Investigating 3D Visual Speech Animation Using 2D Videos

Rabab Algadhy

Supervisor: Dr. Yoshihiko Gotoh and Dr. Steve Maddock

Department of Computer Science

The University of Sheffield

This dissertation is submitted for the degree of

Doctor of Philosophy

Visual Computing Research Group July 2020

To the soul of my beloved sister Khadiga. May winds of heaven blow softly and whisper
in your ear how much I love you and miss you and wish that you were here.
To the sources of my strength, my parents, my beloved husband, and my uncle Fawzi.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Rabab Algadhy
July 2020

Acknowledgements

Through this PhD journey, I have been able to meet many challenges and achieved my aims through the invaluable support afforded by different people. This thesis would not have been accomplished without the gaudiness of my supervisors Dr. Yoshi Gotoh and Dr. Steve Maddock, who always supported my ideas, which made me a better person, student and researcher. I would also like to thank my secondary supervisor, Prof. Jon Barker for his helpful comments and suggestions regarding my work through each panel meeting. I am thankful to my examiners Dr. Heidi Christensen and Dr. Moi Hoon Yap for the valuable comments and useful discussion during the final viva of my defence.

My sincere thanks go to my colleagues in the Visual Computing Lab and the Speech and Hearing Lab for their continuous assistance, especially Robert Chisholm, Peter Heywood, Matthew Leach, James Pyle, Jack Deadman and Gerardo Roa Dabike.

My friends also deserve many thanks for their incredible support, particularly Mashael, Sadeen, Najwa, Basma, Mariam, Taghreed, Manal, Lubna, Eidah, Fatema, Dalia and Rebeeca.

I would never have reached this stage without my family. Thank you for keeping me strong throughout this journey. My beloved husband Adel, your unconditional love, support and understanding underpinned my persistence on this journey, which made the completion of this thesis possible. My dear parents, thanking you is not enough. I would be neither who nor what I am without you. Thank you to my beloved brothers and sisters, who never hesitate to offer help and warm welcomes, especially my caring sister Ranya for her help and advice, even with her busy schedule. I also feel incredibly blessed for the love and support that I continuously receive from my uncle Fawzi.

Finally, the work reported in this thesis would not have been possible without the financial support of the Libyan Ministry of Higher Education and Omar Al-Mukhtar University, to which I am grateful.

Abstract

Lip motion accuracy is of paramount importance for speech intelligibility, especially for users who are hard of hearing or foreign language learners. Furthermore, generating a high level of realism in lip movements is required for the game and film production industries. This thesis focuses on the mapping of tracked lip motions of front-view 2D videos of a real speaker to a synthetic 3D head. A data-driven approach is used based on a 3D morphable model (3DMM) built using 3D synthetic head poses. The 3DMMs have been widely used for different tasks such as face recognition, detect facial expressions and lip motions in 2D videos. However, investigating factors such as the required facial landmarks for the mapping process, the amount of data for constructing the 3DMM, and differences in facial features between real faces and 3D faces that may influence the resulting animation have not been considered yet. Therefore, this research centers around investigating the impact of these factors on the final 3D lip motions.

The thesis explores how different sets of facial features used in the mapping process influence the resulting 3D motions. Five sets of the facial features are used for mapping the real faces to the corresponding 3D faces. The results show that the inclusion of eyebrows, eyes, nose, and lips improves the 3D lip motions, while face contour features (i.e. the outside boundary of the front view of the face) restrict the face's mesh, distorting the resulting animation.

This thesis investigates how using different amounts of data when constructing the 3DMM affects the 3D lip motions. The results show that using a wider range of synthetic head poses for different phoneme intensities to create a 3DMM, as well as a combination of front- and side-view photographs of real speakers to produce initial neutral 3D synthetic head poses, provides better animation results compared to ground truth data consisting of front- and side-view 2D videos of real speakers.

The thesis also investigates the impact of differences and similarities in facial features between real speakers and the 3DMMs on the resulting 3D lip motions by mapping

between non-similar faces based on differences and similarities in vertical mouth height and mouth width. The objective and user test results show that mapping 2D videos of real speakers with low vertical mouth heights to 3D heads that correspond to real speakers with high vertical mouth heights, or vice versa, generates less good 3D lip motions. It is thus important that this is considered when using a 2D recording of a real actor's lip movements to control a 3D synthetic character.

Contents

List of Figures xv

List of Tables xxiii

1 Introduction 1

1.1 Motivation 3

1.2 Problem Statement 5

1.3 Thesis Aims 6

1.4 Contributions 8

1.5 Thesis Structure 11

2 Background 13

2.1 Speech Production 15

2.1.1 Anatomy of the Oral Cavity 15

2.1.2 Speech Production Phases 18

2.1.3 Vocal Tract Acoustics 20

2.1.4 Visual Phonetics Articulatory 22

2.1.5 Coarticulation 24

2.2 Speech Perception 26

2.3 Facial Animation 27

2.3.1 3D Face Reconstruction 28

2.3.2 Parameterising Face Models 32

2.3.3 Modelling Techniques 33

2.4 Visual Speech Animation . . . 34
2.4.1 Viseme Driven Approaches . . . 35
2.4.2 Data Driven Approaches . . . 36
2.4.3 Deep Learning Based Approaches . . . 37
2.5 Evaluation Approaches for Visual Speech Animation Quality . . . 40
2.6 Summary . . . 41

3 Mapping Process 43
3.1 Introduction . . . 43
3.2 Constructing the 3D Morphable Model . . . 44
3.2.1 Collecting Head Data . . . 44
3.2.2 3D Morphable Model . . . 46
3.3 Facial Features Tracking . . . 51
3.4 Mapping 2D to 3D . . . 55
3.4.1 2D Facial Landmarks Tracking . . . 58
3.4.2 Pose Estimation . . . 60
3.4.3 Shape Fitting . . . 62
3.5 A 2D Dataset for Creating, Animating and Evaluating the 3DMM . . . 62
3.6 Investigating Facial Feature Landmarks in the Mapping Process . . . 68
3.6.1 Evaluation . . . 69
3.6.2 Results and Discussion . . . 72
3.7 Summary . . . 74

4 3D Visual Speech Animation Using 2D Videos 77
4.1 Introduction . . . 77
4.2 The Data Sets of 3DMM . . . 79
4.3 Front-view Evaluation of 3D Lip Motion . . . 81
4.3.1 Evaluation . . . 81
4.3.2 Results and Discussion . . . 83
4.4 Side-View Evaluation of 3D Lip Motion . . . 91

4.4.1 Evaluation 92
4.4.2 Results and Discussion 95
4.5 Summary 101

5 Mapping Non Similar Faces 105
5.1 Introduction 105
5.2 Facial Features Classification 106
5.2.1 Background 106
5.2.2 Chosen Approach 108
5.2.3 Basic Statistics and Indices for Each Speaker 111
5.3 Mapping between Non Similar Faces 112
5.4 Objective Evaluation 114
5.5 Subjective Evaluation 137
5.6 Summary 146

6 Conclusions 149
6.1 Original Contributions 151
6.2 Limitations and Future Work 153

Bibliography 155

Appendix A Mapping 2D to 3D 175

Appendix B 3D Visual Speech Animation Using 2D Videos 179

List of Figures

2.1 The external anatomical review of the lips. 16

2.2 The speech organs, reproduced under Creative Commons Licence [Flemming]. 20

2.3 The vowel chart of the International Phonetic Alphabet (IPA). 22

2.4 Viseme classes extracted for a speaker of the GRID corpus [8]. 23

3.1 Schematic view of the proposed methodology to animate a 3D talking head. 45

3.2 Wireframe view of a neutral 3D head pose, generated using FaceGen software. 47

3.3 Generating initial 3D head pose using FaceGen software. 48

3.4 Viseme poses generated using FaceGen software. 49

3.5 An example of different intensities of viseme ah, generated using FaceGen software. 50

3.6 Visualisations of the mean head pose and the first three principal components of head shape for the 3DMM that contains 161 head poses. Each visualised as additions and subtractions from the mean. 52

3.7 A comparison of facial features landmarks tracking systems. The facial landmark points labelled on a real speaker for each tracking system. . . 53

3.8 Consecutive frames of the phoneme /ih/ during utterance of the word "bin" for a real speaker (ID: S48) (first row), and the corresponding 3D heads that are animated using Faceware landmark set (second row), Face Plus Plus landmark set (third row) and R-CR-R landmark set (forth row). 57

3.9 Consecutive frames of the phoneme /ih/ during utterance of the word "bin" for a real speaker (ID: S17) (first row), and the corresponding 3D heads that are animated using Bas et al's method (second row) and Huber et al's method (third row). 59

