Strategies for Implementing an Optimal ASR System for Quranic Recitation Recognition

Mohamed O. M. Khelifa, Mostafa Belkasmi
TES Research Team
ENSIAS School of Engineering
Mohammed V University of RABAT
Morocco

Yahya O. Mohamed Elhadj
Doha Institute, Doha, Qatar &
SAMoVA Research Team
IRIT of Toulouse
Paul Sabatier University
France

Yousfi Abdellah
ERADIASS Research Team
FSJES of Souissi
Mohammed V University of RABAT
Morocco

ABSTRACT
With the help of automatic speech recognition (ASR) techniques, computers become capable of recognizing speech. The Quran is the speech of Allah (The God); it is the Holy book for all Muslims in the world; it is written and recited in Classical Arabic language, the language in which it was revealed by Allah to the Prophet Muhammad. Knowing how to pronounce correctly the Quranic sounds and correct mistakes occurred in reading is one of the most important topics in Quranic ASR applications, which assist self-learning, memorizing and checking the Holy Quran recitations. This paper presents a practical framework for development and implementation of an optimal ASR system for Quranic sounds recognition. The system uses the statistical approach of Hidden Markov Models (HMMs) for modeling the Quranic sounds and the Cambridge HTK tools as a development environment. Since sounds duration is regarded as a distinguishing factor in Quranic recitation and discrimination between certain Quranic sounds relies heavily on their durations, we have proposed and tested various strategies for modeling the Quranic sounds' durations in order to increase the ability in distinguishing them properly and thus enhancing their overall recognition accuracy. Experiments have been carried out on a particular Quranic Corpus containing ten male speakers and more than eight hours of speech collected from recitations of the Holy Quran. The implemented system reached (99%) as average recognition rate; which reflects its robustness and performance.

General Terms
Automatic Speech Recognition (ASR), Computational Linguistics (CL), Artificial Intelligence (AI).

Keywords
Quranic recitation, Quranic sounds, Classical Arabic Language, Hidden Markov models, Hidden semi-Markov Models, Duration modeling.

1. INTRODUCTION
Speech is one of the most vital and natural means of human communication. Thoughts and humans' ideas are exchanged through speech. The purpose of an automatic speech recognition system is enabling a computer to recognize and act on a natural human language, where utterances uttered by a person, using particular algorithms. The matching techniques which are the basis of an ASR system consist of comparing a sound wave to a set of its samples usually compounds of phonemes, that are a basic linguistic unit in a language. An ordinary ASR system usually consists of a microphone unit, a voice recognition engine, a computer and some form of audio/visual /action output. ASR is a branch of artificial intelligence (AI) and is federated to various fields of knowledge, including computer science, computational linguistics and pattern recognition [1], [2]. Research efforts in ASR field have fascinated the public and many scientists around the globe. In its infancy, the expectations on its applications were very optimistic: what more natural than talking directly into computers without having troubles caused by keyboard manipulation? Unfortunately, despite the enormous advances in computer technologies, ASR field remains a topic of an active research and results obtained still far from the perfect [3]. However, if a perfect ASR system does not yet exist, concrete applications are emerging gradually. ASR systems have started equipping mobile phones or GPS by identifying certain keywords to accomplish the desired mission; various IT applications and IT-solutions have appeared as automatic translation systems (ATS), handicapped people's help, speakers and languages identification, authentication and information retrieval. By using speech as input, ASR applications reduce the use of traditional manual input techniques via keyboards and mouse, which render them useful as an alternative input technique for people with disabilities. In spite of the extensive use of ASR technologies in foreign languages, the Arabic language still suffers from the rarity of efficient ASR applications, especially for learning and evaluation purposes. Noting that the enormous progress that has been achieved in ASR framework is made by the use of statistical approach, namely Hidden Markov Models (HMM), which is the most predominant technique in this field [4].

One prominent application of Arabic ASR is the teaching of the holy Quranic sounds system and this done by following the learner through reading with a professional manner and correcting his mistakes. Quranic recitation has a set of rules known as Tajweed to ensure its correct pronunciation and readings. The Quran has been traditionally taught by Sheikhs (skilled religious teachers). Usually, these teachers will listen to the learners' recitation and noted their mistakes. However, the traditional method which requires the presence of these skilled teachers has barriers to support self-learning environment; these systems can play the role of Sheikhs in their absence; this surely helps self-learning and memorization the holy Quran and checking the probity of Tajweed by avoiding errors in reading. Moreover, this can open the door for various classes of Islamic and Arabic applications. Islam is the second largest religion in the world in terms of the number of believers; their number reached a Billion and a half people, representing more than 23% of the world's population. This makes the development of this genre of applications of paramount importance in view of the numerous targets and the need to meet their demands.

