Construction of Audio Corpus of Non-Native English Dialects -Arabs speakers-

Sara Chellali[1], Soumaya Al-Maadeed[2], Muhammad Asim[2], Maamar Ahfir[3], Walid Khald Hidouci[1]
[1] Laboratory LCSI, National School of Computer Science, ESI, Alger, Alegria
[2] Dep of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar
[3] Dep of Computer Science, University Ammar Telidji, Laghouat, Algeria
sa_chellali@esi.dz, s_alali@qu.edu.qa, muhammad.asim@qu.edu.qa, m.ahfir@mail.lagh-univ.dz, w_hidouci@esi.dz

ABSTRACT

Speech recognition for the academic and standard languages has achieved a great development, but in the face of the multiple dialects with their differences, there is still a noticeable lack of recognition rates, especially if these dialects are for non-native speakers of the target language. The challenges facing the recognition or the identification of non-native dialects are numerous, among them the lack of sound databases, whether approved or not.

This article presents part of our work on creating a sound database of non-native English dialects for Arabs speakers. This database will be used later by our Automatic System to Help Learning English as Foreign Language (ASHLEFL). It contains initially three non-native English dialects respectively Qataris, Egyptians, and Pakistanis speakers.

KEYWORDS

Nonnative English Dialect; Sound Database; Arab speakers.

1. INTRODUCTION

This article is part of a project to develop an automatic system to help to learn English as a foreign language. The role of this system is to detect, then to correct the common errors of the pronunciation of the determined populations, where we begin with the identification of their dialects. The objective of this paper is the presentation of the construction of an audio data base of non-native English dialects of these populations.

2. ENGLISH IN THE WORLD ARAB

The Arab world is the Middle East and North Africa (MENA), divided into 22 countries, 10 of which are African. The Arab region has an area of about 14 million km^2, equivalent to 10.2% of the world's area and contains about 6% of the world population. where Arabic is the official language of these countries.

The status of the English language in the Arab world differs between a Second Language (ESL[1]) and Foreign Language (EFL[2]).

In its report of 2016, the organization EF "Education First" of the language training, educational travel, academic certification and cultural exchange program [1], revealed that MENA countries have the lowest level of English proficiency with the average EF EPI[3] equal 44.92 in the world (72 countries were involved in the tests, including 10 Arab countries). The Arab countries were ranked in the "very low" efficiency group with the exception of the United Arab Emirates and Morocco, which were classified as "low", as shown below Figure 1.

This weakness is attributed to several reasons: historical, religious, cultural, economic, educational, and sometimes political, without forgetting the great difference between the Arabic and English

[1] English Second Language
[2] English Foreign Language
[3] English Proficiency Index

language systems, which greatly affects the pace of learning as English is ESL or EFL [2] and [3]. To improve the efficiency of the Arabs in mastering the English language requires the development of the educational system and attention to language learning from the first levels of education.

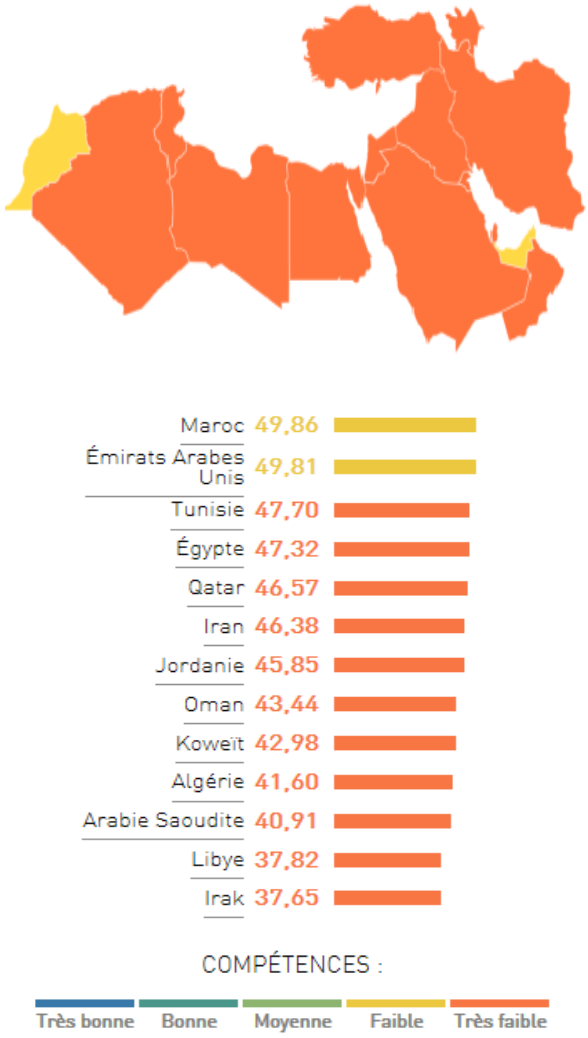

Figure.1 Regional classification according to the EF EPI [1]

