A MANUAL SYSTEM TO SEGMENT AND TRANSCRIBE ARABIC SPEECH

M. Alghamdi[1], Y. O. Mohamed El Hadj[2], M. Alkanhal[1]

[1]Email: {mgamdi,mkanhal}@kacst.edu.sa
King Abdulaziz City for Science and Technology
PO Box 6086, Riyadh 11442, Saudi Arabia
[2]Email: yelhadj@ccis.imamu.edu.sa
Imam Med Bin Saud Islamic University
PO Box 8488, Riyadh 11681, Saudi Arabia

ABSTRACT

In this paper, we present our first work in the "*Computerized Teaching of the Holly Quran*" project, which aims to assist the memorization process of the Noble Quran based-on the speech recognition techniques. In order to build a high performance speech recognition system for this purpose, accurate acoustic models are essentials. Since annotated speech corpus of the Quranic sounds was not available yet, we tried to collect speech data from reciters memorizing the Quran and then focusing on their labeling and segmentation.
It was necessarily, to propose a new labeling scheme which is able to cover all the Quranic Sounds and its phonological variations. In this paper, we present a set of labels that cover all the Arabic phonemes and their allophones and then show how it can be efficiently used to segment our Quranic corpus.

***Index Terms*—** Quran; Arabic; transcription; speech; recognition

1. INTRODUCTION

Human machine interaction is switching from buttons and screens to speech. Speech recognition is an important element in this interaction. However, to build a speech recognition system a speech database is needed. A speech database is essential not only to build a speech recognition system but also to build other systems such as speaker verification and speech syntheses. This is one of the reasons that speech databases have been collected for many languages, for example: English [1], Spanish [2], Dutch [3], Mandarin [4], French [5] and Arabic [6] among others.
Although recited Quran is not used in communication, it is important in teaching the pronunciation of Classical Arabic sounds in addition to the fact that it is indispensable in Islamic worshiping such as prayers. Teaching how to recite the Quran has been through teachers who pronounce the Quranic sounds accurately. Such method has been practiced since the revelation of the Quran.
This paper is part of a project to build a speech recognition system that would be able to teach learners how to pronounce its sounds and correct them when they make mistakes. However, before building the system a speech database of the recited Quran is needed where the sounds are labeled and segmented.
Recent speech databases possess transcription at different levels. These levels range from the phonemes to intonations. In addition to transcribing the speech, the transcription is aligned with the speech acoustic signal [7, 8]. The transcription and alignment can be done manually, automatically or both where the manual transcription is done for verification of the automatic transcription [7, 9].
This paper presents a new transcription labels that are more convenient to the transcribers and appropriate for speech recognition tools such as Hidden Markov Toolkit (HTK) [10]. At the same, they cover all Arabic sounds including that of the Modern Standard Arabic, Arabic dialects and Classical Arabic.

2. SOUND LABLES

The 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 [11]. However, they are not familiarly used in speech databases for the reason that most language programs and speech tools such as Hidden Markov Toolkit do not recognize them. On the other hand, language orthography does not represent all the sound of its language, therefore, it is not used by itself for transcription. So, other symbols available on the keyboard are used for transcription such as @, >, in addition, combinations of two characters such

as the English letters and Arabic numerals were used in other speech databases [8, 12, 13, 14, 15].
Moreover, different sets of symbols have been created to transcribe speech databases. One of them is the Speech Assessment Methods Phonetic Alphabet (SAMPA) [16] which has been used for English and other European languages [7, 17]. Another set is the British English Example Pronunciations (BEEP) [18].
However, these sets are not sufficient to cover the sounds of a European language such as Icelandic [19]. The Arabic sound system is even more remote to be covered by these sets of sounds. For example, there are 13 phonemes that do not have symbols in the Roman alphabet let alone the geminates and other allophonic variations [20].

3. METHODS

Our aim in this work is to create a set of labels that cover all the Arabic phonemes and their allophones. The set needs to include 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 (Figure 1). The first two are letters that represent the Arabic phonemes which are taken from KACST Arabic Phonetic Database [21]. The third character is a number which symbolizes sound duration including geminates. The fourth character is another number that represent the allophonic variations.

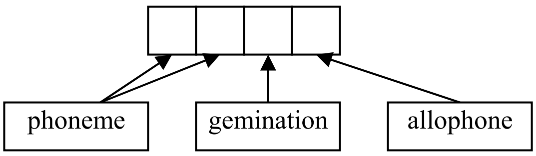

Figure 1. The function of the characters in each label.

