ISSN No. 0976-5697
Volume 2, No. 1, Jan-Feb 2011
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Performance Analysis of MFCC & DTW for Isolated Arabic Digit

Santosh Gaikwad*
Research Student
Department of Computer Science & Information Technology
Dr.B.A.M.University Aurangabad
Aurangabad, Maharashtra, India
santosh.gaikwadcsit@gmail.com

Bhart Gawali
Associate Professor
Department of Computer Science & Information Technology
Dr.B.A.M.University Aurangabad
Aurangabad, Maharashtra, India
bhart_rokade@yahoo.co.in

Pravin Yannawar
Assistant Professor
Department of Computer Science & Information Technology
Dr.B.A.M.University Aurangabad
Aurangabad, Maharashtra, India
pravinyannawar@gmail.com

Abstract: This paper describes a performance analysis of MFCC and DTW. Speech processing domain perform much more application in Real life such as speech based telephone dialing, airline reservation etc.Arabic language is Semitic language and differ from European languages. We describe comparative result of MFCC and DTW these were implemented on speech sample from Arabic digit corpus. The main aim of this paper is to compare the sigficance of these records. The MFCC based recognition system achieved 97.66 with multiple speaker where as DTW based system achieved 98.97 correct digit recognition.

Keywords: DTW, MFCC, sampling rate, speakers, performance.

I. INTRODUCTION

Arabic is a Semitic language & it is one of the oldest languages in the world. Arabic phonemes contain two distinctive classes, which are named pharyngeal and emphatic phonemes [1]. There is Modern standard Arabic (MSA) which has basically 34 phonemes. Out of these six are basic vowels and 28 consonants [2]. A phoneme is the smallest element of speech units that indicates a difference in meaning, word or sentence. Arabic language has few vowels than English language. Automatic Speech Recognition (ASR) was dedicated to dialectal and colloquial Arabic within the 1997 by NIST benchmark evaluations [3, 4]. Arabic digit zero to nine (Sefar, Wahid, Ithnan, Thelathe Arbea, Khams, Sitte, Seba, Theman Tisa,) polysyllable words except the first one, zero, which is monosyllable word. Table 1 shows the ten Arabic digits along with the way of how to pronounce them in modern Standard Arabic (MSA), number and types of syllables in every spoken digit.

Figure 1: Arabic Digit with pronunciation

Much of research work is done in automatic speech recognition various languages like English, but Arabic language had limited number of research efforts [5,6]. The Flow chart of Arabic speech Recognition system is presented in Figure 2.

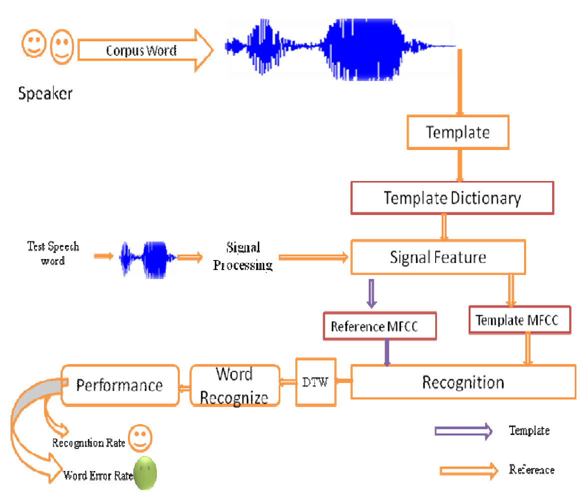

Figure 2: Flow chart of Arabic Speech Recognition System

The paper is organized into five sections, Section 1, gives Introduction, Section 2 deals with details of creating Arabic speech database, Section 3, focuses on Recognition of isolated digits using MFCC and DTW; Section 4, covers results and conclusion followed by Section 5 with the References