This paper summarizes the research efforts that we have undertaken which aimed at building an accurate and robust ASR-based system for Quranic sounds recognition. The first stage of these efforts has devoted to the development of a baseline HMM-based system for basic Quranic sounds. The second one aimed at enhancing the performance of this baseline system by addressing various strategies of sounds' durations modeling, in this context, an explicit duration model is proposed instead of the implicit one of the standard Markov. The role of this proposed explicit model is to enhance the durational behavior of the standard HMM model. Three different distributions were tested in modeling the sounds' durations; those are the Gaussian, the Gamma, and the Poisson distributions. We presented in this paper a practical framework for implementing an optimal system for Quranic recitation recognition with the help of speech recognition techniques to assist self-learning and memorization of the holy Quran. The proposed system is able of recognizing the Quranic sounds with a high accuracy. The system adopts the MFCC algorithm for feature extraction and both HMM/GMM & HSMM/GMM models for feature classification.

The rest of the paper is organized as follows: section 2 gives a brief description of the literature review; section 3 describes the Quranic speech corpus used in experiments, section 4 presents both baseline system and the proposed strategies for its improvement, experimental procedures and obtained results are listed in sections 5 and 6. The paper ended by giving a conclusion

2. LITERATURE REVIEW

Quran is the holy book for Muslims; it is the focus of every Muslim in the world. Practicing five daily prayers is one of the five pillars of Islam and every prayer involves certain Quran recitation. Hence, every Muslim is involved in certain Quran memorization in order to recite some Ayat ("sentences" in Quran) during his/her daily prayers. The Quran was revealed in Classical Arabic language from Allah (the God) to Prophet Muhammad through Angel Gabriel. Quran describes all things about Muslims day life and plays a central role in it. The Arabic language is one of the world's major languages with over 400 million people in various Arab countries who use it as a mother tongue. More than a billion and a half Muslims in places such as India, Pakistan, Indonesia, and Tanzania study Arabic as a foreign or second language for religious purposes. The Arabic language can be categorized into two main variants: Classical Arabic (CA) and Modern Standard Arabic (MSA). The CA is an old literary form of Arabic, which is the most formal type and is the language of the Holy Quran and the old Arabic poetry. The MSA is the current standard form of Arabic, which is utilized in official communications in Arabic countries, broadcast news, formal speeches, etc. Although there's no huge difference between today's Arabic and that spoken by the early Arabs, due to the fact that Arabic is one of the most stable languages throughout history, yet there are some idiosyncrasies as to the way of pronunciation. Teaching how to pronounce correctly the Quranic sounds and correcting mistakes occurred while reading them, is an important and central topic in Islamic applications involving ASR technologies. Reading the Quran is not like other readings, it must be subject to the rules of Tajweed (the rules of perfect reading of the holy Quran), although reading varies across reciters (speakers), perfect application of Tajweed rules during the Holy Quran recitation may lead to a great similarity of pronunciations. Many research efforts have been offered addressing the development of ASR applications for Quranic recitation and correction of mistakes in reading. In [5], the authors have proposed an ASR system for the purpose of helping the students' self-learning of the holy Quran, they confirmed that their systems capable of checking and pointed out the mismatch between the students' recitation and that made by the experienced teachers stored in their speech database, however, they have presented no information about the level of accuracy achieved and the amount of audio data used in their experiments. In [6], the authors proposed the design of a J-QAF learning system for automated Tajweed rules checking based on speech recognition techniques. They proved the accuracy of their system in meeting the learners' needs with an acceptable level of accuracy. In [7], the authors proposed a virtual learning system (Miqra'ah) for Quran recitations for sighted and blind students, their system has administrators which create several virtual learning rooms and register the licensed scientists, the learner can interact directly by voice with the teacher for correcting his errors, however, their system has limitation in term of the recognition accuracy achieved and the ability of correcting the bugs occurred in reading. In [8], the authors investigated the building of a Quranic reader controlled by speech. Their system is based on the open source CMU Sphinx toolkit, which represents an HMM speech recognition toolkit built for the English language, and they tuned it to support Arabic, they have collected a Quranic speech corpus contained some verses to train and test their statistical models. They affirmed the encouragement of the results obtained while regarding the amount of audio data used.