3. PROBLEMATIC

The need of audio databases in the domain of Natural Language Processing (language identification, speech synthesis, speakers' identification, speech recognition...) is unavoidable. So, the first step the researchers make is to either provide the database or create it. Despite three decades of work in this area, the availability of databases for academic languages as well as their native dialects remains persistent. As for non-native dialects, this is another challenge.

The English has taken the lion's share of interest and research as the most widely spoken language (more than 500 million speakers). [4] presents some of the native English dialects databases that were created are as follows ANDOSL[4], SEC[5], WSJCAMO[6], TIMIT[7]. TIMIT is the famous and the widely used speech database. It contains 630 native speakers of American English (70% male and 30% female). Each speaker reads 10 sentences and takes approximately 30 seconds. [5] provides an overview of databases existing of non-native speakers and notes the total absence of Arab speakers of English (Table 1). We note that free or paid databases are still very limited, especially for non-native English dialects in the Arab world, which sometimes obliges laboratories to create their own database internally. So, for to approach this project we need to create our own audio database of nonnative English dialects by Arab speakers.

Table 1 Overview of non-native English Databases [5]

Corpus & date	Available at	Speak.	Native Language	Utter.
ATR-Gruhn **2004**	ATR	96	Chinese, German, French, Japanese, Indonesian	15000
Berkeley Restaurant **1994**	ICSI	55	German, Italian, Hebrew, Chinese, French, Spanish, Japanese	2500
Cambridge-Witt **1999**	U. Cambridge	10	Japanese, Italian, Korean, Spanish,	1200
Cambridge-Ye **2005**	U. Cambridge	20	Chinese	1600

[4] Australian National Database of Spoken Language
[5] British English, Spoken English Corpus
[6] British English speech corpus
[7] American English, Texas Instrument and Massachusetts Institute of Technology corpus

Children News **2000**	CMU	62	Japanese, Chinese	7500
CLSU **2007**	LDC		22 countries	5000
CMU	CMU	64	German	452
Cross Towns **2006**	U. Bochum	161	English, French, German, Italian, Spanish	72000
Duke-Arslan **1995**	Duke University	93	15 countries	2200
ERJ **2002**	U. Tokyo	200	Japanese	68000
Fraenki	U. Erlangen	19	German	2148
Hispanic **1998**		22	Spanish	
IBM-Fischer **2002**	IBM	40	Spanish, French, German, Italian	2000
ISLE **2000**	EU/ELDA	46	German, Italian	4000
MIST **1996**	ELRA	75	Dutch	2200
NATO HIWIRE **2007**	NATO	81	French, Greek, Italian, Spanish	8100
NATO M-ATC **2007**	NATO	622	French, German, Italian, Spanish	9833
PF-STAR **2005**	U. Erlangen	57	German	4627
Sunstar **1992**	EU	100	German, Spanish, Italian, Portuguese, Danish	40000
TC-STAR **2006**	ELDA	unknown	EU countries	
Verbmobil **1994**	U. Munich	44	German	

4. CONSTRUCTION OF THE CORPUS

Qatar is a peninsula surrounded by the Arabian Gulf, located to the east of the Arabian Peninsula, bordered by Saudi Arabia and has maritime borders with both the UAE and Bahrain. It covers an area of 11,437 square kilometers. It contains a mixture of races with the following ethnic distribution: 40% Arabs, 18% Indians, 18% Pakistanis, 10% Iranians, 14% Others.

A. Workplace

Audio samples were collected at Qatar University in Doha, Qatar. Where audio recordings were made in different places of the campus.

In the collection of samples, we took into consideration all the places in which the learner could be present and the presence of the external influences. In order to obtain a pure and clean recording, we registered in the Virtual Reality Laboratory of the Department of Computer Science, College of Engineering. The recording was also made at the General Library of the University to introduce the reverberation and echo elements in the phonograms as well as recording abroad to obtain a natural environment containing various types of chaos (sounds of nature, voices of other people ...).

