Survey on common Arabic language forms from a speech recognition point of view

Mohamed Elmahdy[1,2], Rainer Gruhn[1,3], Wolfgang Minker[1], Slim Abdennadher[2]
[1] *Faculty of Engineering & Computer Science, University of Ulm, Ulm, Germany*
[2] *Faculty of Media Engineering & Technology, German University in Cairo, Cairo, Egypt*
[3] *Harman/Becker Automotive Systems, Speech Dialog Systems, Ulm, Germany*

Abstract

Though Arabic language is a widely spoken language, research done in the area of automatic speech recognition for Arabic is very limited compared to other same rank languages like Mandarin. Here we are highlighting the main characteristics of different Arabic forms from a speech recognition point of view, mainly on the acoustic and language level. The characteristics discussed are Arabic phonetics, diacritization problem, grapheme-to-phoneme, and morphological complexity. The Arabic forms discussed in this paper are: classical Arabic, modern standard Arabic, and Egyptian colloquial Arabic. The main purpose of this paper is to summarize main problems in Arabic speech recognition in one paper so researchers in this field can use it as one reference.

Introduction

Arabic language is the largest still living Semitic language based on the number of speakers. It exceeds 250 million first language speakers and the number of speakers using Arabic as a second language can reach four times that number. Arabic is the official language in 21 countries and ranked as the 6^{th} most spoken language based on the number of first language speakers and it is one of the six official languages of the United Nations. The Arabic alphabets are used in other languages like Persian and Urdu.

Arabic language has two main forms: Standard Arabic and Dialectal Arabic. Standard Arabic includes *Classical Arabic* and *Modern Standard Arabic (MSA)* while dialectal Arabic includes all forms of currently spoken Arabic in day life and it vary among countries and deviate from standard Arabic to some extent and even within the same country we can find different dialects. While there are many forms of Arabic, there still many common features on the acoustic level and the language level.

For standard Arabic we have chosen both classical and MSA forms and for dialectal Arabic we have chosen Egyptian colloquial Arabic (ECA) (Egyptian colloquial Arabic usually stand for the spoken language in Cairo) as the typical example since it is the most popular dialect among Arabic speakers. The main speech recognition problems for Arabic discussed in this paper are: morphological complexity, existence of many dialectal forms, and text resources are mainly non-diacritized.

Modern Standard Arabic (MSA)

MSA is the current standard form of Arabic. Almost all Arabic written resources are in MSA. It is the formal spoken Arabic language. MSA is used in books, newspapers, news broadcast, formal speeches, movies subtitling, etc. MSA can be considered as a second language for all Arabic speakers. All Arabic speakers can understand spoken MSA as in news broadcast. MSA is the only accepted Arabic form throughout all the native Arabic speakers, that is why many radio and TV channels use MSA in order to target all Arabic speakers. The allowed syllables in Arabic are CV, CVC, and CVCC where C is a consonant and V is a long or short vowel. So Arabic utterances and words can only start with a consonant.

MSA phonetics inventory consists of 38 phonemes. Those phonemes include 29 original consonants, 3 foreign consonants, and 6 vowels (see Table 1). We use SAMPA notation [13]. The foreign consonants are: /g/, /p/, and /v/ and they are rarely found in MSA and appear in loan words. The phoneme /l‘/ is a rare phoneme as it appears only in the word /?al‘l‘a:h/ (The God) and its derivatives.

Usually the duration of long vowels is approximately double the duration of short vowels. Arabic is characterized by the existence of pharyngeal and emphatic phonemes like /X\/, /t‘/, /d‘/, /D‘/, and /s‘/. Those types of phonemes exist only in Semitic languages [2] [13]. Emphatic sounds have significant effect over the whole word containing the emphatic consonant. Previous work showed a smaller difference between F1 and F2 for all vowels of words containing emphatic consonant in almost any position [17]. Some phoneticians add two more diphthongs which are: /ay/ (/a/ followed by /y/) and /aw/ (/a/ followed by /w/) and considering these diphthongs requires more effort in transcription since they may appear or not.

