EasyChair Preprint

№ 82

An Empirical Study of Arabic Continuous Speech Recognition Perfomance

Fawaz Al-Anzi and Dia Abuzeina

April 23, 2018

An Empirical Study of Arabic Continuous Speech Recognition Perfomance

Fawaz S. Al-Anzi
Department of Computer Engineering
Kuwait University
Kuwait City, Kuwait
fawaz.alanzi@ku.edu.kw

Dia AbuZeina
Department of Computer Engineering
Kuwait University
Kuwait City, Kuwait
abuzeina@ku.edu.kw

***Abstract*—Although considerable research has been devoted to English speech recognition, rather less attention has been paid to Arabic speech recognition. The Arabic language is one of the most commonly used languages worldwide that is in need for accurate audio to text converters. In this paper, we evaluate the recognition performance of the Arabic continuous speech using Soundflower Mac utility. That is, Soundflower was employed as a speaker-independent continuous speech recognition system to evaluate the word error rate (WER) and the accuracy of the Arabic speech. The study also contains a comparative study of the speech recognition performance for male and female native speakers. The experiments conducted using a broadcast news modern standard Arabic (MSA) speech corpus of 2.63 hours (10 male and 10 female speakers). The experimental results show that the accuracy is 54.02 %, and the accuracy of the male and female speakers is almost same.**

Keywords—Arabic, Speech, Recognition, Corpus, Soundflower.

I. INTRODUCTION

Automatic Speech recognition (ASR) has recently received significant attention as one of successful trend in information retrieval (IR) and intelligent systems. Converting speech into text is an important since it felicitates deploying online audio contents and make it more accessible. However, developing high-quality speech recognition systems is a challenging task and still is a promising research area. Recently, there has been growing interest in speech recognition for the Arabic language as one of the most common languages worldwide. In fact, there is a real need for software tools to transcribe speech into text. Arabic is the language of the holy writings of Islam that raises the demand for software to dictate such huge speech resources. Reference [1] indicated that ASR research is currently moving from mere speech-to-text systems towards "rich transcription" systems, which annotate recognized text with non-verbal information such as speaker identity, emotional state for customer care purposes.

Nevertheless, speech recognition is not a straightforward task, as it requires dynamic programming algorithms along with different stages for training and decoding. Reference [2] demonstrated why speech recognition is difficult. Therefore, obtaining an accurate freely available software is difficult to achieve. However, free and commercial software tools are available for Arabic speech recognition. In this paper, we consider Soundflower [3] Mac utility that is a free, open-source speech application. The goal is to employ this utility to evaluate the performance of the Arabic speech recognition in terms of word error rate (WER) and the recognition accuracy. The study also aims at comparing the recognition performance of male and female Arabic speakers. Reference [4] indicated that the performance of speech recognizers for female speakers is usually worse than that obtained for male speakers. In fact, the research in speech recognition contains different sources of pronunciation variations such as continuous or isolated speech, age, gender, emotion, dialects, noise, different accents, etc. Reference [5] presented the main phonetic differences between the speech of male and female speakers. The previous studies on Arabic speech recognition has not considered speaker's gender on speech recognition. The little research in this domain motivates the authors to to take over this research to find the effect of gender in speech recognition.

We have organized the rest of this paper in the following way. In the next section, we present the literate review. In section 3, we present the speech recognition overview followed by the male and female speech recognition in section 4. Speech corpus information is presented in section 5. The result presented in section 5. Finally, conclusion and future work presented in section 7.

II. LITERATURE REVIEW

In this section, we survey the contributions of Arabic speech recognition. Because speech recognition is a wide multidiscipline topic that contains vast and diverse subtopics, the literature reviewed in this section is restricted to the software developed for dictating (audio-to-text) Arabic speech. Soundflower [3] is a free audio system extension that allows applications to pass audio to other applications. Soundflower employed as speech to text converter that has the following characteristics [6]: a Mac system extension, easy to use, simply presents itself as an audio device, allowing any audio application to send and receive audio with no other support needed.

Sakhr software company developed a commercial ASR [7] engine that has some features such as noisy environments, speaker independent, high accuracy, supports different Arabic accents. The DARPA-funded Babylon project [8] contains Arabic speech recognition as a part of the developed speech-to-

speech translation systems. Hidden Markov Model Toolkit (HTK) [9] is a portable toolkit for speech recognition research. However, the HTK assumes that the files it is using are written using ASCII rather than Unicode, so if the training input text is stored using the standard Arabic character set then it has to be transcribed to something that the HTK can handle. The obvious thing to use is the Buckwalter transcription [10]. CMUSphinx toolkit [11] is another option in the research community that is used to build speech recognition systems. CMUSphinx is an open source speech software from Carnegie Mellon University (CMU) [12]. Unlike HTK, CMUSphinx does support Arabic language that is used directly within the CMUSphinx components such as phonetic dictionaries and the language models. Choosing either HTK or CMUSphinx depends on some aspects such as implementation structure, supporting mobile platform, programming language, etc. nevertheless, both well-known ASR engine share the theoretical background for training and decoding that should give relatively similar outputs.

