Sound Database with Perfect Reading of the Last Part of the Holly Quran

Y. O. Mohamed El Hadj

Imam Med Bin Saud Islamic University
PO Box 8488, Riyadh 11681, Saudi Arabia

Summary
This paper presents a speech corpus of a perfect recitation of the Holly Quran. We mean by perfect recitation, a slow recitation with complete application of correct pronunciation rules ("Tarteel" reading method).
Sheikh Ali Abdelrahman Alhudhaifi, one of the famous reciters in the Islamic world, is recorded memorizing the last part of the Holly Quran. Then, we concentrated on the segmentation and labeling of the audio corpus on three levels: word, phoneme, and allophone in order to be accurate as much as possible.
A textual version of the Holly Quran fully diacriticized is also prepared in this work and used to construct the text files associated with the sound ones.

Key words:
Speech corpus, Quran recitation, speech recognition, Quran memorization, transcription, labeling.

1. Introduction

Speech corpora constitute an important source for linguistic studies, such as phonetics, phonology, etc. and for building and developing computer-systems that use human voice as an input and/or output. Therefore, many organizations took the initiative to build speech corpora of various types for a wide range of research and teaching purposes in different languages. For example, The Linguistic Data Consortium (LDC), which is a center for supporting and coordinating corpora development activities, issued more than 200 corpora for international organizations in more than 20 languages [1].

The efforts devoted to the Arabic language are still limited and mainly focused on the Modern Standard Arabic (MSA) [2], [3], [4], [5], [6] (which is used as official language in education, newspapers, broadcasting, ...etc.), in addition to some local dialects [7], [8].

The Classical Arabic, which is spoken by the early Arabs and represents the literary form of Arabic used in Holly Quran, does not differ greatly from the MSA due to the fact that Arabic is one of the most stable language throughout history. However, there are some idiosyncrasies as to the way of pronunciation. Take for example, the recitation of the Holly Quran, which includes the degrees of vowel elongation, nasalization, making the sound heavy and making the sound soft, etc.

Although recited Quran is not used in communication, it is important in teaching the Classical Arabic sounds in addition to the fact that it is indispensable in Islamic worshiping such as prayers. Accordingly, there must be speech corpus for the Holly Quran to help studying its sound features and building computational applications aiming at teaching how to recite it correctly. This is to protect it from being influenced by the foreign dialects and languages.

This paper is part of a project aiming to build a computer-based environment of teaching the Holly Quran. One of the main tracks in this project consists of assisting the memorization process of the Noble Quran based-on the speech recognition techniques.

In an early stage of this project, we have proposed a new labeling scheme covering all the Quranic Sounds and their phonological variations [9]. Since annotated speech corpus of the Quranic sounds was not available yet, we next focused on building a sound database for Quranic recitation.

Quranic recitation follows some well-defined pronunciation rules that we call "Fan Tajweed" (laws of perfect reading of the Holly Quran). Tajweed is considered as an art, because not all reciters perform the same way [10]. While obeying to Tajweed rules, recitation can be slow in order to give enough of time for each sound to be pronounced perfectly (in this case it is called "Tarteel" recitation), or it can be more fast by abbreviating duration of some sounds (in this case it is called "Hadr" recitation). For this reason, we decided to include all these ways of recitations, from *hadr* (fast recitation) to *tarteel* (slow

Manuscript received July 5, 2009
Manuscript revised July 20, 2009

recitation), in our Quranic sound database. So, two parts of this database are being developed in this project. One is related to the *hadr* recitation [11], while the other one is concerned with the *tarteel* recitation.

Our focus in this paper is on the second part, which concerns *Tarteel* reading method, in order to obtain a perfect pronunciation of the Quranic sounds. Sheikh Ali Abdelrahman Alhudhaifi, one of the famous reciters in the Islamic world, is recorded memorizing the last part of the Holly Quran. Then, we focused on the segmentation and labeling of this audio corpus on three levels: word, phoneme, and allophone in order to be accurate as much as possible.

Remaining of this paper is organized as follows: section two discusses the reasons for choosing the last part of the Holly Quran for building this corpus. In section three, we specify the design procedures of the Holly Quran Sound database. Section four reports results and statistics related to the developed corpus. Section five concludes this paper and highlights some usage of this corpus. It also clarifies some possible feature works related to this corpus.

