LIP SYNCING METHOD FOR REALISTIC EXPRESSIVE THREE-DIMENSIONAL FACE MODEL

ITIMAD RAHEEM ALI AL-RUBAYE

UNIVERSITI TEKNOLOGI MALAYSIA

LIP SYNCING METHOD FOR REALISTIC EXPRESSIVE THREE-DIMENSIONAL FACE MODEL

ITIMAD RAHEEM ALI AL-RUBAYE

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

MARCH 2016

To the tender fountain who provided me with perseverance, *My Parents*

To the source of kindness who provided me with ambition, *My Sister (Fayza), My Husband (Qays) and My Kids (Farah and Yousif)*

For those who provided me with supervision and constant support, *My Supervisors*

To my virtuous professors who taught me in a truthful, fair, and honorable way

To my colleagues in the Universiti Teknologi Malaysia

To all those who contributed to the success of this research

I dedicate this research to you

ACKNOWLEDGEMENT

Initially, all praise is to Allah, the most kind and the merciful for helping me to accomplish this study. Special appreciation goes to my family, friends, study colleagues and supervisors for standing beside me in the good and bad times spent to complete this research. First, I would like to extend my heartfelt gratitude to my family.

During the time spent to complete this thesis, I dealt with various qualified researchers, academics, experts and practitioners who are specialized in various scientific fields. I would like to show my high appreciation and thank them for their valuable guidance to provide significant remarks that changed the course of the research to the correct direction. In particular, I want to share my truthful gratitude to my supervisor, *Prof. Dr. Ghazali Sulong,* and my co-supervisor *Dr. Hoshang Kolivand,* for their valuable support of productive comments and ideas which have contributed to the success of this research.

In addition, I want to thank my beloved friend *Dr. Mohammed Hazim Ameen Alkawaz* and to my friends in MAGICX for their constant support through good and bad times. Thank you for being the friend that I always wanted, I always needed and I always deserved. Finally, I would like to offer my special thanks to the faculty of computing staff, your help will never be forgotten.

ABSTRACT

Lip synchronization of 3D face model is now being used in a multitude of important fields. It brings a more human and dramatic reality to computer games, films and interactive multimedia, and is growing in use and importance. High level realism can be used in demanding applications such as computer games and cinema. Authoring lip syncing with complex and subtle expressions is still difficult and fraught with problems in terms of realism. Thus, this study proposes a lip syncing method of realistic expressive 3D face model. Animated lips require a 3D face model capable of representing the movement of face muscles during speech and a method to produce the correct lip shape at the correct time. The 3D face model is designed based on MPEG-4 facial animation standard to support lip syncing that is aligned with input audio file. It deforms using Raised Cosine Deformation function that is grafted onto the input facial geometry. This study also proposes a method to animate the 3D face model over time to create animated lip syncing using a canonical set of visemes for all pairwise combinations of a reduced phoneme set called ProPhone. Finally, this study integrates emotions by considering both Ekman model and Plutchik's wheel with emotive eye movements by implementing Emotional Eye Movements Markup Language to produce realistic 3D face model. The experimental results show that the proposed model can generate visually satisfactory animations with Mean Square Error of 0.0020 for neutral, 0.0024 for happy expression, 0.0020 for angry expression, 0.0030 for fear expression, 0.0026 for surprise expression, 0.0010 for disgust expression, and 0.0030 for sad expression.

ABSTRAK

Penyelarasan bibir bagi model muka 3D kini digunakan dalam pelbagai bidang yang penting. Ia memberi sentuhan yang lebih bersifat manusia dan realiti dramatik kepada permainan komputer, filem dan multimedia interaktif, dan berkembang dari segi penggunaan dan kepentingan. Realisme peringkat tinggi boleh digunakan dalam aplikasi yang mencabar seperti permainan komputer dan pawagam. Mencipta penyegerakan bibir dengan pengucapan yang kompleks dan halus masih sukar dan penuh dengan masalah dari segi realisme. Justeru, kajian ini telah mencadangkan kaedah penyegerakan bibir realistik model muka 3D ekspresif. Bibir animasi memerlukan satu model muka 3D yang berkebolehan untuk mewakili pergerakan otot muka semasa pengucapan dan satu kaedah untuk menghasilkan bentuk bibir yang betul pada masa yang sesuai. Model muka 3D tersebut direka berasaskan animasi muka standard MPEG-4 untuk menyokong penyegerakan bibir yang diselaras dengan input fail audio. Ia berubah bentuk menggunakan fungsi Ubah Bentuk Kosinus Yang Dinaikkan yang mana ia dicantum kepada geometri input muka. Kajian ini juga telah mencadangkan satu kaedah untuk menghidupkan model muka 3D dari masa ke masa untuk membuat penyegerakan bibir animasi menggunakan satu set *viseme* berprinsip untuk semua kombinasi pasangan fonem yang dikurangkan yang dipanggil ProPhone. Akhir sekali, kajian ini mengintegrasikan emosi dengan mempertimbangkan kedua-dua model Ekman dan roda Plutchik dengan pergerakan mata beremosi dengan melaksanakan Bahasa Pergerakan Mata Beremosi untuk menghasilkan model muka 3D realistik. Keputusan eksperimen menunjukkan bahawa model yang dicadangkan ini boleh menjana animasi visual yang memuaskan dengan Min Kuasa Dua Ralat adalah 0.0020 untuk neutral, 0.0024 untuk ekspresi gembira, 0.0020 bagi ekspresi marah, 0.0030 untuk ekspresi rasa takut, 0.0026 untuk ekspresi kejutan, 0.0010 untuk ekspresi jijik, dan 0.0030 untuk ekspresi sedih.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
	DECLARATION	ii
	DEDICATION	iii
	ACKNOWLEDGEMENT	iv
	ABSTRACT	v
	ABSTRAK	vi
	TABLE OF CONTENTS	vii
	LIST OF TABLES	xii
	LIST OF FIGURES	xiv
	LIST OF ABBREVIATIONS	xx
	LIST OF SYMBOLS	xxii
	LIST OF APPENDICES	xxiii
1	**INTRODUCTION**	**1**
	1.1 Introduction	1
	1.2 Problem Background	4
	1.3 Problem Statement	9
	1.4 Research Aim	10
	1.5 Research Objectives	10
	1.6 Research Scope	10
	1.7 Significance of the Study	11
	1.8 Thesis Outline	12

2	**LITERATURE REVIEW**			**13**
	2.1	Introduction		13
	2.2	Facial Animation		14
		2.2.1	A Brief History of Facial Animation	15
		2.2.2	Facial Animation in Films and Games	17
		2.2.3	Facial Animation Techniques	18
			A. FACS Approach	19
			B. Interpolation	21
			C. Parameterization	21
			D. Image Morphing	22
			E. Pseudo Muscle Model	22
			F. Physics based Muscle Model	23
			G. MPEG-4 Facial Animation Approach	23
	2.3	Lip Syncing		31
		2.3.1	Lips Motion Synthesis	31
		2.3.2	Effect of Speech Animation	35
		2.3.3	Co-articulation	37
		2.3.4	Tongue Action	38
		2.3.5	Teeth	38
	2.4	Vocal and Structural Properties of English Language		38
		2.4.1	Vocal properties of English Language	39
		2.4.2	Structural properties of English Language	40
		2.4.3	Phonemes and Lip Syncing	40
		2.4.4	Visual Phonemes	41
		2.4.5	Phoneme Classification	43
		2.4.6	Point of Articulation	44
	2.5	Realistic 3D Face Model		46
		2.5.1	Modeling Emotions	46
			A. Ekman Model	47
			B. Plutchick’s Wheel of Emotions	47
		2.5.2	Facial Expression	49
		2.5.3	Eye Behavior	54
			A. Eye Behaviours in Social Encounters	55

			B. Eye Blinking, Manipulators and Social Behaviours	56
			C. Eye Saccades	57
	2.6	Current 3D Face Model Systems		58
		2.6.1	Microsoft Agents	58
		2.6.2	Emotional Chinese Talking Head	59
		2.6.3	GRETA	59
		2.6.4	Xface Engine	60
		2.6.5	LUCIA	61
	2.7	Critical Analysis on Lip Syncing and Realistic Expressive 3D Face Model Systems		62
	2.8	Summary		65
3	**RESEARCH METHODOLOGY**			**67**
	3.1	Introduction		67
	3.2	Research Framework		68
	3.3	Stages of Research Framework		72
		3.3.1	Problem Definition (Stage 1)	72
		3.3.2	Research Methodology (Stage 2)	75
			A. 3D Face Model	75
			B. Formulating Motion Parameters for Lip Region of 3D Face Model	76
			C. Lip Syncing	81
			D. Proposing New Lip Syncing Method	82
			E Realistic 3D Face Model	84
			F. Integrating Facial Expression and Eye Behaviour with 3D face model	86
		3.3.3	System Rendering (Stage 3)	89
			A. Synthesis Face Motion based on MPEG-4 approach	89
			B. Face Interpolation	90
		3.3.4	Research Evaluation (Stage 4)	91
	3.4	Operational Framework		96