So, a phoneme such as the pharyngeal consonant /ʕ/ 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. The set consists of 31 phonemes that represent the single vowels and consonants. As it can be seen, the first number is always “1” which means that the sound is single, and the second number is always “0” which means the sound is a phoneme. 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*. The strong relationship between the Arabic orthography and the phonemic transcription is very clear. The reason for this is that the Arabic alphabet represents the Arabic sounds in most of the cases. Unlike English where /f/, for example, can be represented by different letters such as “f, ph, gh”.

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, 6 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.
A word such as “إنسان” *human* is transcribed *hz11is11ss14ss11as21ns11* at this level.

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

AO	NL	D
	us13	velarized
اِ	is11	plain
	is12	emphatic
	is13	velarized
ء	hz11	plain
ب	bs11	plain
	bs15	released with a schwa
ت	ts11	plain
	ts14	nasalized
	ts15	aspirated
ث	vs11	plain
	vs14	nasalized
ج	jb11	plain
	jb14	nasalized
	jb15	released with a schwa
ح	hb11	plain
خ	xs11	plain
د	ds11	plain
	ds15	released with a schwa
ذ	vb11	plain
	vb14	nasalized
ر	rs11	plain
	rs12	emphatic
	rs14	nasalized
ز	zs11	plain
	zs14	nasalized
س	ss11	plain
	ss14	nasalized

AO	NL	D
	tb15	released with a schwa
ظ	zb11	plain
	zb14	nasalized
ع	cs11	plain
غ	gs11	plain
ف	fs11	plain
	fs14	nasalized
ق	qs11	plain
	qs14	nasalized
	qs15	released with a schwa
ك	ks11	plain
	ks14	nasalized
	ks15	aspirated
ل	ls11	plain
	ls12	emphatic
	ls14	nasalized
م	ms11	plain
ن	ns11	plain
هـ	hs11	plain
و	ws11	plain
	ws14	nasalized
ي	ys11	plain
	ys14	nasalized

AO	NL	D	AO	NL	D
ش	js11	plain			
	js14	nasalized			

These sets of labels shown in Table 1 and Table 2 are being used in the Computerized Teaching of the Holly Quran project. First, we had to create a speech database for Quranic citation then transcribing it. The transcription is made at three levels using the Praat tools (Figure 2) [22]. The first level is at the word level where each word is segmented and labelled. 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. The transcription and segmentation are done manually. To avoid typing errors an interface with all the labels and their meanings is created (Figure 3). Each label is designed as a button that transfers its label to the location defined previously at the transcription interface.

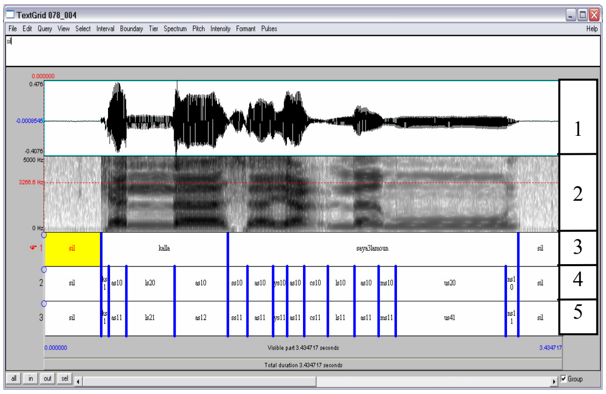

Figure 2. A screenshot of the customized Praat interface: 1) wave, 2) spectrogram, 3) word-level transcription, 4) phoneme-level transcription, 5) allophone-level transcription.

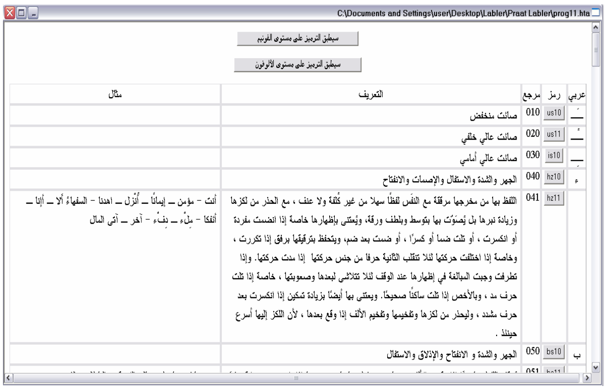

Figure 3. A screenshot of the new transcription with their references and insertion tools.