3.10 Consecutive frames of the phoneme /uw/ during utterance of the word "soon" for a real speaker (ID: S48) (first row), and the corresponding 3D heads that are animated using Bas et al's method (second row) and Huber et al's method (third row). 59

3.11 The facial landmark points labelled on a real speaker (left) and the corresponding 3D head model (right). 61

3.12 Flow diagram of using 2D dataset for creating, animating and evaluating the 3DMM. 64

3.13 The used helmet for recording the front- and side-view videos of real speakers. 64

3.14 Selected examples of front- and side-view frames of the dataset. 65

3.15 Used biometric landmarks of the face (L1 = nasion, L2 = endocanthion, L3 = exocanthion, L4 = alare, L5 = subnasale, L6 = labiale superius, L7 = stomion1, L8= stomion2 L9 = labiale inferius, L10 = gnathion, L11 = cheilion, L12 = zygion) 67

3.16 Measurements taken from each frame. 68

3.17 Sets of landmarks that used to map real faces to the corresponding 3D heads. 70

3.18 Definition of the critical landmarks to measure the width (W) and height (H) of mouth aperture. 71

3.19 Consecutive frames of the phoneme /b/ during utterance of the word bin from sentence "bin white in N 3 now" for a real speaker (ID: S 24) and the corresponding 3D head animated using each set of facial features landmarks. 75

4.1 Schematic view of animating and evaluating 3D lips of each 3DMM. . . . 80

4.2 First row: Front (left) and side (right) photographs of a real speaker (ID: S32); Second row: front and side view of the corresponding 3D heads generated using front photograph only (left) and front and side photographs (right) – the lips are more protruded in the image on the right. 82

4.3 Consecutive frames of the phoneme /w/ during utterance of the letter y from sentence "place green in Y zero again" for a real speaker (ID: S17) and the corresponding 3D head for each data set. 85

4.4 Width and height of mouth trajectories of 2D frames of the real speaker (ID:S17) and the corresponding 3D heads. Top two compare height and width between 17 and 161 poses (both with front- and side-view photos), while the bottom two compare height and width between front- view photo only and front- and side-view photos (both with 161 poses). 86

4.5 Error bars of the width of the mouth aperture parameter (W) of the corresponding 3D heads of each real speaker. 89

4.6 Error bars of the height of the mouth aperture parameter (H) of of the corresponding 3D heads of each real speaker. 90

4.7 Schematic view of the stages of the evaluation process of 3D lip motion from the side-view. 93

4.8 Definition of the critical measurements to determine the lip protrusion parameter in the side-view of a real speaker. The black vertical line represents a reference plane, the green horizontal line represents the measured distance for the lip protrusion, and the red dotted square represents the cropped mouth area. 94

4.9 Cropped mouth area from side view frame of a real speaker (left), and the corresponding binary image and calculated contours with highlighted points for the lip protrusion parameter calculation (right). 95

4.10 Consecutive frames for the side-view of a real speaker (ID: S 32) (a); and the corresponding 3D head (the virtual camera positioned at the centre 0° (b), 12°(c), 15°(d), 17°(e), and 21°(f) respectively). 96

4.11 Consecutive frames of the phoneme /w/ for the utterance of the word "white" from the sentence "bin white at A 9 now" for a real speaker (ID: S17) and the 3D head produced for each data set. 97

4.12 Consecutive frames of the phoneme /w/ for the utterance of the word "white" from the sentence "bin white in N 3 now" for a real speaker (ID: S24) and the 3D head produced for each data set. 98

4.13 Upper lip protrusion trajectories of 2D frames of the side-view video of a real speaker (ID:S17) and the corresponding 3D head, whilst uttering the sentence "bin white at A 9 now". Top: comparison of upper lip protrusion for models constructed with 17 and 161 poses (both with front- and side-view photos) against the real speaker; Bottom: comparison of upper lip protrusion for models constructed using front-view photo only and front- and side-view photos (both with 161 poses) against the real speaker. . . 100

4.14 Error bars of the upper lip protrusion parameter (UL) of the corresponding 3D heads of each real speaker. 102

5.1 Classification of indices for each speaker of the audio-visual Lombard grid speech corpus (80% confidence level for each class of each index). The x axis shows the speaker's ID (where M refers to male speaker and F refers to female speaker) and the y axis shows the indices' number. 113

5.2 An example of the mapping process between 2D video frames of a real speaker (ID: S17) who classified under the high class of index 7, the corresponding 3D head, and the non-corresponding 3D heads. 115

5.3 Consecutive frames of the phoneme /ih/ during utterance of the word "in" from sentence "bin white in O 7 now" for a real speaker (ID: S47) who is classified under the low class of index 7, the corresponding 3D head, the non-corresponding middle and the non-corresponding high 3D heads. . . 117

5.4 Width and height of mouth trajectories of 2D frames of the real speaker (ID: S 31) classified under the low class of index 7, the corresponding 3D head, the non-corresponding middle 3D head (ID: S 32) and the non-corresponding high 3D head (ID: S 19), whilst uttering the sentence "set white at D zero please". 119

5.5 Consecutive frames of the phoneme /b/ during utterance of the word "bin" from sentence "bin white at U three again" for a real speaker (ID: S32) who classified under the middle class of index 7, the corresponding 3D head, the non-corresponding low 3D head and the non-corresponding high 3D head. 122

5.6 Consecutive frames of the phoneme /b/ during utterance of the word "bin" from sentence "bin green by Q zero again" for a real speaker (ID: S22) who classified under the high class of index 7 (first row), the corresponding 3D head (second row), the non-corresponding low 3D head (third row) and the non-corresponding middle 3D head (last row). 124

5.7 Consecutive frames of the phoneme /p/ during utterance of the word "please" from sentence "lay white with A 5 please" for a real speaker (ID: S46) who classified under the high class of index 7 (first row), the corresponding 3D head (second row), the non-corresponding middle 3D head (third row) and the non-corresponding low 3D heads (last two rows). 126

5.8 Width and height of mouth trajectories of 2D frames of the real speaker (ID: S 22) classified under the high class of index 7, the corresponding 3D head, the non-corresponding middle 3D head (ID: S 55) and the non-corresponding low 3D head (ID: S 48), whilst uttering the sentence "set white with S 1 now". 127

5.9 Consecutive frames of the phoneme /ih/ during utterance of the word "bin" from sentence "bin white with M 2 soon" for a real speaker (ID: S16) who classified under the low class of index 10 (mouth width), the corresponding 3D head (second row), the non-corresponding high 3D head (third row). . 130

5.10 Consecutive frames of the phoneme /uw/ during utterance of the word "two" from sentence "bin white in I 2 soon" for a real speaker (ID: S38) who classified under the low class of index 10 (mouth width), the corresponding 3D head (second row), the non-corresponding middle 3D head (third row) and the non-corresponding high 3D heads (last two rows). 131

5.11 Consecutive frames of the phoneme /th/ during utterance of the word "three" from sentence "bin white in N 3 now" for a real speaker (ID: S24) who classified under the middle class of index 10 (mouth width), the corresponding 3D head (second row), the non-corresponding low 3D head (third and forth rows) and the non-corresponding high 3D heads (last row).132

5.12 Consecutive frames of the phoneme /b/ during utterance of the letter "B" from sentence "lay white by B 8 again" for a real speaker (ID: S54) who classified under the high class of index 10 (mouth width), the corresponding 3D head (second row), the non-corresponding low 3D head (third row) and the non-corresponding middle 3D heads (last row). 136

5.13 Structure of stimuli. 139

5.14 Screenshot of the graphical user interface for the naturalness test. 140

5.15 Results for the best, unselected and worst rates for the corresponding 3D heads for the N_e and NN_e groups: (a) Rates for index 7; (b) Rates for index 10. 142

5.16 Results for the best and worst choice rates for the corresponding 3D heads and the non-corresponding 3D heads for each class of index 7 for the N_e and NN_e groups: (a) Rates of the best answer ; (b) Rates of the worst answer. 143

5.17 Results for the best and worst choice rates for the corresponding 3D heads and the non-corresponding 3D heads for each class of index 10 for the N_e and NN_e groups: (a) Rates of the best answer ; (b) Rates of the worst answer. 144

A.1 Consecutive frames of the phoneme /b/ during utterance of the word bin from sentence "bin white with V 7 soon" for a real speaker (ID: S 48) and the corresponding 3D head animated using each set of facial features landmarks. 176

