The Effects of Speakers' Gender, Age, and Region on Overall Performance of Arabic Automatic Speech Recognition Systems Using the Phonetically Rich and Balanced Modern Standard Arabic Speech Corpus

Mohammad A. M. Abushariah
The University of Jordan
m.abushariah@ju.edu.jo

Majdi Sawalha
The University of Jordan
sawalha.majdi@gmail.com

1 Introduction

Arabic is a Semitic language and one of the most widely spoken languages in the world. It is considered as the first language of over 250 million native speakers ranked as fourth after Mandarin, Spanish and English. Arabic language is in three major forms namely Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA), whereby each form has its own distinctive characteristics. Lack of written and spoken corpora is one of the main issues encountered by Arabic Automatic Speech Recognition (ASR) researchers. Spoken corpora are far less compared to written corpora, resulting in a great need for more speech corpora that serve different application domains of Arabic ASR.

Arabic ASR researchers hardly investigated the effects of the speakers' gender, age, and region on the overall systems performance. This is an important issue to be dealt with and there is a need to produce spoken Arabic language resources that provide an adequate representation.

2 Phonetically Rich and Balanced Corpus

Speech corpus is an important requirement for developing any ASR system. The speech corpus contains 415 sentences recorded on 40 (20 male and 20 female) Arabic native speakers from 11 Arab countries representing three major regions. 367 sentences are considered as phonetically rich and balanced, which are used for training Arabic ASR systems. The remaining 48 sentences are created for testing, which are mostly text independent and foreign to the training sentences and there are hardly any similarities in words.

The motivation behind the creation of this phonetically rich and balanced speech corpus is to provide large amounts of high quality recordings of Modern Standard Arabic (MSA) making it suitable for the design and development of any speaker-independent, continuous, and automatic Arabic ASR system.

3 Arabic Automatic Speech Recognition

In order to evaluate the speech corpus, the Arabic ASR system is developed using the Carnegie Mellon University Sphinx 3 tools. The speech engine uses 3-emitting state Continuous Density Hidden Markov Model for tri-phone based acoustic models. The language model contains uni-grams, bi-grams, and tri-grams.

The major implementation requirements and components for developing the Arabic speaker independent automatic continuous speech recognition system consist of feature extraction, Arabic phonetic dictionary, the acoustic model training, and the statistical language model training, which are identified in the HMM-based architecture of the system as shown in Figure 1.

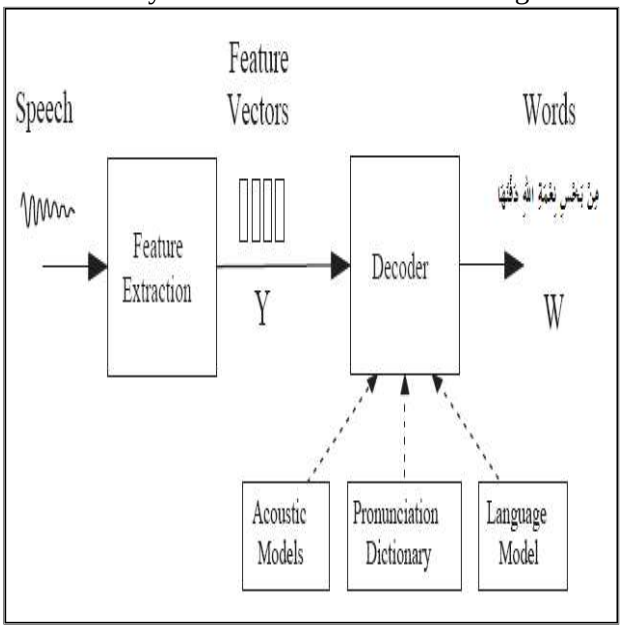

Figure 1: Architecture of the HMM-based Arabic Speaker Independent Automatic Continuous Speech Recognition System

The impact of speakers' variabilities such as gender, age, country and region are examined in this work. Based on 38 hours of training speech data, the acoustic model is composed of 64 GMD and the state distributions tied to 350 senones. Using three different data sets, this work obtains 95.56% and 6.14% average word recognition correctness rate (WRCR) and average word error rate (WER), respectively for the same speakers with different sentences. For different speakers with same sentences, this work obtains 98.81% and 1.81% average WRCR and average WER, respectively, whereas for different speakers with different sentences 95.39% and 6.39% average WRCR and average WER, respectively are achieved.