	3.5	Summary	98
4	**DETAILED DESIGN AND IMPLEMENTATION OF PROPOSED METHODS**		**99**
	4.1	Introduction	99
	4.2	Formulating Motion Parameters of 3D Face Model	100
		4.2.1 Description of Parameters using Xface Engine (Input Stage)	104
		4.2.2 Calculation of FAPU(Process Stage)	112
		4.2.3 Generation of Polygonal Mesh (Process Stage)	113
		4.2.4 Subdivision of the Proposed Face Model (Process Stage)	117
		4.2.5 Reduction Parameters using PCA method(Process Stage)	118
		4.2.6 3D FaceModel (Output Stage)	124
	4.3	Proposing a New Lip Syncing Method	125
	4.4	Integrating Facial Expression and Eye Behaviours with the Proposed 3D Face Model	135
	4.5	Synthesize Face Motion based on MPEG-4 Approach	148
	4.6	Face Interpolation	150
	4.7	Summary	152
5	**IMPLEMENTATION, RESULT AND DISCUSSION**		**153**
	5.1	Introduction	153
	5.2	Analysis of 3D Face Model	154
	5.3	ProPhone Lip Syncing Method	157
	5.4	Integrating Facial Expression and Eye Behaviours with the Proposed 3D Face Model	161
	5.5	Research Evaluation	166
		5.5.1 Real Benchmarking Evaluation	166
		5.5.2 Objective Evaluation	169
		5.5.3 Analysis of the Displacements of Motion	175

		Parameters	
	5.6	Discussion	180
	5.7	Summary	181
6	**CONCLUSION AND FUTURE OUTLOOK**		**182**
	6.1	Overall Conclusion	182
	6.2	Research Achievement	184
	6.3	Future Outlook	186
REFERENCES			**188**
Appendix A			203

LIST OF TABLES

TABLE NO	TITLE	PAGE
2.1	FAP groups in MPEG-4 (Deng and Neumann, 2008)	26
2.2	The description and value of FAPUs	30
2.3	English phonemes (Kalwick, 2006)	41
2.4	The basic and advance emotions of Plutchik's wheel of emotion	49
2.5	A critical analysis on lip syncing and realistic expressive 3D face model	65
3.1	Operational framework of the proposed system	97
4.1	Set of Frequency English phonemes that generated by SAPI (TTS Microsoft Speech API 5.4) for 200 sentences (TRueSpel, 2001; Brett *et al.*, 2002; James, 2006)	128
4.2	Selected emotion from Raouzaiou *et al.* (2009) and Nishida and Faucher (2010)	141
4.3	The total of FAPs used to describe the six basic emotions	143
4.4	The profile the range of variation of the proposed hybrid synthesis method.	144
5.1	Comparison between existing 3D face models with the proposed system	155
5.2	The result of visual phonemes of the lip syncing method	157

5.3	Expression generated from the facial expression synthesis	162
5.4	Screenshot of the upper face for multiple type of expressive speech real human and proposed system	168
5.5	Evaluation results for Xface	170
5.6	Evaluation results for GRETA	170
5.7	Evaluation results for the proposed 3D face model	171
5.8	Number of implemented FAPs and number of FAPs that are has been given a certain score	175
5.9	Average score for each of the faces with and without non-executed FAPs.	175
5.10	Scores given to all individual FAP movements to all faces. For bidirectional FAPs, scores are given for respect the negative and the positive value. PS, LU XF, and GR are short for Proposed Model, LUCIA, Xface, and GRETA. Red cells are the situations in which the Proposed Model performs worse than at least one of the other faces.	176
5.11	Scores given to all individual FAP movements to all faces. For bidirectional FAPs, scores are given for respect the negative and the positive value. PS, LU XF, and GR are short for Proposed Model, LUCIA, Xface, and GRETA. Red cells are the situations in which the Proposed Model performs worse than at least one of the other faces.	177

LIST OF FIGURES

FIGURE NO	TITLE	PAGE
1.1	Examples on Lip Syncing of 3D Face Models (a) Wei and Deng, 2015 (b) Xu *et al.*, 2013 (c) Maya Lip Syncing 2014.	2
1.2	Synthesized facial expressions based on emotion of talking Avatar (Zhang *et al.*, 2010)	4
1.3	Visual phoneme (Balci *et al.*, 2007)	6
1.4	Previous researches concerning facial expressions, eye movements, and speech systems	8
2.1	Implementation of the FACS onto a 3D character. (a) AU2 (outer brows raised), (b) AU2+4 (outer brows raised and brows lowered), (c) AU R6 (right cheek raise) and (d) AU4+6 (brows lowered and cheeks raised) (Platt and Badler, 1981; Morten, 2012)	16
2.2	Animated picture from Pixar Animation Studios (Morten, 2012) showing two characters of Merida in 'Brave' (2012) and Carl Fredricksen in'Up' (2009)	17
2.3	Skull of Shape	19
2.4	Various AUs major codes on the face indicating movements directions of the AU movement (red arrows) as center of mass (red dots) (Morten, 2012)	20
2.5	MPEG-4 Feature Points	25
2.6	The FPs in MPEG-4 FAP (Shapiro, 2011)	26
2.7	MPEG-4 FAP file format (Queiroz *et al*., 2009)	28

2.8	MPEG-4 FAPUs (Queiroz *et al*., 2009)	29
2.9	Mandarin vowels classification procedure (shih *et al*., 2010)	33
2.10	Visual phonemes of (Kalwick, 2006)	42
2.11	Shorter list of another visual phoneme of (Kalwick, 2006)	43
2.12	The place of articulation (TRueSpel, 2001) (Brett Kessler and Rebecca Treiman, 2002), (James M. Scobbie, 2006)	44
2.13	Plutchik's Wheel of Emotion (Plutchik, 1980)	48
2.14	Six essential emotions from left to right (from the top, left to right) anger, fear, and disgust, surprise, happiness and sadness (Ekman, 1979)	52
2.15	The emotion face appearance wheel of Plutchik (Mac Kim, 2011)	52
2.16	Emotion face appearances (Kalwick, 2006)	54
2.17	3D Face Model of GRETA	60
2.18	3D Face Model of Xface	60
2.19	SMIL-Agent script processing (Balcı *et al*., 2007)	61
2.20	3D face model of LUCIA	62
3.1	Research Framework	70
3.2	Research Methodology	71
3.3	SmartBody 3D face model from USC Institute for Creative Technologies(Feng and Shapiro, 2013)	76
3.4	Sample face reduction using PCA method from (Wei and Deng, 2015) (a) Among the 102 parameters, used the 39 green parameters. (b) Illustration of the average face parameters.	79
3.5	Overview of the basic speech animation	82
3.6	Screenshots of expressive 3D face model of (Paul, 2010). (a) joy (b) sad (c) anger (d) fear (e) disgust (f) surprise.	85
3.7	The direction of eye movements	86
3.8	Setting FAPUs of MPEG-4 Standard with seven	90

	points	
3.9	Talking Head Systems (a) Greta((Pasquariello and Pelachaud, 2002), (b) Xface (Balcı, 2007), and (c) Lucia (Leone *et al.*, 2012)	91
4.1	MPEG-4 FPs for human head	101
4.2	The steps of developing proposed 3D face model	103
4.3	Selecting a specific FP	105
4.4	Selecting a specific FAP	105
4.5	Illustration of the vertices from one region to the application of a Raised Cosine Deformation Functions (RCDF) (a) before and (b) after applying RCDF (Queiroz *et al.*, 2009)	107
4.6	Deformation within a region of influence (a) null intensity, (b) medium intensity, and (c) maximum intensity	108
4.7	Two skin regions and FPs	108
4.8	Fixing the weights and deformation functions for the 3D model	110
4.9	The action of enhanced values of FAP 31, FAP 33, and FAP35	111
4.10	The influence of increased FAP31, FAP33, and FAP35 to entire face	111
4.11	FAPU of MPEG-4 approach	112
4.12	Texture map, mesh polygonal, and the internal anatomic components of the proposed 3D face model, where (a) is texture skin256 x 256, (b) mesh polygonal, (c) Tongue, (d) Sock, (e) Teeth and (f) Eyes	113
4.13	Proposed 3D model (a) front view, (b) side view	114
4.14	Polygon mesh of mouth of 3D face model	115
4.15	Polygon mesh of eye of 3D model (wireframe shade visualization)	115
4.16	Forehead of 3D model having polygon meshes	116
4.17	Polygonal mesh structure in the nasolabial furrow of	116

	3D face model	
4.18	Subdivision of proposed 3D model in “two specific areas”	117
4.19	50 FPs of the proposed 3D face model	119
4.20	FP’s Reduction (a) The facial mesh (b) facial mesh of the proposed system (c) application of PCA on the proposed system	121
4.21	The proposed 3D Face Model	124
4.22	Schematic diagram of ProPhone lip syncing method	126
4.23	Classification of English phonemes (TRueSpel, 2001; Brett *et al.*, 2002; James, 2006)	127
4.24	The speech signal of the sentence “when you happy in your inside, the natural beauty will shine in your inside”	129
4.25	Screen capture showing the speech signal and the phoneme-word audio segmentation of the neutrally spoken sub-sentence “the natural beauty”	129
4.26	Phoneme timeline resulted from SAPI engine	131
4.27	Example on construction of the viseme for the word ”the”	133
4.28	Schematic block diagram of facial expressions synthesis	136
4.29	The angular of Paul Ekman (Paul Ekman, 1999)	137
4.30	The facial expressions synthesize method showing the relations between emotion concepts	138
4.31	Archetypal emotions of facial expressions synthesize method	139
4.32	A situation in which a new profile (a_I) is calculated which lies in between two different archetypal emotions	142
4.33	The eye movements framework of EEMML (Li and Mao, 2011)	147
4.34	Eye motions for principal emotion of EEMML (Li and	147

Mao, 2011)

4.35 MPEG-4 animated face (a) FPs of MPEG-4 Standard and (b) FAPUs scale factor, and (c) FPs of the proposed work 148

4.36 The FPs of the inside of the lips, their animation parameters in the direction of the axes x and y and their corresponding FAPUs. Animation parameter animation in red corresponds to the parameters that were included in MPEG-4 (black and red). 149

5.1 Comparison of the proposed 3D face animation model with and without tongue in the phoneme position /th/. (a) Without incorporating the tongue movement. (b) Presence of tongue movement is included in the proposed 3D face model 156

