127
W3	0.3767	0.3352	0.034	0.2086	0.1215	0.1017	0.1744	0.1361	0.345	0.1225
W4	0.1681	0.1266	0.2086	0.0252	0.0871	0.1069	0.0342	0.0725	0.1364	0.0861
W5	0.2552	0.2137	0.1215	0.0871	0	0.0198	0.0529	0.0146	0.2235	0.0484
W6	0.275	0.2335	0.1017	0.1069	0.3018	0.0198	0.0727	0.0345	0.2433	0.0208
W7	0.2023	0.1608	0.1744	0.0342	0.0529	0.0727	0.049	0.0383	0.1706	0.0519
W8	0.2406	0.1991	0.1361	0.0725	0.0146	0.0345	0.0383	0.0673	0.2089	0.2089
W9	0.0317	0.0106	0.345	0.1364	0.2235	0.2433	0.1706	0.2089	0.0098	0.2225
W10	0.2542	0.2127	0.1225	0.0861	0.0538	0.0208	0.0519	0.2089	0.2225	0

TABLE 5
DISTANCE MATRIX FOR MFCC COEFFICIENT FOR CHARACTERS

	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**
C1	0.1049	0.1804	0.2115	0.1481	0.5583	0.5485	0.1896	0.2972	0.3608	0.1729
C2	0.1804	0.0075	0.0755	0.0324	0.3779	0.3681	0.0091	0.1168	0.1804	0.1322
C3	0.1049	0.0755	0.0431	0.0458	0.4534	0.4436	0.0846	0.1923	0.2558	0.068
C4	0.1481	0.0324	0.0431	0.0012	0.4102	0.4005	0.0415	0.1491	0.2127	0.0248
C5	0.5583	0.3779	0.4534	0.4102	0.0098	0.3525	0.3687	0.2611	0.1975	0.3854
C6	0.5485	0.3681	0.4436	0.4005	0.2338	0.0098	0.359	0.2513	0.1878	0.3757
C7	0.1896	0.0332	0.0846	0.0415	0.3687	0.359	0.0091	0.1076	0.1712	0.0167
C8	0.2972	0.1168	0.1923	0.1491	0.2611	0.2513	0.1076	0.0634	0.0636	0.1243
C9	0.3608	0.1804	0.2558	0.2127	0.1975	0.1878	0.1712	0.0636	0.2386	0.1879
C10	0.1729	0.1461	0.068	0.0248	0.3854	0.3757	0.0167	0.1243	0.1879	0.0075

TABLE 6
DISTANCE MATRIX FOR MFCC COEFFICIENT FOR SENTENCES

	S1	**S2**	**S3**	**S4**	**S5**
S1	0.007	0.029	0.0079	0.0182	0.0406
S2	0.029	0.0211	0.0304	0.022	0.0696
S3	0.0079	0.0211	0	0.0584	0.0485
S4	0.007	0.022	0	0.0378	0.0475
S5	0.102	0.0696	0.0485	0.0475	0.0406

B. Dynamic Time Warping (DTW)

The time alignment of different utterances is the core problem for distance measurement in speech recognition. A small shift leads to incorrect identification. Dynamic Time Warping is an efficient method to solve the time alignment problem. DTW algorithm aims at aligning two sequences of feature vectors by warping the time axis repetitively until an optimal match between the two sequences is found. This algorithm performs a piece wise linear mapping of the time axis to align both the signals [11,12].

DTW algorithm is based on Dynamic Programming. This algorithm is used for measuring similarity between two time series which may vary in time or speed.[13] This technique is also used to find the optimal alignment between two times series if one time series may be “warped” non-linearly by stretching or shrinking it along its time axis. This warping between two time series can then be used to find corresponding regions between the two times, for this application is shown in Figure3.

A speech signal is represented by a series of feature vectors which are computed every 10ms. A whole word will comprise dozens of those vectors, and we know that the number of vectors (the duration) of a word will depend on how fast a person is speaking. In speech recognition, we have to classify not only single vectors, but sequences of vectors. Let’s assume we would want to recognize a few command words or digits [14,15].