MSA speech corpora are mainly available in the domain of news broadcast at relatively low price and those corpora can be used to build speaker independent acoustic models [10] [8].

In MSA transcription usually foreign phonemes are not treated as an extra sounds and they are grouped with the closest phoneme, for example /f/ and /v/ are grouped together and treated as the same phoneme, the same approach is applied for /b/ and /p/, and also applied

for /g/, /Z/, and /dZ/.

The reason for this grouping is mainly because foreign phonemes are rarely used in MSA compared to original sounds and due to that the standard Arabic letters does not have any standard letter assigned for foreign sounds. Some efforts were done to differentiate between foreign sounds and original sounds by using non-standard Arabic letters like using the letter پ for the phoneme /p/, using the letter ڤ for the phoneme /v/, and using the letter ڭ for the phoneme /g/, but these conventions are not standard and even standard Arabic keyboard layout does not show these letters and also standard Arabic character sets as the ISO 8859-6 does not include these extra letters [6].

IPA	SAMPA	Description
b	b	Plosive, voiced bilabial
t	t	Plosive, voiceless dental plain
d	d	Plosive, voiced dental plain
tˤ	t‘	Plosive, voiceless dental emphatic
dˤ	d‘	Plosive, voiced dental emphatic
k	k	Plosive, voiceless velar
g	g	Plosive, voiced velar
q	q	Plosive, voiceless uvular
ʔ	?	Plosive, voiceless glottal
p	p	Plosive, voiceless bilabial
f	f	Fricative, voiceless labio-dental
v	v	Fricative, voiced labio-dental
θ	T	Fricative, voicelees interdental plain
ð	D	Fricative, voiced interdental plain
ðˤ	D‘	Fricative, voiced interdental emphatic
s	s	Fricative, voiceless alveolar plain
z	z	Fricative, voiced alveolar plain
sˤ	s‘	Fricative, voiceless alveolar emphatic
ʃ	S	Fricative, voiceless postalveolar
ʒ	Z	Fricative, voiced postalveolar
ʤ	dZ	Affricative, voiced postalveolar
x	x	Fricative, voiceless velar
ɣ	G	Fricative, voiced velar
ħ	X\	Fricative, voiceless pharyngeal
ʕ	?\	Fricative, voiced pharyngeal
h	h	Fricative, voiceless glottal
r	r	Trill, alveolar
l	l	Liquid, dental plain
lˤ	l‘	Liquid, dental emphatic
w	w	Approximant(semi vowel), bilabial
j	j	Approximant(semi vowel), palatal
m	m	Nasal, bilabial
n	n	Nasal, alveolar
i	i	Short vowel, close front unrounded
a	a	Short vowel, open front unrounded
u	u	Short vowel, close back rounded
iː	i:	Long vowel, close front unrounded
aː	a:	Long vowel, open front unrounded
uː	u:	Long vowel, close back rounded

Table 1: Phonemes of Modern Standard Arabic

Classical Arabic

The classical Arabic is the most formal and standard form of Arabic and it is the language of the Koran (the holy book for Muslims). Classical Arabic script represents almost completely the phonetics of the word because the script is fully vowelized and includes diacritic symbols that are usually omitted in MSA.

Arabic language is characterized of the presence of the emphatic consonant /d‘/ and Arabs believes that this phoneme as pronounced in classical Arabic is exclusively appearing in Arabic and not in any other language, that is why Arabic is usually defined as *The language of /d‘/* [11].