5.2 Sound signal of the audio for the sentence “the natural beauty...” 160

5.3 English visual phoneme articulations with normal expression for the sentence “the natural beauty”. 160

5.4 The result of facial expression method for six basic emotions 161

5.5 Result of the facial expression in the presence of eye behavior 163

5.6 The sound signal of the audio for the sentence “I wanna discuss all this behavior” 164

5.7 20 frames for the English phonemes with angry emotion for the sentence “I wanna discuss all this behavior” 164

5.8 The sound signal of the audio for the sentence “unbelievable” 165

5.9 Significant 8 frames for the English phonemes with surprise emotion for the word “unbelievable” 165

5.10 The sound signal of the audio for the sentence “Ha u r so funny” 165

5.11 Important 9 frames for the English phonemes with 165

	happy emotion and eye blink for the sentence "you are so funny"	
5.12	The real actor compared with the proposed system where (a) the real actor and (b) the proposed system	167
5.13	3D expressive talking system used to evaluate the proposed system. From left to right: (a) LUCIA (Leone *et al.*, 2012), (b) Xface (Balcı, 2007), and (c) GRETA (Pasquariello *et al*., 2001)	169
5.14	The relative change for happy expression to the upper lip	172
5.15	The relative change for anger expression to the upper lip	172
5.16	The relative change for sad expression to the upper lip	173
5.17	The relative change for disgust expression to the upper lip	173
5.18	The relative change for fear expression to the upper lip	174
5.19	The relative change for surprise expression to the upper lip	174
5.20	The new archetypal profiles displayed in the proposed system, LUCIA, Xface, and GRETA. Expressions for both subjects (male and female) from left to right: happy, sad, angry, fear, disgust, and surprise	178
5.21	The gap in displacements of the eyelids and lip corners of the Xface.	179

LIST OF ABBREVIATIONS

3D	-	Three Dimension
APML	-	Affective Presentation Markup Language
AU	-	Action Unit
BEAT	-	Behavior Expression Animation Toolkit
BML	-	Behavioural Markup language
CCA	-	Canonical Correlation Analysis
CFDF	-	Raised Cosine Deformation Functions
CML	-	Cognitive Modelling Language
ECA	-	Embodied Conversational Agent
EEMML	-	Emotional Eye Movement Markup Language
FACS	-	Face Action Coding System
FaceGen	-	Face Generation
FAP	-	Face Animation Parameter
FAPU	-	Face Animation Parameter Unit
FAT	-	Face Animation Tables
FDP	-	Feature Description Parameter
FP	-	Feature Points
HML	-	Virtual Human Markup Language
LDA	-	Linear Discriminant Analysis
MPEG-4	-	Moving Picture Experts Group
MPML	-	Multimodal Presentation Markup Language
OCC	-	Ortony Clore and Collins Model for emotions
PAD	-	Pleasure, Arousal and Dominance
PC	-	Principle Components

PCA	-	Principle Components Analysis
ProPhone	-	Priority Phone
SAPI	-	Speech Application Program Interface
SMIL	-	Synchronized Multimedia Integration Language
TTS	-	Text To Speech
VC	-	Virtual Character
VHML	-	Virtual Human Markup Language
VW	-	Virtual World

LIST OF SYMBOLS

CC	-	Correlation Coefficient
D	-	Dimension
γ	-	Dimension of the ground truth
υ_j	-	The set of vertices
υ	-	Average standard deviation
ρ	-	Empowered constant

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	MPEG-4 FA Facial Animation Parameters	203

CHAPTER 1

INTRODUCTION

1.1 Introduction

Face is the central element for expressing human emotion and personality (Xu *et al.*, 2013). Lip syncing is a process of speech assimilation with the lip motions of a 3D face model. A talking model is a challenging task because it should provide control of all articulatory movements and must be synchronized with the speech signal. Many different types of vital information are detectable through lip syncing. Lately, many applications of computer facial animation in the foundation of 3D face model with diverse facial expressions are used in entertainment and other fields. An attractive application can advance the interaction between users and devices via interactive virtual speech, and thereby attract users by providing a pleasant interface. With the advent of computer-aided technologies, animated virtual characters are widely used in movies, games and embodied conversational agents (ECAs) to provide an effective and realistic human computer interaction.

Since the early 1990's researchers have focused on developing Embodied Conversational Agents (ECA) or Virtual Character (VC) for interactions with humans in social situations. The introduction of these models created a ripple effect in several fields such as filming and animation. Creation of interactive graphical user interfaces has been one of the long term goals in this area of research and has become one of the most significant branches of Human Computer Interaction research (HCI). Figure 1.1 shows some examples on lip syncing of 3D face model.

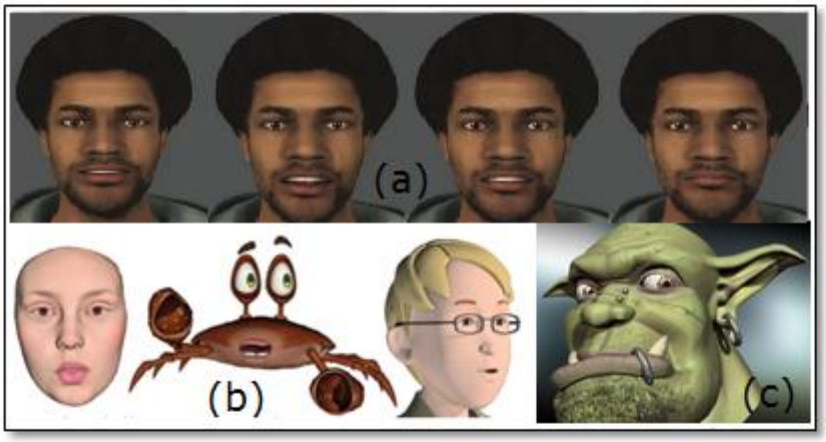

Figure 1.1 Examples on Lip Syncing of 3D Face Models (a) Wei and Deng, 2015, (b) Xu *et al.*, 2013, (c) Maya Lip Syncing 2014.

Recently lip syncing has gained wide acceptance in the field of Human-Computer Interaction. These models show a realistic human face and talk in social situations with emotions such as in e-learning system, healthcare system, e-retail environments and games. Lip syncing usually has an interface which is backed up by a suitable dialogue manager and knowledge base. The 3D face model are capable of using voice; animated dialogue with lips and face appearances; eye, head, and body movements to comprehend signals; expressions of emotions; and execute actions or exhibit hearing or imagining postures (Xu *et al.*, 2013). The presence of these 3D face models has had a positive impact on user experience as reported in previous works (Cassell, 2000; Oyarzun *et al.*, 2010; Yang *et al.*, 2011; Stevens *et al.*, 2013; Ding *et al.*, 2013)

Revelation of human emotion via facial animation is definitely an intricate subject matter in computer graphics, artificial intelligence, physiology, and communication (Cai *et al.*, 2010) (Lee *et al.*, 2011). Creation of a pragmatic face expression is a tricky task because of all-inclusive facial structure. In addition, humans are exceptionally recognizable with facial motions and can easily recognize microscopic details that are deviant or contradictory in an animated face. Truly, the amalgamation between eye movements, lip syncing, motions, appearances of emotions on face, and body direction provide clues regarding flow of thoughts, sequences of thoughts in decision making and depth of understanding and knowing. The alignment of audio with the utterance is so important to get realistic lip syncing. The gaze and saccadic eye movements that speak a lot about the thinking process in human mind are often referred as "window to the mind". Eye movements blended with the expressive gaze convey significant nonverbal information and emotional intentions when a person speaks.

The efficacy of an agent banks on a major factor which is credibility. 3D face models have been made more believable by incorporating emotions. It is very important for these characters to have emotions because they will absolutely increase the user understanding and experience (Cassell, 2000; Deng and Neumann, 2008; Mlakar and Rojc, 2011; Zhao *et al.*, 2013). One of the first major 3D face models that hit the market was the Microsoft Agent which was introduced to help the users of Microsoft products. Microsoft Agent used simple 2D cartoon characters called Merlin, Peedy, Genie and Robby. They were simple speakable characters with a Speech API (SAPI) text to speech engine.

This research focuses on alignment of the lips with an input text or audio file. The generation of emotional 3D realistic talking face model offers more engagement of the users. This motivates one to create a system by improving the weight formula to get smooth movement of facial animation parameters (FAPs) for the lip region. Then, a new lip syncing method is proposed that aligns lips movements with the input text or audio file followed by integrating emotion and eye expression to the 3D face model to acquire realism. In this chapter a rationale of the research is developed and an attempt is made to argue why extensive investigation in the cited thesis topic is absolutely indispensable. In addition, the problem statement, objectives, scope of studies, contribution, research significance and a brief thesis organization are underscored.

1.2 Problem Background

Over the last few years, Virtual Worlds (VWs) particularly 3D graphics contents have advanced significantly. Technological advances in the hardware and software sectors have resulted in tremendous advancement in the technology including animation, display, graphics, distribution network, and mobile communications. Such progress has allowed nearly all users to have access to different tools and applications for the virtual worlds. Creating a 3D face model requires the contribution of several skills for the precise integration of lip syncing with eye gaze and facial expression. Figure 1.2 displays several important facial features that exemplify much of the verbal and non-verbal information required for such combination (Zhang *et al.*, 2010).

Figure 1.2 Synthesized facial expressions based on talking 3D face model (Zhang *et al.*, 2010)

Queiroz *et al.* (2009) introduced a technique to develop a working, extendable, and stable facial animation platform that can easily animate the MPEG-4 parameterized faces. This work uses sophisticated parameters to describe the face actions and properties by rendering interactive platform between user and 3D face model. However, the limitation of this research is in the use of restricted parameters in the face. Bailly *et al.* (2010) acknowledged the process of creating interesting 3D face model by examining audio-visual opposite communication between human to human and a human to virtual conversational agent. The central aim was to configure the mutual gaze patterns in the interaction using innovative instruments. Some measures of the effects of thinking states on communication functioning are demonstrated. A limitation of this work is the weak reproduction of the face deformations around the eyes of the Embodied Conversation Agents (ECA) during eye gaze deviation from the direction of head.

