VOL. 11, NO. 5, MARCH 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

AN ACCURATE RECOGNIZER FOR BASIC ARABIC SOUNDS

Yahya O. M. Elhadj[1], Mohamed O. M. Khelifa[2], Yousfi Abdellah[3] and Mostafa Belkasmi[2]
[1]Doha Institute, Doha, Qatar and SAMoVA Research Team, IRIT, Paul Sabatier University, Toulouse, France
[2]Telecommunications and Embedded Systems Team, ENSIAS, Mohammed V University, Rabat, Morocco
[3]Faculty of Juridical, Economic, and Social Sciences, Mohammed V University, Rabat, Morocco
E-Mail: yelhadj@irit.fr

ABSTRACT

This paper is part of an ongoing work aiming to build an accurate Arabic sounds recognizer for teaching and learning purposes. Early phases of this work were dedicated to the development of a particular sound database from recitations of the Holy Quran to cover classical Arabic sounds; speech signals of this sound database were manually segmented and labelled on three levels: word, phoneme, and allophone. Next, two baseline recognizers were built to validate the speech segmentation on both phoneme and allophone levels and also to test the feasibility of the sounds' recognition intended target. This current phase considers the development of an elaborated recognizer, by considering the basic sounds and looking for their distinctive features (e.g. duration, energy, etc.) to determine which ones will be particularly helpful to identify the phonological variation of the basic sound. Here, we present the first results of the basic sounds recognition obtained so far.

Keywords: speech recognition, sound databases, hidden markov models, speech segmentation, pronunciation errors detection.

INTRODUCTION

Automatic Speech Recognition (ASR) technology allows a machine to identify the textual content of a pronounced speech; depending on the type of application, the textual content of the speech might be further processed to be suitable for a specific task [1-4]. Early ASR applications were limited to handle relatively simple tasks, such as for example speaker-dependent isolated keywords recognition. Nowadays, a myriad type of ASR applications appeared, covering a wide range of tasks such as for example, remote control using phones, helping the disabled and persons with special needs, speaker identification, language identification, archiving, search and retrieval, language acquisition, and so on.

Despite the large use of speech recognition technology in foreign languages, the Arabic language is still suffering from scarcity of mature ASR-based applications, especially for language learning and evaluation.

One Distinguished application of Arabic Speech Recognition is the teaching of the Classical Arabic (CA) sound system. Although classical Arabic is not used in the daily communication, it is required to learn the Holy Quran and the old poetry heritage. Moreover, it can open the door for several kinds of Islamic applications.

This work is a continuation of previous efforts targeted to develop a high performance recognizer for classical Arabic sounds to be used for teaching and learning purposes [5]. First stages of these efforts were limited to the preparation of an appropriate sound database to support the ultimate goal [6-9]. Thus, ten recitations of a well-chosen part of the holy Quran were recorded and manually segmented and annotated on three levels: word, phoneme, and allophone. To validate this particular sound database and to test the feasibility of the goal, two baseline recognizers for phonemes and allophones were developed [10, 11].

This current phase aims to develop an accurate recognizer, by firstly considering the basic sounds and then exploring different features of each basic sound separately to determine the most pertinent ones for the identification of its phonological variation. We mean by the basic sounds the basic phonemes without any phonological variation and even without considering the phonemes germination (the doubling).

In this paper, we present the results of the basic sounds recognition obtained so far. We will follow the same methodology employed in the previous baseline recognizers to be able to make appropriate comparisons and to give pertinent suggestions and recommendations for our future steps. Thus, Hidden Markov Models Toolkit (HTK) is used as development environment of the recognizer as performed in the previous works; each basic sound is modelled by 3-emitting states HMM with a mixture of Gaussian models [4, 12].

The rest of the paper is organized as follows: section 2 gives a quick overview of the previously developed sound database (we will call it "CA Sound Database"); section 3 presents current adaptation of the "CA Sound Database" in order to be annotated in terms of basic sounds; section 4 is dedicated to the development of the recognizer; section 5 discusses the results and the future improvements; the last section concludes the paper and highlights the perspectives.

Overview of the "CA Sound Database"

One of the major obstacles facing the development of Arabic speech recognition applications is the scarcity of appropriate sound databases usually needed for training and testing statistical models. This problem is seriously encountered when we consider the classical Arabic language, as the majority of available corpora (even few) are mainly directed to what is known as Modern Standard Arabic (MSA) and its related dialects.

To contribute overcoming such a problem and to support the development of sound-related applications for classical Arabic, the "CA Sound Database" was built based on Quranic recitations to extract Classical Arabic sounds. Due to the complexity of developing sound databases, only a part of the holy Quran was considered. Recitations of ten reciters (readers) were recorded in an appropriate environment under the control of a specialist of the holy Quran pronunciation rules (Tajweed); more than eight hours of speech were obtained [6-8]. Table-1 shows, for each reciter, the number of sound files, their size and duration.