II. ARABIC SPEECH DATABASE

Automatic recognition of spoken digits is a challenging task in the field of ASR. For accuracy in the speech recognition, we need a collection of utterances [7], which are required for training and testing. The collection of utterances in proper manner is called the database. The generation of a corpus for Arabic digits as the collection of speech data is described below. The age group of speakers selected for the collection of database ranges from 22 to 35. Mother tongue of all the speakers was Arabic. The total number of speakers was 30 out of which 10 were Females and 20 were Males. The vocabulary size and system parameters are shown in table

Table 1 Table System Parameters

Parameter	Value
Sampling Rate	11025
Database	Isolated 10 Arabic Digits
Speakers	30
Condition of Noise	Normal
Accent	Saudi
Preemphased	1-0.97z--1
Window type	Hamming ,25 milliseconds
Window step size	20 millisecond

A. *Acquisition setup*

To achieve a high audio quality the recording took place in the normal room without noisy sound and effect of echo. The sampling frequency for all recordings was 11025 Hz at 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 [8]. 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 an input/output recording device for a PC, which has special features for reliable acoustic measurements [9,10]. A setup of CSL is shown in Figure.

Figure 3: Setup Computerised Speech Lab

B. *The digitization of Arabic digit speech signal*

The digitization of Arabic digit speech signal is shown in fig 4, fig 5, fig 6 and fig 7.These speech signal is slowly varying over time and it is called quasi stationery

