AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer and Systems Engineering

A Robust Audio-Visual Speech Recognition using Improved Features

A Thesis submitted in partial fulfillment of the requirements of
Doctor of Philosophy in Electrical Engineering
(Computer and Systems Engineering)

by

Ali S. Saudi
Master of Science in Computer Engineering
(Computer Engineering)
Faculty of Engineering, Arab Academy for Science, Technology and Maritime Transport, 2012

Supervised By

Prof. Hazem M. Abbas

Dr. Mahmoud I. Khalil

Cairo, 2019

Ain Shams University
Faculty of Engineering
Computer and Systems Engineering Department

Approval Sheet

Name: Ali Salih Mahmoud Saudi

Thesis: A Robust Audio-Visual Speech Recognition using Improved Features

Degree: Doctor of Philosophy in Electrical Engineering (Computer and Systems Engineering)

Examiners' Committee

Name and Affiliation **Signature**

1- Prof. Mohsen Abd-Elrazek Rashwan
Professor of Electronics and Communication Engineering ……………….
Faculty of Engineering, Cairo University, Egypt

2- Prof. Hazem Mahmoud Abbas ……………….
Professor of Computer Engineering
Faculty of Engineering, Ain Shams University, Egypt

3- Prof. Mohamed Watheq El-Kharashi ……………….
Professor of Computer Engineering
Faculty of Engineering, Ain Shams University, Egypt

4- Assoc. Prof. Mahmoud Ibrahim Khalil ……………….
Professor of Computer Engineering
Faculty of Engineering, Ain Shams University, Egypt

Examination Date: … / … / ……

Statement

This thesis is submitted as a partial fulfillment of Doctor of Philosophy in Electrical Engineering in Computer and Systems Engineering, Faculty of Engineering, Ain shams University. The author carried out the work included in this thesis, and no part of it has been submitted for a degree or a qualification at any other scientific entity.

Student Name
Ali Salih Saudi

....

Date:

Researcher Data

Name: Ali Salih Saudi
Date of Birth: 15/10/1980
Place of Birth: Cairo, Egypt
Last academic degree: Master of Science in Computer Engineering
Field of specialization: Computer Engineering
University issued the degre: Arab Academy for Science, Technology and Maritime Transport (AASTMT)
Date of issued degree: July 2012
Current job: Assistant Lecturer at German University in Cairo

Thesis Summary

This research investigates the enhancement of a speech recognition system that uses both audio and visual speech information in noisy environments by presenting contributions in two main system stages: front-end and back-end. The double use of Gabor filters is proposed as a feature extractor in the front-end stage of both modules to capture robust spectro-temporal features. The Gabor features simulate the underlying processing chain that occurs in the Primary Audio Cortex (PAC) in conjunction with Primary Visual Cortex (PVC). We named it GAF and GVF. The performance obtained from the resulted GAF and GVF is compared to the performance of other traditional features such as MFCC, PLP, RASTA-PLP audio features, and DCT2 visual features. The experimental results show that a system utilizing GAF and GVF has attained a 98.89% and 69.23% recognition accuracy, respectively, which significantly outperforms the traditional audio and visual features, especially in a low-Signal to Noise Ratio (SNR) scenario.

To improve the back-end stage, a complete framework of synchronous Multi-Stream Hidden Markov Model (MSHMM) is used to solve the dynamic stream weight estimation problem for Audio-Visual Speech Recognition (AVSR). To demonstrate the usefulness of the dynamic weighting in the overall performance of AVSR system, we empirically show the preference of Late Integration (LI) compared to Early Integration (EI) especially when one of the modalities is corrupted. The results confirm that the proposed AVSR-LI model that utilize the dynamic weighting scheme outperforms the AVSR-EI model by a large difference by improving the average recognition accuracy from 90.65% to 92.83% with approximately 23.33% relative improvement.

Prompted by the great achievements of deep learning in solving AVSR problems, we propose a deep AVSR model based on Long Short-Term Memory Bidirectional Recurrent Neural Network (LSTM-BRNN). The proposed deep AVSR model utilizes the Gabor filters in both the audio and visual front-ends with LI scheme. This model is termed as Gabor LSTM-BRNN$_{av}$-LI model. The experimental results show that the deep Gabor (LSTM-BRNNs)-based model achieves superior performance when compared to the (GMM-HMM)-based models which utilize the same front-ends. The results show that the proposed Gabor LSTM-BRNN$_{av}$-LI model outperforms the Gabor HMM$_{av}$-LI model by a large difference by improving the average recognition accuracy from 92.83% to 94.15% with approximately 18.39% relative improvement. Furthermore, the use of GAF and GVF in both audio and visual front-ends attain significant improvement in the performance compared to the traditional audio and visual features.

To demonstrate the effect of dynamic weighting on improving the AVSR performance in low SNR scenarios, we propose a set of experimental comparisons between the LI and

EI schemes. The results confirm that the proposed Gabor LSTM-BRNN$_{av}$-LI model that utilize the dynamic weighting scheme outperforms the Gabor LSTM-BRNN$_{av}$-EI model by a large difference by improving the average recognition accuracy from 92.18% to 94.15% with approximately 25.19% relative improvement. All of these models were trained and tested using clean and noisy recordings from CUAVE corpus.

Keywords:
Gabor Filters, Visual Feature Extraction, Audio-Visual Speech Recognition, Synchronous Multi-Stream Hidden Markov Model, Audio-Visual Integration, Stream Weight, Reliability Measures, Audio-Visual Databases, Bidirectional Recurrent Neural Network.

Acknowledgment

Ali Salih Saudi
Computer and Systems Engineering
Faculty of Engineering
Ain Shams University
Cairo, Egypt
March 2019

Contents

List of Figures