A.2 An example of texture distortion around the mouth area, due to using the set F3 for animating the 3D lips. 177

B.1 First row: Front (left) and side (right) photographs of a real speaker (ID: S17); Second row: front and side view of the corresponding 3D heads generated using front photograph only (left) and front and side photographs (right) – the lips are more protruded in the image on the right. 180

B.2 First row: Front (left) and side (right) photographs of a real speaker (ID: S48); Second row: front and side view of the corresponding 3D heads generated using front photograph only (left) and front and side photographs (right) – the lips are more protruded in the image on the right. 181

B.3 First row: Front (left) and side (right) photographs of a real speaker (ID: S24); Second row: front and side view of the corresponding 3D heads generated using front photograph only (left) and front and side photographs (right) – the lips are more protruded in the image on the right. 182

B.4 Consecutive frames of the phoneme /b/ during utterance of the word bin from the sentence "bin white in N 3 now" for a real speaker (ID: S24) and the corresponding 3D head for each data set. 183

B.5 Consecutive frames of the phoneme /uw/ during utterance of the word soon from the sentence "bin white with V 7 soon" for a real speaker (ID: S48) and the corresponding 3D head for each data set. 184

B.6 Width and height of mouth trajectories of 2D frames of the real speaker (ID:S32) and the corresponding 3D heads. Top two compare height and width between 17 and 161 poses (both with front- and side-view photos), while the bottom two compare height and width between front- view photo only and front- and side-view photos (both with 161 poses). 185

List of Tables

2.1 Muscles of the lips (The table is adapted from Edge [78]). 17

2.2 Place of Articulation (The table is adapted from Moore [182]). 20

3.1 Description of landmarks for each tracking system. 54

3.2 The RMS error averaged over frames of different phonemes for width (W) and height (H) of the mouth of the real speakers and their corresponding 3D heads that were mapped to each other using different facial features tracking systems. Values in bold means the decreased RMS error. 56

3.3 The RMS error averaged over frames of different phonemes for width (W) and height (H) of the mouth of the real speakers and their corresponding 3D heads that were mapped to each other using two different approaches. Values in bold means the decreased RMS error. 58

3.4 The structure of the GRID sentences [59]. 64

3.5 Metrical features (indices) of the audio-visual lombard grid speech corpus's speakers. 69

3.6 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers and their corresponding 3D heads that were mapped to each other using different sets of facial features landmarks. Values in bold means the decreased RMS error. Width and height error=±0.001. 73

4.1 17 or 161 poses in combination with front-view only or front- and side-views photos are used to prepare four data sets. 81

4.2 The RMS error averaged over frames of spread phonemes (/ey/, /iy/, and /s/) and protruding rounded phonemes (/w/ and /ow/) for width (W) and height (H) of the mouth of a real speaker (ID: S17) and the corresponding 3D heads created using Dataset 2 and Dataset 3 during utterance of sentence "place green in Y zero again". Values in bold indicate the lowest RMS errors for width and height in each row. 87

4.3 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers and their corresponding 3D heads. Values in bold indicate the lowest RMS errors for width and height in each row. Width and height error=±0.001. 88

4.4 The RMS error averaged over 4 sentences for upper lip (UL) protrusion of the real speakers and their corresponding 3D heads. Values in bold indicate lowest RMSE. UL error=±0.001. 101

5.1 Basic descriptive statistics for the indices. 112

5.2 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers classified under the low class of index 7 (vertical mouth height), their corresponding 3D heads and the non-corresponding 3D heads. Values in bold means the decreased RMS error. The last column shows p value of the t-test results between each corresponding 3D head and the non-corresponding 3D heads for width and height. 118

5.3 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers classified under the middle class of index 7 (vertical mouth height), their corresponding 3D heads and the non-corresponding 3D heads. Values in bold means the decreased RMS error. The last column shows p value of the t-test results between each corresponding 3D head and the non-corresponding 3D heads for width and height. 121

5.4 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers classified under the high class of index 7 (vertical mouth height), their corresponding 3D heads and the non-corresponding 3D heads. Values in bold means the decreased RMS error. The last column shows p value of the t-test results between each corresponding 3D head and the non-corresponding 3D heads for width and height. 125

5.5 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers classified under the low class of index 10 (mouth width), their corresponding 3D heads and the non-corresponding 3D heads. Values in bold means the decreased RMS error. The last column shows p value of the t-test results between each corresponding 3D head and the non-corresponding 3D heads for width and height. 129

5.6 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers classified under the middle class of index 10 (mouth width), their corresponding 3D heads and the non-corresponding 3D heads. Values in bold means the decreased RMS error. The last column shows p value of the t-test results between each corresponding 3D head and the non-corresponding 3D heads for width and height. 133

5.7 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers classified under the high class of index 10 (mouth width), their corresponding 3D heads and the non-corresponding 3D heads. Values in bold means the decreased RMS error. The last column shows p value of the t-test results between each corresponding 3D head and the non-corresponding 3D heads for width and height. 135

Chapter 1

Introduction

Over the last few decades, facial animation has received significant attention by researchers and become an active pursuit in a variety of applications of human-computer interaction. The lips play a major role in the portrayal of expressions through facial animation. In addition, they play an instrumental role in speech intelligibility, especially for hard-of-hearing individuals and foreign language learners. Therefore, generating visual speech animation with the behaviour of real people needs an extremely good presentation of lip motion and deformation. This can be achieved when the face image is linked with a 3D face model. There are two main traditional approaches used for reconstructing 3D face models. The first is based on capturing performance of real faces using RGB or RGB-D cameras and then reconstructing the 3D face model using the captured data. The other approach involves scanning and then blending 3D real faces using a linear model [161].

Comprehensive research has been conducted on reconstructing 3D face models from optical sensor measurements of a subject's performance. Face performance can be captured from the subject and represented in a 3D domain either based on illumination data only [103, 167] or with the aid of markers [25, 123]. These techniques involve capturing the face using a camera and then reconstructing the face geometry either via triangulation or colour and depth values based on the type of camera used. In the past, calibrated dense camera arrays with complex indoor lighting setups, which are expensive to set up and operate, were used [19, 96, 123]. Recently, low-cost devices such as monocular RGB and RGB-D cameras have been used for offline and online monocular face reconstruction and tracking [100, 240, 241]. However, the quality of these techniques is affected by lighting conditions that may produce undesired pixels with noisy depth values. This makes capturing faces more complicated. In addition, these techniques are highly challenging since they are based on forming the image by convolving multiple physical dimensions in a single colour measurement. Therefore, current state-of-the-art approaches employ face models and statistical analysis of the distribution of 3D facial shapes.

A large body of research has been conducted on modelling the structure and expression of faces based on low-dimensional subspaces. In some works, a blendshape expression model based on a set of 3D face models, each representing a particular expression, has been used. Facial animation can be achieved by morphing between the neutral face and a specific expression, or by morphing between various expressions. Several studies used delta variations to linearly add each expression to the neutral face [37, 101, 162, 241, 253]. Blendshape models can be constructed using multi-monocular cameras in general surroundings [101], monocular RGB cameras [241] or RGB-D sensor devices [37]. Although these methods can achieve globally pleasant results with regards to the static realism of the face rendering, all of these approaches require professional camera setups for gathering data to train the blendshape model, and the lip shapes that can be acquired during speech are still not fully included.

There is another body of research based on the most commonly used prior, presented by Blanz and Vetter [32], which constructs a 3D face model by learning from a low-dimensional face subspace created from high-resolution laser scans of real faces with neutral expressions (3D morphable model (3DMM)). Geometry and the illumination-corrected textures of the faces are included in such models. These methods applied this reconstruction for 2D face recognition, pose normalisation and illumination[125, 153], face reanimation [30] and facial expression tracking [12] in 2D images, but they were rarely extended to track lip motions during speech [67, 186] due to the expensive devices that are required for gathering the data from real speakers, in addition to the complexity of techniques for preprocessing the gathered data. These methods lack person-specific facial characteristics. In addition, they ignore anatomical and physical plausibility in the reconstructions. Some work has been conducted to personalize the model, though, either by increasing the number of 3D laser scans to model the skin reflectance, gender and age variations [34, 35, 125, 153], or by fitting person-specific shape correctives to the generic face models [37, 162], where performance of the resulting animation is mainly limited to noise levels and resolution of the input device. However, the key problem with these methods is that they are based on using 3D scans of real faces, which require controlled lighting conditions and high-cost devices during the scanning process, as well as complex techniques for preprocessing the scanned data.