As existing literature shows, little work devoted to serve the Arabic language compared to the English language. Dragon [13] is an example of software that is used to convert audio text for English. The developer [13] claimed that Dragon is the fastest and most accurate way to interact with your computer. Gotranscript [14] provides speech recognition service for English. They listed some features of the product such as uncompromising quality, rates within the budget, highly accurate transcripts, timely and convenient delivery. Google [15] cloud speech application program interface (API) enables developers to convert audio to text by applying powerful neural network models in an easy to use API. Reference [16] lists the best 2016 voice recognition software for English. Reference [17] compared the performance of three commercially available continuous speech recognition software packages for the English language. The packages include the IBM software that was found to have the lowest mean error rate (7.0 to 9.1 percent) followed by the L&H software (13.4 to 15.1 percent) and then Dragon software (14.1 to 15.2 percent).

III. Speech Recognition Overview

Speech recognition mainly contains two stages, training and decoding. The training stage requires two datasets: a set of speech files and a set of files containing the phonetic transcriptions of the speech files. There are various ways of getting phonetic transcriptions. The easiest is to use phonetic dictionary in combination with the training textual transcription. Some ASR engines such as HTK have a tool for doing this, or it can be prepared manually. Writing a phonetic dictionary is hard, and if the vocabulary has many words then it will be quite time-consuming. For Arabic, it is reasonable to approximate each Arabic character to a single phoneme. So, for instance, assuming that the phonetic transcription of "kataba" is "k a t a b a", Buckwalter transcription [6]. This method of transcription has two advantages, namely that everyone uses it, so that data can easily be made available to other people and it let the researchers to use other people's data; and that it uses one Roman character for each Arabic character, which is helpful, and which most of the other options don't do. There is, however, a problem, which is that it uses a number of non-alphabetic characters that have a reserved meaning in some ASR engines. Another option to represent words in the phonetic dictionary is by using Arabic characters such as "كَتَبَ" with the the phonemes "K AE T AE B AE", as an example. Reference [18] has more information of how generate phonemes for Arabic words. Of course, there are other ways to generate phonetic dictionary for better performance. Linguistic scholars and phonetic specialists might help to in this regards.

In addition to the phonetic dictionary, the training stage also contains declaring language models that is also called grammars. There are all sorts of kinds of grammars to use. The choice of the grammar is, indeed, the key to the performance of the recognizer. The more of constrains in the range of possible utterances, the more accurate the recognizer will be. There are, in particular, two grammars that one can extract from a set of training textual transcription. One says that the target utterance may be an arbitrary sequence of words drawn from the training textual transcription (in short "any word" grammar); the other says that it must be one of the training textual transcription. The first is almost entirely not constraining, and leads to very poor accuracy (but lets you experiment with the effects of different transcriptions, because it relies entirely on the acoustic model); the other is very tightly constraining, and often leads to 100% accuracy. There are other options to write grammars such as probabilistic N-Grams.

Using the phonetic transcriptions of the textual versions of the training speech, the wave files, and the list of phonemes, we can start training using the desired machine-learning tool such as hidden Markov models (HMMs). The output of the training stages is the acoustic models that are used for testing, also called decoding process. The grammars are required throughout testing process. The testing stage employs a dynamic programming algorithm such as Viterbi algorithm to find the most likely textual words sequence of the spoken words. In fact, speech recognition is a complicated process that need to handle different aspects such as Gaussian mixtures model, speech features such as Mel-frequency cepstral coefficients (MFCCs), Baum–Welch algorithm, triphone, pruning, etc.

MacOS recently introduced dictation (speech-to-text) as a feature usable in any application that takes text as input [19]. Reference [19] presented some technical issues that help to run Soundflower application. Figure 1 shows the Soundflower starting page.

Figure 1. A snapshot of the Soundflower speech application

IV. Male and Female Speakers

One of goals of this work is to investigate the speech recognition performance of male and female Arabic speakers. The research on Arabic speech recognition has tended to focus on mixed male-female speech recognition rather than on gender based speech recognition. That is, the training corpus has mixed male and female speech that ignore the acoustic differences between female and male voices. Vogt in reference [20] indicated that the differences in speech features for male and female speakers are a well-known problem and the gender-dependent emotion recognizers perform better than gender-independent ones. Reference [21] separated the training dataset based on the gender. This separation yielded gender dependent HMMs that found significantly improve the word recognition accuracy over the gender independent method. Reference [4] indicated that separating training corpora into male and female acoustic-phonetic models is a common solution to enhance the speech recognition performance.