Table-1. Sound files and their duration by reciters.

Reciter Number	Reciter Initials	Number of Sound Files	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**

Each sound file represents an Ayah or part of it for the long Ayahs where the reciter needs to take a long breathing.

To have a useful database for multi purposes, the speech signals were manually and accurately segmented on three levels: word, phoneme, and allophone. A new labelling scheme was proposed to annotate the speech segments [9] as the available speech labelling schemes (e.g. IPA , SAMPA , BEEP , etc.) were either not appropriate or not able to cover all Arabic sounds; Figure-1 gives the labelling scheme used for the transcription.

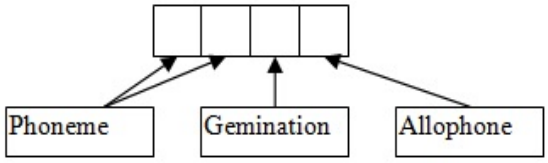

Figure-1. The function of the characters in each label.

The first two positions are letters representing the Arabic phonemes, which are taken from KACST Arabic Phonetic Database [13]; the third position is a number symbolizing the sound duration (1 for basic phonemes, 2 for germination); the fourth position is another number representing the allophonic variation: it is always 1 to represent the single allophones; however, it can be 2 to represent the geminate consonants and vowels or 4, 6 or 8 to represent the longer vowel duration "mudoud المـدود"; note that "0" means no variation is considered.

As an example, a word such as "العنـــبر" ambergris is transcribed as hz10as10ls10cs10as10ns10bs10as10rs10 at the phoneme level; whereas, a word such as "إنســان" human is transcribed as hz11ss14ss11as21ns11 at the allophone level. The strong relationship between the Arabic orthography and the phonemic transcription is very clear.

The transcription was done using a very popular tool called Praat; each sound file has a counterpart (textgrid file) containing its annotated segments; it has also another counterpart (text file) containing the corresponding text (ayah or part of it). These three files have the same name in the "CA Sound Database" but with different extensions (.wav, .textgrid, .txt). The file naming scheme uses this format: SSS_XXX_YYY_ZZ, where SSS represents initial letters of the reciter's name; XXX_YYY represents Surah and Ayah numbers (Quran chapters and their parts); ZZ represents pause numbers inside Ayahs. For example AAZ_000_001_00 means the first Ayah of the Surah 000 (البســـملة) by the reciter AAZ; no breath was marked here (00). Note also that words are coded using a similar format: XXX_YYY_W, where W means the word number (incremental number: 1, 2, etc.) of the Ayah number YYY in the Surah number XXX. Figure-2 gives an overview of the customized Praat interface with the annotation levels and the naming scheme.

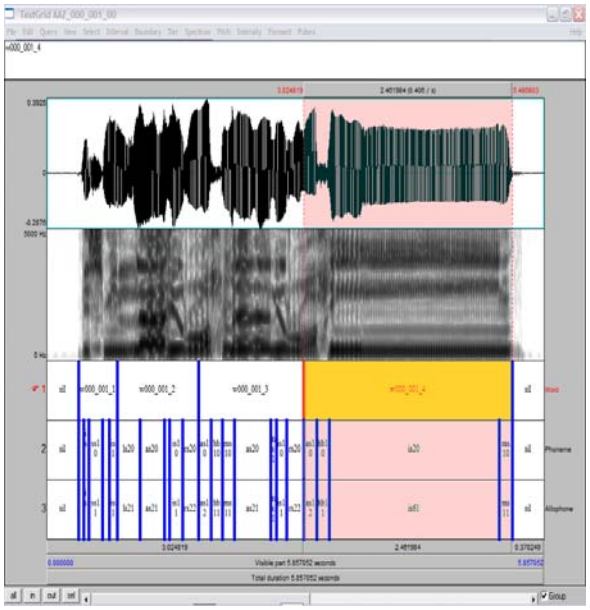

Figure-2. The customized Praat interface: 1) wave, 2) spectrogram, 3) word-level transcription, 4) phoneme-level transcription, 5) allophone-level transcription.

Adaptation of the "CA Sound Database"

As we evoked before, the "CA Sound Database" is annotated at the phoneme and allophone levels. Since

we are considering -in this current work- only the basic phoneme sounds, it was needed to modify the annotation to produce a new one for the basic phonemes in order to be able to train and test our intended recognizer. The list of basic sounds and their associated codes are shown in the Table-2.

Table-2. List of basic phonemes and their codes.