To remedy these shortcomings, an approach for visual speech animation that uses tracked lip motion in front-view 2D videos of a real speaker to drive the lip motion of a synthetic 3D head is proposed. Inspired by the parametric face model presented by Blanz and Vetter [32], a 3D morphable model built using 3D synthetic head poses that are based on photographs of a real person can alleviate the shortcoming of preprocessing 3D laser scans of real head poses [23, 132, 228, 212] and personalising the generated 3D models [34, 35, 125, 153]. The 3D synthetic heads share vertex correspondences, which simplify tracking and analysing the lip motions over time at a detailed level. The generated 3D head model will be animated using 2D videos of a real speaker, which will remedy the limitations of using different input devices that distort the resulting animation due to their noise level and resolution [100, 240, 241]. Furthermore, such a technique does not require manual labelling or training as in [74, 100, 101, 129]. In addition, using 2D images or videos opens the door to representing historical or recently-deceased people for films and games.

1.1 Motivation

Over recent years, visual speech animation has received significant attention, and actively investigated in several applications of speech intelligibility enhancement such as assistive technology for foreign language learners and support for the hard of hearing individuals. Because of the major role that the external articulators (i.e. the lips, teeth, and tongue) especially the lips play for providing a significant proportion of the visual speech signals perceived from the face [178, 226]. This integral information improves the intelligibility of speech in adverse listening conditions either external adversity such as noisy environments [85, 169, 180, 223] or internal adversity such as hard of hearing people [139, 210, 221, 244] or non-native listeners [114, 115]. Many algorithms have been proposed to represent talking heads, for example, a synthesised talking tool (Baldi) presented by Massaro [172] that has been utilised for training vocabulary either for foreign students [173] or deaf students [17, 46] , where different versions of this head have been proposed for Italian [64] and Arabic [191].

However, it is still a big challenge to achieve optimal integration between the auditory signals and lip-sync, where any slight inconsistency between the two signals can lead to an illusion even for native speakers (the McGurk effect [179]). Building a synthetic visual speech system can be based on three main stages: reconstructing a 3D face model, parametrising the 3D face model, and animating the 3D face model using modelling techniques. Reconstruction of a 3D face model can be achieved by using RGB or RGB-D cameras that may produce undesired pixels which lead to a noise animated signal, face rigs created by an artist that should be mapped to the best matched character in facial features to avoid any potential imperfect animation as in [235], or 3D scanning of real faces that require high cost devices for gathering and complex techniques for preprocessing the gathered data. The parametrising stage links between reconstructing the 3D face model and the modelling techniques, where the face can be animated by interpolating the motion of these landmarks. Therefore such landmarks should be carefully selected to ensure animating the parts of the face involved in speech production to achieve high realism.

For modelling the 3D face, two types of information have been utilised to represent visual speech animation: phonetic information extracted from audio signals which are then classified into the corresponding mouth shapes (i.e. the position of the tongue, teeth and lips during utterance of a particular sound (viseme) [92]) (viseme driven approaches) [57], and motion capture data collected from real speakers which are then organised based on phonetic information [236] or processed using statistical models [101] (data driven approaches). Data driven approaches are more favored due to providing synthesised signals in accordance with visual signals of a real speaker which increases the realism. However, these approaches require a large amount of data that include all possible phonetic contexts for training which is a high cost and time consuming to be gathered. Therefore, it is highly desirable to build a cheap and reliable a 3D visual speech system that mimics real speech motions.

synchronising audiovisual sequences in the visual speech animation domain is fundamental for various applications including that of human computer interaction. Realistic 3D mouth motions promote the reliance and emotional of people toward machine by 30% over using text only [189]. Therefore, mapping visual speech signals extracted from a character to a 3D face rig is an urgent need to avoid any ambiguity in the final animation, where such ambiguous animated signal leads to loss of the convergence between customers and 3D animation. Such representation can provide a basis for entertainment applications such as games and films.

1.2 Problem Statement

Realising realism in visual speech animation is fundamental because people can spot any subtle abnormalities in the animated signal. Thus, creating natural-looking mouth animation remains a major challenge for developers aiming to animate a 3D talking head. One of the major challenges is including all variations of visemes during real speech motions, which is termed coarticulation [22], where the current viseme is influenced by the surrounding visemes. To comprise these effects in the synthesized system, a 3D talking head should be driven by motion data captured from a real speaker. This thesis, therefore, investigates how well can a 3DMM that created using synthetic 3D head poses efficiently produce visual speech, driven by tracked lip motions in 2D videos of a real speaker. This includes investigating the required facial features landmarks labelled on both real faces and 3D heads, and the required data for reconstructing and training the 3DMM to achieve a sufficient animation. Furthermore, investigating the impact of differences in facial features between real faces and 3D faces on the final animation.

1.3 Thesis Aims

This thesis investigates animating 3D lips using 2D videos with the aid of a 3DMM. In driving the 3D lips process with a captured 2D data from videos of a real speaker, a key element is to achieve 3D lip motions as smooth as a real speaker's. The goal is to detect the lip motions during speech in 2D videos, without any aid from the user for labelling or training as in the performance capture using illumination data approaches [74, 100, 101, 129]. This thesis uses a data-driven approach that maps tracked lip motions in 2D videos of a real speaker to corresponding 3D landmarks labelled on a 3DMM built using 3D synthetic head poses. To achieve this aim, firstly data of 3D head poses is required to build the 3DMM. The used 3DMM in this thesis is based on 3D synthetic head poses generated using commercial software (FaceGen [1]) to train the model, which saves time and effort. Furthermore, the FaceGen models have corresponding vertex data in each head, which simplifies creating a large number of face poses to train the model. Then, front-view 2D videos of a real speaker are mapped to the corresponding 3D head using Huber et al's [126] method that reconstructs 3D faces from 2D images and videos using a 3DMM which was created using neutral poses of real faces.

The main concept of the mapping between real faces and 3D faces is based on minimising the differences between the 2D landmarks detected from the input 2D frame and the corresponding 3D landmarks [32, 135]. Different number of facial landmarks have been used to animate faces, for example 74 landmarks [42], 75 landmarks [41], or 83 landmarks [232]. However, none of the previous studies has investigated the impact of each set of facial landmarks on the resulting animation. Therefore, it is valuable to test the functionality of each set of facial feature landmarks in the mapping process. This requires investigating different sets of facial feature points. The main research question from conducting such task, then: *Which facial landmarks should be used in the mapping process, and how do they influence the resulting 3D lip motions?*

In order to effectively mimic human lip motion during speech through fitting a 3DMM to video streams of the real speaker, the main visual appearance of mouth poses (visemes) during speech need to be included in the 3DMM. Most of the previous researches have been extensively studied using 3DMMs that are based on different numbers of 3D scans of real faces in a 2D face recognition domain for normalising pose and illumination in 2D images [32, 34, 125, 153]. Some of the models were extended to include facial expression poses [50, 126] and mouth shape poses [30] to reanimate faces [30], or detect facial expressions [12] or lip motions during speech [67, 186] in 2D images. However, these techniques do not give any thought to investigate the impact of using different amounts of data during constructing and training the models on the resulting animation. This thesis investigates how using two sets of photographs to create the initial neutral 3D head pose (i.e. front- and side-view photographs), and different datasets of viseme intensities for training the 3DMM influence the resulting 3D lip motions. The resulting animation is measured against the ground truth data that contains front- and side-view 2D videos of real speakers [10]. This investigation will answer the research question of: *How does using different amounts of data during constructing (Front-view or front- and side-view photos) and training (different intensities of the same viseme shape) the 3DMM affect the 3D lip animation results?*

This thesis also explores the impact of spatial relationships between facial features of real speakers on the resulting 3D lip motions by mapping between a real speaker's face and a non-corresponding 3DMM. Doing so will provide a better understanding of the impact of similarities and differences in facial features between real faces and 3DMMs on the 3D animation. Consequently, criteria of facial feature classification that enables animating 3D lips of a talking head using videos of different non-corresponding real speakers' videos can be defined. These criteria can be followed to animate historical or recently-deceased people for films and games. This is achieved by firstly classifying facial features of the real speakers, and then mapping 2D front-view videos of the real speakers to their corresponding 3D heads and 3D heads that correspond to different real speakers who are classified under different classes. This investigation will answer the research question of: *How does the differences between the facial features of the real speakers in the 2D videos and the 3D heads affect the resulting 3D lip animation?*

1.4 Contributions

The novel contributions of this thesis are:

- **Animating 3D lips using 2D videos with the aid of a 3DMM:** The major theme of the thesis is driving a 3D talking head using extracted information from 2D video frames of a real speaker. The presented method proposes generating accurate speaker-specific lip representations that retain the original characteristics of a speaker's lips via fitting a 3DMM to front-view video frames of a real speaker. This is achieved by following (implementation) Huber et al's [126] method, which reconstructs 3D faces from video frames via a 3DMM. In the current study, the 3DMM is built using synthetic 3D head poses that are generated using photographs of the corresponding real speaker. Then the 3DMM is mapped to front-view 2D videos of the corresponding real speaker. This work is presented in Chapter 3.