Koran phonetics (according to Hafs's narration from Assim) include the sounds of MSA (except foreign phonemes) plus some extra sounds, the following summarize the main extra sounds that can appear in Koran recitation:

- Vowel prolongation with duration of 4, 5 ,or 6 short vowels.
- Necessary prolongation of 6 short vowels.
- Obligatory prolongation of 4 or 5 short vowels.
- Permissible prolongation of 2, 4, or 6 short vowels.
- Nasalization (ghunnah) with duration of 2 short vowels.
- Emphatic pronunciation of the consonant /r/ may happen depending on the context of the consonant /r/.
- Echoing sound in unrest letters (qualquala) for the consonants /q/, /t‘/, /b/, /dZ/, and /d/ may happen depending on the context of those consonants.

Using Koran in speech recognition is currently limited to recitation learning applications as in [14] and [16]. In such applications the acoustic model is trained with Koran recitation and the user is asked to utter a specific verse and then the application identifies any recitation mistakes.

Dialectal Arabic

MSA is not the natural spoken language for native Arabic speakers while colloquial (or dialectal) Arabic is the natural spoken Arabic in everyday life. Colloquial Arabic is not used as a standard form of Arabic in writing or publishing. There are many Arabic dialects and almost every country has its own colloquial form. Even within the same country we can find different dialects. Dialectal Arabic can be divided into two groups: Western Arabic and Eastern Arabic. Western Arabic can be subdivided into Moroccan, Tunisian, Algerian, and Libyan dialects. While Eastern Arabic can be subdivided into Egyptian, Gulf, Damascus, and Levantine. The Damascus Arabic is considered the closest dialect to MSA. Speakers with different dialects usually use MSA to communicate.

Because Arabic dialects are not used in written form, preparing adequate speech corpora for dialectal Arabic for the purpose of acoustic modeling is very costly but still feasible, and preparing large text corpora for dialectal Arabic that can be used in large vocabulary statistical language modeling is more difficult since in large vocabulary language modeling we need a corpus that contains words in the order of millions which is still not feasible. The available corpora for dialectal Arabic are expensive compared to MSA and they are either low quality telephony conversations corpora as the CALLHOME Egyptian Arabic Speech [1], or high quality read speech corpora but with very limited vocabulary as in the Orientel project [4].

Egyptian Colloquial Arabic (ECA)

The main phonetics characteristics of ECA phonetics compared to MSA are:

- /t/ and /s/ are used instead of /T/. e.g. /Tala:Tah/ (three) in MSA is transformed to /tala:tah/ in ECA.
- /g/ is used instead of /Z/ and /dZ/. e.g. /Zami:l/ (beautiful) in MSA is transformed to /gami:l/ in ECA.
- /?/ is used instead of /q/. e.g. /qabl/ (before) in MSA is transformed to /?abl/ in ECA.
- The existence of the mid front unrounded long vowel /E:/ and short vowel /E/ (they do not exist in MSA). e.g. /?iTnayn/ (two) in MSA is transformed to /?itnE:n/ in ECA.
- The existence of the open back unrounded long vowel /A:/ and short vowel /A/ (they do not exist in MSA). e.g. /?arba?\ah/ (four) in MSA is transformed to /?ArbA?\Ah/ in ECA.
- The existence of the mid back rounded long vowel /O:/ (it does not exist in MSA). e.g. /jawm/ (day) in MSA is transformed to /jO:m/ in ECA [15].

On the vocabulary level, some words exist only in ECA and not in MSA like: /t‘ArAbE:zA/ (table) in ECA while it is /t‘awila/ in MSA. The sentence structure in ECA tends to VSO while in MSA it tends to SVO.

The transcription of ECA is difficult because people are so influenced by MSA and always write the MSA word instead, for example the ECA word /tamanjah/ (eight) is usually wrongly transcribed as /tama:njah/ or /Tama:njah/ and keep including the long vowel /a:/ as in MSA while it is replaced in ECA by the short vowel /a/.

Script Diacritization

The fact that there is a strong grapheme-to-phoneme (almost one to one mapping) relation is only true for diacritized Arabic script. Diacritics appear only in sacred books and Arabic language teaching books (see table 2 and 3). MSA is usually written with the absence of the diacritic symbols and the reader infers missing diacritics from the context.