2. Reasons of choosing the last part of the Holly Quran in building the corpus

The corpus of the Quranic recitations was assembled within the framework of a project concerned with a computerized teaching of the Holly Quran. So, it was implausible to work on the whole text of the Quran simply because this would require more time and effort. Therefore, we selected the last part of the Holly Quran for the following reasons:

(a) Most of the Tajweed rules are available in this part.
(b) This part contains long and short surahs (chapters).
(c) This part contains some chapters that were revealed in Makka and others revealed in Madinah.
(d) It is the first part that generally taught in schools.

For these reasons, we think this part fits properly in our project. However, we hope to enlarge the corpus to include the whole text of the Holly Quran later on.

3. Specification and design of the Quran sound database

Four main tasks are generally involved in the development of speech corpora: preparation of the text, selection of speakers, speech recording, and transcription. Each one of these tasks needs to be carefully performed in order to obtain a high quality speech database.

The following subsections describe the above tasks in the context of our work.

3.1. Preparation of a textual version of the Holly Quran

Although many electronic versions of the Holly Quarn are available on the net, we failed to get an official authentic version of it. Specialized authorities in Saudi Arabia, such as Ministry of Islamic Affairs, King Fahd Complex for Holly Quran Printing, have been contacted, but they claim that they have no authentic textual version. We had nothing to do other than to take one of the available versions and ask those who are specialized in the Quran and its sciences to proofread it. This has been done by some scholars affiliated to some official authorities[1].

Notice that the Holly Book is available in slightly-different versions according to the narration from the Prophet Muhammad PBH. This affects the textual version and the recited one. In this work, we considered the narration of Hafs from 'Asim, which is largely used in the Islamic world.

We revised the version once more to make sure that everything is correct and is diacriticized properly. It is worth mentioning that we focused mainly on the phonetic aspects of this corpus, particularly segmentation and duration of phonetic units. The textual version can be easily replaced without any further changes contrary to the sound part which will require re-segmentation and re-labeling if it is replaced.

3.2. Labeling scheme

The most appropriate symbols for accurate speech transcription are those of the International Phonetic Alphabet (IPA) for the fact that they represent the speech sounds of all languages and their dialects [12]. However, they are not familiarly used in speech databases for the reason that most language programs and speech tools do not recognize them. Other sets of symbols, such as Speech Assessment Methods Phonetic Alphabet (SAMPA) [13], British English Example Pronunciations (BEEP) [14], etc., have been created and used for transcribing many European languages.

However, these sets are not sufficient to cover the Arabic sound system. For example, there are 13 Arabic phonemes

[1] Imam University, Ministry of Islamic Affairs, and King Fahd Complex for Holly Quran Printing.

that do not have symbols in the Roman alphabet [15]. For this purpose, we have proposed a new labeling scheme covering all the Arabic phonemes and their allophones [10]. The set of proposed symbols includes the sound system of the Classical Arabic (CA) and that of the Modern Standard Arabic (MSA) in addition to be flexible to include the sounds found in the Arabic dialects. The labels are consistent in terms of the number of characters. Each label consists of four characters (see figure 1). The first two are letters that represent the Arabic phonemes which are taken from King Abdulaziz City for Sciences and Technology (KACST) Arabic Phonetic Database [16]. The third character is a number which symbolizes sound duration including geminates. The fourth character is another number that represents the allophonic variations.

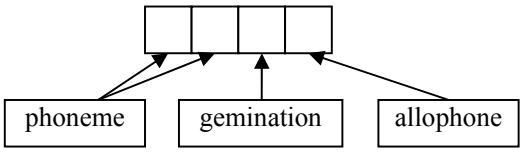

Figure1: composition of each label

A phoneme such as the pharyngeal consonant /M/ is represented as "cs10", where "1" means single (not geminate) and "0" represents its phonemic status. The complete set of the sound system of the CA at the phonemic level is shown in Table 1. To represent the geminate counterparts of these phonemes, the first number must be "2". The labels of the single and geminate phonemes can be used to transcribe CA speech at the phoneme level. A word such as "العنبر" the ambergris is transcribed as hz10as10ls10cs10as10ns10bs10as10rs10.