According to Balci *et al.* (2007), it is easy to extend research on virtual face by using Xface Open Source Project and SMIL-Agent Scripting Language. Figure 1.3 illustrates the creation of animated 3D face model. MPEG-4 cannot render an animated language but only a set of low- and high-level considerations. Though Xface is a dominant system in face animation it further needs higher levels abstraction, time control and incident organization. Another limitation of Xface is the non-implementation of various Facial Animation Parameters (FAPs), (such as the FAPs 14, 15, 23-30, 35, 36, 39-47). Therefore, this research implemented more FAPs than was implemented in Xface to give a more realistic motion and smooth blending.

Lee *et al.* (2010) developed realistic expressive 3D face model of real human. This was implemented to analyse the efficiency of expressive model. This method did not take into account in the pilot study human emotion recognition during temporal changes and or verbal clues. Gillies *et al.* (2010) introduced a real-time multimodal interaction to the 3D face model animation system in virtual reality situation. Shapiro (2011) achieved a high level of realism and control by describing a system for virtual characters motions, where a set of significant features of simulated character models and games are included. Xu *et al.* (2013) improved the work of Shapiro in practical terms by demonstrating a lip syncing method without addressing the issue of emotional content during speech. Čereković *et al.* (2010) applied pre-processed sets of realized behaviours to virtual character modules using Back-propagation Neural Network (BNN). This work was further improved by Čereković and Pandžić (2011) by innovating multi-platform Real Actor animation system for realizing real-time behaviour of Embodied Conversational Agents (ECA). The work relied on a solution for gestures and speech synchronization using neural networks.

Leuskia *et al.* (2014) demonstrated 3D face model assisted portable private health-care system, which is a model interface for user-level medical diagnoses gadgets with various subjects.

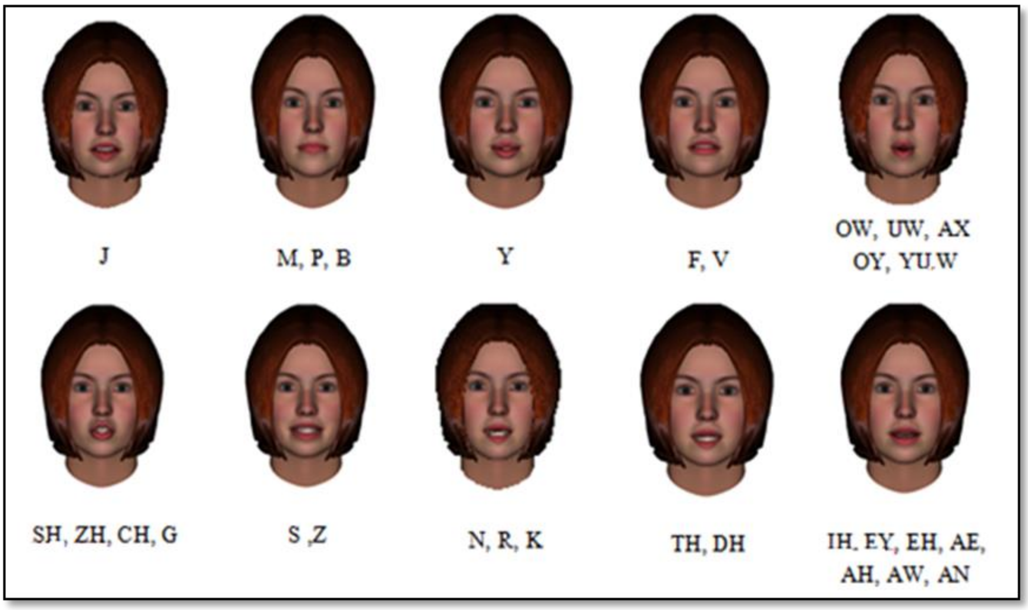

Figure 1.3 Visual phoneme (Balci *et al.*, 2007)

It is established that the lip syncing is the essential component in human interaction. It plays a key role in practical conversation between humans and 3D face model. The coherence in expressive face appearance is significant to enhance the realism of the 3D face model. The voice conveys much emotional information and the dialogue reveals a sequence of phones, where each phone is connected with an ocular depiction of the phoneme viseme. The animated visemes in a given phoneme are related to the pose of lip, jaw and tongue location. These methods are fashionable for creating real-time speech animation system including personality and interactive emotions. Shih *et al.* (2010) designed a mobile device for real-time voice driven lip shape character generation system that synchronizes the lip shape with the corresponding speech with the exception of emotion and eye movements.

Making an accurate synchronization with the event expected for normal behaviour is the main challenge in designing real-time voice driven lip shape character generation system. Li Zheng and Mao (2011) proposed an Emotional Eye Movement Markup Language (EEMML) as a script instrument. This aids in designing eye movements in face to face conversation that describes non-verbal interactions and intention of emotions. Based on lively talking, a real-time pragmatic talking animation is created by decomposing lower face movements and ending with applying motion blending. The issue of this work is a very high missing rate because it used a limited size of the captured dataset (Wei and Deng, 2015).

To achieve a realistic 3D face model, the combination of eye gaze, lip shapes, and expressions should be taken into consideration (Zhang *et al.,* 2010; Shapiro, 2011). The lip motions and voice should be synchronized to give realistic lip synchronization animation. To obtain the abilities used in spoken dialog, higher-level synchronization between two modules is necessary (Wei and Deng, 2015). Major systems use text-to-speech engine (TTS) activated visemes sets (Chuensaichol *et al.*, 2011; Xu *et al.,* 2013), where TTS engine converts a text format sound into a series of phonemes (Tao *et al.*, 2011; Xu *et al.*, 2011). This procedure is used to create a pragmatic speech animation without having to manually set the positions for visemes sets (Lee *et al.*, 1995). Serra *et al*. (2012) developed a visual dialogue simulation module in an attempt to accelerate and assess the quality of phonemes to viseme mappings device for English. Figure 1.4 summarizes chronologically some of the relevant researches focusing on eyes, lips, and facial expressions.

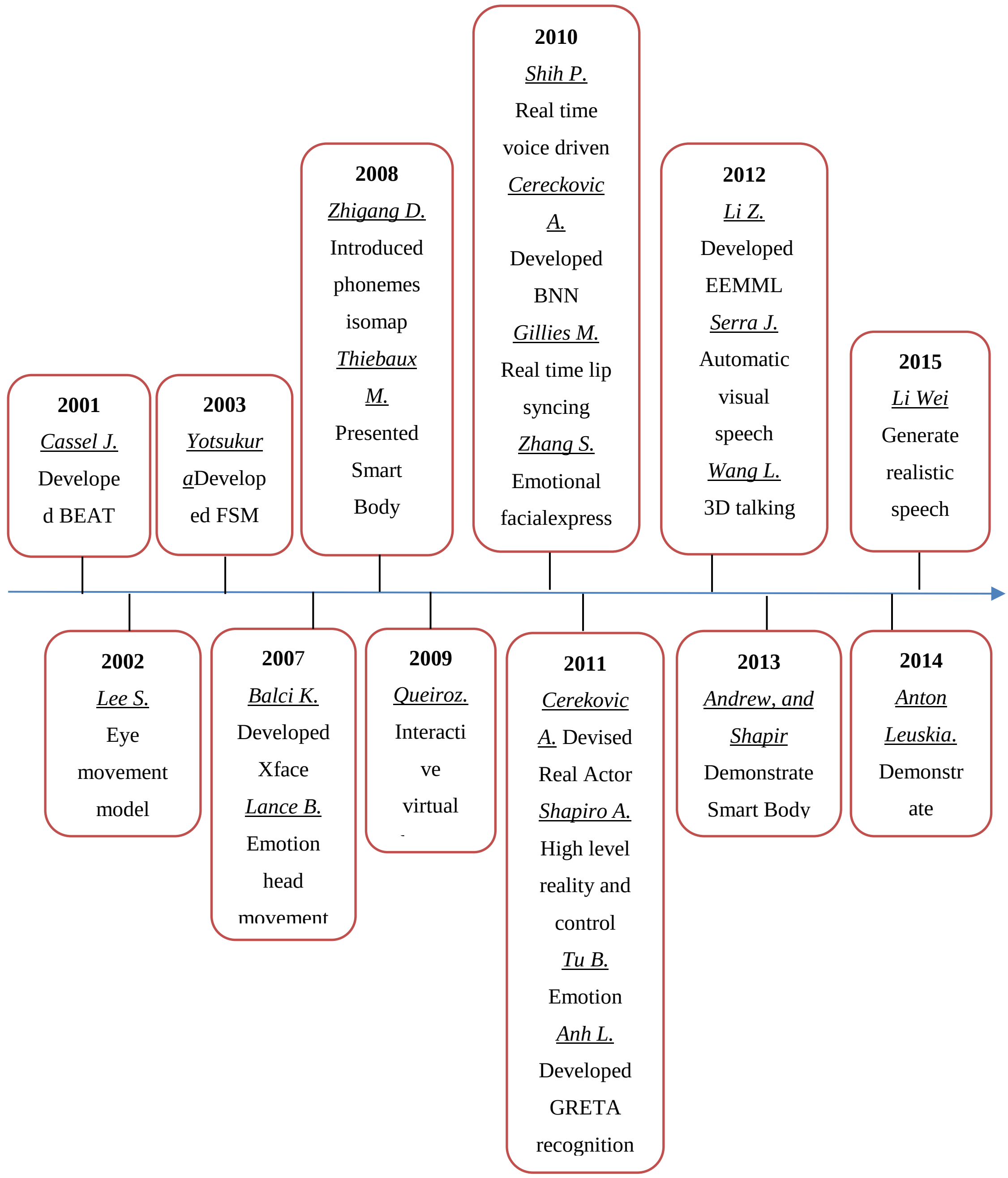

Figure 1.4 Previous researches concerning 3D face models, facial expressions, and eye movements.

