International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:6, No:10, 2012

On Developing an Automatic Speech Recognition System for Standard Arabic Language

R. Walha, F. Drira, H. El-Abed, and A. M. Alimi

***Abstract*—**The Automatic Speech Recognition (ASR) applied to Arabic language is a challenging task. This is mainly related to the language specificities which make the researchers facing multiple difficulties such as the insufficient linguistic resources and the very limited number of available transcribed Arabic speech corpora. In this paper, we are interested in the development of a HMM-based ASR system for Standard Arabic (SA) language. Our fundamental research goal is to select the most appropriate acoustic parameters describing each audio frame, acoustic models and speech recognition unit. To achieve this purpose, we analyze the effect of varying frame windowing (size and period), acoustic parameter number resulting from features extraction methods traditionally used in ASR, speech recognition unit, Gaussian number per HMM state and number of embedded re-estimations of the Baum-Welch Algorithm. To evaluate the proposed ASR system, a multi-speaker SA connected-digits corpus is collected, transcribed and used throughout all experiments. A further evaluation is conducted on a speaker-independent continue SA speech corpus. The phonemes recognition rate is 94.02% which is relatively high when comparing it with another ASR system evaluated on the same corpus.

***Keywords*—**ASR, HMM, acoustical analysis, acoustic modeling, Standard Arabic language

I. Introduction

THE most simple, faster and natural manner widely used by human societies to communicate has always been the spoken language rather than the writing one. Thus, researchers and industrialists are interested in developing applications that use speech as a mean of human-machine interaction. The ASR is considered as an important branch of this interaction. Despite the very important recent advances in the ASR field, current systems have not yet achieved the human speech precision and delicacy which makes the ASR an active research topic.

In fact, an ASR system is generally intended for a given language. Unfortunately, and unlike other languages such as English and French, Arabic language still remains very little approached in ASR field despite it is the fourth most widely spoken language in the world. Furthermore, researches are mainly concentrated on SA which is a formal linguistic standard used throughout the Arabic-speaking world, employed in the media, taught in schools, and spoken in the formal framework. During the past few years, some recent research works on Arabic ASR have been dedicated to single phonemes [1, 2], and others to single words [3, 4].

R. Walha is with the National School of Engineers of Sfax, 3038 Sfax, Tunisia (phone: +(216)25-25-29-21; e-mail: walha.rim@gmail.com).
F. Drira, is with National School of Engineers of Sfax, 3038 Sfax, Tunisia (phone: +(216)21-41-31-36; e-mail: fadoua.drira@gmail.com).
H. El-Abed is with the Institute for Communications Technologie, 38106 Braunschweig, Germany (e-mail: elabed@tu-bs.de).
A. M. Alimi is with the National School of Engineers of Sfax, 3038 Sfax, Tunisia (e-mail: adel.alimi@ieee.org).

Ejbali *et al.* [5] have worked on continue SA speech. However, recognition rates of these systems are still far from the perfection.

Our main contributions in this study are twofold. In the first instance, we develop and study an ASR system. The second contribution is to collect a transcribed multi-speaker connected-digits corpus dedicated for SA.

This paper is organized as follows. Section II summarizes the main characteristics of the SA language. Section III and Section IV describe respectively the proposed system and the corpora used in this study. Section V presents and discusses experimental results. Section VI concludes and gives some perspectives of this work.

II. Standard Arabic Language

SA language is a Semitic language composed of 34 phonemes, of which 6 are basic vowels and 28 are consonants. Among these consonants, 3 (ا, و ,ي) are either consonants or long vowels according to their appearance context in the word. The Arabic phonetics originality is mainly based on the lengthening relevance in the vocalic system and on the presence of emphatic and geminated consonants.

Arabic vowels have not the same temporal duration. The vocalic system has 3 short vowels (/a/, /i/, and /u/) and 3 long vowels (/a:/, /i:/, and /u:/). Their phonetic realization is highly variable and depends on the consonant environment and the place of vowel in the word.

Emphatic consonants are achieved in the rear part of the oral cavity. During their production, the root of the tongue is carried against the pharynx. Arabic language has 4 emphatic consonants: 2 plosives: /ṭ/, /ḍ/, and 2 fricatives: /ð /, /ṣ/. In the example of the two words /naṣaba/ (imputed) and /nasaba/ (erected), an emphatic versus nonemphatic opposition is observed on /s/ [6].