- **Identifying a set of facial features landmarks for achieving desired 3D lip motions**: Introducing a study that investigates the functionality of each set of facial features landmarks in the mapping process between a 3DMM and front-view video frames of the corresponding real speaker. Five sets of facial features landmarks were used to map a 3DMM to front-view 2D videos of the corresponding real speaker. To evaluate the resulting animation, two geometric articulatory measurements which are width and height of the mouth aperture were calculated from 2D videos of both the real speaker and the corresponding 3D animation and then compared against each other. The test results showed that all facial features points that include eyebrows, eyes, nose, and lips should be used to produce the best performance of 3D lip motions while adding the contour (i.e. boundary of the outside edges of the face) landmarks restricts the face's mesh, which leads to undesired lip motions. This work is also presented in Chapter 3.

- **Identifying the required amount of data to construct and train the 3DMM for producing efficient 3D lip motions** : Introducing a study that investigates how using front- and side-view photographs, rather than just a front-view photograph for construction of the initial neutral 3D head pose enhances the animation results. Furthermore, it investigates how using different intensities of the same viseme shape when training the 3DMM produces better animation results. For each real speaker, different 3DMMs were created using the above differentiating factors which are using two sets of photographs to generate the neutral 3D head pose and using different numbers of viseme intensities to train the 3DMM. Then front-view videos of the real speaker were mapped to each corresponding 3DMM. The resulting 3D lip motion was evaluated in comparison with front-view videos of the real speaker, using two geometric articulatory measurements (width and height of the mouth aperture). The results indicate that using both front- and side-view photographs to create the initial 3D head pose, as well as using a large number of 3D head poses to train the 3DMM, provides the best performance of the 3D lip motions. This study is presented in Chapter 4 and has been published as:

  – R. Algadhy, Y. Gotoh, and S. Maddock, "3D Visual Speech Animation Using 2D Videos", in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

- **Evaluation of 3D lip motions from the side-view:** Conducting a side-view evaluation of performance of the resulting 3D lip animation in comparison with ground-truth data which is side-view videos of the corresponding real speaker. This investigates whether or not the 3D lips are protruded adequately when only front-view videos are used in the mapping process. One geometric articulatory measurement was used to test the performance of the animated 3D motion, which is the upper lip protrusion parameter. The results confirmed that using both front- and side-view photographs to create the initial 3D head pose, and using a large number of 3D head poses to train the 3DMM, give the best performance of the 3D lip motions. This work is presented in Chapter 4.

- **First study of investigating the influence of differences in facial features between real faces and 3D faces on the final animation:** A study that investigates the impact of the spatial relations of real speakers' facial features on the animated 3D lip motion is presented in this thesis. Furthermore, a specific criteria of facial features' classification is defined which guides the solution to fit a 3DMM to 2D video of a non-corresponding real speakers without influencing the final 3D lip motion. Facial features of the real speakers of the Audiovisual Lombard Speech Grid corpus [10] were classified into three categories which are low, middle, and high. In this thesis, two facial features were tested that related to the vertical height and width of the mouth. Based on the classes of each feature, front-view 2D videos of each real speaker were mapped to the corresponding 3D head and the non-corresponding 3D heads that relate to the real speakers who classified under the other two classes. The resulting 3D lip motions were tested objectively and subjectively. The objective test results indicate that mapping between videos of real speakers that have low vertical mouth height and 3DMMs that correspond to real speakers that have high vertical mouth height, or vice versa, provides undesired 3D lip motions. The results were mixed for the mapping between non-similar faces based on the difference in mouth width, which confirms that this feature does not have a significant impact on the resulting 3D lip motions. User evaluation results found that the 3D heads that correspond to real speakers that have high vertical mouth height or mouth width, have the worst lip motions when they were mapped to real speakers that have low mouth features. This work is presented in Chapter 5.

1.5 Thesis Structure

The remainder of this thesis is presented in Chapters 2 to 6. The content of these chapters can be summarised as follows:

- **Chapter 2: Background.** This chapter presents the anatomical structures and behaviour relevant to the processes that underlie speech perception and production. A review of facial animation techniques and visual speech animation techniques are also presented in this chapter.

- **Chapter 3: Mapping Process.** This chapter presents the process of constructing the 3DMM and the framework for mapping between the 3D heads and the corresponding real faces used in this thesis. This chapter starts with an explanation of how the 3DMM is constructed using the synthetic 3D head poses, followed by a description of the technical implementation of the fitting 3D heads to their corresponding real faces. A study that investigates the role of each set of the facial features in the mapping process is presented as well in this chapter. The experimental evaluation of the animated 3D lip motion resulting from using different sets of facial features landmarks in the mapping process is then presented.

- **Chapter 4: 3D Visual Speech Animation Using 2D Videos.** First, the chapter illustrates the 3D data set that used to construct the 3DMMs, covering the set of 3D head poses used to train the 3DMMs, and the photographs of the real speaker used to reconstruct the initial 3D head pose. Evaluation of the resulting 3D lip motion is then presented by illustrating the geometric articulatory measurements and the methodology that is used to test the performance of the 3D lip motion from front- view only, followed by a presentation of the results and a discussion. In addition to, evaluation of the resulting 3D lip motion from side-view, followed by a presentation of the results and a discussion.

- **Chapter 5: Mapping Non Similar Faces.** This chapter starts with a review of facial features techniques, followed by an experimental design of fitting 3D heads to similar and non similar faces. Chapter 5 also demonstrates an experimental evaluation of the resulting 3D lip motion, which compares the performance of non similar faces with similar faces.

- **Chapter 6: Conclusions.** The final chapter concludes the thesis and highlights possible directions for future work.

Chapter 2

Background

Speech synthesis does not require only a knowledge of modelling and encoding organs and tissues of the vocal tract, but also an understanding of how these organs interact physically with each other for speech production and how the produced speech is perceived. The production of speech passes through different stages, starting from communicative intent (the speaker's thoughts), through a linguistic format that is articulated by the vocal tract's resonance and the physical movements of the external articulators, to finally produce speech sounds [105]. There is a link here between the process of real speech production and a sequence of questions regarding the representation of speech synthesis: what does the speaker want to say, how is this represented internally, and how a vocal tract model should be controlled to produce proper movements of the speech articulators? In the same way, the effectiveness of any synthetic speech signal is measured by comparing it with a ground truth signal, therefore, the perceived signal can affect the design of such artificial speech systems.

Speech communication is a harmonic process between speaker and listener. The speaker transforms thoughts and emotions into a linguistic format such as words, which are then converted to motor signals by controlling the movements of the vocal tract articulators. As the speaker adopts the speech production by forcing the air through the vocal tract cavity, causes changes in positions of the speech articulators, which results in producing speech sounds. The listener receives the speech signals by hearing and vision, where the brain processes these audiovisual signals to be decoded into the corresponding linguistic format for speech perception [105].

The process of speech communication as described above is a complex interaction of physical movements of the speech articulators and the linguistic format of a language for producing and perceiving speech. The speech system consists of production mechanisms (anatomy, acoustics, etc) and perception (audiovisual speech signals). For this reason, the first two sections of this chapter review the background of speech production and perception relevant to the speech synthesis systems.

Visual speech animation combines the techniques of facial animation and speech synthesis. Facial animation techniques involve modelling and encoding expressions of the face [196], while speech synthesis techniques involve modelling mouth shapes in synchronisation with acoustic speech signals. Visual speech animation is sourced from the strong correlation between the two techniques in a graphical and acoustical realistic speech synthesis manner, which includes both lips, tongue and teeth movements and facial expressions [171].

The rest sections of this chapter present a review of facial animation techniques, visual speech animation techniques, as well as evaluation approaches of visual speech animation. Section 2.3 presents a general review of the main procedures for animating faces. The procedures are broken down into 3D face reconstruction, parameterising face models, and modelling techniques. The first two procedures that are relevant to the scope of the thesis are addressed in more detail. Section 2.4 reviews visual speech animation techniques by exploring their processes. This review aims to highlight the main techniques and evidence from the literature that refers to the impact of using different sources of audiovisual data for feeding the synthesis engines on the resulting animation. Also, how the amount of training dataset can influence the final animation (Section 2.4.3). Section 2.5 covers approaches that are used to measure the effectiveness of visual speech animation techniques.