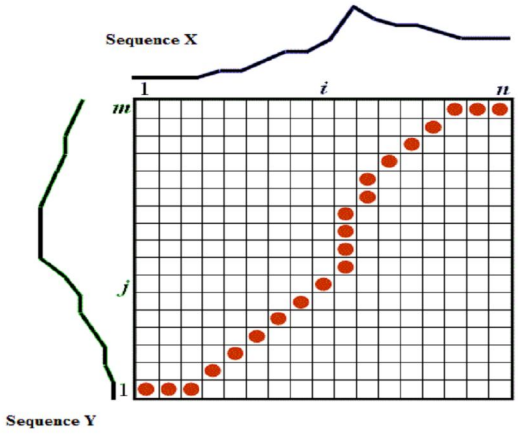

Fig.3 Distances between two segments.

TABLE 7
THE DISTANCE MATRIX OF ARABIC WORD FOR SAME SUBJECT USING DTW.

	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
W1	15.57	23.50	33.44	23.68	30.60	39.39	26.95	29.29	23.12	34.47
W2	23.50	10.83	34.85	20.72	25.01	33.85	27.53	29.39	15.55	22.58
W3	33.44	34.85	9.28	23.38	22.46	20.77	14.22	14.91	30.79	12.48
W4	23.68	20.72	23.38	7.08	17.49	25.06	18.13	17.89	19.17	20.89
W5	30.60	25.01	22.46	17.49	8.94	25.87	23.27	21.70	26.50	18.59
W6	39.39	33.85	20.77	25.06	25.87	21.60	23.62	23.96	32.790	18.91
W7	26.95	27.53	14.22	18.13	23.27	23.62	5.47	6.40	16.98	18.59
W8	29.29	29.39	14.91	17.89	21.70	23.96	6.72	7.33	21.12	17.43
W9	23.12	15.55	30.79	19.17	26.50	32.79	16.98	21.12	9.72	27.02
W10	34.47	22.58	12.48	20.89	18.59	18.91	18.59	17.43	27.02	5.74

TABLE 8
THE DISTANCE MATRIX OF ARABIC DIGIT FOR SAME SUBJECT USING DTW.

	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9
D0	9.77	30.12	22.70	22.14	22.87	21.37	18.53	17.38	22.35	18.737
D1	30.12	8.74	27.34	29.68	32.48	32.64	30.86	28.74	21.98	32.18
D2	22.70	27.34	15.14	20.56	24.68	25.41	25.55	24.38	14.57	27.33
D3	22.14	29.68	20.56	14.05	25.47	21.86	23.56	18.74	16.04	25.73
D4	22.87	32.48	24.68	25.47	10.27	25.24	26.81	15.08	23.98	20.64
D5	21.37	32.64	25.41	21.86	25.24	11.97	19.92	22.04	25.96	18.40
D6	18.53	30.86	25.55	23.56	26.81	19.92	11.49	20.72	26.78	20.40
D7	17.38	28.74	24.38	18.74	15.08	22.04	20.72	8.60	18.00	15.97
D8	22.35	21.98	14.57	16.04	23.98	25.96	26.78	18.00	6.32	24.86
D9	18.73	32.18	27.33	25.73	20.64	18.40	20.40	15.97	24.86	8.56

TABLE 9
THE DISTANCE MATRIX OF ARABIC CHARACTER FOR SAME SUBJECT USING DTW.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
C1	11.10	32.93	30.56	29.28	35.34	38.53	30.02	34.46	32.22	32.31
C2	32.93	10.23	11.02	19.68	31.38	14.92	18.59	22.19	24.73	24.83
C3	30.56	11.02	13.02	12.22	32.95	18.20	15.58	20.50	22.84	24.29
C4	29.28	19.68	12.22	13.81	38.10	23.52	12.00	24.88	26.51	23.80
C5	35.34	31.38	32.95	38.10	17.54	23.52	36.15	25.41	26.66	40.54
C6	38.53	14.92	18.20	23.52	23.52	12.95	21.11	28.79	30.49	23.64
C7	30.02	18.59	15.58	12.00	36.15	21.11	19.69	25.13	5.99	24.81
C8	34.46	22.19	20.50	24.88	25.41	28.79	25.13	4.86	5.99	35.58
C9	32.22	24.73	22.84	26.51	26.66	30.49	5.99	5.99	6.97	37.74
C10	32.13	24.83	24.29	23.80	40.54	23.64	24.81	35.58	37.74	21.82

TABLE 10 THE DISTANCE MATRIX OF ARABIC SENTENCES FOR SAME SUBJECT USING DTW.