V. The Speech Corpus

The speech corpus used in this work is an in-house corpus that contains of 275 wave files recorded by 20 Arabic native speakers (10 male and 10 female). Each male speaker utter 15 speech items, while some of female speakers utter less than 15 speech items (see Table 3). The speech files mainly contains local and international news recorded from Al-Sabah TV channel in Kuwait. The modern standard Arabic (MSA) is the language that used by all speakers. The speech file were prepared to have a fixed length between 30-60 seconds. The speech items were sampled at 16 kHz and sum up to 2.63 hours of speech. The training textual transcription of the speech files were prepared by transcribing the wave file according to speakers' utterance. Table 1 composed of the corpus information.

Table 1. The Corpus Information

#	Gender	Number of Speakers	Number of speech files	Length (hour)	Number of Unique words
1	Male	10	150	1.53	5,149
2	Female	10	104	1.10	3,738
	Total	**20**	**254**	**2.63**	**8,887***

*the unique words in the entire corpus is 7,386

VI. Experimental Results

Using the speech corpus described in the previous section, we evaluated the performance for three cases; male only, female only and mixed case (male and female) speech files. The accuracy was used to measure the accuracy that is based on WER. The WER is measured using the following formula [22]: WER= (D+S+I)/N, where D is the deletion errors, S is the substitution errors, I is the insertion errors, and N is the total number of labels in the reference (actual) transcriptions. The accuracy is expressed as:

$$\text{Accuracy} = (1\text{-WER}) \times 100\%$$

Figure 2 shows an example of the Soundflower output of a particular speech file after recognition process. This textual output is aligned with the actual transcription to find D, S, I, N, to be used for calculating the WER according to what we have recognized, either for a single speech file or for the entire speech files collection.

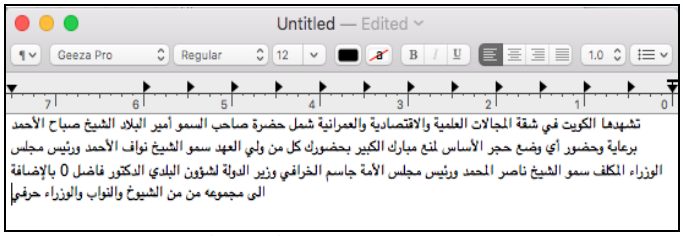

Figure 2. An example of Soundflower output

In the first case, the performance was measured using the male speech files. That is, Soundflower employed to measure the accuracy of 150 speech files that belong to 10 male speakers. Table 2 shows the achieved results of each speaker. The table also shows the range of accuracy [42.26%, 70.39%]. The difference in the scored accuracy is related to several factors such as speaker's anatomy of vocal tract, the speed of the speech, and the accent.

Table 2. WER for Male Only speech

#	Male Speakers	Number of speech files	Length (min:sec)	Accuracy (%)
1	Speaker 1	15	9:29	70.39
2	Speaker 2	15	10:12	48.80
3	Speaker 3	15	9:47	64.93
4	Speaker 4	15	8:46	59.07
5	Speaker 5	15	8:59	42.26
6	Speaker 6	15	9:55	54.67
7	Speaker 7	15	8:12	57.87
8	Speaker 8	15	8:30	55.16
9	Speaker 9	15	9:36	44.58
10	Speaker 10	15	8:54	55.66
	Total	150	92:20	Avg WER=55.33%

For female speakers, 104 speech files were used to evaluate the accuracy. Table 3 shows the accuracy of each person of 10 female speakers. The accuracy range was [46.52%, 68.73%]. This range is close to what we achieved for male speakers. This reveals that the male and female speech recognition is very close in case of using Soundflower tool. This result calls for more research to find the effect of acoustic differences between male and female speakers on Arabic speech recognition.

Table 3. WER for Femal Only speech

#	Female Speakers	Number of speech files	Length (min:sec)	Accuracy (%)
1	Speaker 1	3	2:00	60.51
2	Speaker 2	15	9:42	68.73
3	Speaker 3	15	10:34	57.19
4	Speaker 4	7	5:12	52.07
5	Speaker 5	15	8:00	56.53
6	Speaker 6	15	9:15	50.85
7	Speaker 7	15	8:45	46.89
8	Speaker 8	2	1:27	62.83
9	Speaker 9	2	1:29	67.63
10	Speaker 10	15	9:56	46.52
	Total	104	66:20	Avg WER=56.97%

The average of accuracy for the previous two cases indicates that the female speech recognition outperforms the male speech recognition. The third case separate the corpus for male and female speech to find the accuracy separately. Finally, we evaluated for the mixed male and female case all speech files combined. Table 4 shows the results of the mixed case.

