



**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

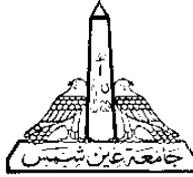
Supervised By

**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 degree:** 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<sub>av</sub>-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<sub>av</sub>-LI model outperforms the Gabor HMM<sub>av</sub>-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<sub>av</sub>-LI model that utilize the dynamic weighting scheme outperforms the Gabor LSTM-BRNN<sub>av</sub>-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

<b>Contents</b>	<b>xii</b>
<b>List of Figures</b>	<b>xvi</b>
<b>List of Tables</b>	<b>xviii</b>
<b>Abbreviations</b>	<b>xix</b>
<b>Symbols</b>	<b>xxii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Bimodal Systems and Signals . . . . .	1
1.2 Motivation and Overview . . . . .	2
1.3 Aim of the Research . . . . .	6
1.4 Original Contributions of Thesis . . . . .	8
1.5 Publications Resulting from Research . . . . .	9
1.6 Outline of the Thesis . . . . .	10
<b>2 Audio-Visual Speech Processing</b>	<b>12</b>
2.1 Introduction . . . . .	12
2.2 Audio-Visual Speech Processing by Humans . . . . .	12
2.2.1 The Speech Chain . . . . .	13
2.2.2 The Process of Speech Production . . . . .	13
2.2.3 Phonemes and Visemes . . . . .	16
2.2.4 Audio-Visual Speech Perception . . . . .	16
2.3 Audio-Visual Speech Recognition . . . . .	17
2.4 Audio-Visual Databases Review . . . . .	20
2.5 CUAVE Database . . . . .	26
2.6 Chapter Summary . . . . .	30
<b>3 Feature Extraction</b>	<b>33</b>
3.1 Introduction . . . . .	33
3.2 Audio Front-End . . . . .	33
3.2.1 Audio Pre-Processing . . . . .	34
3.2.2 Audio Feature Extraction . . . . .	36
3.2.2.1 Mel-Frequency Cepstral Coefficients (MFCC) . . . . .	37
3.2.2.2 Perceptual Linear Prediction (PLP) . . . . .	41
3.2.2.3 RelAtive SpectTrA-PLP (RASTA-PLP) . . . . .	43

3.2.2.4	Energy and Time Derivative Features . . . . .	45
3.3	Visual Front-End . . . . .	46
3.3.1	Visual Pre-Processing . . . . .	46
3.3.2	Visual Feature Extraction . . . . .	47
3.3.2.1	Appearance-based Approaches . . . . .	47
3.3.2.2	Contour-based Approaches . . . . .	49
3.3.2.3	Hybrid Approaches . . . . .	50
3.3.2.4	Choosing a Visual Feature Extraction Method . . . . .	50
3.3.2.5	DCT2 Visual Features . . . . .	52
3.4	Gabor Filters . . . . .	52
3.5	The Proposed Gabor Audio and Visual Front-Ends . . . . .	53
3.5.1	Gabor Audio Features (GAF) . . . . .	53
3.5.2	Gabor Visual Features (GVF) . . . . .	56
3.6	Chapter Summary . . . . .	58
<b>4</b>	<b>Integration Schemes for Audio-Visual Speech Recognition</b>	<b>59</b>
4.1	Introduction . . . . .	59
4.2	HMMs, Training and Integration Models . . . . .	60
4.2.1	Hidden Markov Models . . . . .	60
4.2.2	The Hidden Markov Model Toolkit (HTK) . . . . .	62
4.2.3	Single-Stream HMM <sub>av</sub> -EI AVSR Model Setup . . . . .	63
4.2.4	Multi-Stream HMM <sub>av</sub> -LI AVSR Model Setup . . . . .	64
4.3	Stream Weight Estimation . . . . .	66
4.4	The Proposed Signal-based Measures . . . . .	67
4.4.1	Time-domain Features . . . . .	68
4.4.1.1	Short-Time Energy (STE) . . . . .	68
4.4.1.2	Zero-Crossing Rate (ZCR) . . . . .	68
4.4.1.3	Energy Entropy (EE) . . . . .	69
4.4.2	Frequency-domain Features . . . . .	69
4.4.2.1	Spectral Centroid (SC) . . . . .	69
4.4.2.2	Spectral Flux (SF) . . . . .	70
4.4.2.3	Spectral Roll-off (SR) . . . . .	70
4.5	The Proposed Stream Weighting Scheme . . . . .	71
4.6	DNNs, Training, and Integration Models . . . . .	72
4.6.1	Deep Neural Networks . . . . .	72
4.6.2	Gabor Filters Utilization in DNNs . . . . .	74
4.6.3	Recurrent Neural Networks . . . . .	75
4.6.4	Connectionist Temporal Classification . . . . .	76
4.6.5	Deep BRNN <sub>av</sub> -EI AVSR Model Setup . . . . .	78
4.6.6	Deep BRNN <sub>av</sub> -LI AVSR Model Setup . . . . .	79
4.7	Chapter Summary . . . . .	82
<b>5</b>	<b>Experimental Work and Results</b>	<b>83</b>
5.1	Data Corpus . . . . .	83
5.2	HMM Experimental Results . . . . .	83
5.2.1	Audio-only and Visual-only Speech Recognition . . . . .	83
5.2.2	Audio-Visual Speech Recognition with Early Integration . . . . .	85