All consonants of the Arabic language can be geminated. Arabic grammarians consider that the termination feature is a duplication of the consonant. It is caused by the extension and strengthening of the consonant articulation without changing the position of phonation organs.

The allowed syllable structures in Arabic are CV, CVC, and CVCC where V indicates a (long or short) vowel while C indicates a consonant. Arabic utterances can only start with a consonant [7].

III. System Description

The proposed ASR system is based on a statistical approach introduced by F. Jelinek [8]. It includes five modules: acoustical analysis module, modeling module, transcription module, training module and decoding module. Fig. 1 illustrates an overview of the proposed system.

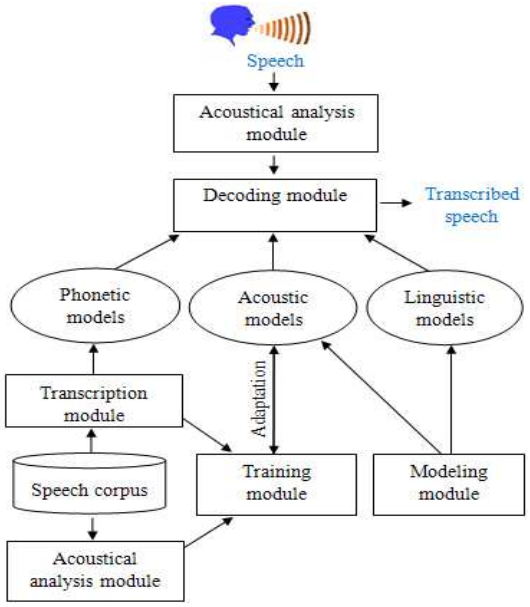

Fig. 1 Overview of the proposed ASR system

A. Acoustical analysis module

Acoustical analysis module includes pre-treatments such as recording, digitalization, pre-emphasis, blocking into frames, and frame windowing by using Hamming window. It includes also the extraction of features; performed to give an observation vector of the acoustic parameters for each frame.

This module is one of the most complex steps in the development of an ASR system. Thus, the acoustic parameters choice conditions the system performances. In order to guarantee enough informative observation vectors, we made experiments related to frame windowing (size and period) and feature extraction methods traditionally used in ASR such as MFCC (Mel-scale Frequency Cepstral Coefficients) and PLP (Perceptual Linear Prediction). A detailed discussion of these experiments is given later in section V.

B. Modeling module

This module includes both of linguistic and acoustic modeling modules.

For the first modeling process, we used a simple word grammar to describe the sequence of words successfully recognized by the system. This grammar can be depicted through network transitions as it is illustrated in Fig.2.

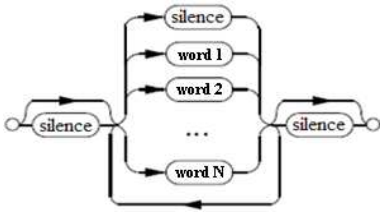

Fig. 2 Description of the grammar by network transitions

Concerning the acoustic modeling module, the choice of the speech recognition unit is very important.

In a first time, we used the phoneme as an acoustic unit (34 phonemes allow to describe a standard spoken Arabic). According to their performances and popularity [9], acoustic units are modeled by continuous-density HMM. To model phoneme, we choose a simple topology 'left-right' having three active states authorizing the looping to the current state and the passage to the following state. Indeed, the proposed topology is well-adapted in automatic continuous speech recognition [10].

In a second time, we developed a new acoustic modeling for Arabic language based on phoneme and diphoneme used to take into account the coarticulation's effects. We didn't use diphonemes as they are used by classic ASR systems. Indeed, they consider the diphoneme model as a phoneme that can be a consonant or a vowel followed by a single neighbor phoneme representing a consonant or a vowel. Based on the Arabic language specificities, the proposed diphoneme model is used to represent a consonant followed by a vowel (short or long) and the phoneme model is used to represent consonant located at the end of closed syllable. The proposed acoustic modeling for standard Arabic language generates 196 models.

To model diphoneme, we choose the same topology of phoneme model but with four active states. As it is defined, this model can be interpreted as the fusion result of two successive phoneme models; the state modeling the creation of the second phoneme coincides with the state modeling the realization of the first phoneme.

A comparative study between these two acoustic modeling is given later in section V. A silence model was also used to model non-speech acoustic artifacts.