TABLE 4. WER FOR MALE AND FEMALE SPEECH

Gender	Total number of speakers	Number of speech files	Length (min:sec)	Accuracy (%)
Male	10	150	92:20	54.66
Female	10	104	66:20	55.17
Male & Female	20	254	158:40	54.02

Figure 3 shows the information provided in Table 4 as a bar chart graph. The figure shows that the accuracy for Arabic speech is relatively low as the maximum scored of accuracy was 54.02%. This result motivates the research to enhance the performance of Arabic speech recognition.

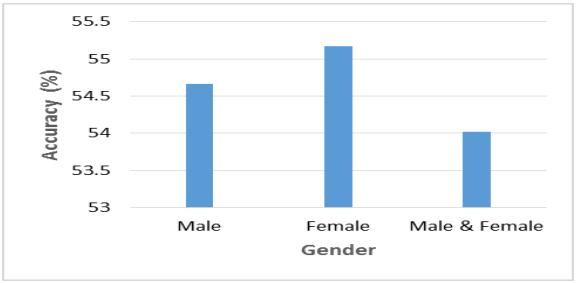

Figure 3. The accuracy of different testing cases

Even though gender is an important factor that has to be considered in speech recognition. However, the experimental evaluation did not show clear performance differences using the prepared corpus and the Soundflower tool. Despite we expect to have less accuracy in the case of female speech, as reported in some literature such as [4], it was found that the female speech perform better than male speech.

VII. CONCLUSISON AND FUTURE WORKS

The study demonstrated the performance of speaker independent Arabic continuous speech recognition. A free MAC software tool used to find the recognition accuracy. It was found that the maximum accuracy scored 54.02% of mixed speech of male and female. The experimental results did not show obvious difference between the accuracy based on the gender. As a future work, we propose more investigation of the effect of gender on Arabic speech recognition.

ACKNOWLEDGMENT

This work is supported by Kuwait Foundation of Advancement of Science (KFAS), Research Grant Number P11418EO01 and Kuwait University Research Administration Research Project Number EO06/12.

We would like to acknowledge the convenience that the researchers got from using Al-Sabah TV in Kuwait as a source of speech collections.

REFERENCES

[1] Metze, Florian, et al. "Comparison of four approaches to age and gender recognition for telephone applications." Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. Vol. 4. IEEE, 2007.

[2] Forsberg, Markus. "Why is speech recognition difficult." Chalmers University of Technology (2003).

[3] http://soundflower.en.softonic.com/mac

[4] R. Vergin, A. Farhat and D. O'Shaughnessy, "Robust gender-dependent acoustic-phonetic modelling in continuous speech recognition based on a new automatic male/female classification," Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, Philadelphia, PA, 1996, pp. 1081-1084 vol.2.

[5] Simpson, Adrian P. "Phonetic differences between male and female speech." Language and Linguistics Compass 3.2 (2009): 621-640.

[6] https://code.google.com/archive/p/soundflower/

[7] http://www.sakhr.com/index.php/en/solutions/speech-technologies

[8] Waibel, Alex, et al. "Speechalator: two-way speech-to-speech translation in your hand." Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations-Volume 4. Association for Computational Linguistics, 2003.

[9] http://htk.eng.cam.ac.uk/

[10] http://www.qamus.org/transliteration.htm

[11] http://cmusphinx.sourceforge.net/wiki/tutorialoverview

[12] http://www.speech.cs.cmu.edu/

[13] http://shop.nuance.co.uk/store/nuanceeu/en_GB/DisplayHomePage

[14] https://gotranscript.com/

[15] https://cloud.google.com/speech/

[16] http://voice-recognition-software-review.toptenreviews.com/

[17] Devine, Eric G., Stephan A. Gaehde, and Arthur C. Curtis. "Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports." Journal of the American Medical Informatics Association 7.5 (2000): 462-468.

[18] Ali, M., Elshafei, M., Al-Ghamdi, M. and Al-Muhtaseb, H. 2009. Arabic Phonetic Dictionaries for Speech Recognition. Journal of Information Technology Research. 2, 4 (2009), 67-80.

[19] http://telestreamblog.telestream.net/2013/12/using-dictation-to-turn-recorded-audio-to-text-2/

[20] Vogt, Thurid, and Elisabeth André. "Improving automatic emotion recognition from speech via gender differentiation." Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa. 2006.

[21] Abdulla, W. H., N. K. Kasabov, and Dunedin–New Zealand. "Improving speech recognition performance through gender separation." changes 9 (2001): 10.

[22] Alghamdi, M., Elshafei, M. and Al-Muhtaseb, H. 2007. Arabic broadcast news transcription system. *Int J Speech Technol*. 10, 4 (2007), 183-195.