



Bahria University
Discovering Knowledge

FINAL YEAR PROJECT REPORT
REAL TIME SPEECH DRIVEN FACE
ANIMATION SYSTEM

By

IZHAR US-SALAM KAYANI	(36567)
MUHAMMAD AHMER ASIM	(36581)
HASSAN SHAHAB ABBASSI	(36563)
RAKESH KUMAR BHIMANI	(36598)

SUPERVISED BY

MADAM ASIA SAMREEN

BAHRIA UNIVERSITY (KARACHI CAMPUS)

2017

REAL TIME SPEECH DRIVEN FACE ANIMATION SYSTEM

ABSTRACT

ACKNOWLEDGEMENTS

We would like to thank everyone who had contributed to the successful completion of this project. We would like to express our gratitude to our research supervisor, Madam Asia Samreen for her invaluable advice, guidance and her enormous patience throughout the development of the research.

In addition, we would also like to express my gratitude to our loving parent and friends who had helped and given us encouragement.

TABLE OF CONTENTS

REAL TIME SPEECH DRIVEN FACE ANIMATION SYSTEM

DECLARATION	ii
APPROVAL FOR SUBMISSION	iii
ACKNOWLEDGEMENTS	iv
ABSTRACT	vii
TABLE OF CONTENTS	viii
LIST OF FIGURES	x
LIST OF TABLES	xi

ABSTRACT

We have opted a paper “LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING” as a base paper of our Final Year Project.

Lip-reading is the task of decryption text from the movement of a speaker’s mouth. Ancient approaches separated the issue into 2 stages: planning or learning visual options, and prediction. Newer deep lip-reading approaches are end-to-end trainable (Wand et al., 2016; Chung & Zisserman, 2016a). However, existing work on models trained end-to-end perform solely word classification, instead of sentence-level sequence prediction. Studies have shown that human lip-reading performance will increase for extended words (Easton & Basala, 1982), indicating the importance of options capturing temporal context in an ambiguous communication. Intended by this observation, our project presents, a model that maps a video frames to text, creating use of spatial-temporal convolutions, a neural network, and therefore the connection temporal classification loss, trained entirely end-to-end. End-to-end sentence-level lip-reading model that at the same time learns spatial-temporal visual options and a sequence model.

DESIGN AND METHODOLOGY

2.1 Design

2.1.1 Work Breakdown Structure

TABLE OF CONTENTS

	DECLARATION	ii
	APPROVAL FOR SUBMISSION	iii
	ACKNOWLEDGEMENTS	vi
	ABSTRACT	vii
	TABLE OF CONTENTS	viii
	LIST OF TABLES	x
	LIST OF FIGURES	xi
	LIST OF SYMBOLS / ABBREVIATIONS	xii
	LIST OF APPENDICES	xiii
CHAPTER		
1	INTRODUCTION	1
	1.1 Background	1
	1.2 Problem Statements	2
	1.3 Aims and Objectives	3
	1.4 Scope of Project	3
2	LITERATURE REVIEW	4
	2.1 Lip Reading	4
	2.2 Real Time Communication	5
	2.2.1 WEB RTC	5
3	DESIGN AND METHODOLOGY	8
	3.1 Design	8
	3.1.1 Work Breakdown Structure	8

		ix
	3.1.2 Actor Use Cases	9
3.2	Methodology	10
	3.2.1 Lip Reading	10
	3.2.2 WebRTC	12
4	IMPLEMENTATION	17
4.1	WebRTC	17
	4.1.1 Code Implementation of WebRTC	18
4.2	Lip-Reading	19
	4.2.1 Code Implementation of Lip-reading	22
5	RESULTS AND DISCUSSIONS	27
5.1	WebRTC	27
	5.1.1 Results of Video Calling Application	28
5.2	Lip-Reading	29
6	CONCLUSION AND RECOMMENDATIONS	31
6.1	Conclusion	31
6.2	Future Work	31
	REFERENCES	32
	APPENDICES	35