The audio record included various levels available between teachers, staff, students, and researchers (09 female and 25 male).

B. Dialects concerned on sampling collect

The Qatari and Egyptian speakers (whom their native language is Arabic) were chosen along with the Pakistani speakers (for whom Arabic is a religious language for the Pakistani Muslims, - who present about 95% of the Pakistani -)

We begin with these three non-native English dialects namely the Qatari, the Egyptian and the Pakistani (as a neutral dialect for comparison) dialects. The database would then be expanded to include the rest of the non-native English dialects of Arab speakers.

5. CREATE THE AUDIO DATABASE

A. Processing of Voice Data collected

ISBN: 978-1-941968-43-7©2017 SDIWC

The voice recording has been performed over several sessions in in real acoustic environment using Sony Dictaphone.

we asked each speaker to read five times in one session the first ten numbers, twelve isolated words divided into five groups, five short sentences, and a paragraph of 69 words. it wasn't a spontaneous read but prepared read by each speaker.

The length of the recordings ranged from 2 to 6 minutes (some speakers did not respect the number of repetition of reading). After we received all the recordings, we processed them. Where we started the process of cutting into recordings of length ranging from one second to 29 seconds using the free demo program Power MP3 Cutter Professional Version 6.2. (Figure 2). Each record contains one utterance: the ten numbers, one group of isolated words, one sentence, or the paragraph

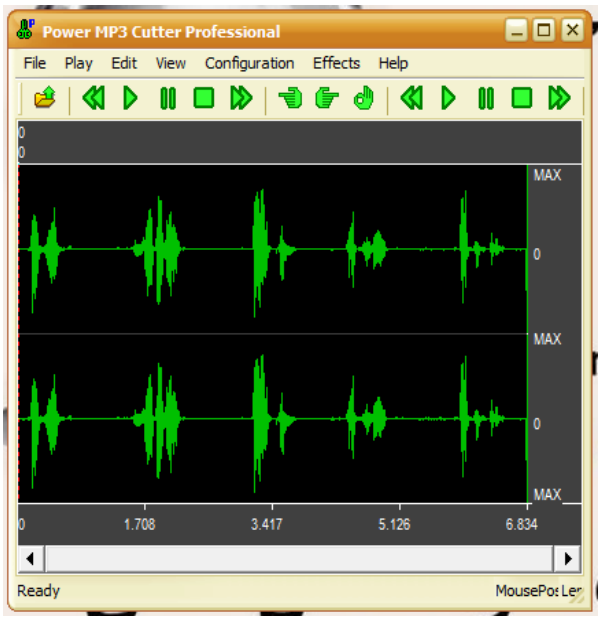

Figure.2 Segmentation of audio recording with Power MP3 Cutter Pro

After the two recording was canceled for two speakers (one Qatari's speaker and one Egyptian's speaker) because they could not read in English and deleted some utterances for a bad recording, we got a corpus comprised of 1535 audio utterances from 34 people.

TABLE 2 DIVIDE FOR SPEAKERS

	FEMALE	MALE	TOTAL
EGYPT	04	03	07
PAKISTANI	02	13	15
QATARI	03	09	12
TOTAL	09	25	**34**

B. Codification

After cutting the recording into audio files. We moved to the coding process where we gave each audio file a nine-digit code with information about the country, city, speaker, and number of utterances

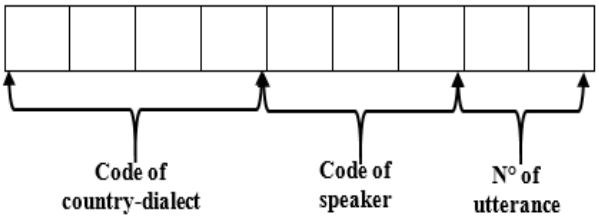

Figure.3 Code of recording sound

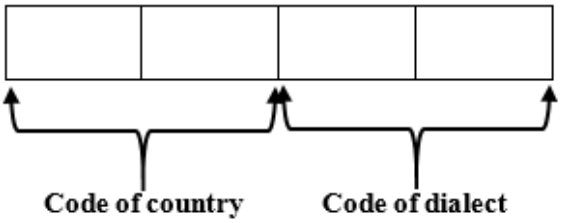

Figure.4 Code of country-dialect

Code of country: we need a code of 2 numbers for coded the 22 countries in the world Arab,

e.g.: 01: Egypt, 02: Pakistani, 03: Qatar.