C. Transcription module

For the transcription module, the speech recognition word vocabulary and the audio corpus are specified in terms of the basic recognition units. The first output of this module is an audio corpus which is orthographically and phonetically transcribed. The second output is a pronunciation dictionary containing phonetic models. The phonetic transcription is a work of interpretation which requires a scrupulous attention. As an Arabic word may be pronounced by various manners, according to its position in the sentence, its morphological variability, or simply according to the habits of speakers, we can integrate phonetic variants to relax the pronunciation and take into account the speech variations. Thus, every Arabic word could have several phonetic transcriptions in the pronunciation dictionary.

D. Training module

The training of acoustic models is realized under HTK toolkit by using embedded training method based on the Baum-Welch algorithm [11]. Several experiments were designed to evaluate the effect of varying the number of embedded re-estimations of the Baum-Welch algorithm and the effect of varying the number of Gaussian Mixtures.

E. Decoding module

Decoding module is also realized under HTK toolkit [11]. Decoding is controlled by a recognition network deduced from the grammar, the pronunciation dictionary and the acoustic models.

This network can model a set of linguistic constraints by which the recognition will be guided. It is composed by a set of nodes, which represent words, connected by arcs. Each node is itself a network denoting the phonetic model which is composed by the phonetic units modeled by HMMs. Thus, once fully compiled, recognition network ultimately consists of HMM states connected by transitions. This hierarchy is illustrated in Fig. 3 which exposes three different levels: word, model and state.

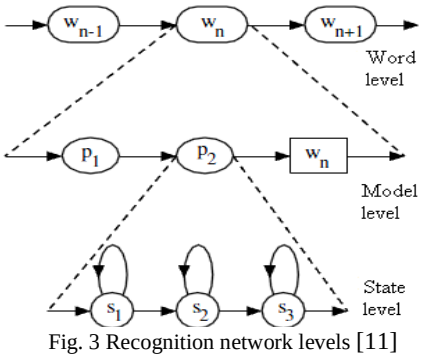

Fig. 3 Recognition network levels [11]

For an unknown input utterance with T frames, every path from the start node to the exit node of the network which passes through exactly T emitting HMM states is a potential recognition hypothesis. The role of the decoder is to assign a probability for each of these paths and to find, through the network, paths that have highest probabilities. This process is provided by the Viterbi algorithm.

IV. CORPORA

We have evaluated the performances of the proposed ASR system on two SA corpora.

The first corpus, collected by ourselves, is a multi-speaker (i.e., the same set of speakers was used in both of the training and testing phases) connected-digits (sequence of 1 to 10 digits) database. It comprised a small vocabulary of ten digits (from zero to nine). The training data, spoken by 41 speakers (18 males and 23 females), contains 513 connected-digits utterances. The test data was spoken by 24 speakers (8 males and 9 females) including 17 speakers having participated in the training data construction and 7 speakers not involved in the training data construction. The 105 connected-digits utterances formed the test data. Our corpus was recorded in a normal office environment and sampled at 16 kHz sampling rate and digitized to 16 bit resolution. The phonetic transcriptions associated to the audio data were realized on the basis of Arabic phonemes and well checked to be reflected in best the acoustic context.

The second corpus, already defined by R. Ejbali et al. [5], is a speaker-independent (i.e., speakers used for the training phase are different from those used for the testing phase) continue speech database. It contains the pronunciation of 20 lists. Each list consists of 10 phonetically balanced Arabic sentences [12]. Training data which is about 1 hour and 10 minutes was spoken by 13 speakers (7 males and 6 females). The test data which is about 7 minutes was spoken by 2 speakers (1 male and 1 female) not involved in the training data construction. All data are sampled at 16 kHz sampling rate and digitized to 16 bit resolution. Each audio file is associated with a transcription text file.

V. EXPERIMENTS AND RESULTS

A. Evaluation Criteria

The proposed system performances are evaluated by the recognition percentage defined by the following formula:

$$\%\ \text{Recognition} = (N - O - S - I)/N * 100 \quad (1)$$

where O, S, I, N are respectively deletions, insertions, substitutions and the total number of speech units of the reference transcription.

B. Acoustic Analysis experiments

The acoustic analysis module is evaluated against various points such as the choice of the acoustic analysis method, the number of acoustic parameters describing each frame, the size and the period of frame windowing.

In these first series of experiments, the test was performed by training continuous-density single Gaussian Mixture (GM) phoneme models using the connected-digits corpus collected for SA language.

Training acoustic models is a key step in any ASR system. That's why, results were observed according to the number of embedded re-estimations of the Baum-Welch Algorithm (up to 20 embedded re-estimations).