4. CONCLUSION AND FUTURE WORK

The method for transcription has been applied to Quranic recitation to collect a sufficient Quranic speech database for training and testing. The database will be used to build the Computerized Teaching of the Holly Quran system in the HTK environment. The initial results are encouraging but not enough to be reported here. We hope to report the results of this project in another paper when adequate results are available.

5. ACKNOWLEDGMEN

This paper is supported by KACST (AT-25-113).

6. REFERENCES

[1] TIMIT: Acoustic-Phonetic Continuous Speech Corpus. DMI. 1990.

[2] Moreno, P., O. Gedge, H. Heuvel, H. Höge, S. Horbach, P. Martin, E. Pinto, A. Rincón, F. Senia, R. Sukkar. SpeechDat Across all America: SALA II. Project website: www.sala2.org.

[3] The Spoken Dutch Corpus: http://www.elis.rug.ac.be/cgn/.

[4] Tseng, C., Y. Cheng, W. Lee and F. Huang Collecting Mandarin Speech Databases for Prosody Investigations, The Oriental COCOSDA. Singapore. 2003.

[5] Langmann, D., R. Haeb-Umbach, L. Boves and E. den Os. FRESCO: The French Telephone Speech Data Collection - Part of the European SpeechDat(M) Project. FRESCO. The Fourth International Conference on Spoken Language Processing. Philadelphia. 1: 1918-1921. 1996.

[6] Appen: http://www.appen.com.au

[7] Auran, Cyril, Caroline Bouzon and Daniel Hirst. The Aix-MARSEC project: an evolutive database of spoken British English. International Conference: Speech Prosody 2004. Nara, Japan. March 23-26, 2004.

[8] Hansakunbuntheung, Chatchawarn, Virongrong Tesprasit and Virach Sornlertlamvanich. Thai Tagged Speech Corpus for Speech Synthesis. Proceedings of The Oriental COCOSDA 2003, Singapore, 1-3 October 2003.

[9] Demuynck, Kris, Tom Laureys and Steven Gillis. Automatic Generation of Phonetic Transcriptions for Large Speech Corpora. International Conference on Spoken Language Processing. 1: 333-336. 2002.

[10] Young, S., G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version 3.2)," Microsoft Corporation, Cambridge University Engineering Department, December 2002.

[11] http://www.arts.gla.ac.uk/ipa/ipa.html

[12] Jande, Per-Anders. Automatic detailed transcription of speech using forced alignment and naive pronunciation rules. Speech Recognition Course, Course Project Report. The Royal Institute of Technology (Kungliga Tekniska Högskolan). 2004.

[13] Curl, Traci S. The phonetics of sequence organization: an investigation of lexical repetition in other-initiated repair sequences in American English. Unpublished Ph. D. thesis, University of Colorado. USA. 2002.

[14] Frankel, Joe. Linear dynamic models for automatic speech recognition. Unpublished Ph. D. University of Edinburgh, UK. 2003.

[15] Lesaffre1, Micheline, Koen Tanghe, Gaëtan Martens, Dirk Moelants, Marc Leman, Bernard De Baets, Hans De Meyer and Jean- Pierre. The MAMI query-by-voice experiment: collecting and annotating vocal queries for music information retrieval. Martens 44th International Conference on Music Information Retrieval, Baltimore, Maryland, USA, October 27-30, 2003.

[16] http://coral.lili.uni-bielefeld.de/Documents/sampa.html

[17] Barrobes, Helenca Duxans. Voice conversion applied to text-to-speech systems. Unpublished Ph. D. thesis. Universitat Politecnica. Barcelona, Spain. 2006.

[18] Donovan, Robert Edward. Trainable speech synthesis. Unpublished Ph. D. thesis. Cambridge University. UK. 1996.

[19] Kristinsson, Björn. Towards speech synthesis for Icelandic. MA thesis. University of Iceland. 2004.

[20] Alghamdi, Mansour. Algorithms for Romanizing Arabic Names. Journal of King Saud University: Computer Sciences and Information. 17: 1-27. 2005.

[21] Alghamdi, Mansour. KACST Arabic Phonetics Database, The Fifteenth International Congress of Phonetics Science, Barcelona, 3109-3112. 2003.

[22] http://www.fon.hum.uva.nl/praat/