2.1 Speech Production

Speech production is a process that conveys information and expresses thoughts and emotions into speech. This contains word selection, organisation of grammatical forms, and then production of sounds by physical movements of the vocal articulators (lips, teeth, tongue, etc.). This section presents a review of the anatomy of the oral cavity associated with speech production with an emphasize on the lip shapes and motions, the physical structure of the main speech articulators, the classification of speech utterance, and speech motor control, converting the utterances of speech to low-level muscular control regarding a series of articulatory movements and the produced audio signals (coarticulation (Section 2.1.5)).

2.1.1 Anatomy of the Oral Cavity

The oral cavity is bounded between the external border that includes the lips and cheeks, and the internal border that includes the oropharynx. The lips surround the oral cavity and separate it from the external environment. They are formed anatomically by connective tissues and muscular skeleton, which are covered by the skin externally and the labial mucosa internally. The red part of the lip is termed vermilion. The vermilion border is demarcated by the sharp junction between the skin and the vermilion. The nasolabial grooves separate the upper lip from the cheeks, while the lower lip is separated from the chin by the labiomental groove. The midline of the upper lip is a vertical groove, which is termed a philtrum. Giving a bow shape to the lip which is known as a cupid's bow [44, 116]. The middle of the lips thicken and then thin to the corners where they are joined at the oral commissure [20, 44]. Figure 2.1 shows the external view of the lips anatomy.

There is a variety of lip shapes between people due to age, sex, and ethnicity effects. Aged people have a smaller lip area and thinner lips than young people, and females have thinner lips than males [68], especially for the upper lip [104]. Hashim et al. [113] reported Saudi males have a longer lower lip (i.e., the distance between the midpoint of the

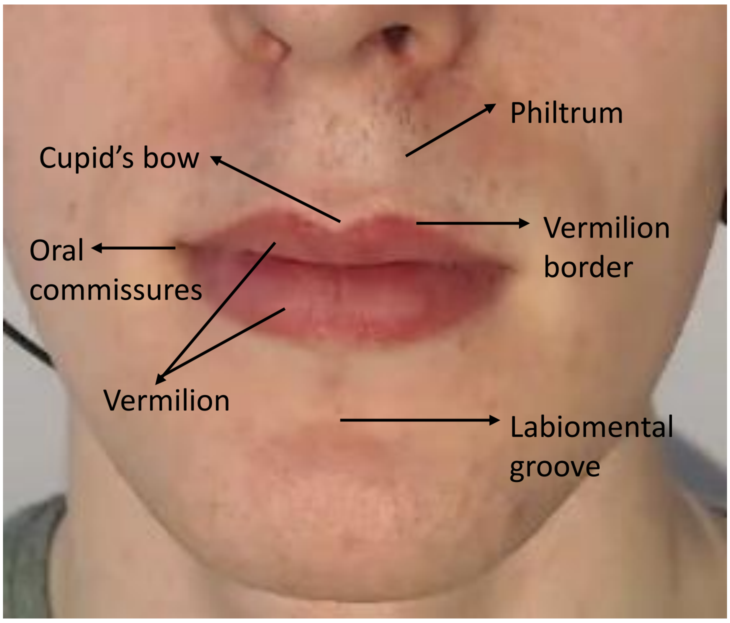


Figure 2.1 The external anatomical review of the lips.

inner contour of the lower lip and the bottom of the chin) and more protruded lips than Saudi females. Statistical comparisons of face dimensions between African-Americans and Caucasians showed that African-Americans have a significantly greater measurement of lip length (i.e., the distance between the corners of the lips) [177, 267, 268]. In a comparative study of Turkish and North-American adults, Uysal et al. [245] reported a more significant protrusion and shortness in the upper and lower lips in Turkish subjects for males only. Another comparative study between Korean and European-American adults showed that Korean subjects have more protruded lips [128].

The lips play a vital role in providing a significant proportion of the visual speech cues perceived from the face [121, 178, 179, 225, 227]. Most of the speech gestures are controlled by the lips (i.e., lip motions). For example, pressing the lips together produces the bilabial plosive sounds, touching the lower lip with the upper teeth produces the labiodental sounds, and rounding the lips produces some vowel sounds [250]. Table 2.1 shows the muscles of the lips and their actions. The lip shapes affect the sound radiation from the mouth, and they can influence the length of the vocal tract, consequently

Name	Action
Buccinator	Compresses the cheek against the teeth, and retracts the corner of the lip down.
Depressor anguli oris	Draws the corner of the lip downward.
Depressor labii inferious	Depresses the lower lip.
Incisive inferior	Pulls the lower lip toward the teeth.
Incisive superior	Pulls the upper lip toward the teeth.
Levator anguli oris	Moves the corner of the mouth up.
Levator labii superioris	Raises the upper lip.
Levator labii superioris alaeque nasi	Raises the upper lip and nostril.
Mentalis	Raises and protrudes the upper lip.
Obicularis oris	Closes the lips, compresses the lips, and protrudes the lips.
Platysma	Pulls the corner of the mouth down and back.
Risorius	Pulls the corner of the mouth back.
Zygomaticus major	Draws the corner ofthe mouth laterally and upward.
Zygomaticus minor	Draws the outer part of the upper lip upward, laterally and outward.

Table 2.1 Muscles of the lips (The table is adapted from Edge [78]).

affect the vocal tract resonance frequencies [117]. More detailed discussion on the muscle structure and function during speech can be found in [234].

The roof of the mouth is formed by the palate, which separates the oral cavity from the nasal cavity. It is divided into the hard palate and the soft palate (velum). The hard palate is located at the front and forms two-third of the roof of the mouth. It forms the maxilla with a bone surface that covered by a mucosal tissue and holds the upper teeth and the alveolar ridge that shapes the back arch of the upper incisors [20, 44]. The alveolar ridge plays a role in producing consonants, particularly for fricatives such as /s/ [117].

The soft palate originates from the bone palate that forms the hard palate ends. The soft tissues of the soft palate separate the nasal cavity from the oral cavity. The palatoglossus and palatopharyngeal folds run on the sides of the soft palate and cover the palatoglossus and the palatopharyngeus muscles that surround the tonsillar fossa. The velum is located at the posterior part of the soft palate, where it is responsible for raising and lowering the soft palate, opening and closing the passage to the nasal cavity.

The muscle that responsible for these actions is the levator veli palatini [168]. The free edge of the velum in the midline is called the uvula [117].

The tongue is formed by extrinsic and intrinsic muscles. The extrinsic muscles are connected to cartilage and bones and responsible for changing the tongue position, while the intrinsic muscles start and end within the tongue's soft tissue and are responsible for changing the shape of the tongue [111]. The tongue plays a vital role in speech production, where it is responsible for producing most of the speech sounds, except the glottal and bilabial consonants. Furthermore, it is an essential part of the oral anatomy that mainly contributes to forming the acoustic cavity for producing normal speech by restricting the air passage through the oral cavity [156].

Most of the speech sounds are produced by a combination of mouth articulators such as lips, teeth, tongue, etc. The following section describes the kinematics of these articulators to produce speech sounds.

2.1.2 Speech Production Phases

Most speech sounds are produced by forcing the air stream from the lungs towards the trachea, the nasal cavity and the oral cavity, which includes the hard and soft palates, the tongue, the teeth, and the lips (see Section 2.1.1). Figure 2.2 illustrates the involved organs in the speech production process. Some modifications are made by the vocal tract organs on the air passage, resulting in production of different sounds. The speech production process involves four phases: initiation, phonation, oro-nasal process and articulation. The following points summarise each process according to Giegerich [106] and Trujillo [243]:

- Initiation starts when the air is exhaled from the lungs and forced through the vocal tract and the oral and nasal cavities. This process is required for producing all speech sounds in the English language, since they result from a pulmonic egressive air stream, while in some other languages, speech sounds result from the opposite air stream direction (ingressive).

- Phonation occurs when the air passes through the larynx, in which the gap between the vocal folds (glottis) is opened and closed, causing vibrations of the vocal cords. A variety of different speech sounds can be produced when the shape of the glottis is altered by the vocal folds' movements: slightly opened glottis produces voiced sounds (increased vibration); widely opened glottis produces voiceless sounds (reduced vibration). The vocal cords' movements determine other acoustic features, which are the: Fundamental Frequency (F0) - determined by the number of the vocal folds' movements in one second; Intensity (loudness of sound) - determined by the energy of the vocal folds' movements; and the Quality - determined by patterns of the folds' movements.

