A Comparison of DHMM and DTW for Isolated Digits Recognition System of Arabic Language

Z.HACHKAR[1,3], A. FARCHI[2] , B.MOUNIR[1], J. EL ABBADI[3]
[1]Ecole Supérieure de Technologie, Safi, Morocco.
zhachkar2000@yahoo.fr.
mounirbadia@yahoo.fr.
[2]Faculté des sciences et Techniques, Settat, Morocco.
direction@setronic.ma
[3]Ecole Mohammadia d'Ingénieurs, Rabat, Morocco.
elabbadi10@yahoo.fr

***Abstract*— Despite many years of concentrated research, the performance gap between automatic speech recognition (ASR) and human speech recognition (HSR) remains large. Especially for Arabic language, research efforts are still limited in comparison with other languages such as English or Japanese. In this work, we have use two algorithms to implement a system of Automatic Recognition of isolated Arabic Digits: Dynamic Time Warping (DTW) and Discrete Hidden Markov Model (DHMM). The endpoint detection, framing, normalization, Mel Frequency Cepstral Coefficient (MFCC) and vector quantization techniques were used to process speech samples to accomplish the recognition.**

The better recognition accuracy of about 92% was obtained with DHMM-based system. In noisy environment, the recognition performances for the two ASR are worse but the pattern recognition using HMM is better than the pattern using DTW.

Keywords ; Arabic language, Speech recognition, DTW, MFCC, DHMM

I. INTRODUCTION

Automatic Speech Recognition (ASR) Technology has made enormous advances in the last 20 years. Current ASR systems work well when the test and training conditions match. In real world environments, there is often a mismatch between testing and training conditions [1]. Various factors like additive noise, acoustic echo, and speaker accent, affect the speech recognition performance. The challenge remains for ASR to approach human performance. The simplicity of models used in current ASR, when compared to the complex, nonlinear processing done by humans, suggests that there remain many ways to improve ASR. At each level (i.e., feature estimation, temporal and acoustic modeling, language modeling, search, decision making), compromises have been made in ASR to have simple and fast processing, at the expense of lower accuracy.

A speech recognition system basically contains extraction of features and classification of an utterance. The measurements made on the speech waveform include energy, zero crossings, extrema count, formants, LPC cepstrum [2, 3] and the Mel Frequency Cepstrum Coefficient (MFCC) [4]. The LPC method provides a robust, reliable and accurate method for estimating the parameters that characterize the linear, time-varying, system which is recently used to approximate the nonlinear, time-varying system of the speech waveform. The MFCC method uses the bank of filters scaled according to the Mel scale to smooth the spectrum, performing a processing that is similar to that executed by the human ear. For recognition, the performance of the MFCC was better than the LPC cepstrum [4]. As to classification of an input utterance, the most successful speech recognition methods are the pattern matching using Dynamic Time Warping (DTW) [5, 6], vector quantization (VQ) [7, 8] and hidden Markov model (HMM) [9, 10]. Since the same word uttered by the same speaker may have different duration of the same phoneme, the DTW process nonlinearly expands or contracts the time axis to match the same phoneme or landmark positions between the input speech and reference templates. The VQ is an information theoretic data compression principle introduced by Shannon [7]. When it is applied to speech compression, a training sequence is used to generate a set of reproduction vectors (codeword), called the codebook of the speech. In general, the selection of a perceptually meaningful distortion measure inclustering and the construction of an optimal codebook are difficult. It is also difficult to apply the VQ to a large vocabulary because the computational cost is still high in clustering. The theory of HMM was published by Baum et al. [9], but widespread understanding and applications of the theory of HMMs to speech processing have occurred only within the past 10 years. The HMM technique has significantly reduced the computational cost and has been used for large vocabulary connected and continuous speech recognition applications.