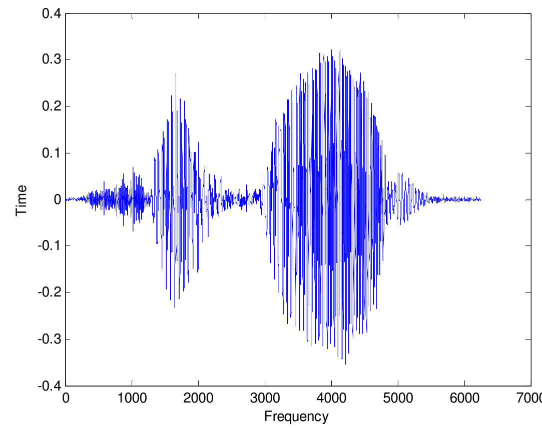

Figure 4: Pronunciation of Sefer

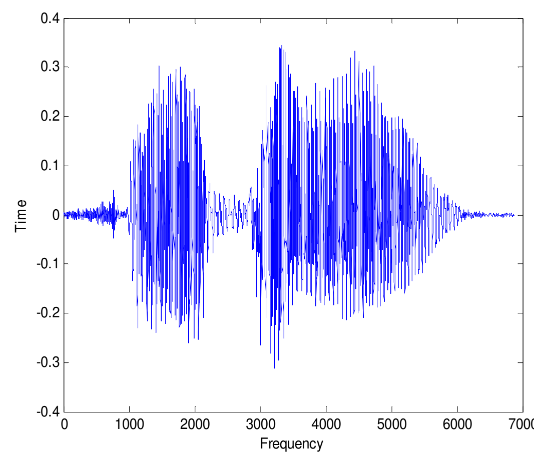

Figure 5: pronunciation of Seba

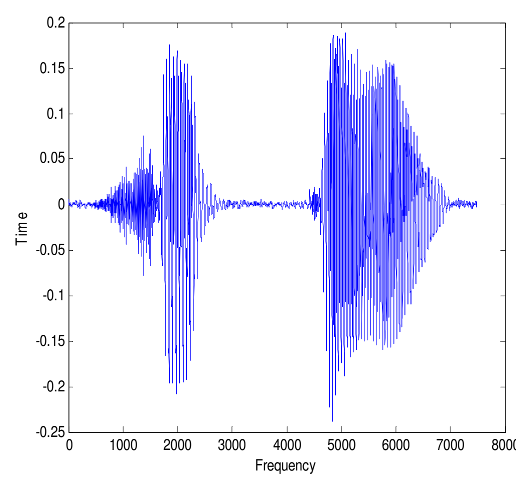

Figure 6: Pronunciation of Site

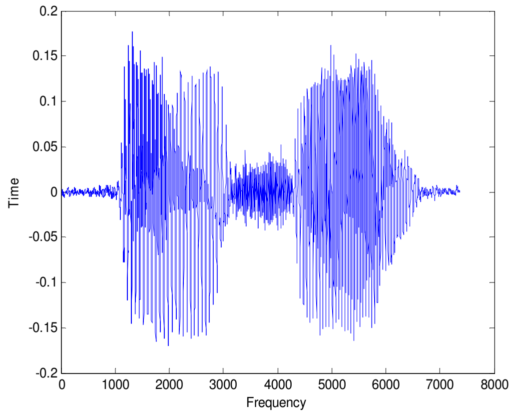

Figure 7: Pronunciations of Khems

Above plot shows the speech word sefer, seba.sitte.khems respectively the recordings were digitized at f samples is equal to 11,025 samples per second and at 16 bits per sample. All speakers having a Saudi accent .Time goes from left to right and amplitude is shown vertically. When the speech signal is examined over a short period of time such as 5 to 100 milliseconds, the signal is reasonably stationery, and therefore this signals are examine in short time segment, short time segments is referred to as a spectral analysis. This means that the signal is blocked into 20-30 milliseconds of each frame. And to avoid the loss of any information due to windowing adjacent frame is overlap with each other by 20 percent to 40 percent.

III. ISOLATED ARABIC DIGIT RECOGNITION SYSTEM

In this paper we are focusing on the use of Mel frequency Cepstral Coefficients (MFCC) and dynamic time warping (DTW) for automatic speech recognition system. For recognition system we collect 300 sample in database. In Feature extraction from MFCC & DTW are being recognised by traning the system. The system is trained for MFCC and DTW feature. For traning 120 sample are used for testing 180 sample. We applying both feature extraction technique on collected database. In MFCC feature extraction we applying steps of and find the 13 feature in which energy is first feature.in traning database we store a 13 feature for each frame and applying euclidian distance to compare a test word feature to train word feature.in DTW we align a shortest path between test sample to refrence sample.

A. Feature extraction

In order to extract feature of isolated words there are several methods and algorithm have been reported in literature. In this paper the most prominent methods that are MFCC and DTW are used to extract the feature of isolated words [11].

[a] Feature Extraction using MFCC

As we are characterizing the signal in terms of the parameters of such a model, we must separate source and the model (filter). In ASR the source (fundamental frequency and details of glottal pulse) are not important for distinguishing different phones [12,13]. Instead, the most useful information for phone detection is the filter, i.e. the exact position and shape of the vocal tract. If we knew the shape of the vocal tract, we would know which phone was being produce to separate the source and filter (vocal tract parameters) efficient mathematical way is cepstrum. The cestrum is defined as the inverse DFT of the log.[14]

The ceptral property have been extremely useful where the variances of different coefficients are tends to be uncorrected. The cepstral coefficients have the extremely Useful property that variance of the different coefficients tends to be uncorrelated [15]. This is not true for the spectrum, where spectral coefficients at different frequency bands are correlated. The fact that cepstral features are uncorrelated means that the Gaussian acoustic model doesn't have to represent the covariance between all the MFCC features, which hugely reduces the number of parameters.[16].

$$c[n] = \sum_{n=0}^{N-1} log\left(\left|\sum_{n=0}^{N-1} x[n]e^{-j\frac{2\pi}{N}kn}\right|\right) e^{j\frac{2\pi}{N}kn}$$

Where c(n) is cepstral coefficient and x(n) is the input signal. Since the MFCC is the most popular feature extraction technique for ASR [17], the steps involved in extraction of MFCC is shown in fig 8 and fig 9.