In summarizing these works, it is concluded that 3D face models carry remarkable information about an individual. It is evident from the wide range of applications of 3D model. Indeed, this is the motivating factor to choose a 3D face model as the research domain. Actually, facial animation is a wide research area focusing on very important aspects of combination between eye behaviours, emotional facial expression, and lip syncing according to the input text with optimal real-time interaction.

1.3 Problem Statement

Despite continuous research progression and model development the animation of realistic facial appearances remains a critical challenge (Yu *et al.*, 2014). The most sophisticated component in the face is the lip movements (Fratarcangeli, 2013). As mentioned above, this is mainly due to depiction of emotions which is a vital course of action in human intellect (Ichim, 2015). Therefore, it is imperative to produce 3D face model in terms of practical utterances with correct lip synchronized face animation involving close-to-nature lip movements and to integrate face expressions with eye movements to improve the realism of the 3D model (Kang *et al.*, 2015). The emerging issues can be stated as follows based on the problem background and research questions:

1. The motion parameters of 3D face model have been the focal issue of generating an efficient 3D model in computer graphics for a few decades. Much effort has been put into creating models that can heighten the believability of animated models in social science, yet this issue has not been solved (Pasquariello and Pelachaud, 2002; Tao and Tan, 2004; Balci, 2007; Paul, 2010; Shapiro, 2011; Leone *et al.*, 2012) (Cai *et al.*, 2010) (Lee *et al.*, 2011).

2. Inaccurate alignment in lip syncing is an important issue that must be addressed. Lip shapes for many phonemes are modified based on phoneme. Although previous efforts focused on lip syncing, this issue has not been solved yet (Queiroz *et al.*, 2009; Lee *et al.*, 2011; Serra *et al.*, 2012; Taylor *et al.*, 2012; Leuski and Richmond, 2014; Wei and Deng, 2015).

3. Realistic 3D model issue is insufficient in facial animation area. Inspection of the eye movements with expressions to the 3D face model will increase the realism (Bailly *et al.*, 2010; Fagel and Bailly, 2011). Previous efforts (Lee *et al.*, 2010; Moussa *et al.*, 2010; Tinwell *et al.*, 2011; Mlakar and Rojc, 2011; Pelachaud, 2011; Ibbotson and Krekelberg, 2011; D'Mello *et al.*, 2012; Zhao *et al.*, 2012; Sun *et al.*, 2014) offered a series of eye behaviours, which are combined with diverse emotions, but also have not solved the realism issue.

1.4 Research Aim

The aim of this research is to improve the motions of the parameters in the lip regions and to propose a new lip syncing method that aligns the text or audio file with the lip movements of the 3D face model and to improve the realism of the proposed 3D face model by integrating facial expression and eye behaviours.

1.5 Research Objectives

To achieve the aim, the following objectives need to be followed:

1. To formulate the motion parameters of the lip region to get believable 3D face model.
2. To propose a new lip syncing method for aligning the lip movements with the input text or audio file.
3. To integrate facial expressions and eye behaviours with 3D face model to improve the realism of the 3D face model.

1.6 Research Scope

To accomplish the proposed research objectives, the following research scopes are set:

1. MPEG-4 facial animation approach is used to animate the feature points in the 3D face model.
2. The proposed lip syncing method used 15 phonemes of English Language.
3. EEMML is used to implement the eye movements and combine with the proposed 3D model by SMIL language.

1.7 Significance of the Study

The virtual reality 3D face model provides accurate interaction between users and the applications of the 3D face model. Users need to stimulate more in the communication domain so that the synchronization between users and 3D face model allows more realistic, interesting, understandable, and interactive relationship. This thesis suggests a model which can be used to build facial animation engine and create an integrated model. This will create MPEG-4 animation automatically using facial action scripts with lip synchronization and eye movements even in the absence of key-framing meshes. In this model, interactive character actions can be defined via exterior controls. Some simple applications will be developed using combined computer vision algorithms to obtain communication between users and 3D model. This model following the MPEG-4 face animation standard will be able to illustrate a succession of various kinds of high-level facial action including talking, emotional expressions, and eye behaviour. These face actions being independent of each other will be processed using the proposed animation engine via different modules including lip harmonization and emotional face appearance. The main contribution would be the construction of a comprehensive animation system using open-ware resources. This will further integrate some popular simulation models that can create varieties of realistic face behaviour interactively or through a script based way with superior quality.

1.8 Thesis Outline

This thesis consists of six chapters. Chapter 2 provides a thumb nail sketch of the recent relevant literatures. Some of the exhaustively reviewed topics are facial animation, facial animation techniques, lip syncing, vocal and structural properties of English language, realistic 3D face model consist of emotion modelling, facial expression, and eye behaviour, current 3D face model systems and vital analysis on realistic expressive 3D face model systems.

Chapter 3 describes the detailed research methodology. It highlights the proposed model including research framework, stages of research methodology, literature review and problem definition, system design and implementation, system motion and rendering, evaluation and operational framework. Chapter 4 describes the proposed methodology in detail with respect to formulating motion parameters of 3D face model, proposing new lip syncing method, improving the realism by integrating facial expressive synthesise and eyes movements method then synthesizing facial motion with MPEG-4 approach, interpolation and smoothing, employing interpolation and smoothing rendering and SMIL scripting language.

Chapter 5 describes the experimental results based on the newly formulated model. The implementation and benchmarking of the model are also highlighted. This chapter renders experimental setting, details of the conducted performance evaluations, and the implementation results of analysis of the facial model, the result of lip syncing method, the result of integration, animation results, evaluation of the experimental results, real benchmark, objective evaluation, and analysis of the displacements of motion parameters.

Chapter 6 concludes the thesis with future outlook. The successful fulfilment of all the proposed research objectives and the remaining unresolved issues are systematically discussed. The major contributions are also emphasized.

REFERENCES

Albrecht, I., Haber, J., and Seidel, H. P. (2002). Speech Synchronization for Physics-based Facial Animation. *Proceedings of WSCG,* 9–16.

Argyle, M., Cook, M. (1976). Gaze and Mutual Gaze. *Cambridge University Press. London.*

Arsov, I., Jovanova, B., Preda, M., and Preteux, F. (2010). On-Line Animation System for Learning and Practice Cued Speech. *ICT Innovations*, 315-325. Springer Berlin Heidelberg.

Asadpour, V., Homayounpour, M., and Towhidkhah, F. (2011). Audio–visual Speaker Identification using Dynamic Facial Movements and Utterance Phonetic Content. *Applied Soft Computing*, 11(2), 2083-2093.

Avaro, O., Eleftheriadis, A., Herpel, C., Rajan, G., and Ward, L. (2000). MPEG-4 Systems: Overview. *Signal Processing Image Communication*, 15(4), 281–298.

Bailly, G., Raidt, S. and Elisei, F. (2010). Gaze, Conversational Agents and Face-To-Face Communication. *Speech Communication*, 52(6), 598–612.

Balci, K. (2004). Xface: Mpeg-4 based Open Source Toolkit for 3D Facial Animation. *Proceedings of the working conference on Advanced visual interfaces*, 399-402. ACM

Balcı, K., Zancanaro, M. and Pianesi, F. (2007). Xface Open Source Project and SMIL-Agent Scripting Language for Creating and Animating Embodied Conversational Agents. *Proceedings of the 15th International Conference on Multimedia*, 1013–1016. ACM.

Bao, C., Ong, E. P., Niswar, A., and HUANG, Z. (2011). A Facial Animation System for Generating Complex Expressions. 1-7. *APSIPA ASC 2011*.

Berger, M. A., Hofer, G., and Shimodaira, H. (2011). Carnival—Combining Speech Technology and Computer Animation. *Computer Graphics and Applications*, 31(5), 80-89. IEEE

Blanz, V., and Vetter, T., (1999). A Morphable Model for the Synthesis of 3D Faces. *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*. 187–194. ACM.

Brand, M. (1999). Voice Puppetry. *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, 21-28. ACM.

Bregler, C., Covell, M., and Slaney, M. (1997). Video rewrite: Driving visual speech with audio. *Proceedings of the 24th annual conference on Computer graphics and interactive techniques.* 353-360. ACM Press/Addison-Wesley Publishing Co..

Brett, K and Rebecca, T. (2002). Syllable Structure and the Distribution of Phonemes in English Syllables. *Journal of Memory and Language*. 37(3), 295-311.

Cai, Q., Gallup, D., Zhang, C., and Zhang, Z. (2010). 3D Deformable Face Tracking with a Commodity Depth Camera, *Computer Vision–ECCV*, 229–242. Springer Berlin Heidelberg

Cao, Y., Tien, W. C., Faloutsos, P., and Pighin, F. (2005). Expressive Speech-Driven Facial Animation. *ACM Transactions on Graphics*, 24(4), 1283–1302.

Cassell, J. (2000). Embodied Conversational Agents. *MIT press*.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T. andStone, M. (1994). Animated Conversation: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, 413- 420. ACM.

Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. (2004). Beat: The Behavior Expression Animation Toolkit. *Life-Like Characters*,163-185. Springer Berlin Heidelberg

Cassell, J., Bickmore, T., Campbell, L., and Vilhjdlmon, H. (2000). Human conversation as a system framework: designing embodied conversational agents, Embodied conversational agents, *MIT Press, Cambridge*, MA, USA. 29-63. ACM

Čereković, A., Pejša, T., and Pandžić, I. S. (2010). A Controller-Based Animation System for Synchronizing and Realizing Human-Like Conversational

Behaviors. *Development of Multimodal Interfaces: Active Listening and Synchrony*, 80-91. Springer Berlin Heidelberg.

Čereković, A., and Pandžić, I. S. (2011). Multimodal Behavior Realization for Embodied Conversational Agents. *Multimedia tools and applications*, 54(1), 143-164.

Chelali, F. Z., and Djeradi, A. (2011). Audiovisual Speech/Speaker Recognition, Application to Arabic Language. *Procedding of Multimedia Computing and Systems*, 1-7. IEEE.

Chiu, C., and Marsella, S. (2011). How to Train Your Avatar : A Data Driven Approach to Gesture Generation, 127–140.

Chou, Y.-F., and Shih, Z.-C., (2010). A Nonparametric Regression Model for Virtual Human's Generation. *Multimedia Tools and Applications*, 47(1).163–187.

Chuensaichol, T., Kanongchaiyos, P., and Wutiwiwatchai, C. (2011). Lip Synchronization from Thai speech. *Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry*. 355-358. ACM.

Cohn, J. F. (2010). Advances in Behavioral Science Using Automated Facial Image Analysis and Synthesis. *Signal Processing Magazine*, 27(6). 128–133. IEEE

Colburn, A., Cohen, M. F., and Drucker, S. (2000). The Role of Eye Gaze in avatar Mediated Conversational Interfaces. Sketches and Applications, SIGGRAPH.

Collins, B., and Mees, I. M. (2003). The Phonetics of English and Dutch. *Brill Academic Pub*.

Cosker, D., Marshall, D., Rosin, P. L., and Hicks, Y. (2004). Speech Driven Facial Animation Using a Hidden Markov Coarticulation Model. *In Pattern Recognition, ICPR 2004. Proceedings of the 17th International Conference*, 1, 128-131. IEEE.

D'Mello, S., Olney, A., Williams, C., and Hays, P. (2012). Gaze Tutor: A Gaze-Reactive Intelligent Tutoring System. *International Journal of Human-Computer Studies*, *70*(5), 377–398.

Darwin, C. (1965). The Expression of the Emotions in Man and Animals (Second edi.). *Aylesbury, UK: Hazell, Watson and Viney*.

Darwin, A. (2005). The phage-shock-protein response. *Molecular microbiology*, 57(3), 621-628.

DeCarlo, D., Metaxas, D., and Stone, M. (1998). An Anthropometric Face Model using Variational Techniques. *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 67-74. ACM.

Deena, S. P. (2012). Visual Speech Synthesis by Learning Joint Probabilistic Models of Audio and Video. *The University of Manchester*.

Deng, Z., and Neumann, U. (2006). eFASE : Expressive Facial Animation Synthesis and Editing with Phoneme-Isomap Controls, *Proceedings of the 2006 ACM SIGGRAPH/ Eurographics Symposium on Computer Animation, 251-260. Eurographics Association*

Deng, Z., and Neumann, U. (2008). Data-driven 3D Facial Animation. *1st edition Springer*.

Ding, Y., Pelachaud, C., and Arti, T. (2013). Modeling Multimodal Behaviors from Speech Prosody. *In Intelligent Virtual Agents*, 217–228. *Springer Berlin Heidelberg*

Dulguerov, P., Marchal, F., Wang, D., and Gysin, C. (1999). Review of Objective Topographic Facial Nerve Evaluation Methods. *The American Journal of Otology and Neurotology*, 20(5), 672-678.

Dupont, S., Aubin, J. and Ménard, L. (2005). A study of the McGurk Effect in 4-and 5-year-old French Canadian Children. *ZAS Papers in Linguistics*, 40. 1-17.

Egges, A., Kshirsagar, S., and Magnenat-Thalmann, N. (2003). A model for personality and emotion simulation. In *Knowledge-based intelligent information and engineering systems*. 453-461. Springer Berlin Heidelberg.

Ekman, P., and Friesen, W. V. (1978). Manual for Facial Action Coding System. *Consulting Psychologists Press*.

Ekman, P., and Friesen, W. V. (2003). Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues. *Ishk Englewood Cliffs New Jersey: Prentice-Hall*.

Ekman P. (1999). Basic Emotions. *University of California, San Francisco, Handbook of Cognition and Emotion, Chapter 3.*

Ekman P. (2004). Emotional and Conversational Nonverbal Signals. *Language, Knowledge, and Representation*, 39-50. Springer Netherlands

Ekman P. (1979). Emotional and Conversational Signals. *Human ethology. Cambridge University Press.* (pp. 169–202). Cambridge University Press.

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. *Pattern Recognition*, 44(3). 572–587.

Essa, I. and Pentland, A. P. (1995). Facial Expression Recognition using a Dynamic Model and Motion Energy. *Proceedings of Computer Vision, Fifth International Conference*, 360-367. IEEE.

Ezzat, T., Geiger, G. and Poggio, T. (2002). Trainable Videorealistic Speech Animation. *SIGGRPAH*, 21(3). 388–398. ACM

Face FX. (2015). Retrieved from https://www.facefx.com/

FaceGen. (2014). Retrieved from http://facegen.com/

Fagel, S., and Bailly, G. (2011). Speech , Gaze and Head Motion in a Face-to-Face Collaborative Task. *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces, Theoretical and Practical Issues*, 256–264. Springer Berlin Heidelberg

Fasel, B. and Luettin, J. (2003). Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, 36(1). 259-275.

Feng A. and Shapiro A. (2013). SmartBody. *Institute for Creative Technologies University of Southern California*

Feng, G. C., and Yuen, P. C. (2000). Recognition of Head-and-Shoulder Face Image using Virtual Frontal-View Image. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions*, 30(6), 871-882.

Fratarcangeli, M. (2013). Computational Models for Animating 3D Virtual Faces. *Linköping Studies in Science and Technology Thesis, No. 1610, Division of Image Coding, Department of Electrical Engineering, Linköping University*, 58-83.

Fu, Y., Tang, H., Tu, J., Tao, H., and Huang, T. S. (2010). Human-Centered Face Computing in Multimedia Interaction and Communication. *Proceeding of Intelligent Multimedia Communication: Techniques and Applications Springer Berlin Heidelberg*, 280. 465-505.

Garau, M., Slater, M., Bee, S., and Sasse, M. A. (2001). The impact of eye gaze on communication using humanoid avatars. In Proceedings of the SIGCHI conference on Human factors in computing systems. 309-316. ACM.

Gibert, G., Leung, Y., and Stevens, C. J. (2013). Control of Speech-Related Facial Movements of an Avatar from Video. *Speech Communication*, 55(1). 135-146.

Gillies, M., Pan, X., and Slater, M. (2010). Piavca: A Framework for Heterogeneous Interactions with Virtual Characters. *Virtual Reality*, 14(4), 221–228.

Gonseth, C., Vilain, A., and Vilain, C. (2013). An Experimental Study of Speech/Gesture Interactions and Distance Encoding. *Speech Communication*, 55(4). 553-571.

Guenter, B., Grimm, C., Wood, D., Malvar, H., and Pighin, F., (1998). Making Faces. *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*. 55-66. *ACM*

Gunes, H., and Pantic, M., (2010). Automatic, Dimensional and Continuous Emotion Recognition. *International Journal of Synthetic Emotions*, 1(1). 68–99.

Hjortsjo, C. H. (1969). Man's Face and Mimic Language. Lund, Sweden.

Hong, P., Wen, Z., and Huang, T. S. (2002). Real-Time Speech-Driven Face Animation with Expressions using Neural Networks. *IEEE Transactions on Neural Networks / a Publication of the IEEE Neural Networks Council*, *13*(4), 916–27

Ibbotson, M., and Krekelberg, B. (2011). Visual Perception and Saccadic Eye Movements. *Current Opinion in Neurobiology*, *21*(4), 553–8.

Ichim, A. E. (2015). Dynamic 3D Avatar Creation from Hand-held Video Input. *Transactions on Graphics*. 34(4). 45. ACM

ISO/IEC 14496. (2014). MPEG-4. Retrieved from http://mpeg.chiariglione.org/standards/mpeg-4

Jaimes, A., and Sebe, N. (2007). Multimodal human–Computer Interaction: A Survey. *Computer Vision and Image Understanding*, 108(1), 116–134.

James, M. Scobbie, O. B. G. and B. M. (2006). QMUC Speech Science Research Centre Working. *Queen Margaret University College*, 7. 3–30.

Jung, Y., Kuijper, A., Fellner, D., Kipp, M., Miksatko, J., Gratch, J., and Thalmann, D. (2011). Believable Virtual Characters in Human-Computer Dialogs. *Eurographics*. 75-100.

Kalberer, G. A., Mueller, P., and Van Gool, L. J. (2003). A visual speech generator. *Electronic Imaging 2003*, 46-53. International Society for Optics and Photonics.

Kalra, P., Mangili, A., Magnenat-Thalmann, N., and Thalmann, D. (1991). Smile: A Multilayered Facial Animation System. *In Modeling in Computer Graphics*, 189-198. Springer Japan.

Kalwick, D. (2006). Animating Facial Features and Expressions (Graphics Series). *Charles River Media*. 51-67.

Kang, S. H., Feng, A. W., Leuski, A., Casas, D., and Shapiro, A. (2015). Smart Mobile Virtual Humans: “Chat with Me!”. *Intelligent Virtual Agents*, 475-478. Springer International Publishing

Kessous, L., Castellano, G., and Caridakis, G. (2010). Multimodal Emotion Recognition in Speech-Based Interaction using Facial Expression, Body Gesture and Acoustic Analysis. *Journal on Multimodal User Interfaces*, 3(1-2), 33-48.