Table 1. Arabic orthography (AO) and the new labels (NL).

AO	NL	AO	NL	AO	NL
ـَ	as10	ذ	vb10	ف	fs10
ـُ	us10	ر	rs10	ق	qs10
ـِ	is10	ز	zs10	ك	ks10
ء	hz10	س	ss10	ل	ls10
ب	bs10	ش	js10	م	ms10
ت	ts10	ص	sb10	ن	ns10
ث	vs10	ض	db10	هـ	hs10
ج	jb10	ط	tb10	و	ws10
ح	hb10	ظ	zb10	ي	ys10
خ	xs10	ع	cs10		
د	ds10	غ	gs10		

Although the labels in Table 1 and their geminate counterparts are sufficient for the transcription at the phoneme level, they do not discriminate between allophones at the phonetic level transcription. But the label sets are flexible to contain the allophonic variations. Table 2 shows the CA allophones of the single phonemes. The letters are the same as of those in Table 1. The first number is always 1 to represent the single allophones. However, it can be 2 to represent the geminate consonants and vowels or 4, 7 or 8 to represent the longer vowel duration "mudoud". The second number is always 1 or higher to cover the allophones not only in the CA but also that of MSA.

Table 2. Arabic orthography (AO), the new symbols (NS) and the phonetic description (D).

AO	NL	D	AO	NL	D
ـَ	as11	plain	ص	sb11	plain
	as12	emphatic		sb14	nasalized
	as13	velarized	ض	db11	plain
	as16	centralized		db14	nasalized
ـُ	us11	plain	ط	tb11	plain
	us12	emphatic		tb14	nasalized
	us13	velarized		tb15	released with a schwa
ـِ	is11	plain	ظ	zb11	Plain
	is12	emphatic		zb14	Nasalized
	is13	velarized	ع	cs11	Plain
ء	hz11	plain	غ	gs11	Plain
ب	bs11	plain	ف	fs11	Plain
	bs15	released with a schwa		fs14	Nasalized
ت	ts11	plain	ق	qs11	Plain
	ts14	nasalized		qs14	Nasalized
	ts15	aspirated		qs15	released with a schwa
ث	vs11	plain	ك	ks11	Plain
	vs14	nasalized		ks14	Nasalized
ج	jb11	plain		ks15	Aspirated
	jb14	nasalized	ل	ls11	Plain
	jb15	released with a schwa		ls12	Emphatic
ح	hb11	plain		ls14	Nasalized
خ	xs11	plain	م	ms11	Plain
د	ds11	plain	ن	ns11	Plain
	ds15	released with a schwa	هـ	hs11	Plain
ذ	vb11	plain	و	ws11	Plain
	vb14	nasalized		ws14	Nasalized
ر	rs11	plain	ي	ys11	Plain
	rs12	emphatic		ys14	nasalized
	rs14	nasalized			
ز	zs11	plain			
	zs14	nasalized			
س	ss11	plain			

AO	NL	D	AO	NL	D
	ss14	nasalized			
ش	js11	Plain			
	js14	nasalized			

A word such as "إنسان" human is transcribed at this level as: hz11ss14ss11as21ns11.

3.3. Collecting the audio corpus

Remember that our main objective in this work is to build a speech database of Quranic recitations with a perfect application of Tajweed rules, in such a way that can be considered as a reference for any phonetic or acoustic study on the Quran or the Classical Arabic language in general.

We contacted King Fahd Complex for Holly Quran Printing in Almadinah to get clear and professional recordings of one of its authentic reciters. Having got permission, we then chose Sheikh Ali Abdelrahman Alhudhaifi who is distinguished with his beautiful recitation and his ability to:

(a) Apply the Tajweed rules perfectly,
(b) Consider the places of articulation,
(c) Keep a stable level of recitation.
(d) He also does not attach the verses while reciting, i.e pauses at the end of every single verse.
(e) In addition, he considers all pauses within verses.
(f) Finally, he rarely makes incidental sudden pauses during recitation which happen when the reciter experiences shortness of breath at positions where pauses are impermissible. Such incidental pauses make the reciter resume the utterance by repeating the last independent meaningful words.