3. OVERVIEW OF THE QURANIC SPEECH CORPUS

Arabic is one of the most widely spoken languages in the world. Statistics show that it is the first language (mother tongue) of more than 400 million native speakers ranked fourth after Mandarin, Spanish and English. It is the religious language of all Muslims, more than one billion and a half Muslims in the globe uses it as a religious language, regardless of their native language. The Quran is revealed in Classical Arabic language which is the old literary form of Arabic. Despite its great importance, the research effort on speech recognition technologies for the Arabic language unfortunately remains insufficient. It should be noted that a lack of Arabic speech corpus, collected in a qualitative and professional way, is observable. This may explain why ASR technologies in Arabic are not yet so developed in comparison with other languages of similar importance, such as English. This lack is seriously encountered when dealing with classical Arabic language since most Corpora (corpuses) currently available are specifically oriented towards what is known as modern standard Arabic (MSA) and its dialects.

To address this issue and in order to contribute to the development of ASR applications for the Quranic recitation and other Islamic ASR applications, the Islamic University of Imam Muhammad ibn Saud (in Kingdom of Saudi Arabia) has developed a speech corpus for the Holy Quran. It is the first of its kind, to the best of our knowledge. This corpus [10] is built as a first stage towards developing a speech recognition tool for the purpose to assist self-learning and memorization of Holy Quran. The corpus is part of a research project [11] aimed at developing a computerized environment for learning the Holy Quran and its sciences, namely Hadith and Fiqh (Hadith: Prophet Mohammed's speech explaining Quran, Fiqh: Islamic jurisprudence). The said corpus contains more than eight hours of speech collected from the Quranic recitation of ten different chosen male reciters (speakers)

reciting the thirtieth part of the holy Quran with regard to Tajweed rules. Due to the difficulty of developing this kind of Corpora, only a part of the Holy Quran was considered. The rules of Tajweed are the rules of perfect reading of the Holy Quran. Table 1 shows for each speaker the number of audio files, their volumes and their durations.

Table 1: Sound files and their duration by reciters

Reciter Number	Reciter Initials	Sound Files Numbers	Duration (minutes)	Size (MB)
1	AAH	600	49.36	249
2	AAS	590	52.09	261
3	AMS	612	45.78	229
4	ANS	597	49.72	250
5	BAN	585	54.75	276
6	FFA	578	44.11	220
7	HSS	601	49.76	251
8	MAS	580	46.24	232
9	MAZ	608	51.47	258
10	SKG	584	44.29	220
Total		**5935**	**487.53 (8h, 8m)**	**2446**

The corpus speech signals were segmented manually and precisely in three levels: words, phonemes, and allophones. A labeling system has been proposed to annotate the speech segments of because the available labeling systems are inappropriate because they do not cover the sounds of classical Arabic. In this paper, we limit ourselves to the level of the phoneme; the investigation in the level of Allophones will be subject to another publication. The corpus contains total of (60) phonemic units and their geminated parts, for which a fictitious unit has been added to designate the silence independently of its occurred place.

Arabic has (34) phonemes: (28) consonants, three long vowels, and three short vowels. The difference between short and long vowels resides in the duration of their acoustic realization. In fact, the phonetic system of Arabic is characterized by the relevance of the elongation in the vocalic system and by the presence of geminated consonants. Short vowels are indicated by diacritical marks, which are short lines placed above or below the preceding consonant. Another diacritic is also used to represent the doubling of the consonant (gemination), which is also an important characterization of Arabic. Diacritical marks are very important for correct pronunciation and for phonetic transcription due to the fact that Arabic has a strong grapheme to phoneme dependency. However, the Arabic texts are generally not diacterized but this is not the case for the Holy Quran. Noting here that the said Quranic speech corpus was used in a previously funded project, for the development of two baselines HMM-based systems for phonemes and allophones respectively [12- 13]. These baselines systems were built for the purpose to examine the feasibility of developing such a recognition system with an acceptable level of accuracy. The authors reported in their published papers [12-13] that the average recognition rates obtained for these recognizers were respectively 92% and 88% for phonemes and allophones. In a previous stage of this work [14], we have adapted the Quranic corpus for the aim to be annotated in term of basic Quranic sounds. We mean by the basic Quranic sounds, the basic phonemes (single phonemes) without any phonological variation and even considering neither the phonemes geminating nor others. In this context, the new version of the corpus has been utilized in various research efforts, which includes an enhanced Arabic phonemes recognizer using duration modeling techniques [16], an accurate HSMM-based system for Arabic phonemes recognition [17] and a helpful statistics in recognizing basic Arabic phonemes [18]. Table 2 lists the basic Quranic sounds employed in the present work.