Kim, Y., and Neff, M. (2012). Component-based Locomotion Composition. In Proceedings of the *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 165-173. Eurographics Association.

King, S. A. (2001). *A Facial Model and Animation Techniques for Animated Speech*. Doctoral dissertation. The Ohio State University.

Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H. and Vilhjálmsson, H. (2006). Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. *Intelligent virtual agents*. 205-217. Springer Berlin Heidelberg.

Kowler, E. (2011). Eye Movements: The Past 25 Years. *Vision Research*, 51(13). 1457-1483.

Lance, B., and Marsella, S. C. (2007). Emotionally Expressive Head and Body Movement during Gaze Shifts. *Intelligent Virtual Agents*. 72-85. Springer Berlin Heidelberg

Lee, C., Lee, S., and Chin, S. (2011). Multi-layer Structural Wound Synthesis on 3D Face. *Computer Animation and Virtual Worlds*. 22(23), 177-185.

Lee, S., Carlson, G., Jones, S., Johnson, A., Leigh, J., and Renambot, L. (2010). Designing an Expressive Avatar of a Real Person. *Intelligent Virtual Agents*, 64–76. Springer Berlin Heidelberg

Lee, W. S. and Magnenat-Thalmann, N. (2000). Fast Head Modeling for Animation. *Image and Vision Computing*, 18(4). 355-364.

Lee, Y., Terzopoulos, D., and Walters, K. (1995). Realistic Modeling for Facial Animation. *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques* SIGGRAPH 95, 95(1). 55-62. ACM

Leone, G. R., Paci, G., and Cosi, P., (2012). LUCIA: An Open Source 3D Expressive Avatar for Multimodal hmi. *Intelligent Technologies for Interactive Entertainment*, 193-202. Springer Berlin Heidelberg

Leuski, A., and Richmond, T. (2014). Mobile Personal Healthcare Mediated by Virtual Humans. *Proceedings of the Companion Publication of the 19th International Conference on Intelligent User Interfaces*, 21-24. ACM

Lewis, J. P., Anjyo, K., Rhee, T., Zhang, M., Pighin, F., and Deng, Z. (2014). Practice and Theory of Blendshape Facial Models.

Li, B., Zhang, Q., Zhou, D., and Wei, X. (2013). Facial Animation Based on Feature Points. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 11(3), 1697-1706

Li, Z., and Mao, X. (2011). EEMML: The Emotional Eye Movement Animation Toolkit. *Multimedia Tools and Applications*, *60*(1), 181–201.

Lien, J. J., Kanade, T., Cohn, J. F., and Li, C. C., (1998). Automated Facial Expression Recognition based on FACS Action Units. *Automatic Face and Gesture Recognition, 1998. Proceedings Third IEEE International Conference*, 390-395. IEEE.

Liu, C. (2009). An Analysis of the Current and Future State of 3D Facial Animation Techniques and Systems, *Doctoral dissertation, School of Interactive Arts and Technology-Simon Fraser University Canada*.

López-Colino, F., and Colás, J. (2012). Spanish Sign Language Synthesis System. *Journal of Visual Languages and Computing*, 23(3). 121–136.

Mac Kim, S., (2011). Recognising Emotions and Sentiments in Text. *University of Sydney*.

Maestri, G. (2011). FaceFX Studio 2010. *Computer Graphics World*, 34(4), 44-45

Malatesta, L., Raouzaiou, A., Karpouzis, K., and Kollias, S. (2009). MPEG-4 Facial Expression Synthesis. *Personal and Ubiquitous Computing*, 13(1), 77-83.

Martino, J. M. De. (2007). Benchmarking Speech Synchronized Facial Animation Based on Context-Dependent Visemes.

Mass 3 effect. (2012). *Electronic art*. Retrieved from http://www.masseffect.com

Massaro, D. W., and Cohen, M. M., (1983). Evaluation and Integration of Visual and Auditory Information in Speech Perception. *Journal of Experimental Psychology*. *Human Perception and Performance*, 9(5). 753–71.

Massaro, D. W., Cohen, M. M., Tabain, M., Beskow, J., and Clark, R. (2012). Animated Speech: Research Progress and Applications. *Processing of Auditory-Visual Speech*, 1-24.

McAllister, D. F., Rodman, R. D., Bitzer, D. L., and Freeman, A. S. (1997). Lip Synchronization for Animation. *ACM SIGGRAPH 97 Visual Proceedings: The art and Interdisciplinary Programs of SIGGRAPH'97*, 225. ACM

Microsoft (2015). *TTS Engine Vendor Porting Guide (SAPI 5.3)*

Mlakar, I., and Rojc, M. (2011). Towards ECA's Animation of Expressive Complex. *Analysis of Verbal and Nonverbal Communication and Enactment*. 185–198. Springer Berlin Heidelberg

Morten, H. L. (2012). Partially Automated System for Synthesizing Human Facial Expressions in Interactive Media. *Aalborg University*

Moussa, M. B., Kasap, Z., Magnenat-Thalmann, N., Chandramouli, K., Haji Mirza, S. N., Zhang, Q., and Daras, P. (2010). Towards an Expressive Virtual Tutor: an Implementation of a Virtual Tutor based on an Empirical Study of Non-Verbal Behaviour. *Proceedings of the 2010 ACM workshop on Surreal Media and Virtual Cloning*. 39-44. ACM

Naturwissenschaften, B. (2013). Affective and Attentive Interaction with Virtual Humans in Gaze-based Settings. *Doctoral dissertation, Augsburg, Universität Augsburg.*

Nicolaou, M. A., Gunes, H., and Pantic M. (2011). Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space. *Affective Computing Transactions on IEEE*, 2(2) .92-105.

Niewiadomski, R., Obaid, M., Bevacqua, E., Looser, J., Anh, L. Q., and Pelachaud, C. (2011). Cross-Media Agent Platform. *Proceedings of the 16th International Conference on 3D Web Technology*, 11-19. ACM.

Nishida, T., and Faucher, C. (2010). Modelling Machine Emotions for Realizing Intelligence: Foundations and Applications. *Springer Science and Business Me*dia, 1.

Noh, J., and Neumann, U. (1998). A survey of facial modeling and animation techniques. *USC Technical Report*. 99–705.

Oyarzun, D., Mujika, A., Álvarez, A., Legarretaetxeberria, A., Arrieta, A., and del Puy Carretero, M. (2010). High-realistic and Flexible Virtual Presenters. *Articulated Motion and Deformable Objects*, 108-117. Springer Berlin Heidelberg.

Pai, N. S., and Chang, S. P. (2011). An Embedded System for Real-Time Facial Expression Recognition based on the Extension Theory. *Computers and Mathematics with Applications*, 61(8), 2101-2106.

Pandzic, I. S., and Forchheimer, R. (2003). MPEG-4 Facial Animation: The Standard, Implementation and Applications. *John Wiley and Sons*.

Parke, F. I. (1974). A Parametric Model for Human Faces. *University of Utah. Computer Science*.

Parke, F. I. (1982). Parameterized Models for Facial Animation. *IEEE Computer Graphics and Applications*, 9(2), 61–68 .

Parke, F. I., and Waters, K., (2008). Computer Facial Animation. 2nd editionCRC *Press*, 289. 291-328.

Pasquariello, S., and Pelachaud, C. (2002). Greta: A Simple Facial Animation Engine. *Soft Computing and Industry*, 511-525. Springer London

Patel, N., and Zaveri, M. (2010). 3D Facial Model Construction and Expressions Synthesis using a Single Frontal Face Image. *International Journal of Graphics*, 1(1), 1-18.

Paul, R. (2010). Realization and High Level Specification of Facial Expressions for Embodied Agents. Master thesis. *Electrical Engineering, Mathematics and Computer Science.*

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11). 559–572.

Pelachaud, C. (2009). Studies on Gesture Expressivity for a Virtual Agent. *Speech Communication*, 51(7), 630-639.

Pelachaud, C. (2011). Expressive Gesture Model for Humanoid Robot. *Affective Computing and Intelligent Interaction*, 224–231. Springer-Verlag Berlin Heidelberg

Pelachaud, C., Badler, N. I., and Steedman, M. (1996). Generating facial expressions for speech. Cognitive science, 20(1), 1-46.

Pelachaud, C., Badler, N. I., and Steedman, M. (1991). Linguistic Issues in Facial Animation. *Proceeding of the 91 Computer animation*, 15-30. Springer Japan.

Petajan, E. (1999). Very Low Bitrate Face Animation Coding in MPEG-4. *Encyclopedia of Telecommunications*, 17, 209-231.

Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., and Salesin, D. H. (2006). Synthesizing Realistic Facial Expressions from Photographs. ACM SIGGRAPH 2006 Courses, 2(3). 19. ACM

Platt, S. M., and Badler, N. I. (1981). Animating Facial Expressions. *ACM SIGGRAPH Computer Graphics ACM*, 15(3), 245-252. ACM.

Plutchik, R. (1980). Emotion: A Psychoevolutionary Synthesis.

Plutchik, R., and Kellerman, H. (Eds.). (2013). The Measurement of Emotions. Academic Press. 4. Elseiver

Preda, M., and Jovanova, B. (2013). Avatar Interoperability and Control in Virtual Worlds. *Signal Processing: Image Communication*, 28(2), 168-180

Queiroz, R. B., Cohen, M., and Musse, S. R. (2009). An Extensible Framework for Interactive Facial Animation with Facial Expressions, Lip Synchronization and Eye Behavior. *Computers in Entertainment (CIE)*, 7(4), 58

Quené, H., Semin, G. R., and Foroni, F. (2012). Audible Smiles and Frowns Affect Speech Comprehension. *Speech Communication*, 54(7), 917-922.