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

We developed scripts and programs to read the code for each speech segment at the phoneme level from the "CA Sound Database" and then to transform the geminated phonemes to their basic ones; this means that, we look at each code, if it was "??20" it is transformed to "??10", otherwise it is kept as is. These changes are done in the transcription files (textgrids). So, the 5935 files were examined and modified to produce new ones with the appropriate codes. These data are used to build the recognizer as we will explain in the following section.

Basic sounds recognizer

The methodology employed to build the previously mentioned baseline recognizers (for phonemes and allophones) is completely followed here. Thus, the basic sound units are modelled by 3-emitting states HMM with a mixture of Gaussian models.

Models and data preparation

We have 31 basic phonemes (see table 2) for which a special unit is added to denote the silence independently of its occurring place, either at the beginning or inside or at the end of each Ayah. An HMM with three emitting states is considered as a model template for all basic phoneme units; Gaussian Mixture Models (GMMs) are associated with each state of the template HMM model to identify the characteristics of the sound portion at this state.

A vector of 39 coefficients is used to represent speech parameters extracted from the sound files for each 10 ms Hamming window; these coefficients are: 1) the first twelve MFCC (Mel Frequency Cepstral Coefficients) representing the static features of the signal portion, 2) their first and second derivatives (velocity and acceleration) to capture the dynamic features, and 3) the energy of the signal portion and also its derivatives (1st and 2nd). Coefficients extraction are performed by the HTK command "HCopy" with appropriate configuration file.

Notice that all basic phoneme units have the same initial model with same parameters, but they become different after training, as each one is tuned on its related portions of signals.

The transcription files (textgrids) are used to extract the lexicon of pronunciations; a chunk of it is shown below in the Table-3.

Table-3. Part of phoneme-based pronunciations lexicon.

عَمَّ	cs10 as10 ms10 as10
يَتَسَاءَلُونَ	ys10 as10 ts10 as10 ss10 as10 hz10 as10 ls10 us10 ns10
عَنِ	cs10 as10 ns10 is10
النَّبَإِ	ns10 as10 bs10 as10 hz10 is10
الْعَظِيمِ	ls10 cs10 as10 zb10 is10 ms10

Notice that the majority of words are pronounced the same way across all the reciters (readers); the variability of pronunciations is visible at low levels, such as allophones.

For the recognition, we used the following flat language model (Table-4) to allow all pronunciation possibilities (any sound can appear after any other one).

Table-4. Flat language model for basic phonemes.

```
$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 > )
```

This format is transformed to an internal HTK representation by the command "HParse".

Experimentations

For the initialization and training, we used the following combination of HTK tools "HInit + HRest + HERest", which was proven in the previous works [10, 11] as the best combination for HTK training tools.

For the appropriate number of GMMs, we conducted a lot of experimentations varying their numbers from 1 to 16 on specific groups of training & testing sets defined as follows: the sound database is divided into ten groups of training and testing sets; all of them are used 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 ayah of each Surah from it; the remaining ayahs of this reciter as well as all ayahs of the other reciters are used for training. Since our sound database contains 38 Surahs and 572 ayahs, so the training and testing sets in each group are respectively composed of 534 and 38 ayahs. This means that in each group, about 93% of the corpus is used for training and 7% is used for testing. Notice that there is nothing in the literature to indicate which number of GMMs is the best for a given context, and thus the optimal number has to be determined by experimentation.

The results are reported in the Table-5 and depicted in the Figure-3 to be more readable and analysable.

Table-5. Average recognition rates (1 to 16 GMM) (%).

GMMs	AAH	AAS	AMS	ANS	BAN	FFA	HSS	MAS	MAZ	SKG
1	89.23	89.69	81.12	87.23	85.90	84.04	90.08	86.97	82.85	81.25
2	93.88	92.04	88.43	90.16	93.35	86.97	94.39	94.28	92.42	90.03
3	96.81	94.39	88.83	94.95	95.88	91.49	96.48	94.28	94.02	93.35
4	97.07	97.78	90.29	96.01	96.14	95.33	97.65	95,21	95.35	95.21
5	97.87	98.43	93.88	96.81	96.54	95.74	97.78	96.28	95.74	95.74
6	97.87	98.43	94.55	97.07	96.54	95.48	98.17	96.41	96.81	97.07
7	98.14	98.30	94.68	98.01	97.34	95.35	98.04	96.28	96.94	97.07
8	98.01	98.83	95.35	97.61	98.14	96.14	98.04	96.01	96.41	96.94
9	97.87	98.69	95.61	97.87	97.74	95.74	98.69	96.68	96.94	97.47
10	97.74	99.09	95.35	97.87	97.87	95.88	98.96	96.94	97.21	97.74
16	98.54	99.48	96.81	98.67	97.61	97.47	98.43	97.07	98.27	96.94