All existing speech recognition system methods were widely used for certain languages such as English or Japanese. For Arabic language, research efforts remain limited. H. Bahi et al. [11] proposed an Arabic numeral recognition technique that made use of vector quantization and HMM. Their proposed approach made use of LP and LP Cepstral coefficients. They trained and tested their system on a locally generated database of 1500 (the utterances were divided into training and test datasets) utterances produced by 50 speakers. The recognition accuracy was 91% and 95% for previously unseen and seen speakers during training, respectively. In 2002, W. Alkhaldi et al. [12] proposed HMM based Arabic numeral recognition technique. They created two systems: one that made use of Wavelet Cepstral coefficients and the other that made use of Mel frequency Cepstral Coefficients (MFCCs). The systems were trained on 500 utterances produced by 29 male and 21 female speakers. They tested their system on 130 utterances produced by male speakers. For different numerals the recognition accuracy ranges from 61% to 92% for MFCC based and 76% to 92% for Wavelet based systems. Elmisery et al. [13] proposed another HMM based technique for numeral recognition. They devised the HMM algorithms for Field Programmable Gate Arrays (FPGAs). They also developed two systems for comparative study based on LPC and MFCC features. They used a database of 200 utterances for codebook generation produced by a single male speaker. They used 40 utterances for training. The recognition accuracy ranges from 91% to 96% for LPC based and 95% to 98% for MFCC based systems.

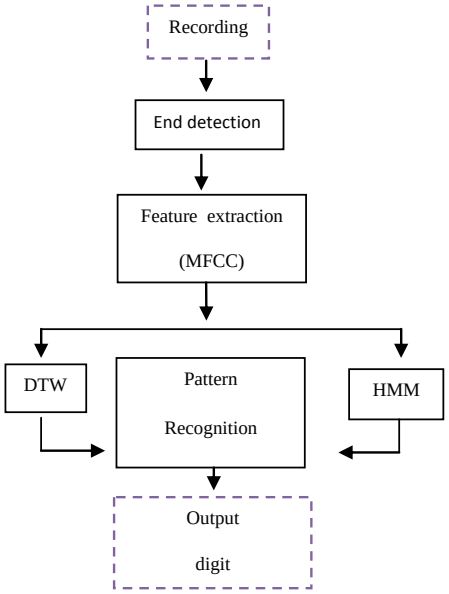

Figure 1. Bloc diagram of Arabic digit recognition using DTW and HMM

This paper presents development of an experimental isolated word recognizer for **Isolated Arabic Language**. Research is further extended to comparison of speech recognition system for small vocabulary of speaker dependent isolated spoken words using the Discrete Hidden Markov Model (DHMM) and Dynamic Time Warping (DTW) technique (Figure 1). The presented work emphasizes on template-based recognizer approach using MFCC with dynamic programming computation and vector quantization with Hidden Markov Model based recognizers in isolated word recognition tasks, which also significantly reduces the computational costs. The analysis, design and development of the two automation systems are done in MATLAB 6.5, using [14, 15]

II. ARABIC LANGUAGE

Arabic is a Semitic language, and it is one of the oldest languages in the world. Arabic is the first language in the Arab world. Arabic phonemes contain two distinctive classes, which are named pharyngeal and emphatic phonemes. These two classes can be found only in Semitic languages like Hebrew [16, 17]. Arabic digits (figure 2) zero to nine are polysyllabic words except the first one, zero, which is a monosyllable word [16]. 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

Digits	Arabic Digits	Transcription
0	صفر	SIFAR
1	واحد	WAHID
2	اثنان	ITHNAN
3	ثلاثة	THALATHA
4	اربعة	ARBAA
5	خمسة	KHAMSA
6	ستة	SITA
7	سبعة	SABAA
8	ثمانية	THAMANIYA
9	تسعة	TISAA

Figure 2. Arabic digits

III. DTW

Dynamic Time Warping algorithm (DTW) [18] is an algorithm that calculates an optimal warping path between two time series. Suppose we have two numerical sequences (*a*1, *a*2, ..., *an*) and (*b*1, *b*2, ..., *bm*). The algorithm starts with local distances calculation between the elements of the two sequences using different types of distances. The most frequent used method for distance calculation is the absolute distance between the values of the two elements. That results in a matrix of distances having *n* lines and *m* columns of general term:

$$d_{ij} = |a_i - b_j|, i = 1,\dots,n, j = 1,\dots,m$$

Starting with local distances matrix, the minimal distance matrix between sequences is determined using a dynamic programming algorithm and the following optimization criterion:

$$a_{ij} = d_{ij} + \min \quad (a_{i-1,j-1}, a_{i-1,j}, a_{i,j-1})$$

where a*ij* is the minimal distance between the subsequences (*a*1,*a*2, ..., *ai*) and (*b*1, *b*2, ..., *bj*).