4 Effects of Speakers' Gender, Age, and Region on the Overall Performance of the ASR Systems

It is found that utterances collected from female speakers achieve better performance than that of the male speakers. This is due to the fact that male and female speakers obviously differ in features and characteristics of the voice. For different speakers with same sentences (speakers independent with text dependent), female speakers obtain an average WRCR of 99.06% and an average WER of 1.33%, whereas male speakers obtain an average WRCR of 98.56% and an average WER of 2.29%. On the other hand, for different speakers with different sentences (speakers independent with text independent), female speakers obtain an average WRCR of 95.78% and an average WER of 5.29%, whereas male speakers obtain an average WRCR of 95.01% and an average WER of 7.48%. From technical perspective, the Mel Frequency Cepstral Coefficients (MFCCs) rely heavily on the energy or loudness and the frequency as extracted from the input speech signals. In order to verify this point, one sound file is randomly selected for the first training sentence for all speakers. Figure 2 shows the gender comparison based on the mean values for the loudness (dB) and the frequency (Hz) using PRAAT.

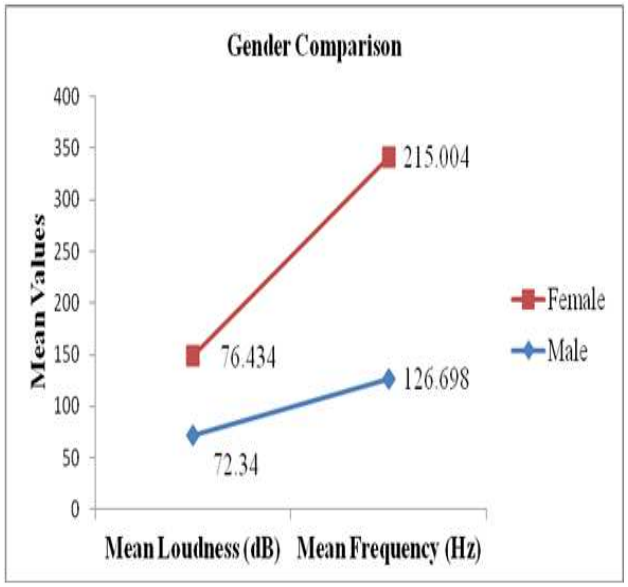

Figure 2: Gender Comparison based on Mean Values for the Loudness and Frequency

Based on Figure 2, it is clearly shown that female speakers have higher means of loudness (dB) and frequency (Hz) than the male speakers. As a result, the feature vectors produced using the MFCC are expected to be of better quality and more intelligible than the male ones, which positively reflects on the WRCR and WER for the female speakers.

In addition, the speakers' age was also examined in this work. It is found that speakers that are less than 30 years old outperformed speakers that are 30 years old and above. This is due to vocal characteristics, whereby as the speaker grows older the vocal characteristics change and that obviously affects the speech recognition systems' performance. It is noticed that younger speakers have better vocal characteristics than the older speakers. Results will be shown in the final manuscript.

The effects of speakers' country and region have also been examined in this research work. Speakers living in the Levant region outperform the speakers living in Gulf and Africa regions although all of them have recorded in the MSA. Speakers from Africa region are influenced by their dialects even though they were asked to record in MSA, but the dialect is really influential especially for speakers from Sudan and Egypt. Therefore, the region where the speaker is located can affect the speech recognition systems' performance. Results will be shown in the final manuscript.

Experimental results will be reported in detail in the final manuscript. However, they show that the developed systems are speaker independent and text independent, and are highly comparable and better than many reported Arabic ASR research efforts.

The final Arabic ASR system is able to successfully recognize speech from speakers with different variabilities. It is also able to narrow down differences with respect to gender, age, country, and region of the speakers. In conclusion, the MSA phonetically rich and balanced speech corpus has shown excellent performance in this research work

References

Mohammad Abd-Alrahman Mahmoud Abushariah, Raja Noor Ainon, Roziati Zainuddin, Assal Ali Mustafa Alqudah, Moustafa Elshafei Ahmed, Othman Omran Khalifa (2012). Modern Standard Arabic Speech Corpus for Implementing and Evaluating Automatic Continuous Speech Recognition Systems. Journal of the Franklin Institute, Elsevier, Vol. 349, No. 7, pp. 2215 – 2242.

Mohammad A. M. Abushariah, Raja N. Ainon, Roziati Zainuddin, Moustafa Elshafei, and Othman Khalifa (2012). Arabic Speaker-Independent Continuous Automatic Speech Recognition Based on a Phonetically Rich and Balanced Speech Corpus. The International Arab Journal of Information Technology, Vol. 9, No. 1, pp. 84 – 93.

Mohammad A. M. Abushariah, Raja N. Ainon, Roziati Zainuddin, Moustafa Elshafei, and Othman O. Khalifa (2012). "Phonetically Rich and Balanced Text and Speech Corpora for Arabic Language". Language Resources and Evaluation Journal, Springer, Vol. 46, Issue 4, pp. 601 - 634.