Non-diacritized script leads to lots of ambiguities for the pronunciation and meaning, for instance the non-diacritized word كـتـب can have different possible diacritizations and in each case a different pronunciation and a different meaning as shown in table 4. Another problem is that the vowel diacritic case marker on the last letter in a word is determined by the word position in the sentence for example whether the word is subject or object. Furthermore, the speaker is free to choose whether to pronounce or to omit the vowel case marker.

Actually the problem of missing diacritics is not only a speech recognition problem but it is more crucial in Arabic speech synthetic Text-To-Speech (TTS) systems. Estimation of the correct diacritization will reduce ambiguity in all Arabic speech and language processing tasks. The majority of Arabic corpora available for the task of acoustic modeling have non-diacritized transcription. That is why most of previous work in Arabic acoustic modeling as in [3] include an important stage for automatic script diacritization which is usually built on statistical language modeling to estimate the most probable diacritics for a word given the context where the word appear. Automatic diacritization is a very important area in Arabic speech recognition since it was proven that a fully vowelized Arabic script improves accuracy over non-vowelized script (grapheme-based) [9].

Currently available commercial applications for automatic Arabic diacritization as Fassieh [5] still need manual review to correct errors which is a time consuming and costly operation especially for preparing large diacritized corpora.

Grapheme	Phoneme	Grapheme	Phoneme
أ	?	س	s
إ	?	ش	S
ؤ	?	ص	s‘
ئ	?	ض	d‘
ء	?	ط	t‘
آ	?a:	ظ	D‘
ى	a:	ع	?\
ا (wasla)	?	غ	G
ا	a:	ف	f
ب	b	ق	q
ت	t	ك	k
ث	T	ل	l
ج	Z	م	m
ح	X\	ن	n
خ	x	ه	h
د	d	و	w
ذ	D	ي	j
ر	r	ة	t
ز	z		

Table 2: Arabic grapheme to phoneme conversion

Diacritic	Phoneme	Diacritic Description
بَ	b a	Fatha (short vowel a)
بِ	b i	Kasra (short vowel i)
بُ	b u	Damma (short vowel u)
بْ	b	Sukun (no vowel)
بَّ	b b a	Shadda (double consonant) & Fatha
بِّ	b b i	Shadda & Kasra
بُّ	b b u	Shadda & Damma
بًا	b a n	Tanween (Nunation)[1]Fatha
بٍ	b i n	Tanween Kasra
بٌ	b u n	Tanween Damma
بًّا	b b a n	Shadda, Tanween Fatha
بٍّ	b b i n	Shadda, Tanween Kasra
بٌّ	b b u n	Shadda, Tanween Damma
بَا	b a:	Alif Al-Madd (long vowel a:)
بِي	b i:	Yaa Al-Madd (long vowel i:)
بُو	b u:	Waaw Al-Madd (long vowel u:)

Table 3: Arabic diacritics to phoneme conversion, example with the consonant /b/

Word	SAMPA	English meaning
كَتَبَ	/kataba/	He wrote
كُتِبَ	/kutiba/	It was written
كَتَّبَ	/kattaba/	He dictated
كُتِّبَ	/kuttiba/	It was dictated
كُتُبْ	/kutub/	Books

Table 4: Possible diacritizations for the word كتب with English meaning

Morphological Complexity

Arabic is a morphological very rich language, and hence dealing with Arabic as morphemes rather than words will limit the size of dictionary and drastically decrease number of Out-Of-Vocabulary (OOV) words. For instance a lexicon of 65,000 words in the domain of news broadcast leads to OOV rate of 4% in Arabic while in English it leads to OOV rate of less than 1% [7]. Most words in Arabic have a root that consists of three consonants called radicals (rarely two and four). A large number of affixes (prefixes, infixes, and suffixes) can be added to the three consonant radicals to form patterns. Arabic is a high inflected language with gender, number, tense, person, and case. A single Arabic word can represent a whole English sentence like: وَبِاسْتِطَاعَتِهِم /wabi?stit'a:?\atihim/ (and with their ability) [12].