Code of dialect: we find several dialects in the same Arab country. So, we gave a different code to each dialect.

e.g.: 00 (not identified)
01 (dialect of Cairo)
02 (dialect of Alexandria)

Example:

0	1	0	2	0	2	5	1	0

The utterance n°10 of a speaker n°25 who speak the dialect of Alexandria (Egypt).

ISBN: 978-1-941968-43-7©2017 SDIWC

C. Text and the transcription phonetic

In the choice of text; we tried to take into account the above characteristics, the text used:

i. Digit: One, Two, Three, Four, Five, Six, Seven, Eight, Nine, Ten

/wən/, /tú/, /θrí/, /fɔr/, /fájv/, /síks/, /sɛvən/, /ét/, /nájn/, /tɛn/

ii. Isolated words:

1. Public, Jupiter, Parking, Pepsi, Pakistan

/pəblɪk/, /dʒúpətər/, /párkɪŋ/, /pɛpsi/, /pækəstæn/

2. Victoria, Vacation, Vegetable, Verb, Video

/vɪktɔ́riə/, /vekéʃən/, /vɛ́dʒtəbəl/, /vɜ́rb/, /vídio/

3. Kick, Key, Keyboard, King, Kite

/kík/, /kí/, /kíbɔ̀rd/, /kíŋ/, /kájt/

4. Jupiter, Jug, Jumping, John, Jogging

/dʒúpətər/, /dʒə́g/, /dʒə́mpɪŋ/, /dʒán/, /dʒágɪŋ/

iii. Prepared sentences:

1. I like to eat vegetables
   /áj lájk tú ít vɛ́dʒtəbəlz/
2. Victoria is jumping
   /vɪktɔ́riə íz dʒə́mpɪŋ/
3. John is drinking Pepsi
   /dʒán íz dríŋkɪŋ pɛ́psi/
4. The car is clean
   /ðə kár íz klín/

iv.Paragraph:

For the selected paragraph, the same paragraph used in the website "the speech accent archive " has been adopted [5] :

“Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station”.

/plíz kól stɛ́lə ǽsk hər tú bríŋ ðíz θíŋz wíð hər frə́m ðə stɔ́r síks spúnz ə́v frɛ́ʃ snó píz fájv θík slǽbz ə́v blú ʧíz ǽnd mébi é snǽk fɔ́r hər brə́ðər báb wí ɔ́lso níd é smɔ́l plǽstɪk snék ǽnd é bíg tɔ̀j frág fɔ́r ðə kídz ʃí kǽn skúp ðíz θíŋz ìntú θrí rɛ́d bǽgz ǽnd wí wíl gó mít hər wɛ́nzdi ǽt ðə trén stéʃən/.

6. Conclusion

The database created is a gain in a sense that there exists no database for nonnative English dialects for Arabic speakers (to be completed). However, it is only the first step in our work. The next step is to conduct experiments on the human experts for to validate the hypothesis of their ability of identification of non-native dialects.

Acknowledgment

This paper was made possible by a QUCP award QUCP-CENG-CSE-15-16-1] from the Qatar University. The statements made herein are solely the responsibility of the authors.

References

[1] http://www.ef.dz/epi/

[2] S.S. Sabbah “Negative Transfer: Arabic Language Interference to Learning English” English Language Center, Community College of Qatar, Arab World English Journal (AWEJ) Special Issue on Translation No.4 May, 2015 Pp. 269-288.

[3] M. Elkhair, H. Idriss, “Pronunciation Problems: A Case Study of English Language Students at Sudan University of Science and Technology”, English Language and Literature Studies; Vol. 4, No. 4; 2014, ISSN 1925-4768 E-ISSN 1925-4776, Published by Canadian Center of Science and Education.

[4] M. Alghamdi, F. Alhargan, M. Alkanhal, A. Alkhairy, M. Eldesouki, A. Alenazi, “Saudi Accented Arabic Voice Bank”, Computer and Electronic Research Institute, King Abdulaziz City for Science and Technology (KACST), Riyadh, Saudi Arabia, 25/06/2008

[5] M. Raab, R. Gruhn, E. Noeth, “NON-NATIVE SPEECH DATABASES”, Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2007

[6] http://accent.gmu.edu/browse_language.php?function=detail&speakerid=145