- Oro-nasal starts after the phonation process, in which air passes through the nasal cavity or the oral cavity. In normal breathing the air goes through the nasal cavity, while in speech sounds production it is directed to the oral cavity, since the nasal cavity is closed by the velum. The state of the velum can regulate the production of speech sounds: oral sounds such as /v/, /f/ and /l/ are produced when the velum is raised; nasal sounds such as /m/ and /n/ are produced when the velum is lowered.

- Articulation shapes part of the vocal tract above the vocal cords to produce more distinguishable speech sounds. This process concerns the movements of the articulators (tongue, teeth, lips, etc.) to direct the air into either the oral or the nasal cavities. The articulators can be active, such as the velum and the tongue, or passive, such as the alveolar ridge and the hard palate as presented in Table 2.2. The passage of air is constrained by articulatory movement that regulates either contact between passive and active articulators (e.g. contact between the tongue and the teeth), or airflow through the nasal cavity or the oral cavity (e.g. lowered or raised velum as explained in the previous process).

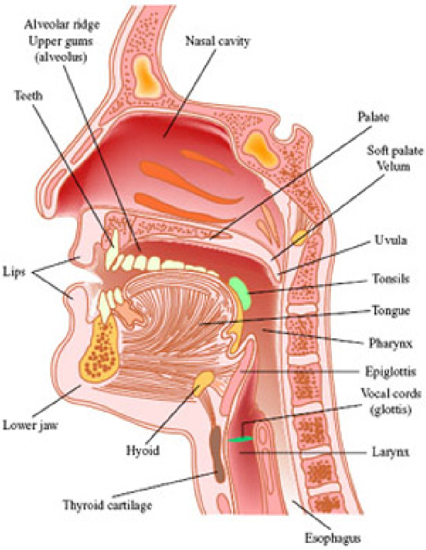


Figure 2.2 The speech organs, reproduced under Creative Commons Licence [Flemming].

Place of Articulation	Active Articulator	Passive Articulator	Example Sounds
Bilabial	upper and lower lips	none	[p b m]
Labiodental	lower lip	upper front teeth	[f v]
Dental	tongue tip	upper front teeth	[θ ð]
Alveolar	tongue tip or blade	alveolar ridge	[t d n l s z]
Postalveolar	tongue tip or blade	rear of alveolar ridge	[ɹ ʃ]
Retroflex	tongue tip	hard palate	[ʈ ɖ ɳ]
Palatal	tongue front	hard palate	[j ɲ]
Velar	tongue back	soft palate	[k g ŋ]
Uvular	tongue back	uvula	[q ɢ]
Pharyngeal	tongue root	rear wall of pharynx	[ħ]
Glottal	vocal folds	none	[h ʔ]

Table 2.2 Place of Articulation (The table is adapted from Moore [182]).

2.1.3 Vocal Tract Acoustics

As is apparent from the summary of speech production phases given above, different speech sounds can be produced by articulatory movement that regulates the passage of the air through the oral or nasal cavities. In the following, phonetic terminology that is utilised to classify the acoustic signals is presented, followed by a classification of the visual signals.

Phonetic Terminology

Phonemes are the basic units of speech sounds, which form words when they are combined together, and they can be classified into one of the following categories [206]:

- Vowel: is produced when the vocal tract is opened, and is defined by the position of the tongue (location and height) and the roundedness of the lip. A vowel can be formed by a single sound (monophthong) such as /i/ in a word (h*i*t), or by a combination of two vowels (diphthongs) such as /oy/ in a word (b*oy*). A standard transcription of all possible vowel sounds has been established by the International Phonetic Association (IPA) (Figure 2.3). In the (IPA) chart, the position of a vowel reflects the position of the tongue during production of that vowel. The upper-left point on the chart presents that the tongue is closer to the front of the mouth, and it is closer to the back of the mouth on the upper right of the chart. The roundedness of the lip is presented in this chart as follows: symbols on the left of the chart represent unrounded vowels, and symbols on the right presents the rounded vowels.

- Consonant: is produced when the vocal tract is completely or partially closed. A consonant can be classified into: nasal, which is produced when the air passage is directed to the nasal cavity by the velum (e.g.: /m/ and /n/); fricative, which is produced when the air passes through a narrow exit (e.g.: /f/ and /s/); affricative, which is produced when the air passage is constricted and then released (e.g.: /ch/ and /jh/); and plosive, which is produced when the air passage is completely stopped and then released (e.g.: /b/ and /d/).

- Semi-vowel: is a sound that has a phonetic nature of vowel sounds and is produced by the same manner of consonant sounds such as /w/ and /y/.

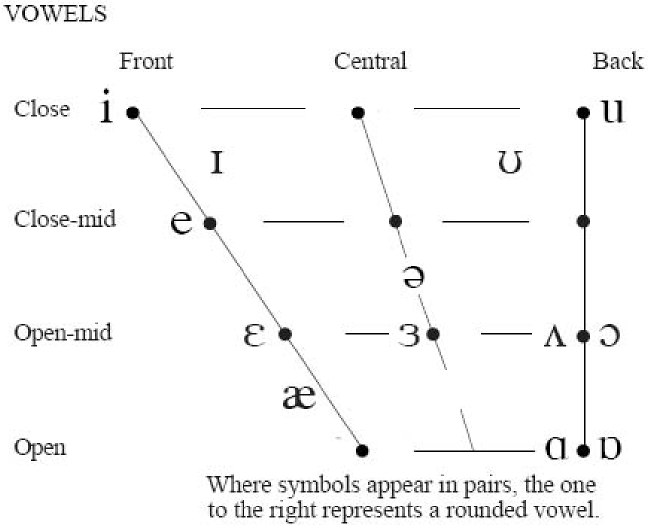

VOWELS
Front
Central
Back
Close
Close-mid
Open-mid
Open
i
u
ɪ
ʊ
e
ə
ɛ
ɜ
ʌ
ɔ
æ
ɑ
ɒ
Where symbols appear in pairs, the one to the right represents a rounded vowel.

Figure 2.3 The vowel chart of the International Phonetic Alphabet (IPA).

2.1.4 Visual Phonetics Articulatory

As described in the previous sections, articulatory phonetics relates the production of different sounds by physical movements of the vocal articulators (lips, teeth, tongue, etc.). The visual extent of these articulatory movements can enhance the intelligibility of speech [219, 223].

Given that the intelligibility of speech communication is increased when visual signals are combined with audio signals, it is therefore sensible that such visual signals are classified. This classification involves recognising speech elements based on the articulatory configuration of a specific phoneme (viseme). In other words, phonemes of speech are classified into visual units called visemes. Figure 2.4 shows viseme classes extracted for a speaker of the GRID corpus [8]. The term viseme was introduced by Fisher [92] to identify visually perceived consonants. Several viseme-to-phoneme mappings based on a many-to-one relationship have been proposed in the literature, but there is no consensus or standard classification system. Two approaches can be distinguished for building a map. The first approach involves defining viseme classes based on the linguistic knowledge of a language and articulatory rules (e.g. the lips' position and

