A STUDY OF ARABIC SPEECH RECOGNITION ACCURACY USING A FULLY DIACRITIZED CORPUS (BUKHARI CORPUS)

Ali Zahran Al-Bimani

Abstract

The thesis focuses on the study of speech recognition for the Arabic language. The aim is to emphasize the importance of using a fully diacritic[1] Arabic corpus in Arabic speech recognition systems rather than a Romanized[2] one. To achieve this aim a sample corpus of an old Arabic book called Sahih Al-Bukhari, was used. The sample corpus contains a fully diacritic Arabic text of the sayings of Prophet Mohammed, PBUH[3], and that is the reason, therefore, that it was chosen for the study. For its analysis, the study used an open source speech recognition engine called Sphinx 3 which is free, and is readily available online for researchers with all its documentation. After comparing the accuracy of the Arabic speech recognition system using the two types of Arabic corpuses -- Fully diacriticized and Romanized text -- the study showed that better accuracy was achieved with fully diacriticized Arabic corpus. This conclusion can be used as a first step towards finding other solutions for the challenges facing Arabic speech recognition, to encourage researchers to develop more studies in the Arabic speech recognition field, and to build more corpuses of Arabic language using old diacritic Arabic books. The findings of this study will also hopefully encourage Arab speaking universities to take more interest in establishing Arabic language resource centers that will provide more digital data, both audio and text, to be used in the studies of Arabic speech and text processing.