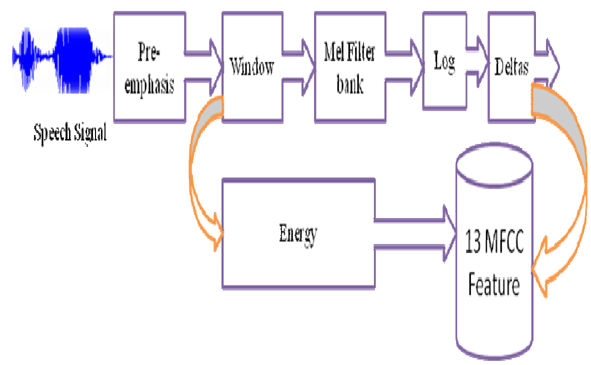

Figure 8: Steps for extracting a sequence of 13MFCC feature vectors from waveform

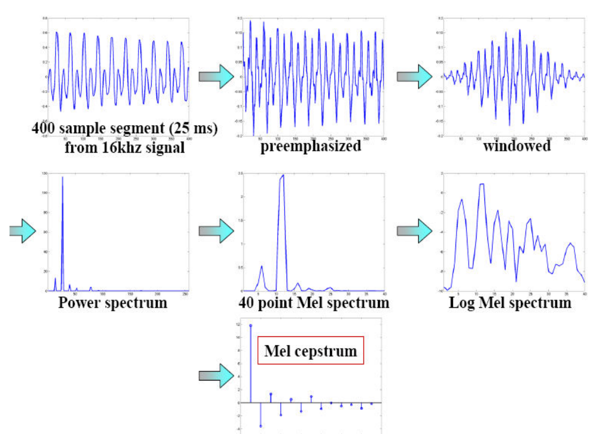

Figure 9: Extraction of MFCC Feature for a Frame

[b] Feature Extraction using DTW

Dynamic time warping is a simplest way to recognize an isolated word. Actually it used as a template matching technique. In DTW we compare sample to stored word template and determine best match. This goal is complicated by a number of factors. First, different samples of a given word will have somewhat different durations. This problem can be eliminated by simply normalizing the templates and the unknown speech so that they all have an equal duration. However, another problem is that the rate of speech may not be constant throughout the word; in other words, the optimal alignment between a template and the speech sample may be nonlinear [18, 15]. Dynamic Time Warping (DTW) is an efficient method for finding this optimal nonlinear alignment [19]. DTW is an instance of the general class of algorithms known as dynamic programming. Its time and space complexity is merely linear in the duration of the speech sample and the vocabulary size. The algorithm makes a single pass through a matrix of frame scores while computing locally. Optimized segments of the global alignment path are as shown in figure 10[20].

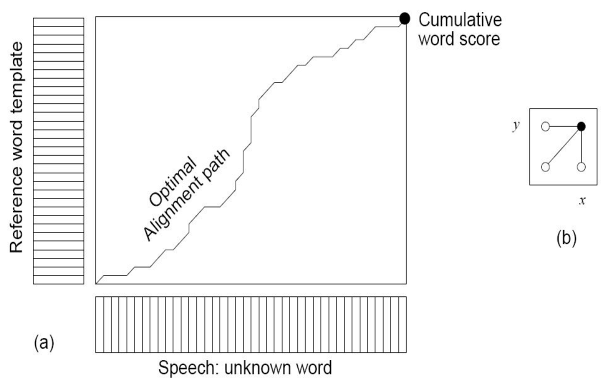

Figure 10: alignment of shortest path in DTW

[c] Template matching using Euclidian Distance

In the speech recognition phase, an unknown speech voice is represented by a sequence of feature vector {x1, x2 ….xi), and then it is compared with the features codebooks from the train database. In order to identify the unknown speech, this can be done by measuring the distortion distance of two vector sets based on minimizing the Euclidean distance. The Euclidean distance is the "ordinary" distance between the two points that one would measure with a ruler, which can be proven by repeated application of the Pythagorean Theorem. The Euclidean distance between two points p,[20,21]

P = (p1, p2…pn) and Q = (q1, q2...qn), The speech with the lowest distortion distance is chosen to be identified as the unknown speech.