---

5.2.3	Audio-Visual Speech Recognition with Late Integration . . . . .	87
5.2.4	The Evaluation of the Proposed Front-End . . . . .	91
5.2.5	The Evaluation of the Proposed Dynamic Weighting Scheme . . . . .	91
5.3	BRNN Experimental Results . . . . .	95
5.3.1	Audio-only and Visual-only Speech Recognition . . . . .	95
5.3.2	Audio-Visual Speech Recognition . . . . .	97
5.3.3	Performance Evaluation . . . . .	100
5.4	Chapter Summary . . . . .	100
<b>6</b>	<b>Conclusions and Future Work</b>	<b>102</b>
6.1	Conclusions . . . . .	102
6.2	Future Work . . . . .	103
	<b>Bibliography</b>	<b>104</b>

# List of Figures

2.1	The sagittal section of the vocal apparatus. . . . .	15
2.2	The general structure of an AVSR system. . . . .	19
2.3	Sample images from various audio-visual databases. . . . .	26
2.4	Sample frame from the normal-individual part of the CUAVE database. . . . .	30
3.1	Example of the time representation before and after pre-emphasis of the digit '9' uttered by a CUAVE speaker in clean environment. . . . .	35
3.2	Frame blocking pre-processing step of the speech signal. . . . .	35
3.3	Audio signal windowing. . . . .	36
3.4	Weighting functions for Mel-scale filtering. . . . .	38
3.5	Relation between Hz and Mel scales. . . . .	38
3.6	MFCC Calculation. . . . .	40
3.7	PLP Calculation. . . . .	42
3.8	RASTA-PLP Calculation. . . . .	44
3.9	Configuration of an acoustic feature vector including the static spectral coefficients and energy with the deltas and acceleration features. . . . .	46
3.10	Visual pre-processing step output for subject 's04f' from CUAVE database (a) face detection, (b) face cropped, (c) mouth detection, (d) mouth cropped, (e) mouth region resize by 128×128, and (f) mouth region as gray-scale. . . . .	47
3.11	Different 2D GBFB Filters Shapes. . . . .	55
3.12	Gabor Audio Features (GAF) Calculation. . . . .	55
3.13	Gabor wavelets at five scales and eight orientations (a) The real parts of the Gabor filters, (b) The magnitudes of the Gabor filters. . . . .	57
4.1	HMM diagram. . . . .	60
4.2	The proposed HMM <sub>av</sub> -EI AVSR model structure. . . . .	63
4.3	Various Multi-Stream HMM architectures. Note, the audio densities are shown in blue and visual in red. . . . .	66
4.4	The block diagram for the proposed Multi-Stream HMM <sub>av</sub> -LI AVSR model. . . . .	73
4.5	The standard RNN and the unfolding in time. . . . .	75
4.6	The BRNN structure. . . . .	76
4.7	The proposed BRNN <sub>av</sub> -EI AVSR model structure. . . . .	79
4.8	The block diagram for the proposed deep BRNN <sub>av</sub> -LI AVSR model. . . . .	81
5.1	The best performance of HMM <sub>a</sub> ASR using MFCC, PLP, RASTA-PLP and GAF as audio features at different SNR levels. . . . .	85
5.2	Performance of the HMM <sub>av</sub> -EI AVSR using GAF as audio features with GVF, and DCT2 as visual features at different SNR levels. . . . .	86