Faculty of Engineering
Communication & Electronic Dept.

جـــامعة الفـــيوم
Fayoum University

Enhancement Quality and Accuracy of Speech Recognition System Using Multimodal Audio Visual Speech Signal

By

Eslam Eid Ali Mohammed El Maghraby

Under the supervision of

Prof. Dr.

Amr Mohamed Refaat Gody

Professor of Digital Signals
Electronics and Communication
Department,
Faculty of Engineering
Fayoum University

Prof. Dr.

Mohamed Hesham Farouk El-Sayed

Professor in
Engineering Math. and Physics
Department,
Faculty of Engineering,
Cairo University

Faculty of Engineering
Fayoum University

2020

Faculty of Engineering
Communication & Electronic Dept.

جـــــامعة الفـــيوم
Fayoum University

Enhancement Quality and Accuracy of Speech Recognition System Using Multimodal Audio Visual Speech Signal

A Thesis

Submitted in fulfillment of degree of

Doctor of Philosophy in Electronics and Electrical Communications Engineering

Submitted by

Eslam Eid Ali Mohammed El Maghraby

Supervised by

Prof. Dr. Amr Mohamed Refaat Gody

Professor of Digital signals, Electronics and Communication Department
Faculty of Engineering, Fayoum University

Signature:…….

Prof. Dr. Mohammed Hisham Farouk

Professor of Engineering Math. & Physics Dept.
Faculty of Engineering, Cairo Univ

Signature:…….

Faculty of Engineering
Fayoum University
2020

Faculty of Engineering
Communication & Electronic Dept.

جـــامعة الفـــيوم
Fayoum University

Approval Sheet

Enhancement Quality and Accuracy of Speech Recognition System Using Multimodal Audio Visual Speech Signal

By

Eslam Eid Ali Mohamed El Maghraby

This thesis for ph.D degree has been

Approved by:

Prof.Dr. Amr Mohamed Refaat **(Main Advisor)**

Signature:……

Prof.Dr. Mohammed Hisham Farouk **(Advisor)**

Signature:……

Prof. Dr. Rania Ahmed Abdel Azeem Abul Seoud **Examiner**

Signature……

Prof. Dr. Muhammad Fathi Abu Al Yazid **Examiner**

Signature……

Fayoum University
2020

ABSTRACT

Multimodal speech recognition is proved to be one of the most promising solutions for robust speech recognition, especially when the acoustic signal is corrupted by noise. The visual signal can be used to obtain more information to enhance the speech recognition accuracy in noisy system because it is not affected by the acoustic noise. In the situations when the SNR of acoustic signals is low, the video cues can compensate the acoustic signals, and thus their method significantly improve the recognition accuracy

The critical stage in designing robust speech recognition system is the choice of reliable classification method from large variety of the existing classification techniques. This research introduces an Audio-Visual Speech Recognition (AVSR) model using both audio and visual speech modality to improve recognition accuracy in a clean and noisy environment.

The two main contributions of this work are:

- First, In order to choose the most effective visual extraction methods which give the enhancement for the speech recognition system, we propose to compare different visual features like Discrete Cosine Transform (DCT), blocked DCT, and Histograms of Oriented Gradients with Local Binary Patterns (HOG+LBP), then Principle Component Analysis (PCA) have been used for dimensionality reduction purpose to extract the effective features vector length. These features are then early integrated with audio features obtained by using Mel Frequency Cepstral Coefficients (MFCCs) and feed to the classification process.
- Second: The Classification process is performed by using one of the main Deep Neural Network (DNN) architecture, Bidirectional Long-Short Term Memory (BiLSTM) have been validated against the Convolution Neural Network (CNN) and the traditional Hidden Markov Models (HMMs).

The effectiveness of the proposed model is demonstrated on two multi-speakers AVSR benchmark datasets named AVletters and GRID on different Signal to Noise Ratios (SNR). The experimental results

show that early integration between audio and visual features achieved an obvious enhancement in the recognition accuracy and prove that BiLSTM is the most effective classification technique when compared to CNN and HMM. In case of GRID, using integrated audio-visual features achieved highest recognition accuracy of 99.13% and 98.47% with enhancement up to 9.28% and 12.05% over audio-only for clean and noisy data respectively. For AVletters, the highest recognition accuracy is 93.33% with enhancement up to 8.33% over audio-only. The obtained results show a performance enhancement compared to previously obtained audio-visual recognition accuracies on GRID and AVletters and prove the robustness of our BiLSTM-AVSR model when compared with CNN and HMM. This is because BiLSTM considers the sequential characteristics (Temporal behaviors) of the speech signal.