	S1	S2	S3	S4	S5
S1	17.2841	49.1675	53.4164	53.9998	50.7215
S2	49.1675	171935	50.0854	43.6064	51.2795
S3	53.4164	50.0854	56.971	53.0345	16.4854
S4	53.9998	43.6064	53.0345	16.4989	51.1744
S5	50.7215	51.2795	56.971	51.1744	20.3975

C. Pattern Matching Using Neural Networks

Artificial neural networks (ANNs) are intelligent systems that are related in some way to a simplified biological model of the human brain. They are composed of many simple elements, called neurons, operating in parallel and connected to each other in the forward path by some multipliers called the connection weights. Neural networks are trained by adjusting values of these connection weights between the network elements. Neural networks have self-learning capability, are fault tolerant and noise immune, and have applications in system identification, pattern recognition, classification, speech recognition, image processing, etc.[16,17].

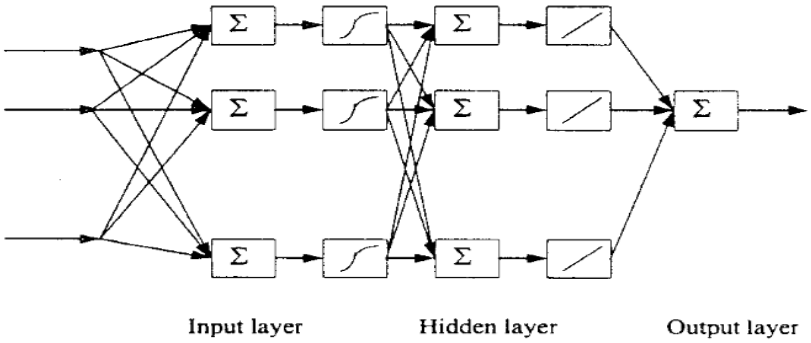

Fig.6 Architecture of Neural Network

To solve a problem steps required are :

- Present the neural network with a number of inputs (Vectors each representing a pattern)
- Check how closely the actual generated output matches with desired output.
- Change the neural network parameters (weights) to better approximate the outputs. The three layer feed-forward neural network architecture for this application is shown in Figure 6.

The ANNs were trained with three voice samples recorded at different instants of time of 10 different speakers uttering the same phrase at all times. An initial leaning rate, an allowable error and the maximum number of training cycles are the parameters that specified during the training phase. The neural network is constructed in the MATLAB environment. Figure 7 shows a plot of the sum squared error versus the number of epochs during the training phase. The sum squared error goal was reached in just 2 epochs.

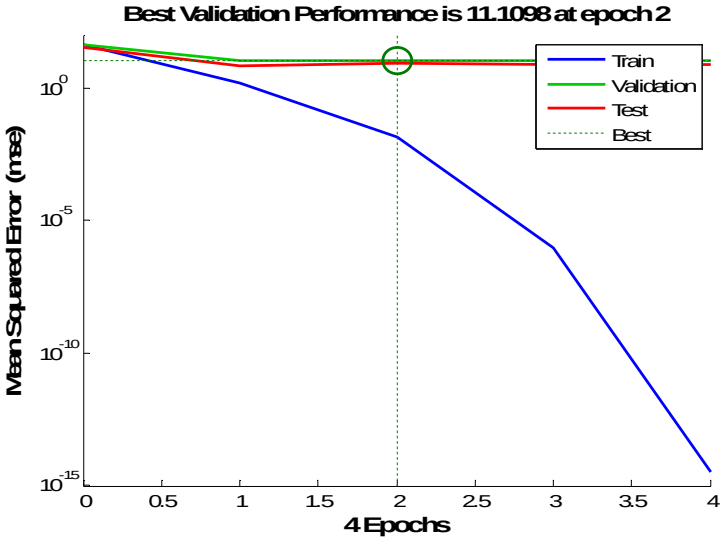

Fig. 7 sum squared error versus the number of epochs during the training phase.

IV. RESULT AND DISCUSSION

The objective of this research is to compare the performance of MFCC, DTW and ANN. The test pattern is compared with the reference pattern to get the best match. Table 3, 4, 5 and 6 represents the MFCC distance Matrix of Arabic digit, word, character and sentences same subject. Table 7, 8, 9 and 10 provides information regarding accuracy in the form of distance Matrix of Arabic Digit Same Subject, Table 11 provides Comparative Recognition Accuracy for MFCC, DTW and ANN. A Digit, Word, Character and sentences at some exceptional case makes system distance about speech from distance Matrix we derived overall accuracy by

$$Accuracy = \frac{N-C}{N} * 100$$

Where N is Number of total token and C is a Number of token not match.