1. Effect of varying frame windowing

The first experiment has been conducted to examine the effect of varying the size and period of frame windowing on phoneme recognition performance. Each frame was represented by 12 acoustic parameters augmented by the corresponding delta and delta-delta coefficients.

According to the MFCC (respectively PLP) coefficients, we expose in Table I (respectively Table II), the number of embedded re-estimations of the Baum-Welch Algorithm for which the phonemes recognition rate is maximum for each test.

TABLE I
VARIATION EFFECT OF THE FRAME WINDOWING ON PHONEMES RECOGNITION BY USING MFCC COEFFICIENTS

Frame Windowing	% Phoneme recognition	Number of embedded re-estimations
20 ms every 10 ms	90.23	20
25 ms every 10 ms	91.51	18
30 ms every 15 ms	92.61	20
34 ms every 16 ms	93.01	20
38 ms every 18 ms	**93.68**	**14**
42 ms every 20 ms	93.32	16

Using either MFCC or PLP features extraction method, Table I and Table II show that the best phonemes recognition rate is reached by applying 38 ms window size every 18 ms.

TABLE II
VARIATION EFFECT OF THE FRAME WINDOWING ON PHONEMES RECOGNITION BY USING PLP COEFFICIENTS

Frame Windowing	% Phoneme recognition	Number of embedded re-estimations
20 ms every 10 ms	89.78	18
25 ms every 10 ms	89.42	20
30 ms every 15 ms	93.18	18
34 ms every 16 ms	93.58	18
38 ms every 18 ms	**96.34**	**10**
42 ms every 20 ms	96.27	8

2. Effect of varying the number of acoustic parameters

The second experiment has been conducted to examine the effect of varying the number of acoustic parameters on phoneme recognition performance. Based on the previous experimental results, frame windowing was characterized by 38 ms window size every 18 ms in this experiment and each frame was represented by acoustic parameters augmented by the corresponding delta and delta-delta coefficients.

Fig. 4 (respectively Fig. 5) shows the phoneme recognition rate against the embedded re-estimation number of the Baum Welch algorithm according to MFCC (respectively PLP) coefficients. Fig. 4 illustrates a clear superiority of the curve representing the test using 16 MFCC coefficients achieving 96.5% phonemes recognition for the 18 embedded re-estimations of the Baum-Welch Algorithm. Fig. 5 shows that the maximal phoneme recognition rate is reached by using 16 PLP coefficients and 16 embedded re-estimations.

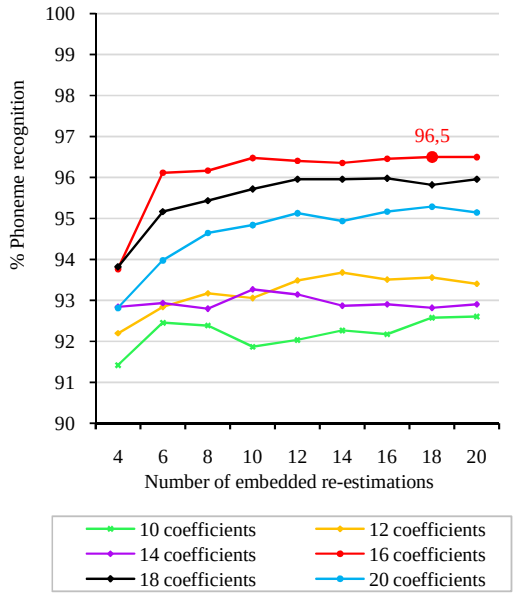

Fig. 4 Variation of phoneme recognition rate according to MFCC coefficient number and embedded re-estimation number

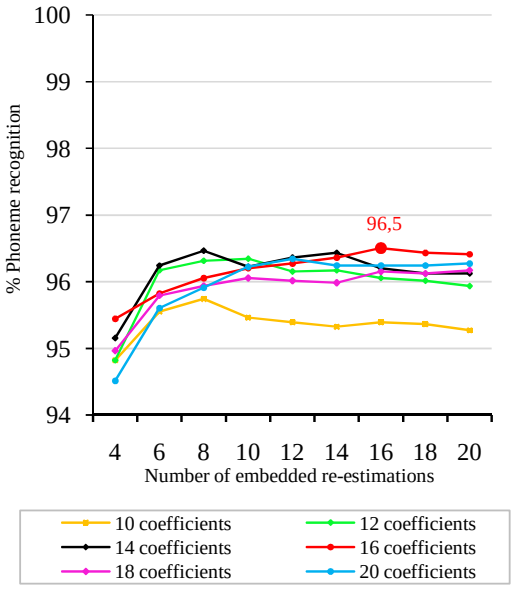

Fig. 5 Variation of phoneme recognition rate according to PLP coefficient number and embedded re-estimation number

