Deep Learning for Acoustic Modeling for Arabic Speech Recognition

By

Engy Ragaei Rady Abdelmaksoud

Lecturer Assistant, Basic Science Department,
Faculty of Computers and Information, Fayoum University

A Thesis
Submitted to Physics Department, Faculty of Science, Fayoum University

In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

In

Experimental Physics

Under the Supervision of

Prof. Arafa Sabry Hassen
Physics Department
Faculty of Science
Fayoum University

Prof. Nabila Mohamed Hassan
Basic Science Department
Faculty of computers and information
Fayoum University

Prof. Mohamed Hesham Farouk El-Sayed
Engineering Math. and Physics Department
Faculty of Engineering
Cairo University

Faculty of Science
Fayoum University
2021

ABSTRACT

In this thesis, we demonstrate the deep learning techniques for performing Arabic automatic speech recognition (AASR) via speech and emotion. It comprises two parts.

The first part is focused on recognizing twenty isolated Arabic words. Two techniques are used during the feature extraction phase; Mel frequency spectral coefficients (MFSC) and Gammatone-frequency cepstral coefficients (GFCC) with their first and second-order derivatives. The two-dimensional convolutional neural network (2-D CNN) is mainly used to execute the feature learning and classification process. CNN achieved performance enhancement in automatic speech recognition (ASR). Local connectivity, weight sharing, and pooling are the crucial properties of CNNs that have the potential to improve ASR. We tested the CNN model using an Arabic speech corpus of isolated words.

The used corpus is synthetically augmented by applying different transformations such as changing the pitch, the speed, the dynamic range, adding noise, and forward and backward shift in time. It was found that the maximum accuracy obtained when using GFCC with 2-D CNN is 99.77 %. The outcome results of this work are compared to previous reports and indicate that CNN achieved better performance in AASR.

In the second part, an Arabic automatic speech emotion recognition from acoustic features of speech is performed. It is focused on identifying hidden emotion in speech. A novel one-dimension convolutional neural network (1-D CNN) for Arabic automatic speech emotion recognition (AASER) is introduced.

We present results for the recognition of the three emotions happy, angry, and surprised. The Arabic natural audio dataset (ANAD) is used. Twenty-five low-level descriptors (LLDs) are extracted from the speech signals. Different combination of extracted features is examined. For the classification stage, a deep feed-forward neural network (DFFNN) and a one-dimension convolutional neural network (1D-CNN) were used. Also, the problem of imbalances samples in the dataset is managed by using the borderline-synthetic minority over-sampling technique (B-SMOTE). It is shown from the results that the best accuracy is obtained when using all the extracted features with 1D-CNN, which is 99.05 %. When the combination of line spectral frequency (LSF) and Mel-frequency cepstrum coefficients (MFCC) is used, the accuracy becomes 98.92 %. This result is not too much different from the accuracy of using all the extracted features. The obtained results showed an improvement compared to previous studies.