[1]Tanween (Nunation) may only appears on the last letter of a word.

Conclusion

Working in Arabic speech recognition requires a good knowledge about the characteristics of Arabic language. In this paper we have shown that Arabic language has many important features on the acoustic and language level that have to be well considered in automatic speech recognition research work.

References

[1] Alexandra Canavan, George Zipperlen, and David Graff, 1997. CALLHOME Egyptian Arabic Speech. Linguistic Data Consortium.

[2] Clive Holes, 2004. Modern Arabic: Structures, Functions, and Varieties. Georgetown University Press.

[3] Dimitra Vergyri and Katrin Kirchhoff, 2004. Automatic diacritization of Arabic for acoustic modeling in speech recognition. In proceedings of COLING.

[4] ELRA: European Language Resources Association. URL: `http://www.elra.info/`.

[5] Fassieh, RDI. URL: `http://www.rdi-eg.com/`.

[6] ISO 8859-6: Information processing – 8-bit single-byte coded graphic character sets – Part 6: Latin/Arabic alphabet, 1987. International Organization for Standardization.

[7] J. Billa, M. Noamany, A. Srivastava, D. Liu, R. Stone, J. Xu, J. Makhoul, and F. Kubala, 2002. Audio Indexing of Arabic Broadcast News. In proceedings of ICASSP.

[8] Linguistic Data Consortium (LDC). URL: `http://www.ldc.upenn.edu/`.

[9] M. Afify, L. Nguyen, B. Xiang, S. Abdou, J. Makhoul, Sept. 2005. Recent Progress in Arabic Broadcast News Transcription at BBN. In Proceedings of InterSpeech.

[10] M. Yaseen, M. Attia, B. Maegaard, K. Choukri, N. Paulsson, S. Haamid, S. Krauwer, C. Bendahman, H. Fersoe, M. Rashwan, B. Haddad, C. Mukbel, A. Mouradi, A. Al-Kufaishi, M.Shahin, N. Chenfour, and A. Ragheb, 2006. Building Annotated Written and Spoken Arabic LRs in NEMLAR Project. In proceedings of LREC.

[11] D. Newman, 2002. The phonetic status of Arabic within the world's languages : the uniqueness of the lughat al-daad. Antwerp papers in linguistics, 100, pp. 65-75.

[12] Riyad Alshalabi, 2005. Pattern-based Stemmer for Finding Arabic Roots. Information Technology Journal 4(1):38-43.

[13] SAMPA symbols for Arabic. URL: `http://www.phon.ucl.ac.uk/home/sampa/arabic.htm`.

[14] Sherif Abdou, Salah Eldeen Hamid, Mohsen Rashwan, Abdurrahman Samir, Ossama Abd-Elhamid, Mostafa Shahin, and Waleed Nazih, 2006. Computer Aided Pronunciation Learning System Using Speech Recognition Techniques. In proceedings of Interspeech.

[15] Virginia Stevens and Maurice Salib, 2005. A Pocket Dictionary of the Spoken Arabic of Cairo. The American University in Cairo Press.

[16] Zaidi Razak, Noor Jamaliah Ibrahim, Mohd Yamani Idna Idris, Emran Mohd Tamil, Mohd Yakub, Zulkifli Mohd Yusoff, and Noor Naemah Abdul Rahman, August 2008. Quranic Verse Recitation Recognition Module for Support in j-QAF Learning: A Review. IJCSNS International Journal of Computer Science and Network Security.

[17] Zeki Majeed Hassan and John H. Esling, 2007. Laryngoscopic (Articulatory) and Acoustic Evidence of a Prevailing Emphatic Feature Over the Word in Arabic. In proceedings of the 16th International Congress of Phonetic Sciences.