The recitation was recorded in the complex in a sound-proof compartment and digitally sampled at a rate of 44100 Hertz and coded in 16-bit.

To prepare the sound files, we divided them into files of short durations in a way that makes every single file represents an audio recording of a Quranic verse, except for the long verses during which the reciter needs to pause for breathing. These long verses were segmented according to the number of pauses. Then the text files of the spoken counterparts were created by copying the relevant content from the textual version we prepared before. Text and sound files are named as follows: XXX_YYY_ZZ, where XXX represents the surah (chapter) number, YYY represents ayah number inside surah, and ZZ represents number of pauses inside the ayah.

3.4. Transcription

Having finished and revised the sound files, we started the most important part related to the phonetic transcription. The transcription and alignment can be done manually, automatically or both where the manual transcription is done for verification of the automatic transcription. However, as far as this work concern the Holly Quran, we preferred to perform a manual transcription in order to be accurate as much as possible. The transcription is made at three levels using the Praat tools [17]. The first level is at the word level where each word is segmented and labelled. Labels of words are composed of surah number, ayah number, and word number joined by underscores.

The second level is at the phoneme level where the labels from Table 1 are used. The third level is the allophone/phonetic level where labels from Table 2 are used. To avoid typing errors, an interface with all the labels and their meanings is created. Each label is designed as a button that transfers its label to the location defined previously in the transcription interface.

The transcription files are extracted from PRAAT and kept in their original format as "txtgrids" files. Each sound file has now a corresponding file that represents its transcription (.textgrid) in addition to the text file previously described.

4. Results and statistics

The final audio corpus contains a total number of 574 sound files coupled with their textual counterparts that represent the texts spoken by the reciter and their transcriptions on three levels (word, phone, allophone). The average duration of sound files is about 6 seconds resulting in almost one hour of speech.

The total number of words in this corpus is 2318 distributed on 1359 unique words with an average of frequency of almost 2 per word (1.7 exactly). These words are arranged in 37 surahs (chapters) of the Holly Quran, which contains 114 surahs. They contain 565 ayahs (verses) from about 6400 ayahs of the whole Quran.

We have conducted some basic statistics on the low levels of segmentation. They show that the corpus contains a total number of 15191 phonemes distributed on 60 unique phonemes (those from table 1 and their geminates). The corpus contains also the same number (15191) of allophones but distributed on 110 unique allophones (those from table 2 with their allophonic variations).
Notice that silence is considered as a phoneme (allophone), and is represented by a special code "sil".

Frequencies as well as minimum, maximum and average durations are computed for each phoneme and allophone in the corpus. Figures 2 and 3 show phonemes that are respectively occurred more that 100 (most frequent) and less than 10 (least frequent).

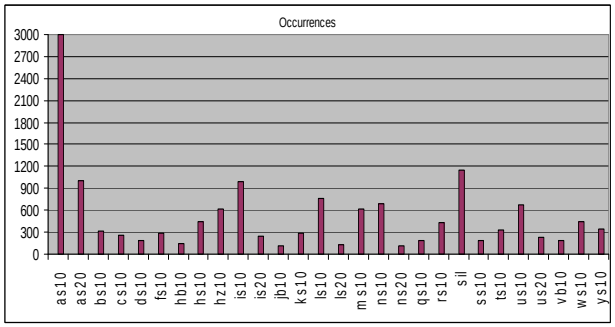

Figure 2: Most frequent phonemes

From figure 2, we remark that 28 phonemes (i.e. ~47%) have been occurred more than 100 times. The vowel Fatha "as10" is occurred alone almost 3000 times (2996).