IV. VQ/HMM SYSTEM

A Hidden Markov Model (HMM) is a statistical model in which is assumed to be a Markov process with unknown parameters. The challenge is to find all the appropriate hidden parameters from the observable states. To illustrate an application of HMMs for speech recognition, we present in Figure 3 our implementation for isolated word recognition
system on Discrete Hidden Markov Models. We have a vocabulary of L words to be recognized, and each word is to be modelled by a distinct HMM. The training sets consist of K utterances of each words. In ordre to obtain a word recognizer, we performed the following steps:

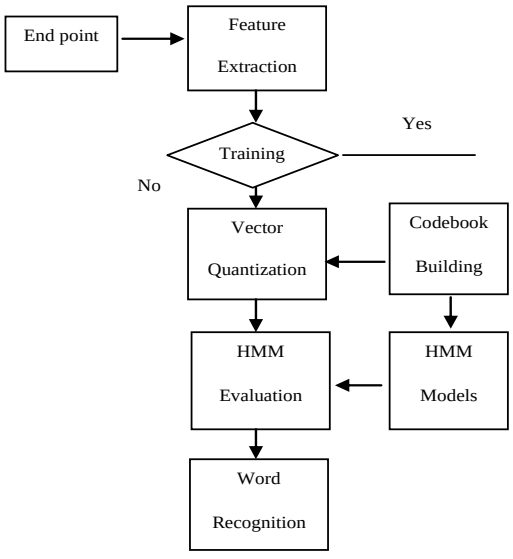

Fig ure 3. The building blocks for Teaching Module speech recognition

A. *VQ Codebook*

In discrete HMM system, the continuous feature space is subdivided by a vector quantized into J non overlapping subsets and each subset is represented with a codeword m_j ($1 \leq j \leq J$).The set of available code words is termed the codebook.The VQ codebook is constructed by an unsupervised cluster algorithm.

B. *Re–Estimation of HMM*

For each word of the vocabulary, we will built a HMM, that is, we estimated the model parameters that optimize the likelihood for the training set of observation sequences. There are many criteria that can be used to this problem. We have used for this problem the Baum-Welch algorithm [19] developed by Baum which is one of the most successful optimization methods

C. *Recognition*

For each unknown word to be recognized, we calculated the likelihood models for all possible models, and selected the models with the highest likelihood. The probability calculation was performed using Viterbi algorithm [20], precisely the logarithm of the maximum likelihood.

V. EXPERIMENTAL SETUP

A. *Database preparation*

An "in-house corpus" was created from all 10 Arabic digits. A number of 5 Moroccan speakers (3 males and 2 females) were asked to utter all digits 10 times. During the recording session, each utterance was played back to ensure that the entire digit was included in the recorded signal.

Speech recognition performance highly depends on endpoint detection accuracy [21], for all experiments; every speeches file from the database was analyzed by an endpoint detection program in order to locate more accurate endpoints

B. *End point Detection*

Every speech recognition system contains a speech/non-speech detection stage. This problem is often referred to as the endpoint location problem. The method proposed in [21] and used in this work uses two measures of the signal: the Zero Crossing Rate (ZCR) and the Energy. This algorithm was performed in Matlab [22], and applied to all corpus.

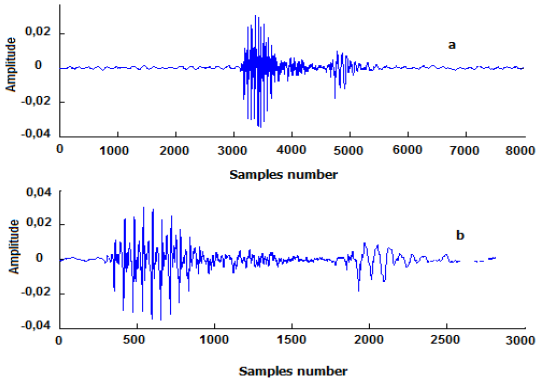

Figure 4. Signal "0: Sifar" recorded before (a) and after (b) application of endpoint detection algorithm