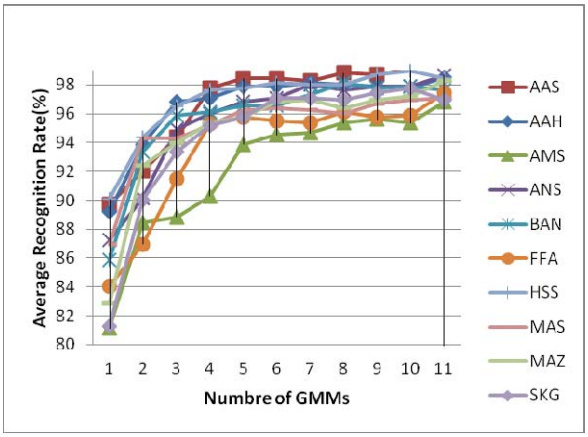

Figure-3. Average recognition rates for GMMs.

DISCUSSION AND COMPARISON OF RESULTS

It is worth to mention (firstly) that the results obtained here for the basic phoneme recognizer are very similar to those obtained for the general phoneme recognizer [10] in terms of homogeneity, correlation with the number of GMMs, etc. Globally, the recognition rates are largely improved here, which seems logic as the number of basic units is less than the previous one; the lowest recognition rate is passed from 88% in the previous recognizer to ~97% in this current recognizer, and the highest one has moved from 94% to 99.5%; while the global average recognition rate is increased from 92% to 98%.

For an in-depth analysis of the obtained results, we looked at the confusion matrices of the basic units' recognition to see how they are distinguished from each other. We noticed the appearance of some kind of insertions, deletions, and substitutions between units. Although that these phenomenon were largely reduced compared to what we have seen in the previous baseline recognizers, they are still representing an important barrier of the accuracy of the recognizer. We think that, these kinds of miss-recognition need more consideration of particular properties of the sound units.

CONCLUSIONS

We presented in this paper the first results of the basic phonemes' recognizer we developed for Classical Arabic Sounds. High recognition rates were obtained given an average of 98% for all reciters. However, an in-depth analysis indicates that we still have an important confusion between some sound units; other sounds, such as “as10, bs10” for example, are simply omitted over several reciters.

We are currently looking at the confusion matrices obtained from the experimentations for all reciters to determine the sound units substituted or deleted. These units will be analysed to see how they can be distinguished from others; based on this analysis, we will propose either, specific HMM models, or incorporation of particular features, such as for example the energy (among other ones), which is a very distinctive cue.

ACKNOWLEDGEMENTS

This work exploits the results (CA Sound Database) of a project previously funded by KACST (www.kacst.edu.sa) under grant number “AT – 25 – 113”.

REFERENCES

[1] L. Rabiner, and B.H. Juang. 1993. Fundamentals of Speech Recognition. Prentice Hall.

[2] F. Jelinek. 1998. Statistical Methods for Speech Recognition. Cambridge, MA: MIT Press.

[3] X.Huang, A. Acero, and H. Hon. 2001. Spoken Language Processing, Prentice Hall PTR.

[4] Daniel Jurafsky, James H. Martin. 2008. Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition. 2nd Edition, Prentice Hall.

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

[6] Y.O.M. Elhadj, M. AlGhamdi, M. AlKanhal, A.M. Alansari. Design and Development of a High Quality Speech Corpus for Classical Arabic. Submitted for publication to the Language Resources and Evaluation Journal (LREV).

[7] Y.O.M. Elhadj, M. AlGhamdi, M. AlKanhal, A.M. 2009. Alansari. Sound Corpus of a part of the noble Quran (in Arabic). Proc. of the International Conference on the Glorious Quran and Contemporary Technologies, King Fahd Complex for the Printing of the Holy Quran, Almadinah, Saudi Arabia.

[8] Y.O.M. Elhadj. 2009. 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) , Rabat, Morocco. pp. 5-8.

[9] M. AlGhamdi, Y.O.M. Elhadj, M. AlKanhal. 2007. A manual system to segment and transcribe Arabic Speech. Proceedings of IEEE ICSPC'07, Dubai, UAE. ISBN 1-4244-1236-6. pp. 233-236.

[10] Y.O.M. Elhadj, M. Alghamdi, M. Alkanhal. 2014. Phoneme-Based Recognizer to Assist Reading the Holy Quran. Recent Advances in Intelligent Informatics, Advances in Intelligent Systems and Computing .Springer. 235: 141-152.

[11] Y.O.M. Elhadj, M. Alghamdi, M. Alkanhal. 2013. 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.

[12] L. Rabiner. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE. 77(2).

[13] Alghamdi, M. 2003. KACST Arabic Phonetics Database. The Fifteenth International Congress of Phonetics Science, Barcelona, Spain. pp. 3109-3112.