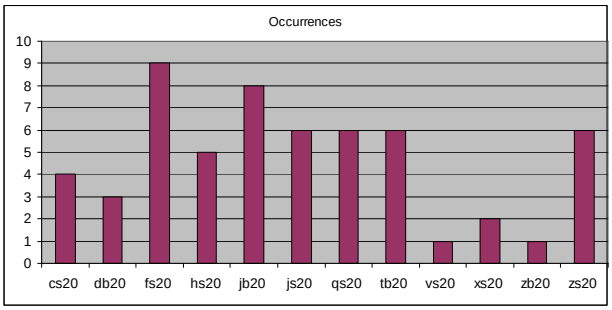

Figure 3: Least frequent phonemes

Figure 3 shows that 12 phonemes have been occurred less than 10 times. From this set, two phonemes (vs20, zb20) are produced only one time.

Figures 4 and 5 represent similar frequencies but for allophones.

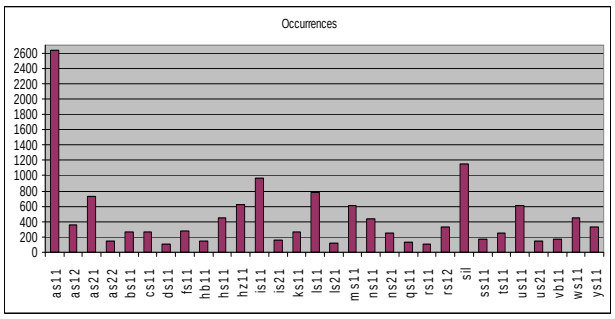

Figure 4: Most frequent allophones

Here, we remark that 30 allophones are occurring more that 100 times. The phoneme as10 is now split into eight variants: as11, as12, as21, as22, as41, as42, as61, as62 to represent different prolongation degrees of this vowel. These are phonological variations of the vowel as10.

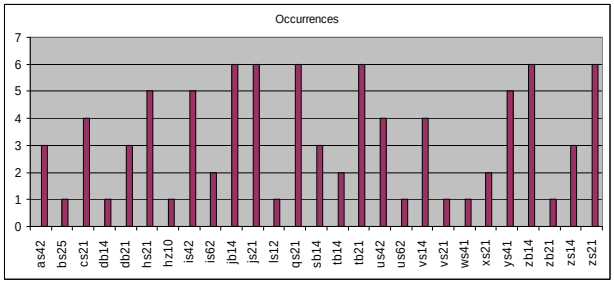

Figure 5: Least frequent allophones

More than 35 allophones are occurred less that 10 times. Figure 5 shows only 27 of them for the sake of readability.

If we look globally at the distribution of these phonemes in the corpus, we found that the short vowels, which are represented in Arabic by diacritic marks (fatha "as10", kasra "is10", damma "us10") represent 4640, i.e. ~31%. With their geminate counterparts (as20, is20, us20) that represent long vowels, they count 6150, i.e. ~41%. If we add other diacritic marks, such as consonant doubling, we obtain more than 50%. This put in evidence the importance of the diacritic marks in the correct pronunciation of Arabic sounds.

5. Conclusion

In this paper, we presented a speech database of the last part of the Holly Quran recited by Sheikh Ali Abdelrahman Alhudhaifi. Although the duration of this sound database is not big enough (about one hour), yet we think it is a significant contribution. As far as we know, this is the first corpus related to the Holly Quran that follows the methodology mentioned above.

As we said before, one of our main goals in the above mentioned project was to develop a speech corpus of the Holly Quran that includes all ways of recitations. By this new contribution, a high quality speech corpus of almost ten hours is now available for the Holly Quran, opening the door for different computer-based studies. It is composed of two parts: one related to "hadr" recitation and this one related to "tarteel". The first one is composed of ten recitations of medium-aged reciters (aged between 16-35) memorizing the same part of the Holly Quran. It still under finalization and is being published separately.

This corpus will represent a platform for various applications aiming at enhancing the computational studies of the Holly Quran. For example, it can be used in building applications that are basically concerned with Tajweed rules. It can also be used to support self learning and memorization of the Holly Quran as well as to teach how to pronounce correctly the classical Arabic Sounds.

Another kind of usage of this corpus may be in the field of sound studies related to the Holly Quran. As an example, identify a reference of vowels' prolongation (almodoud), al ghalgala, etc.

We hope researchers could carry on and enlarge the corpus to include the whole text of the Holly Quran following the same methodology. This could hopefully be extended to the other recitation's narrations.