C. *Feature extraction*

Modelling of the glottal flow is a difficult problem and very few studies attempt to precisely decouple the source tract components of the speech signal. Standard feature [23] extraction method MFCC simply ignores the pitch component and roughly compensates the spectral tilt by applying a pre-emphasis filter prior to spectral analysis. The speech samples (8 kHz sampling rate) are windowed into overlapping 25 ms frames with a frame shift of 10 ms. A 256 point FFT is used to compute the power spectrum that is used in an emulated filter-bank composed of 24 triangular weighting functions on a Mel scale. The natural logarithm is then applied to the 24 filter-bank

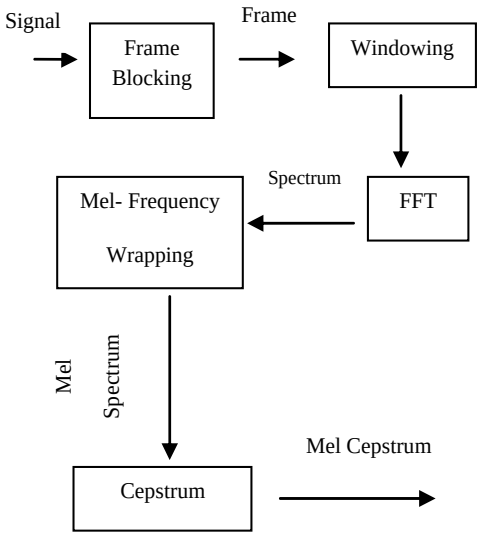

Figure 5. Block diagram of the MFCC processor

energies. The Mel spectrum coefficients are real numbers; we can convert them to the time domain using the Discrete Cosine Transform (DCT)

In this part, a DHMM-based speech recognizer is used for the recognition of isolated words. Here, the DHMM for each word has five states. Transitions between states are allowed only in left-to-right direction with no skipping of states.

VI. RESULTS

A. *DTW recognition*

To search for more effective signal characteristics, we, first, add the log Mel cepstrum energy coefficient to each frame. Then, we obtain 13 coefficients per frame. To account for the temporal properties, these 13 coefficients are derived twice to obtain respectively the vectors ΔMFCC and ΔΔMFCC. The recognition percent is then calculated for each case (table 1).

Table *I*. Recognition percent for different extracted features

Extracted Features Per Frame	**Recognition % for frame and overlap lengths 512*256**
MFCC	77
MFCC+Energy	80
MFCC+Energy+ Δ	83
MFCC+Energy+ Δ Δ	86

The obtained results (Table 1) show that the extracted features "MFCC+Energy+Δ+ΔΔ" per frame give the best system performances. In this case the recognition percent reaches 86 %.

B. *Effects of Vector Quantization Codebook Size (DHMM)*

We varied the speaker codebook size from 8 to 128. The results are shown in Table 3. The WER decreases when the codebook size increases from 80 % to 92 %. This can be explained as the codebook size increases, the distortion (quantization) error decreases.

Table II. Recognition accuracy for varying codebook size

Codebook Size	**Recognition accuracy (in %)**
8	80
16	84
32	91
64	91.7
128	92

The codebook with size of more than 256 can perform better than codebook of size 128, but the average time to recognize one speaker is higher compared to codebook size of 128. This can affect the time response of identification system. Therefore, codebook size of 128 was used throughout the experiment.

C. Effects of additive noise (DHMM and DTW)

In order to study the effect of noise on the two ASR developed; we add Gaussian noise to the original speech signals. Table 3 shows the comparison digit recognition accuracy after the noise with various signals to noise ratios (SNRs). The results show that the recognition performances for the two ASR are worse with the noise but the pattern recognition using HMM is better than the pattern using DTW in noisy conditions.

Table III. Effect of additive noise distortion on the recognition performance

Distorsion (in dB)	Recognition accuracy (in %) For DTW	Recognition accuracy (in %) With codebook size 128
clean	86	92
30 dB	69	73.4
25 dB	45	56.9
20 dB	40	44.1

VII. CONCLUSION

In this study, we have designed DTW and DHMM based systems recognition and evaluated their performance on Arabic Digits.

DTW-based system recognition used the MFCC coefficients as basic characteristics vector lead to recognition accuracy of 77%. This recognition accuracy can achieve 86% by using additional characteristics as power information (energy) and differential information (Δ and ΔΔ). DHMM-based system recognition with codebook size 128 presents interesting recognition accuracy with about 92%. In noisy environment, the recognition performances for the two ASR are worse but the pattern recognition using HMM is better than the pattern using DTW.