3. Combination of MFCC and PLP coefficients with energy

In this experiment, we examined the effect of combining the MFCC and the PLP coefficients with the normalized energy. The combination was made by a simple concatenation of acoustic parameters. Results presented in Table III show that the maximal phoneme recognition rate is obtained by combining 16 PLP coefficients with energy and their corresponding delta and delta-delta coefficients and by using only 12 embedded re-estimations.

TABLE III
COMBINATION EFFECT OF ACOUSTIC PARAMETERS WITH ENERGY ON PHONEME RECOGNITION

Acoustic parameters	% Phoneme recognition	Number of embedded re-estimations
16 PLP+Energy+Δ'+ Δ"	96.5	12
16 MFCC+Energy+Δ'+ Δ"	92.15	14

Based on the previous acoustic analysis experimental results, we used in the following experiments, as the best features representing every frame of the speech signal, 16 PLP coefficients and even the normalized energy coefficient augmented by the corresponding delta and delta-delta coefficients.

C. Acoustic Modeling experiments

In these second series of experiments, the test was performed by training acoustic models using the connected-digits corpus collected for SA language.

1. Effect of varying speech recognition unit

In this experiment, we compared between two acoustic modeling which are described in section III. Let's remind that, the first acoustic modeling is based on phoneme as speech recognition unit. The second acoustic modeling is based on two speech recognition unit: phoneme and diphoneme.

The test was performed by training continuous-density single GM acoustic models using up to 20 embedded re-estimations of the Baum-Welch Algorithm. The following table presents the best recognition rate and the number of embedded re-estimations made for each choice of speech recognition unit.

TABLE IV
EFFECT OF VARYING SPEECH RECOGNITION UNIT ON PHONEME RECOGNITION

Speech recognition unit	% Recognition	Number of embedded re-estimations
Phoneme	96.5	12
Phoneme and diphoneme	96	14

Results, given in Table IV, show that the choice of phoneme as speech recognition unit is better than the choice of a combined phoneme and diphoneme unit. These results could be interpreted by the insufficient speech data available for training acoustic models in the second choice of speech recognition unit.

2. Effect of varying Gaussian number per HMM state

Based on all the previous experiments, the overall recognition performance does not exceed 96.5%. This is due to the fact that the single GM HMMs were not able to provide a good parametric modeling of the acoustic space. Therefore, this experiment examined the effect of varying Gaussian number per HMM state. For each test, Gaussian number were split by a factor of 2 and HMMs parameters were estimated using up to 20 embedded re-estimations of the Baum-Welch algorithm.

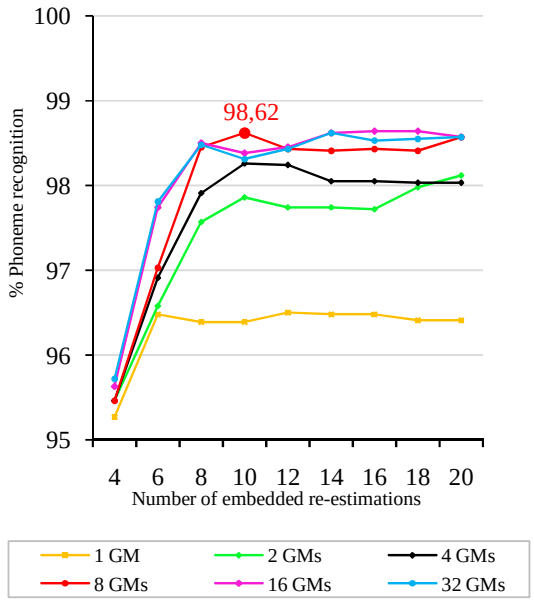

Fig. 6 Variation of phoneme recognition rate according to Gaussian number and embedded re-estimation number

Fig. 6 shows that the phoneme recognition performance increases with Gaussian number per state. We observed that the computational complexity increase exponentially with the Gaussian number.