Table 2: List of basic Quranic sounds and their codes

Arabic Letters	Label	Arabic Letters	Label
فتحة ـَ	as10	صاد ص	sb10
ضمة ـُ	us10	ضاد ض	db10
كسرة ـِ	is10	طاء ط	tb10
همزة ء	hz10	ظاء ظ	zb10
باء ب	bs10	عين ع	cs10
تاء ت	ts10	غين غ	gs10
ثاء ث	vs10	فاء ف	fs10
جيم ج	jb10	قاف ق	qs10
حاء ح	hb10	كاف ك	ks10
خاء خ	xs10	لام ل	ls10
دال د	ds10	ميم م	ms10
ذال ذ	vb10	نون ن	ns10
راء ر	rs10	هاء هـ	hs10
زاء ز	zs10	واو و	ws10
سين س	ss10	ياء ي	ys10
شين ش	js10	صامت	Sil

4. SECTIONS

Our efforts towards developing an optimal ASR system for Quranic sounds recognition have started by building a baseline system for the basic Quranic sounds and then the focus focalized on how improving its performance and robustness. The baseline system is built on the basis of the statistical approach of Hidden Markov Model (HMM) technique. Especially, a left-to-right HMM topology and continuous Gaussian mixture models were utilized. The HMM is the central model widely used in the development of speech recognizers, its internal structure doesn't come from any knowledge of speech. Therefore, the use of HMMs in speech recognition limited in computing quantities related to the speech (a computation model). The system is built for the purpose to be able of accurately recognizing the basic sounds of the Holy Quran. The system uses the previously mentioned speech corpus to train and test statistical models, and the

Cambridge Hidden Markov Model Toolkit (HTK) [18] as a development environment, (HTK) is software written in C programming language which allows building and testing HMM-based systems, it is widely used by the scientific committee in this field. Each Quranic sound was modeled by an HMM with three emitting states to capture its acoustic properties. A Gaussian Mixture Models (GMMs of 16 laws) with diagonal covariance matrices was also associated with each state of the mentioned HMM acoustic model for the purpose to identify the characteristics of the Quranic sound portion at this state. The cepstral features vectors were used; for each Hamming window of 8ms, we have extracted a vector of 39 acoustical. These coefficients are the first twelve MFCCs (Mel Frequency Cepstral Coefficients) plus their first and second derivatives to capture the static characteristics of the Quranic sound portion, which results in a vector dimension of 36. In order to take into account the dynamic characteristics signals, the energy feature plus its first and second derivatives are also extracted and appended to the MFCCs. the results obtained through this baseline system give an excellent recognition accuracy which reached (98%) as average recognition rates for all reciters using 16 GMM of Laws. However, an in-depth analysis of the reported results shows that we still have a considerable confusion between certain Quranic sounds. In order to overcome this misrecognition, we decided to investigate various strategies for modeling the Quranic sounds' duration for the purpose to render the system capable of distinguishing between them accurately and this surely will improve their overall recognition accuracy. In certain languages as Classical Arabic (The language of the Holy Quran); sounds' duration regarded as a distinguishing cue in Quranic phonology. Phonological variation of sounds' occurrences yields to mistakes in their pronunciation and this affects negatively on the system performance. Thus an accurate modeling of sounds' duration can be an essential topic. In literature, the successful applicability of HMMs to various aspects of speech modeling has been approved in various experiments in recent years. These investigations are all based on the assumption that speech signal is a quasi-stationary process whose static intervals can be described by the residence time of a single state of a standard HMM. The duration of phonemes plays an important role in the recognition and understanding of speech. Information on duration is generally ignored in ASR devices due to the use of Markov models (HMMs) that are unable to correctly model the phonemic durations of sounds. Previous studies have shown that the use of different approaches for modeling sounds' duration has remarkably improved the performance of RAP systems. Consider a standard HMM model. In this model, the duration of staying in a state is implicitly controlled (by default) by the probabilities of transitions between states. It is assumed that the transition probabilities are invariable over time. We can, therefore, indicate a_{ii} the probability of looping on the state q_i (self-transition) and therefore $(1 - a_{ii})$ would be the probability of passing to other different HMM states. From this, it can be deduced that the probability distribution of the durations passed in each HMM state is of the form: $p_i(\tau) = {a_{ii}}^{\tau-1}(1 - a_{ii})$. This is a decreasing exponential distribution. It gains its maximum value at the minimum duration $\tau = 1$ and decreases exponentially as τ increases. It has been found that the said distribution can only efficiently describe the average duration. Beyond that, it is unable to model the variations produced in the duration distributions. This rather weak modeling of the state duration is considered one of the major weaknesses of Markov modeling, which subsequently affects the performance of RAP systems that employ them. In literature, one of the solutions proposed to this problem is to explicitly integrate the probability distributions of the state duration in the HMMs models. The resulting model is known as the Hidden Semi-Markov Model (HSMM). Thus, a HMMM is an extended version of the standard HMM model where the residence time in an HMM state can follow any law, unlike the Markovian case, where one is constrained to be an exponential or geometric law. For this reason, the HMSMs models are very general and well adapted to the various applications. In this work, we have implemented an HSMM model which uses three different strategies for modeling the state durations of standard HMM. This is made by using three distributions, namely: Gamma, Gaussian and Poisson. Their mathematical formulas used to estimate the probability of the duration of the state are as follows:

For Gamma distribution:

$p(\tau|\ i) = \frac{{\eta_i}^{v_i}\ \tau^{v_i-1}e^{-\eta_i\tau}}{\Gamma(v_i)}$; where η_i and v_i are parameters of the Gamma distribution having a mean of $\mu_i = v_i{\eta_i}^{-1}$ and a variance $\acute{o}^2_i = v_i{\eta_i}^{-2}, \tau$ denotes the spent duration in the state i and follows this distribution in this investigated case.

For Gaussian distribution:

$p(\tau|\ i) = \frac{1}{\acute{o}_i(2\pi)^{1\backslash 2}}e^{\frac{-(\tau-m_i)^2}{2\acute{o}^2_i}}$; where m_i and $\acute{o}_i$ are the mean and variance of the Gaussian distribution in state i, τ denotes the spent duration in the state i and follows this distribution in this investigated case.

For Poisson distribution:

$p(\tau|\ i) = e^{-k_i}\frac{{k_i}^{\tau}}{\tau!}$; where m_i and $\acute{o}_i$ are the mean and variance of the Gaussian distribution in state i, τ denotes the spent duration in the state i and follows this distribution in this investigated case. The duration spent is stage has an expected value representing the one parameter of the Poisson density.

For each investigated distribution (Gaussian, Gamma and Poisson), the mean and variance of the state duration were estimated for each state in each HSMM model with the optimal state sequence for each training utterance. Their estimation was performed via the Viterbi algorithm after the HMMs were well trained. The standard algorithms of BaumWelch and Viterbi have been modified to incorporate the implemented explicit model into the training and decoding processes of HMMs. This incorporation was implemented in the Cambridge (HTK) tools by adjusting its library functions for BaumWelch and Viterbi. These changes gave rise to a new version of HTK in which the explicit model is integrated; this will be more explained in the experimental section.

5. EXPERIMENTS AND EVALUATION

It should be mentioned that the methodology used to test these various strategies for Quranic sounds' modeling is the same as that of the baseline system with the exception of the use of an explicit duration model (HSMM) instead of implicit one of the ordinary HMM. As we evoked earlier, the BaumWelch and Viterbi library functions in Cambridge HTK tools were modified to integrate the explicit duration model. This means that the HTK libraries functions like (HInit, HRest, HERest and HVite) were modified and the new ones were built for incorporating the explicit model into the HMMs' training and decoding processes. Since this is an explicit model, the diagonal elements of the transition Matrix will be automatically rounded to zero because self-transition is not

allowed in HSMM modeling framework. Thus, the three investigated strategies presented in this paper have given birth to three new systems. For clarity reason; we denote (S0) the baseline system that uses the standard HMMs, in other words, this system uses the implicit geometrical distribution in modeling Quranic sounds' state duration. We denote (S1) the newly implemented system that uses an explicit modeling framework based on Gamma distribution, this means that the duration of Quranic sounds follows Gamma distribution in this case. We denote (S2) the newly implemented system that uses an explicit modeling framework based on Gaussian distribution, this means that the duration of Quranic sounds follows Gaussian distribution in this case. We denote (S3) the newly implemented system that uses an explicit modeling framework based on Poisson distribution, this means that the duration of Quranic sounds follows Poisson distribution in this case. These newly implemented systems were trained and tested using the new version of HTK tools.

The Quranic speech corpus used consists of 5935 audio files over its corresponding MFCCs files, label files, TextGrids files and text files containing the corresponding text (Quranic Ayat or part of it). In addition, it contains an Arabic dictionary, a list of all unique words included in the whole corpus. All Quranic sound models have three emitting states and for each state, the conversion from the original waveform to a series of acoustical vectors is performed with the HTK tool "HCopy". Considering the HMMs' initialization and training, we have employed the following combination of HTK tools "HInit + HRest + HERest". This combination has been proved in previous works as the best HTK's training tools. So, all the experiments were carried out using this combination. Considering the suitable number of Gaussians Mixtures (GMMs), we performed various experiments varying their number from 1 to 16 GMMs on particular sets of training and testing defined as follows: As we have ten reciters, we divided the speech corpus into ten groups of training and testing sets; they are all employed in the experimentations, one at a time, and then a global average is computed. For each group, we consider a particular reciter to construct the testing set by extracting the first Ayat of each Surat from it; the remainders of Ayahs of this reciter, as well as all Ayahs of the others are used for training. Thus, about 93% of the speech corpus is used for training and 7% for testing. It should be mentioned here that there is nothing in the literature indicating which number of GMMs is the best for a given context, and thus their optimal number has to be determined by experiments.