[d] Performance of System

In the presented recognition system for finding a performance a **word error rate (WER)** is used**.** Error rate is a common metric of the performance of a speech recognition or machine translation system. The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence (supposedly the correct one). The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level [22]. This problem is solved by first aligning the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment. Word error rate can then be computed as

$$WER = \frac{S + D + I}{N}$$

Where

[i] S is the number of substitutions,
[ii] D is the number of the deletions,
[iii] I is the number of the insertions,
N is the number of words in the reference

When reporting the performance of a speech recognition system, sometimes word recognition rate (WRR) is used instead.

$$WRR = 1 - WER = \frac{N - S - D - I}{N} = \frac{H - I}{N}$$

Where

[iv] H is N-(S+D), the number of correctly recognized words

Table III Confusion Matrix of MFCC

	Zero	*One*	*Two*	*Three*	*Four*	*Five*	*Six*	*Seven*	*Eight*	*Nine*	*Accuracy (%)*	*Tokens*	*Missed*
Zero	30	0	0	0	00	0	0	0	0	0	100	30	0
One	0	29	0	0	0	0	1	0	0	0	96.66	30	1
Two	1	1	26	0	0	0	2	0	0	0	86.66	30	4
Three	0	0	0	28	0	0	2	0	0	0	93.33	30	2
Four	0	0	0	0	30	0	0	0	0	0	100	30	0
Five	0	0	0	0	0	30	0	0	0	0	100	30	0
Six	0	1	0	0	0	0	29	0	0	0	96.66	30	1
Seven	0	0	0	0	0	0	0	30	0	0	100	30	0
Eight	1	0	0	0	0	0	0	1	28	0	93.33	30	2
Nine	0	1	0	0	0	0	0	0	0	28	93.33	30	2
Average											**97.66**		

Table IV Confusion Matrix of DTW

	Zero	*One*	*Two*	*Three*	*Four*	*Five*	*Six*	*Seven*	*Eight*	*Nine*	*Accuracy (%)*	*Tokens*	*Missed*
Zero	30	0	0	0	0	0	0	0	0	0	100	30	0
One	1	29	0	0	0	0	0	0	0	0	96.66	30	1
Two	0	0	30	0	0	0	0	0	0	0	100	30	0
Three	0	0	0	29	0	0	0	1	0	0	96.66	30	1
Four	0	0	0	0	30	0	0	0	0	0	100	30	0
Five	0	0	0	0	0	30	0	0	0	0	100	30	0
Six	0	2	0	0	0	0	28	0	0	0	93.33	30	2
Seven	0	0	0	0	0	0	0	30	0	0	100	30	0
Eight	0	0	0	0	0	0	0	0	30	0	100	30	0
Nine	1	0	0	0	0	0	0	0	0	29	96.66	30	1
Average											**98.97**		

IV. RESULT AND CONCLUSION

This work is first approch towards compare performance of MFCC(where 13 coefietient are used)& DTW.The speech data used in this experiment are isolated digit of arabic language.for resulting DTW path we compare test pattern to refrence pattern for getting a best match.The symmatric form of DTW algorithm is used to optimally align test and refrence pattern to give average distance associated with optimal path.The recognition system used the utterrances of the spoken digit for traning and remaining utterances for testing.The confusion matrix of MFCC is shown in table 2 where as for DTW shown in table 3. In each test we pass 30 tokens of each word. The performance is calculted by checking how many words the system recognises correctly and also for how many words the system is getting confused. The indivual comparative digit recognition accuracy is shown in table 4.The result obtained from accuracy test is about 97.66% in MFCC. While result obtained for DTW is 98.97. we apply DTW technique on obtained feature of MFCC.Thus result showed promising arabic digit recognition.Furthe the rate of recognition an be increased by using fusion technique of MFCC and DTW.

V. ACKNOWLEDGMENT

The authors would like to thank the university authorities for providing infrastructure to carry out the experiments and DST for supporting the research work.