The performance of the two DTW and DHMM based automatic systems recognition can be further increased with further research by using a larger corpus of Arabic language and a pattern recognition fusion using Dynamic Time Warping (DTW) and Hidden Markov Model (HMM).

References

[1] D.S. Kim, S.Y. Lee, R.M.Kil , "Auditory processing of speech signals for robust speech recognition in real world noisy environments ". IEEE Trans. Speech and Audio Process. 7 (1), 55–69.1999

[2] S.S. McCandless, "An algorithm for automatic format extraction using linear prediction spectra," .IEEE

[3] M.R. Sambur, L.R. Rabiner,"A speaker-independent digit recognition system, " . B.S.T.J. 54 (1) (1975) 84–102.

[4] S.B. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, ".IEEE Trans. Acoust. Speech Signal Process. 28 (4) (1980) 357–366.

[5] Y. Tohkura, "A weighted cepstral distance measure for speech recognition, " . IEEE ICASSP 86, Tokyo, 1986, pp. 761–764.

[6] L.R. Rabiner, S.E. Levinson, A.E. Rosenberg, J.G. Wilson, "Speaker independent recognition of isolated words using clustering techniques, ". IEEE Trans. Acoust. Speech Signal Process. 27 (1979) 336–349.

[7] C.E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," .R.E. Machol (Ed.), Information and Decision Processes, McGraw-Hill, New York, 1960, pp. 93–126.

[8] B.H. Juang, D.Y. Wong, A.H. Gray , "Distortion performance of vector quantization for LPC voice coding, " . IEEE Trans. Acoust. Speech Signal Process. 30 (2) (1982) 294–303.

[9] L.E. Baum, T. Petrie, G.R. Soules, N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, ". Ann. Math. Statist. 41 (1) (1970) 164–171.

[10] K.F. Lee, "Context-dependent phonetic hidden Markov models for speaker independent continuous speech recognition, " . IEEE Trans. Acoust. Speech Signal Process. ASSP-38 (4) (1990) 599–609.

[11] H. Bahi and M. Sellami: "Combination of Vector Quantization and Hidden Markov Models for Arabic Speech Recognition," .Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications, 2001, pp: 96.

[12] W. Alkhaldi , W. Fakhr , N. Hamdy : "Multi-Band Based Recognition of Spoken Arabic Numerals Using Wavelet Transform," .The 19th National Radio Science Conference Alexandria, March 19-21, 2002.

[13] F.A. Elmisery, A.H. Khalil ,A.E. Salama . H.F. Hammed : "A FPGA Based HMM for a Discrete Arabic Speech Recognition System," ICM, Cairo, Egypt, December 9-11, 2003.

[14] MFCC: http://www.freewebs.com/lvtaoran/AuditoryToolbox.rar

[15] HMM: http://www.freewebs.com/lvtaoran/HMMall.rar

[16] M. Alkhouli. "Alaswaat Alaghawaiyah, " .Daar Alfalah, Jordan, 1990 (in Arabic).

[17] M. Elshafei. "Toward an Arabic Text-to-Speech System, " .The Arabian Journal for Science and Engineering, Vol. No. 16, Issue No. 4B, pp. 565-83, Oct. 1991.

[18] H. Sakoe, S. Chiba "Dynamic programming algorithm optimization for spoken word recognition, " .IEEE, Trans. Acoustics, Speech and Signal Proc., Vol. ASSP-26, 1978.

[19] J. C. Segura, A.J. Rubio, A.M. Peinado, P. Garcia and R. Roman "Multiple VQ Hidden Markov Modeling for Speech recognition, " . Speech Communication, vol. 14,pp 163-170,1994.

[20] X.D. huang, H. Hon, M. Hwang and K. Lee, "A comparative study of discrete, semi Continuous and Continuous Hidden Markov Models, " .Computer Speech and Language, voil 7 pp. 359-368, 1993

[21] L.R. Rabiner, and M.R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," . The Bell System Technical Journal, Vol. 54, No. 2, February 1975, pp. 297-315.

[22] http://www.clear.rice.edu/elec301/Projects99/wrcocee/endpt.htm

[23] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont , T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens, "Automatic speech recognition and speech variability: A review, " .Speech Communication 49 763–786 (2007)