Table 3: Flat language model for Quranic sounds

```
$Phon = as10 | bs10 | cs10 | db10 | ds10 | fs10 | gs10 | hb10 |
hs10 | hz10 |is10 | jb10 | js10 | ks10 | ls10 | ms10 |ns10 | qs10
| rs10 | sb10 | sil | ss10 | tb10 | ts10 | us10 | vb10 | vs10 | ws10
| xs10 | ys10 | zb10 | zs10 ;

( < $Phon > )
```

Considering the process of recognition, we have utilized the following flat language model (see Table 3) to allow all pronunciation possibilities (any Quranic sounds can appear after any other one).This format is converted to an internal HTK representation by the tool "HParse". The results obtained through experiments are reported in Table IV and depicted in Fig. 1 to be more readable and analyzable. For comparison purpose, we reported the results in the same table with those of the baseline system.

6. RESULTS AND DISCUSSION

Table 4: Average recognition rates for 1 to 16 GMM Laws (%)

GMMs	Systems	AAH	AAS	AMS	ANS	BAN	FFA	HSS	MAS	MAZ	SKG
1	S0	89.23	89.69	81.12	87.23	85.90	84.04	90.08	86.97	82.85	81.25
	S1	91.40	90.82	82.18	88.63	86.94	85.88	90.21	88.30	84.64	81.91
	S2	91.13	90.24	81.90	87.10	86.15	85.20	90.11	87.27	83.93	81.15
	S3	88.10	87.54	79.76	88.43	82.39	85.86	91.48	84.13	81.82	78.29
2	S0	93.88	92.04	88.43	90.16	93.35	86.97	94.39	94.28	92.42	90.03
	S1	94.15	92.69	88.56	90.82	93.95	88.18	95.86	90.96	93.12	90.82
	S2	93.90	92.10	88.16	90.11	92.12	87.22	95.34	90.64	92.87	90.31
	S3	92.76	90.79	89.26	83.64	90.97	84.36	94.80	88.17	92.69	87.29
4	S0	97.07	97.78	90.29	96.01	96.14	95.33	97.65	95.21	95.35	95.21
	S1	97.61	97.65	90.16	96.48	97.22	95.81	98.33	96.30	94.28	96.37
	S2	97.21	97.52	90.45	96.87	96.89	95.70	98.21	96.65	94.35	96.60
	S3	96.14	95.72	88.79	94.80	94.88	91.82	96.44	90.74	94.58	94.37
6	S0	97.87	98.43	94.55	97.07	96.54	95.48	98.17	96.41	96.81	97.07
	S1	97.47	98.86	95.08	97.47	97.68	95.54	98.14	97.44	97.43	97.98
	S2	96.97	98.19	95.23	97.14	97.42	95.11	98.37	97.05	97.20	97.04
	S3	96.34	96.87	93.80	96.27	92.70	95.81	96.63	94.83	93.54	95.64
8	S0	98.01	98.83	95.35	97.61	98.14	96.14	98.04	96.01	96.41	96.94
	S1	98.20	98.70	96.11	98.31	98.73	97.76	98.17	97.81	97.72	97.28
	S2	97.83	98.29	95.91	97.98	98.68	97.94	97.67	97.39	97.49	97.08
	S3	96.46	93.53	95.47	95.86	94.40	94.85	98.86	94.63	93.64	94.64
10	S0	97.74	99.09	95.35	97.87	97.87	95.88	98.96	96.94	97.21	97.74
	S1	97.89	99.09	96.94	98.67	97.61	96.56	99.15	97.20	98.13	97.21
	S2	97.32	98.76	95.72	97.59	97.93	96.18	98.28	96.36	97.73	97.72
	S3	95.74	97.85	93.94	96.25	94.75	93.70	96.04	93.93	95.36	95.60
16	S0	98.54	99.48	96.81	98.67	97.61	97.47	98.43	97.07	98.27	96.94
	S1	99.61	99.65	98.12	99.17	98.11	99.27	99.29	98.22	98.79	98.14
	S2	99.46	99.16	97.74	98.49	98.36	98.73	98.48	97.36	97.66	97.35
	S3	95.75	96.94	95.72	96.06	94.89	97.19	96.79	96.95	98.07	97.93