VI. REFERENCES

[1] Al-Zabibi, M. "An acoustics phonetic Approach in Automatic Arabic Speech Recognition, the British library in Association with UMI, 1990.

[2] Le, A. 2003 Rich Transcription: spring Speech to Text transcription evaluation results, Proc. RT02 workshop, 2003.

[3] (2002) The IEEE website. [Online]. Available: http://www.ieee.org

[4] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available:http://www.ctan.org/texarchive/macros/latex/contrib/supported/IEEEtrans.

[5] S.A.R. AI-Haddad ,S.A Samad, A.Hussain,K.A. ishak "Isolated Malay Digit recognition using Pattern Recognition Fusion of Dynamic time and Hidden Markov Models" American Journal of Applied sciences 5(6):714-720,2008

[6] Rafik Djemili,Mouldi bedda,and Hocine Bourouba "Recognition of Spoken Arabic Digit Using Neural Predictive Hidden Markov Models "International Arab journal of Information Technology ,Vol.1,No.2,Jully 2004

[7] Yousef Ajami Alotaibi "Comparative Study of ANN and HMM to Arabic Digit Recognition Systems "JKAU: Eng.Sci, Vol.19 No.1, pp:43-60(2008 A.D/1429 A.H.)

[8] Ben J. Shannon,Kuldip k.Paliwal "A Comparative Study of Filter Bank Spacing For Speech Recognition" Microelectronics Engineering research conference 2003

[9] Yousef Ajami Altaibi,mansour alghamdi,Fahad Alotaiby "Speech Recognition System of Arabic Digit based on A Telephony Arabic Corpus"

[10] Santosh Gaikwad,Bharti Gawali "A Review on Speech Recognition" International Journal of Computer application Vol 10 ,Number 3 ,10 Nov 2010[online].

[11] "Isolated word, Speech recognition using Dynamic Time Warping towards smart appliances" online Source http://www.cnel.ufl.edu/~kkale/6825Project.html

[12] Elghonemy, M, Fikri, M "Speaker independent isolated Arabic word recognition system" Acoustic, speech and signal Processing, IEEE International Confrence on ICSSP 2008

[13] Hiromi Sakaguchi and Naoaki Kawaguchi, "Mathematical Modeling of Human Speech Processing Mechanism Based on the Principle of Bain Internal Model of Vocal Tract" Journal of the faculty of Engineering, Shinshu University, No. 75,1995.

[14] Steven W. Smith, "the Scientist and Engineer's Guide to Digital Signal Processing", California Technical Publishing, 1977, page (s): 169-174

[15] F. Jelinek, L.R. Bahl,, and R. L, Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech", IEEE Trans. Informat. Theory, vol. IT-21, PP. 250-250, 1975

[16] Y.YanandE.Bernard,"Anapporchtoautomaticlanguageidentificationbasedonlanguagedepandantphonerecognition",ICASSP'95,vol.5,May.1995p.3511

[17] C.S.MyersandL.R.Rabiner,ALevelBuildingDynamicTimeWarpingAlgorithmforConnectedWordRecognition,IEEETrans.Acoustics,SpeechSignalProc.,ASSP29:284-297,April1981

[18] Spector,SimonKingaandJoeFrankel,Recognition,Speechproductionknowledgeinautomaticspeechrecognition,JournalofAcousticSocietyofAmerica,2006

[19] R.KlevansandR.Rodman,"VoiceRecognition,ArtechHouse,Boston,London1997

[20] L.R.Bahlet.al,AmethodofConstructionofacousticMarkovModelforwords,IEEETransactiononAudio,speechandLanguageProcessing,Vol.1,1999

[21] M.A.Anusuya,S.K.Katti"SpeechRecognitionbyMachine:AReview"InternationaljournalofcomputerscienceandInformationSecurity2009

[22] Word error Rate [online] Viewed on 12 Oct 2010.

[23] Source: en.wikipedia.org/wiki/Word_error_rate