Acknowledgment

This paper is supported by King Abdulaziz City for Science and Technology under the grant number AT-25-113, Riyadh, Saudi Arabia. Our thank goes also to King Fahd Complex for Holly Quran Printing.

6. References

[1] Tan Lee, W.K. Lo, P.C. Ching and Helen Meng,. Spoken language resources for Cantonese speech processing. Speech Communication, Vo1.36, 34, pp 327-342, March 2002.
[2] Alghamdi, Mansour, Moustafa Elshafei and Husni Almuhtaseb. Arabic Broadcast News Transcription System. International Journal of Speech Technology, pp 1572-8110, 2009.
[3] Alghamdi, Mansour, Fayez Alhargan, Mohamed Alkanhal, Ashraf Alkhairy, Munir Eldesouki and Ammar Alenazi. Saudi Accented Arabic Voice Bank. Workshop on Experimental Linguistics. Athens, Greece, 2008.
[4] Mohamed A, Hassan H. S. Development of an Arabic speech database. Information and Communications Technology, 2005.
[5] S. Alansary, M. Nagi, N. Adly. Building an International Corpus of Arabic (ICA): Progress of Compilation Stage.
[6] Katrin K, et al. Novel Approaches To Arabic Speech Recognition: Report From The 2002 Johns-HOPKIN Summer Workshop.
[7] Mahatab N, et al. Network of Data Centers (NETDC) BNSC-An Arabic Broadcast News speech Corpus. ELDA-Evaluation and Language Resources Distribution Agency, Namelar conference, Cairo 2004.
[8] Imed Z, et al. ORIENTEL: SPEECH-BASED INTERACTIVE COMMUNICATION APPLICATIONS FOR THE MEDITERRANEAN AND THE MIDDLE EAST. ICSLP, 2002.
[9] M. AlGhamdi, Y.O. Mohamed El Hadj, M. AlKanhal. A MANUAL SYSTEM TO SEGMENT AND TRANSCRIBE ARABIC SPEECH. Proceedings of IEEE ICSPC'07, pp 233-236, Dubai, UAE, 2007.
[10]T. Hassen, F. Wassim, M.Bassem. Analysis and Implementation of an Automated Delimeter of "Quarnic" Verses in Audio Files using Speech Recognition Techniques. Information and Communication Technologies, ICTTA '06, 2nd, Vol. 2, pp 2979-2984, 2006.
[11]Y.O. Mohamed El Hadj, M. AlGhamdi, M. AlKanhal .Building and preparing a Sound corpus of the Holly Quran. International Conference on the Glorious Quran and Contemporary Technologies, King Fahd Complex for the Printing of the Holly Quran, Al-Madinah Almunawwarah, Saudi Arabia, October 13-15, 2009.
[12]http://www.arts.gla.ac.uk/ipa/ipa.html
[13]http://coral.lili.uni-bielefeld.de/Documents /sampa.html
[14]Donovan, Robert Edward. Trainable speech synthesis. Unpublished Ph. D. thesis. Cambridge University, UK, 1996.
[15]Alghamdi, Mansour. KACST Arabic Phonetics Database. Fifteenth International Congress of Phonetics Science, Barcelona, 3109-3112. 2003.
[16]Alghamdi, Mansour. Algorithms for Romanizing Arabic Names. Journal of King Saud University: Computer Sciences and Information. Vol. 17, pp 1-27, 2005.
[17] http://www.fon.hum.uva.nl/praat/

Yahya O Mohamed El Hadj received his Bachelor's degree in Computer Sciences from Caddi Ayyad University at 1996 in Morocco. He obtained Master's degree and PhD in distributed and parallel processing respectively at 1998 and 2001from a joined European Program with Mohamed I University of Morocco. He performed many research stays in different European laboratories, such as LIP-ENS Lyon, IRIT Toulouse, etc.
Since 2002, he is assistant professor at the College of Computer and Information Sciences, Imam Muhammad Bin Saud University in Saudi Arabia. His research interests include Arabic Language technologies (Text and Speech), parallel and distributed processing.
He participated and directed many research projects sponsored by different scientific institutions, such as King Abdulaziz City for Sciences and Technology (KACST), Imam University, etc.