Since the increase in performance from 8 to 32 Gaussians per HMM state did not compensate for the computational complexity, a decision was taken to use only 8 Gaussians. Hence, we concluded that by using 8 Gaussians and 10 embedded re-estimations of the Baum-Welch algorithm, satisfactory performance (98.62% phonemes recognition) can be achieved with reasonable computational complexity.

D.Experiment for continue SA speech recognition

In this section, we evaluated our phoneme-based ASR system on continue SA speech recognition. The test was performed by training acoustic models using the speaker-independent continue SA speech corpus collected by Ejbali *et al.* [5] and described in section IV. As we noted in section I, Ejbali *et al.* had developed a phoneme-based ASR system which was evaluated on the same corpus.

Table V compares the characteristics and the performances of these two systems. It shows that the phoneme recognition rate reached by our system is about 94% which is higher than that obtained with the Ejbali *et al.*system.

TABLE V
COMPARISON OF OUR SYSTEM WITH THE SYSTEM OF EJBALI *ET AL*

ASR system	Acoustic parameters	Number of Gaussian per HMM state	Number of embedded re-estimations	% Phoneme recognition
System of Ejbali *et al.*	12 PLP+ Energie+ Δ'+ Δ"	64	10	80.36
Our system	16 PLP+ Energie+ Δ'+ Δ"	8	10	**94.02**

VI. CONCLUSION AND PERSPECTIVES

The main contribution of this work is the proposition of a HMM-based ASR system suited for the SA language. The performance of this system has been evaluated using a speaker-dependent SA connected-digits corpus and a speaker-independent continue SA speech corpus. A well-established study was conducted to define the best parameters of a performant ASR system for SA language. For instance, the utilization of 16 PLP coefficients, combined with energy and their corresponding delta and delta-delta coefficients and extracted from each 38 ms frame size every 18 ms frame period, achieve the best informative acoustic parameters representing an audio frame. Moreover, the utilization of 8 Gaussians per HMM state and the application of 10 embedded re-estimations of the Baum-Welch algorithm improved the system's performance. Under these parameter definition, the phoneme recognition rate is about 98.62% using the SA connected-digits corpus and 94.02% using the continue SA speech corpus.

In a future work, we intend improving the proposed system by using a large vocabulary linguistic model which will be automatically generated from SA textual corpus. We will furthermore evaluate our system on a large vocabulary SA speech corpus.

REFERENCES

[1] M. Kabache, and M. Guerti, "Application des réseaux de neurones à la reconnaissance des phonèmes spécifiques à l'Arabe standard", *SETIT 2005*, Tunisia, March 2005.
[2] S.A. Selouani, and J. Caelen, "Recognition of phonetic features using neural networks and knowledge-based system: a comparative study", *International Journal on artificial intelligence tools*, world scientific publishing editors, vol. 8, no. 1, pp. 73–103, 1999.
[3] S. Hazmoune, F. Bougamouza, and M. Benmohammed, "La reconnaissance automatique de la parole par combinaison de classifieurs markoviens", *JEESI'09*, 2009.
[4] Y. A. Alotaibi, "Comparative Study of ANN and HMM to Arabic Digits Recognition Systems", *JKAU: Eng. Sci.*, vol. 19, no. 1, pp. 43–60, 2008.
[5] R. Ejbali, Y. Ben Ayed, and A. M. Alimi, "Arabic continues speech recognition system using context-independent", *Sixth International Multi-Conference on Systems, Signals & Devices*, Tunisia, March 2009.
[6] M. Elshafei, "Toward an Arabic text-to-speech system," *The Arabian Journal for Science and Engineering*, vol. 16, no. 4, pp. 565–583, 1991.
[7] M. Alkhouli, *Linguistic Phonetics*, Daar Alfalah, Swaileh, Jordan, 1990.
[8] F. Jelinek, "Continuous speech recognition by statistical methods", *Proc. IEEE*, vol. 64, pp. 532–556, 1976.
[9] L. R. Rabiner, "A Tutorial on Hidden Markov Models and selected applications in Speech Recognition", *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.
[10] F. Lefevre, *Estimation de probabilité non-paramétrique pour la reconnaissance Markovienne de la parole*, Pierre and Marie Curie University, Jan. 2000.
[11] S. Young *et al.*, *The HTK Book (for HTK Version 3.4)*, Cambridge University, March 2009.
[12] M. Boudraa, and B. Boudraa "Twenty list of ten arabic Sentences for Assessment", *ACUSTICA acta acoustica*. vol. 86, no. 43.71, pp. 870–882, Nov. 1998.