Saisan, P., Bissacco, A., Chiuso, A., and Soatto, S. (2004). Modeling and Synthesis of Facial Motion Driven by Speech. *Computer Vision-ECCV 2004*. 456-467. Springer Berlin Heidelberg

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C., and Narayanan, S. (2013). Paralinguistic in Speech and Language—State-Of-The-Art and the Challenge. *Computer Speech and Language*, 27(1), 4-39.

Seongah Chin, Chung yean Lee, S. L. (2011). Multi-layer Structural Wound Synthesis on 3D face. *Computer Animation and Virtual Worlds*, *22*(2-3), 177–185.

Serra, J., Ribeiro, M., Freitas, J., and Orvalho, V. (2012). A Proposal for a Visual Speech Animation System. *Advances in Speech and Language Technologies for Iberian Languages*. 267–276. Springer-Verlag Berlin Heidelberg

Shapiro, A. (2011). Building a Character Animation System. *Motion in Games*. 98-109. Springer Berlin Heidelberg

Shih, P. Y., Wang, J. F., and Chen, Z. Y. (2010). Kernel-Based Lip Shape Clustering with Phoneme Recognition for Real-Time Voice Driven Talking Face. *Advances in Neural Networks-ISNN 2010*, 516-523. Springer Berlin Heidelberg

Sibbing, D., Habbecke, M., and Kobbelt, L. (2011). Markerless Reconstruction and Synthesis of Dynamic Facial Expressions. *Computer Vision and Image Understanding*, 115(5), 668-680.

Skantze, G., and Al Moubayed, S. (2012). IrisTK: A Statechart-Based Toolkit for Multi-Party Face-To-Face Interaction. *Proceedings of the 14th ACM international conference on Multimodal interaction*, 69-76. *ACM*.

Skantze, G., and Gustafson, J. (2009). Attention and Interaction Control in a Human-Human-Computer Dialogue Setting. *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 310-313. Association for Computational Linguistics

Somasundaram, A. (2006). *A Facial Animation Model for Expressive Audio-Visual Speech*. Doctoral dissertation. The Ohio State University.

Stevens, C. J., Gibert, G., Leung, Y., and Zhang, Z. (2013). Evaluating a Synthetic Talking Head Using a Dual Task: Modality Effects on Speech Understanding and Cognitive Load. *International Journal of Human-Computer Studies,* 71(4), 440-454.

Sun, X., Yao, H., Ji, R., and Liu, X. M. (2014). Toward Statistical Modeling of Saccadic Eye-Movement and Visual Saliency. *Image Processing, IEEE Transactions on*, 23(11), 4649-4662.

Tao, J., and Tan, T. (2004). Emotional Chinese Talking Head System. *Proceedings of the 6th International Conference on Multimodal Interfaces*, 273-280. *ACM*.

Taylor, S. L., Mahler, M., Theobald, B., and Matthews, I. (2012). Dynamic Units of Visual Speech. *Proceedings of the 11th ACM SIGGRAPH/Eurographics Conference on Computer Animation*, 275-284. Eurographics Association.

Theune, M., Meijs, K., Heylen, D., and Ordelman, R. (2006). Generating Expressive Speech for Storytelling Applications. A*udio, Speech, and Language Processing IEEE Transactions*, 14(4), 1137-1144.

Thiebaux, M., Marsella, S., Marshall, A. N., and Kallmann, M. (2008). Smartbody: Behavior Realization for Embodied Conversational Agents. *Proceedings of* *the 7th International Joint Conference*, 1. 151-158. Autonomous Agents and Multiagent Systems

Tinwell, A., Grimshaw, M., Nabi, D. A., and Williams, A. (2011). Facial Expression of Emotion and Perception of the Uncanny Valley in Virtual Characters. *Computers in Human Behavior*, *27*(2), 741–749

Torre, D., F., and Cohn, J. F. (2011). Facial Expression Analysis. *Visual Analysis of Humans*. 377-409. Springer London

TRueSpel. (2001). *English-Truespel (USA Accent) Text Conversion Tool*. Retrieved from http://www.foreignword.com/dictionary/truespel/transpel.htm

Tony de Peltrie (1985). Groupe Multi-média du Canada.

Tu, B. and Yu, F. (2012). Bimodal Emotion Recognition based on Speech Signals and Facial Expression. *Foundations of Intelligent Systems*, 691-696. Springer Berlin Heidelberg.

Van Welbergen, H., Nijholt, A., Reidsma, D., and Zwiers, J. (2005). Presenting in virtual worlds: Towards an architecture for a 3D presenter explaining 2D-presented information. Lecture notes in computer science, 3814, 203.

Vertegaal, R., Slagter, R., Van der Veer, G., and Nijholt, A. (2001). Eye Gaze Patterns in Conversations: There is More To Conversational Agents than Meets the Eyes. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 301-308. ACM.

Vertegaal, R., van der Veer, G., and Vons, H. (2000). Effects of Gaze on Multiparty Mediated Communication. *Graphics Interface*, 95-102.

Vijayarangan, R. (2011). Emotion based Facial Animation using Four Contextual Control Modes. Master Thesis. Faculty of Graduate Studies, Ontario, Canada.

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., and Schröder, M. (2012). Bridging the Gap between Social Animal and Unsocial Machine: A Survey of Social Signal Processing. *IEEE Transactions on Affective Computing*, 3(1). 69-87.

Wang, L., Chen, H., Li, S., and Meng, H. M. (2012). Phoneme-Level Articulatory Animation in Pronunciation Training. *Speech Communication*, 54(7), 845-856.

Waters, K. (1987). A Muscle Model for Animation Three-Dimensional Facial Expression. *ACM SIGGRAPH Computer Graphics*, 21(4). 17-24. ACM

Watson, M. J. (2013). Mass Effect. *Electronic art the Jennifer Hale Handbook*, 308. Retrieved from http://www.masseffect.com

Wei, L. and Deng, Z. (2015). A Practical Model for Live Speech-Driven Lip-Sync. *Computer Graphics and Applications,* 35(2), 70-78. IEEE

Whissell, C. (2010). Whissell’s Dictionary of Affect in Language: Technical Manual and User’s Guide. Laurentian University, http://www. hdcus. com/manuals/wdalman.

Wilson M. H. (2005). Face Robot Facial Animation Technology Pioneers New Approach to Believable Emotional Digital Acting. http://www.gizmag.com/go/4364/ 2005.

Wolff, R., Roberts, D., Murgia, A., Murray, N., Rae, J., Steptoe, W., and Sharkey, P. (2008). Communicating Eye Gaze Across a Distance Without Rooting Participants to the Spot. *Proceedings of the 2008 12th IEEE/ACM International Symposium on Distributed Simulation and Real-Time Applications*. 111-118. IEEE Computer Society

Woodward, A., Delmas, P., Chan, Y. H., Strozzi, A. G., Gimel’farb, G., and Flores, J. M. (2012). An Interactive 3D Video System for Human Facial Reconstruction and Expression Modeling. *Journal of Visual Communication and Image Representation*, 23(7), 1113–1127.

Wu, Z., Zhang, S., Cai, L., and Meng, H. M. (2006). Real-time Synthesis of Chinese Visual Speech and Facial Expressions using MPEG-4 FAP Features in a three-Dimensional Avatar. *INTERSPEECH*.

Xu, M., Ouyang, J., and Huang, Y. (2011). The Continuous Speech Triseme Reccognition Approach and Syncronized Mouth Animation System. *Journal of Computer and Information Technology*. 1(1), 1-5.

Xu, Y., Feng, A. W., Marsella, S., and Shapiro, A. (2013). A Practical and Configurable Lip Sync Method for Games. *Proceedings of the Motion on Games*, 131–140. ACM

Yang, M., Tao, J., Mu, K., Li, Y., and Che, J. (2011). A Multimodal Approach of Generating 3D Human-Like Talking Agent. *Journal on Multimodal User Interfaces*, *5*(1-2), 61–68.

Yu, H., Garrod, O., Jack, R., and Schyns, P. (2014). Realistic Facial Animation Generation Based on Facial Expression Mapping. *Fifth International*

Conference on Graphic and Image Processing. International Society for Optics and Photonics. 906903. 1-5.

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1), 39-58.

Zhang, L., Jiang, M., Farid, D., and Hossain, M. A. (2013). Intelligent Facial Emotion Recognition and Semantic-Based Topic Detection for a Humanoid Robot. *Expert Systems with Applications*, 40(13), 5160-5168.

Zhang, S., Wu, Z., Meng, H. M., and Cai, L. (2010). Facial Expression Synthesis Based on Emotion Dimensions for Affective Talking Avatar. *Modelling Machine Emotions for Realizing Intelligence*, 109-132. Springer Berlin Heidelberg

Zhang, Y., Prakash, E. C., and Sung, E. (2001). Animation of Facial Expressions by Physical Modelling. *Eurographics*. 1017-1027

Zhao, M., Gersch, T. M., Schnitzer, B. S., Dosher, B. A., and Kowler, E. (2012). Eye Movements and Attention: The role of Pre-Saccadic Shifts of Attention in Perception, Memory and the Control of Saccades. *Vision research*, 74. 40-60.

Zhao, X., Dellandréa, E., Zou, J., and Chen, L. (2013). A Unified Probabilistic Framework for Automatic 3D Facial Expression Analysis based on a Bayesian Belief Inference and Statistical Feature Models. *Image and Vision Computing*. 31(3), 231-245.

Zoric, G., Forchheimer, R., and Pandzic, I. S. (2011). On creating multimodal virtual humans—real time speech driven facial gesturing. Multimedia Tools and Applications, 54(1), 165-179.