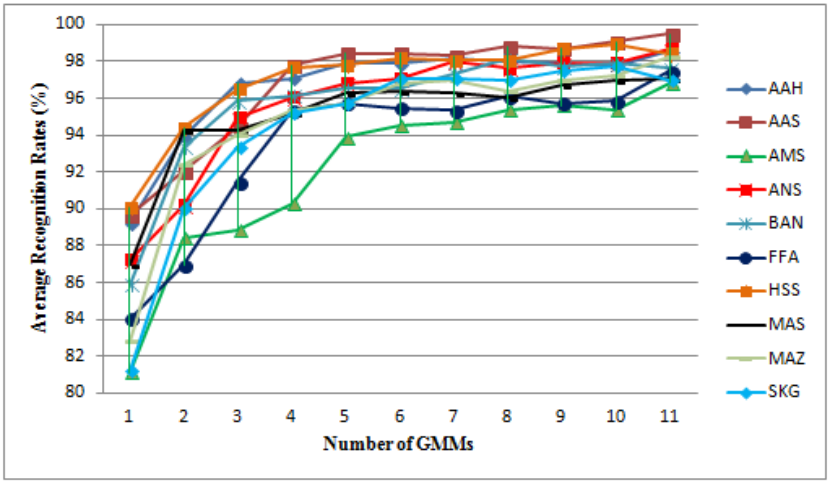

Fig. 1: Average Recognition Rates for GMM Laws

From the results seen above, we clearly discovered the success of the proposed strategies for improving the whole system performance. Both S1 and S2 systems outperform the baseline S0 system in term of global recognition accuracy, while system S3 underperforms the baseline one. By examining the reported results in detail, we observed that an improvement in performance of about 1.5% was reached comparing S0 and S1 in term of the global recognition accuracy; while comparing S0 and S2, this rate was only around 1%. When checking results related to system S3, we found it slightly-underperform that of the baseline system, and this automatically indicates it's weakest in performance comparing to other implemented systems.

The increasements observed in accuracy covered all reciters in the whole corpus, for example, using one GMM Law, for the reciter AMS, the lowest recognition rate passed from 81.12% in system S0 to 82.18% in system S1. While in the case of AAH, this percentage passed from 89.23 in S0 to 91.40 in S1. These improvements were respectively 1% and 2%. It has to be noted that the increasements rate varies across reciters, in some of them; it has to be approaching to 2% in three reciters; while remaining in 1% in all others. These observed variations in improvement may be explained by the fact that some reciters have a speaking rate higher or lower than others and this affects the explicit model proposed which his influence cannot be identical to all reciters. Using 16 GMM Laws, the lowest recognition rate passed from 96.81% in S0 to 98.11% in S1, while the global average recognition rate is increased from 98% to 99% comparing S0 and S1. In spite of its significant role, the explicit model proposed here cannot enhance the system performance as much as expected. Moreover, the reported results show clearly the need of giving special focus to the whole sounds' duration modeling not states. It should also be remarked that the use of GMM models leads to a significant improvement due to their capability of neutralizing and separating the sounds' characteristics, their optimal numbers may depend on the model parameters and the amount of training audio data used. In this application, between 7 and 10 GMM Laws seem to be enough.

7. CONCLUSION

In this article, we have summarized the research efforts that we have undertaken which aimed at building an accurate and robust ASR-based system for Quranic sounds recognition. The first stage of these efforts has devoted to the development of a baseline HMM-based system for basic Quranic sounds. The second one aimed at enhancing the performance of this baseline system by addressing various strategies for modeling the Quranic sounds' durations, in this context, an explicit duration model was implemented instead of the implicit one of the standard Markov. The role of this proposed explicit model is to enhance the durational behavior of the standard HMM model. Three different distributions were tested and integrated into the HMMs' training and decoding procedures for modeling the HMM state durations; those are Gamma, Gaussian and the Poisson distributions. The results reported here show clearly the utility of the proposed strategies and this was positively reflected on the implemented system. Both S1 and S2 systems outperform the baseline system in term of global recognition accuracy, whereas both S1 and S2 systems outperform the baseline system in term of global recognition accuracy, while system S3 underperforms the baseline one, also S2 underperforms S1 and this proves the suitability of Gamma distribution in modeling the Quranic sounds' duration. However, the Gaussian distribution gives slight improvement but it stays anywhere better than the standard geometric distribution. The proposed modeling strategies have greatly succeeded in making the implemented system capable of recognizing and accurately distinguishing Quranic sounds. *Despite the utility of the strategies implemented here in reaching high-performance system, we still have confusion among certain Quranic sounds, which negatively reflects the system's accuracy. The misrecognition reported makes sense because the level of which the durational behavior should be considered is the whole sound segments, not states. to overcome this limitation, Our future steps will investigate the incorporation of supplementary speech features into the system for the purpose to make the acoustic models more accurate and robust to speech variability and thus improve their overall accuracy. Pitch, sounds duration and formants frequencies seem to be the most relevant features for which their incorporation resulting in ASR systems improvement.*

8. ACKNOWLEDGMENTS

The work presented here utilizes the results (Quranic Speech Corpus) of a project previously funded by King Abed Al-Aziz City for Science and Technology (KACST) in Saudi Arabia under grant number "AT – 25 – 113".