The result obtained in this paper motivated to use fusion MFCC,DTW and ANN feature extraction method which produces high accuracy with minimum time effort, this technique in an improvement in continuous speech recognition system .

TABLE 11 COMPARATIVE RECOGNITION ACCURACY FOR MFCC, DTW&ANN

Subject / Technique	Digit	Character	Word	Sentences
MFCC	90	80	80	80
DTW	90	74.28	80	80
ANN	74%	71.79%	73.33%	70.67%

V. CONCLUSION

The paper presents the experimental analysis of Arabic speech recognition using three different methods, MFCC, DTW and ANN. It presents the comparison among these three techniques. From table 11 it is observed that recognition of digits is easier and as the word are increasing the recognition rates found to be decreased. Among all MFCC is found to be efficient. These is a scope in increasing the recognition rate of neural network.

REFRENCES

[1] Chadawan Ittichaichareon, Siwat Suksri and ThaweesakYingthawornsuk, “Speech Recognition using MFCC” International Conference on Computer, Simulation and Modeling (ICGSM'2012) July 28-29, 2012 Pattaya (Thailand)

[2] MSNencarta, “Languages Spoken by More Than 10 Million People" ;http://encarta.msn.com/media_701500404/Languages_Spoken_by_More_Than_10_Million_Pe ple.html, 2007.

[3] Muhammad Alkhouli. “AlaswaatAlaghawaiyah”;DaarAlfalah, Jordan,1990 (in Arabic).

[4]ZaidiRazak,NoorJamilah Ibrahim, emranmohdtamil,mohdYamaniIdnaIdris, MohdyaakobYusoff ",Quranic verse recitionfeature extraction using mel frequency ceostral coefficient(MFCC)",Universiti Malaya.

[5]F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B.Weiss, ".A database of German EmotionSpeech.",INTERSPEECH 2005, September, 4-8, Lisbon,Portugal

[6] RafikDjemili,Mouldibedda,andHocineBourouba"Recognition of spoken Arabic Digit Using Neural predictive Hidden Markov Models "International Arab journal of information technology,Vol.1,No.2,Jully 2004.

[7] Bharti W. Gawali, Santosh Gaikwad, PravinYannawar, Suresh C. Mehrotra “MarathiIsolated Word Recognition System using MFCC and DTW FeaturesACEEEInt". J. on Information Technology, Vol. 01, No. 01, Mar 2011.

[8] The website for The Disordered Voice DatabaseAvailable:http://www.kayelemetrics.com/Product%20Info/CSL%20Family/4500/4500.html.

[9] Jamal Price, sophomore student," Design an automatic speech recognition system using maltab", University of Maryland EsternShore Princess Anne

[10] E.C. Gordon,"Signal and Linear System Analysis".John Wiley &Sons Ltd.,New York, USA,1998.

[11]ZaidiRazak,NoorJamilahIbrahim,emranmohdtamil,mohdYamaniIdnaIdris,MohdyaakobYusoff,'"Quranicverse recitionfeature extraction using mel frequency ceostral coefficient(MFCC)",Universiti Malaya.

[12] http://www.cse.unsw.edu.au/~waleed/phd/html/node38.html,downloaded on 3rd March 2010

[13] Jamal Price, sophomore student, "Design an automatic speechrecognition system using maltab", University of Maryland Estern Shore Princess Anne.

[14] Palden Lama and MounikaNamburu 'Speech Recognition with Dynamic Time Warping using MATLAB'CS 525, SPRING 2010.

[15]ANJALI BALA 'VOICE COMMAND RECOGNITIONSYSTEM BASED ON MFCC AND DTW'AnjaliBala et al. / International Journal of Engineering Science and Technology Vol. 2 (12), 2010, 7335-7342.

[16] KishanMehrotra, Chilukuri K. Mohan, Sanjay Ranka, Elements of Artificial Neural Networks, Penram International, 2007.

[17]Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-Rahman Mohamed et.al. “Deep Neural Networks For Acoustic Modeling In Speech Recognition”, IEEE Signal Processing Magazine, November 2012.

