بسم الله الرحمن الرحيم

∞∞∞

تم رفع هذه الرسالة بواسطة / سامية زكى يوسف

بقسم التوثيق الإلكتروني بمركز الشبكات وتكنولوجيا المعلومات دون أدنى مسئولية عن محتوى هذه الرسالة.

ملاحظات: لا يوجد

Ain Shams University
Faculty of Computer & Information Science
Computer Science Department

Developing High Performance Arabic Speech Recognition Engine

A THESIS

Submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Computer Science

Faculty of Computer and Information Sciences

Computer Science Department

Ain Shams University

By

Hamzah Ahmed Abdurab Alsayadi

Supervisors

Prof. Dr. Zaki Taha Ahmed Fayed

Computer Science Department

Faculty of Computer & Information Sciences

Ain Shams University

Dr. Islam Mohamed El-Sayed Hegazy

Computer Science Department

Faculty of Computer & Information Sciences

Ain Shams University

2022

Ain Shams University

Developing High Performance Arabic Speech Recognition Engine

by

Hamzah Ahmed Abdurab Alsayadi

A THESIS

Submitted in partial fulfillment of the requirements of the degree of
Doctor of Philosophy in Computer Science

Faculty of Computer and Information Sciences
Computer Science Department

under supervision of

Prof. Dr. Zaki Taha Fayed

Dr. Islam Hegazy

Faculty of Computer and Information Sciences
Computer Science Department
Ain Shams University
EGYPT

September 2022

Declaration of Authorship

I, Hamzah A. Alsayadi, declare that this thesis titled, 'Developing High Performance Arabic Speech Recognition Engine' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed: _____

Date: _____

Dedication

To

My Parents

&

My Wife

&

My Children

Acknowledgements

All acknowledgments and gratitude are due to ALLAH for blessing and support to crown this work by success. I would like to express my deepest gratitude and sincere thanks to my supervisor **Prof. Zaki Taha Fayed**, for his excellent guidance, helpful discussions, constructive guidance which was the cause for the entire work to be carried out. And constantly encouraged me during the development of this work. I learned from him a lot of things in my study and live. I always received many useful suggestions from him in our regular meetings, especially when problems occurred and research results were not as promising as expected. It is actually extreme luck to be one of his students. I hope him all the best, and again I am full of gratitude for him.

Also, I would like to thank my co-supervisor **Dr. Islam Hegazy**, for his guidance in conducting this research. He helped me a lot in improving my research skills through his valuable feedback on our publications. I hope them all the best, and again I am full of gratitude for him.

I would like to express my deepest gratitude and sincere thanks to my great teacher **Dr. Abdelaziz A. Abdelhmid**, for his excellent guidance, helpful discussions, constructive guidance which was the cause for the entire work to be carried out. He suggested problems in this thesis, read many of my drafts, and constantly encouraged me during the development of this work, without which I would be lost in this diverse field of study. I am grateful for the time he spared for me, in spite of his tight schedule. **Dr. Abdelaziz A. Abdelhmid** taught me many things in both science and life. I hope him all the best, and again I am full of gratitude for him.

Not forgotten, my deep gratefulness to my father, my mother, my wife, my children, my brothers and my sisters for their support and patience all the time of my study.

Also, I am grateful to faculty of Computer and Information Sciences - Ain Shams University and Ibb University for giving me a chance to study a Ph.D. degree.

Finally, I would like to thank all my colleagues and friends for their support.

Hamzah Ahmed Alsayadi

Abstract

Speech recognition systems play an important role in human–machine interactions. Many systems exist for Arabic speech with modern standard Arabic (MSA), however, there are limited systems for dialectal Arabic speech. Arabic language has a set of sound letters called diacritics, these diacritics play an essential role in the meaning of words and their articulations. The change in some diacritics leads to a change in the context of the sentence. However, the existence of these letters in the corpus transcription affects the accuracy of speech recognition. In addition, the Arabic language comprises many properties, some of which are ideal for building automatic speech recognition systems such as syntax and phonology, while other properties are unsuitable for developing speech systems. Importantly, most data are in non-diacritized form, vary in dialect, and contain morphological complexity. Moreover, the Arabic dialects lack a standard structure. Arabic automatic speech recognition (ASR) methods with diacritics have the ability to be integrated with other systems better than Arabic ASR methods without diacritics. There are two approaches for automatic speech recognition including: i) traditional ASR based on traditional methods; ii) end-to-end ASR based on deep learning methods. In this thesis, we employed a high performance multi Arabic speech recognition system using conventional ASR and end-to-end ASR approaches. We present different Arabic ASR systems for diacritized MSA, non-diacritized modern standard Arabic (MSA), dialectal Arabic. This thesis comprises conventional Arabic ASR and end-to-end Arabic ASR approaches as follows:

Conventional Arabic ASR: in this approach, our overall system is a combination of seven acoustic models based on Gaussian mixture model (GMM), subspace GMM (SGMM), and deep neural network (DNN) for diacritized Arabic. Acoustic features are created using Mel-Frequency cepstral coefficients (MFCC) which is adapted based on linear discriminative analysis (LDA) method. This acoustic features is used to train and evaluate all models. After GMM model training, it is adapted using two adaptation techniques namely maximum mutual information (MMI) and minimum phone error (MPE) to build new models based on main acoustic and GMM features. Then, SGMM is trained based on main acoustic and GMM features. We used one adaptation technique namely boosted MMI (bMMI) to adapt SGMM model in order to produce a new model. Finally, we employ DNN models based on main acoustic and GMM features. After DNN model training, it is adapted using one MPE technique to build a new model.

End-to-end Arabic ASR: in this approach, the application of state-of-the-art end-to-end deep learning approaches are investigated to build robust Arabic ASR systems for diacritized MSA, non-diacritized MSA, and dialectal Arabic. This approach includes two systems namely: i) end-to-end Arabic ASR based encoder-decoder which is state-of-the-art for only diacritized MSA; ii) end-to-end Arabic ASR based on CNN-LSTM with attention-based model, which is state-of-the-art for Arabic and dialectal Arabic ASR. Acoustic features are built based on the MFCC and the log Mel-Scale Filter Bank energies. In end-to-end Arabic ASR based on encoder-decoder, we propose bidirectional long short term memory (BLSTM) and joint connectionist temporal classification (CTC) with attention-based models for an encoder-decoder model. BLSTM is used as an encoder for network training, joint CTC with attention-based models are used as adaptation processes to enhance performance, and BLSTM and joint CTC with attention-based models as decoder for the recognition process. In addition, we build an n-gram Language model (LM) based recurrent neural network (RNN). The decoder is integrated with language model to perform the recognition process. While in the second type of end-to-end ASR, we propose a hybrid model based on convolutional neural network (CNN) and long short term memory (LSTM) models for network training and LSTM with attention-based model are used as decoder for the decoding process. In addition, a word-based language model (LM) is employed as an external LM to achieve better performance and accuracy based on RNN and LSTM. The decoder depends on the trained external LM to improve and enhance the end-to-end ASR performance. The external LM is utilized to enhance the performance of the end-to-end ASR. We employ four acoustic models for diacritized MSA, non-diacritized MSA, augmented non-diacritized MSA, and dialectal Arabic. These acoustic models are built and evaluated separately. Furthermore, there is no prior research that employed data augmentation for CNN-LSTM and attention-based models in Arabic ASR systems. Thus, Data augmentation is applied on the original corpus for increasing training data by applying noise adaptation, pitch-shifting, and speed transformation. This system is considered a multi Arabic ASR system.

To train and evaluate all models, we use the standard Arabic single speaker corpus (SASSC) as MSA data and the third multi-genre broadcast (MGB-3) as dialectal Arabic data. We report word error rate (WER) for all systems. Conventional Arabic ASR is evaluated based on diacritized SASSC and achieved 33.72% as the best WER. End-to-end Arabic ASR based on encoder-decoder is also evaluated based

on diacritized SASSC with 31.10% as WER. The Joint CTC-attention ASR framework reduced WER by 2.62% over conventional Arabic ASR. CNN-LSTM with an attention framework is achieved 28.48%, 14.96%, 10.41%, and 62.02% WER based on diacritized SASSC, non-diacritized SASSC, augmented non-diacritized SASSC, and dialectal MGB-3, respectively. The CNN-LSTM with an attention framework could achieve a WER better than conventional ASR and the Joint CTC-attention ASR by 5.24% and 2.62%, respectively. In addition, WER for non-diacritized data is significantly improved when compared to diacritized data. The achieved average reduction in WER is 13.52%. Results also show that applying data augmentation improved word error rate (WER) when compared with the same approach without data augmentation. The achieved average reduction in WER is 4.55%.

Contents

List of Figures