9. REFERENCES

[1] B. Jacob, M.M Sondhi and H.Yiteng, Springer Handbook of Speech Processing, Springer, (2008).

[2] Jurafsky, D.,Martin, J., Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall, (2009).

[3] X. Huang, A. Acero and H.-W. Hon, Spoken Language Processing: a guide to theory, algorithm, and system development, Prentice Hall, (2001).

[4] M. A. Anusuya and S. Katti, Front end analysis of speech recognition: A review, Int. J. Speech Technology, vol. 14, no. 2, pp. 99–145, (2011).Bowman, M., Debray, S. K., and Peterson, L. L. (1993).

[5] Ahsiah, I., Noor, N. M., Idris, M. Y. I. Tajweed checking system to support recitation. International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp.189–193. (2013).

[6] Noor Jamaliah Ibrahim, Zulkifli Mohd Yusoff, Zaidi Razak and Rosli Salleh, Improve Design for Automated Tajweed Checking Rules Engine of Quranic Verse Recitation: A Review. QURANICA-International Journal of Quranic Research, 1(1), pp. 39–50, (2011).

[7] Mohamed, S.A.E., et al. Virtual Learning System (Miqra'ah) for Quran Recitations for Sighted and Blind Students. Journal of Software Engineering and Applications, 7, 195-205, (2014).

[8] Yekache, Y., Kouninef, B., Mekelleche, Y., Mohamed, S., Building Quranic reader voice interface using sphinx toolkit. Journal of American Science, 9(11), pp. 473–479, (2013).

[9] Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[10] Y.O.M. Elhadj, "Preparation of speech database with perfect reading of the last part of the Holly Quran (in Arabic)". Proc. of the 3rd IEEE International Conference on Arabic Language Processing (CITAL'09), pp. 5-8, May, (2009).

[11] Y.O.M. Elhadj, I.A. Alsughayeir, M. Alghamdi, M. Alkanhal, Y.M. Ohali, and A.M. Alansari. Computerized teaching of the Holy Quran (in Arabic), Final Technical Report, King Abdulaziz City for Sciences and Technology (KACST), Riyadh, KSA, (2012).

[12] Yahya O.M. ElHadj, Mansour Alghamdi, Mohammad Alkanhal , Phoneme-Based Recognizer to Assist Reading the Holy Quran. Recent Advances in Intelligent Informatics, Advances in Intelligent Systems and Computing .Springer. 235: 141-152., (2013).

[13] Yahya O.M. ElHadj, Mansour Alghamdi, Mohamed Alkanhal, Approach for Recognizing Allophonic Sounds of the Classical Arabic Based on Quran Recitations. Theory and Practice of Natural Computing, Lecture Notes in Computer Science. Springer. 8273: 57-67, (2013).

[14] Y.O.M. Elhadj, Mohamed .O.M. Khelifa, A. Yousfi and M. Belkasmi. "An Accurate Recognizer for Basic Arabic Sounds". ARPN Journal of Engineering and Applied Sciences. vol. 11, no. 5, pp. 3239- 3243, Mar. (2016).

[15] Mohamed O.M. Khelifa, Y.O.M. Elhadj, Y. Abdellah and M. Belkasmi, "Enhancing Arabic Phoneme Recognizer using Duration Modeling Techniques,", in proc. of Fourth International Conference on Advances in Computing, Electronics and Communication - ACEC 2016, Dec 15, 2016, Rome-Italy.

[16] Mohamed O.M. Khelifa, Y.O.M. Elhadj, Y. Abdellah and M. Belkasmi, "An Accurate HSMM-based System for Arabic phonemes Recognition," in proc. of The IEEE Ninth International conference on Advanced Computational Intelligence (ICACI 2017), Feb. 2, 2017, Doha, Qatar.

[17] Mohamed O.M. Khelifa, Yousfi Abdellah, Yahya O.M. ElHadj and Mostafa Belkasmi, "Helpful Statistics in Recognizing Basic Arabic Phonemes" International Journal of Advanced Computer Science and Applications(ijacsa), 8(2), (2017). http://dx.doi.org/10.14569/IJACSA.2017.080231.

[18] S.Young. HTK Book (V.3.4). Cambridge University Engineering. Department of Engineering, UK, (2009).