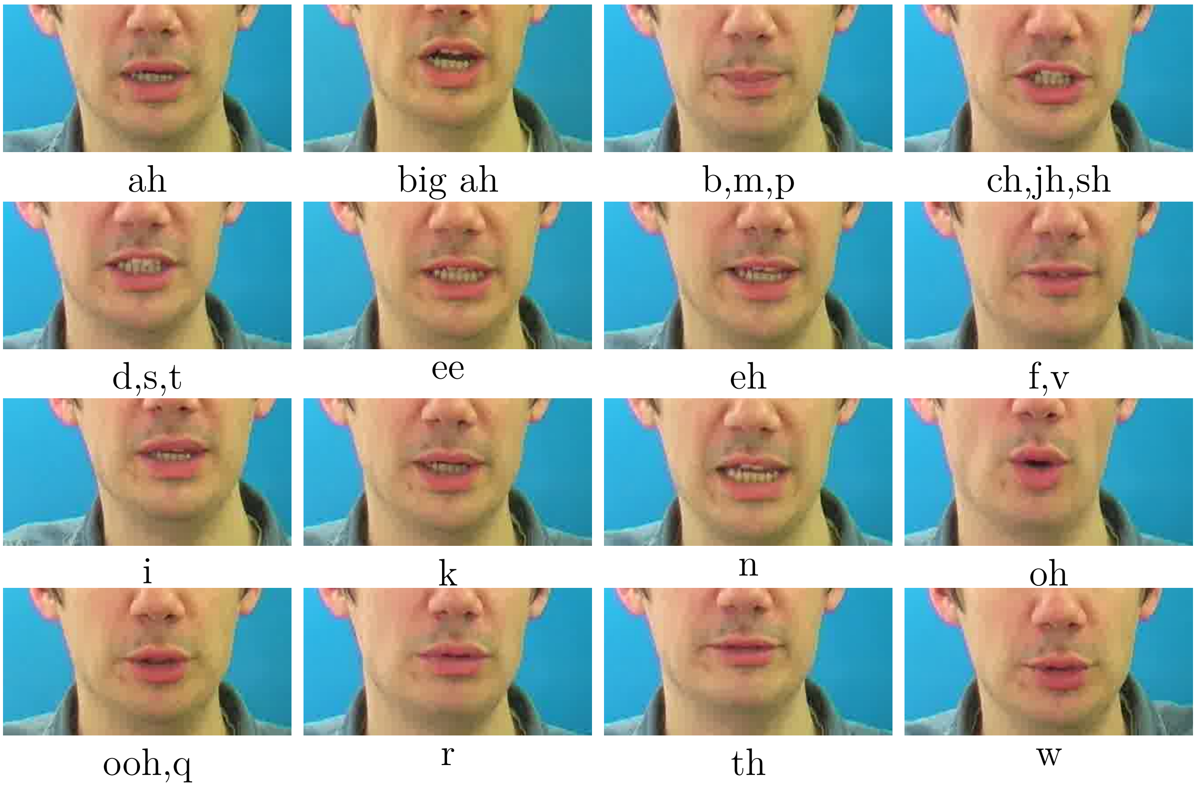


Figure 2.4 Viseme classes extracted for a speaker of the GRID corpus [8].

place of articulation) to predict phonemes that have similar visual appearances [13, 133]. The second approach involves using data-driven approaches to define viseme classes, in which a real speaker's visual speech data are recorded and analysed [26, 77, 181]. However, the many-to-one viseme mapping approach has multiple limitations. First, it does not consider the synchronisation between the audio signals and visual signals of a phoneme, where the two signals do not always correspond to each other. Second, some phonemes can be produced without using visual articulators. Such phonemes could not be classified under the same viseme class; for example, /k/ and /g/ are velar consonant sounds produced at the back of the soft palate. Finally, this approach does not consider coarticulation effects on visual speech. Thus, a many-to-many relationship mapping scheme should be considered instead [131].

2.1.5 Coarticulation

Coarticulation refers to the process that the brain follows to organise sequences of speech sounds (consonants and vowels) into speech units called syllables. In other words, it is the utterance of two or more syllables together that exert on each other. Coarticulation can be identified as backward or forward coarticulation. Backward coarticulation (carry-over coarticulation) occurs when the articulatory gesture of a sound is affected by the previous gesture in the speech sequence (e.g. lip protrusion while uttering the phoneme /s/ in the word "boots"). Forward coarticulation (anticipatory coarticulation) occurs when the articulatory gesture of a particular sound is affected by a gesture of a later sound in the speech sequence (e.g. lip rounding while uttering the phoneme /s/ in the word "stew") [22].

Hence, the visual appearance of a particular phoneme is shaped not only by its articulation properties but also by the neighbouring phonemes in the speech sequence. In other words, a single phoneme can have different visual appearance shapes, which means it can be classified into different viseme classes. Consequently, mapping phonemes to visemes should be a many-to-many relationship [131]. Mattys et al [176] have taken the first step in this direction by redefining some of the phoneme classes presented by Auer and Bernstein [14] and considering the phonetic context of consonants. However, a viseme set generated following this approach is inadequate for defining atomic units of visual speech. Thus, the classification of visual speech units could only be determined by considering the visual features rather than phonemes' segmentation. Hilder et al [118] proposed such an approach by classifying different allophones of a phoneme into a different viseme labels that automatically take the visual coarticulation into account. Although there is no direct mapping between phonemes and visemes, which makes this method far from straightforward for analysing and synthesising visual speech, Taylor et al [236] used such mapping to animate a 3D talking head.

Modelling the coarticulation has received considerable attention from researchers [27, 57, 187, 207]. Öhman presented a numerical model that identifies vocalic and consonant gestures with more cognitive and robust blending. In this model, the behaviour of the tongue muscles during the non-symmetric vowel-consonant-vowel syllable ($V_1\mathrm{C}V_2$) utterance is predicted as follows:

$$s(x;t) = v(x;t) + k(t)[c(x) - v(x;t)]w_c(x) \tag{2.1}$$

where $s(x,t)$ represents the shape of the vocal tract at a point x positioned on the tongue body at a time t between the shapes of the initial vowel V_1 and the final vowels V_2. $v(x,t)$ and $c(x)$ represent the vocal tract shapes of the surrounding vowels and consonants, respectively. $k(t)$ represents the emergence of the consonant, and $w_c(x)$ measures the dominance's amount of a vowel shape that distorts the target consonant shape. $k(t)$ varies from zero for the initial vowels V_1 to one for the consonant and then set to zero again to present the final vowels V_2.

Although this model does not account for more complex coarticulation effects, such as consonant-consonant coarticulation, it is considered the first step toward general modelling of coarticulation and speech synthesis applications. Revéret et.al [207] applied this model to general coarticulation to animate their 3D talking head. Such a model is unsatisfactory because they marked 30 points on one side of the face and then were mirrored for tracking the lip motions. Resulting in a very symmetric animation which is unrealistic, where most people speak asymmetrically [140]. Also, Löfqvist [164] extended this model to general speech, where each articulator is defined by several related dominance functions that simulate the effect of a corresponding viseme on speech production. A speech utterance trajectory can be determined by the shape of the resulting dominance functions. Cohen and Massaro [57] implemented Löfqvist's model of coarticulation for visual speech synthesis, and their model is considered the most commonly applied model for visual speech synthesis [65, 147, 157]. However, the main weakness of the model is the failure to represent lip closure for bilabial sounds [22, 175]. Deng et.al [71] presented a model that learns speech co-articulation from motion captured data and audio for expressive speech animation. The model fails to handle different rates of speech and dynamics.

Another major limitation of this model is that it is relying on predefined labels of key viseme shapes in the training data, which is painstaking manual work that may lead to ambiguities in the training data.

2.2 Speech Perception

Speech is produced as audio waves that are interpreted by a listener. The produced audio signals contain the meaning of the speech, which enable to communicate remotely without the need of face to face communication. However, there are two main information streams assist speech perception which are audio signals including a series of speech sounds, and visual signals including visual articulators of the speaker such as lips [172]. Studies in audiovisual speech intelligibility have proved conclusively that visual information such as the movements of the lips assist the perception of speech significantly [169, 223]. In fact, the visual signals do not just help deaf or hard-of-hearing people to understand what the speaker says, they also improve the intelligibility of speech and enhance sensitivity to acoustic information for hearing people in a noisy environment [169, 188, 217]. Le Goff et al [158] have proved that using degraded audio signals, two-third of the missing auditory intelligibility can be provided by the natural face, half of the missing intelligibility can be provided by their facial model (without tongue movements) and their lip model provided a third of it. For foreign language learners, pronunciation is considered an essential factor for listening and speaking. Intelligible pronunciation does not only aid students to understand and be understood, but also enhances their self confidence to be involved in an engaging manner [47]. In order to achieve correct pronunciation, an extremely accurate position of various articulators such as tongue, teeth and lips is required. In particular, the articulation of rounded vowels requires good perception of lip protrusion to be pronounced correctly. This is often difficult for some English language learners such as Turkish learners, since they are not able to make the essential muscular effort to produce the phonemes /ao/ and /ow/ (IPA notation: / ɔ: / and /oʊ/ respectively) accurately [119].

Extensive research has shown that the perception of speech can be changed when visual signals are combined with incorrect audio signals (McGurk effect) [179], providing compelling evidence of audiovisual speech synchrony for speech perceptibility [213, 224]. For example, when a video of lip movements of "ga" dubbed with the audio signal of "ba", it is perceived as a syllable "da". This suggests that such a presented example of McGurk would not be expected to appear when constructing any speech synthesis system. However, this phenomenon confirms importance of good linking between visual signals and audio signals for speech perception. Incompatibility between these two signals leads to speech disambiguation, which confirms achieving competent models for generating speech movements are essential in animation.

2.3 Facial Animation

Facial animation concerns techniques for modelling and encoding facial expressions. Three procedures are followed to animate faces: designing the 3D mesh of the face, parameterising the 3D facial mesh and animating the 3D mesh in a controllable manner to simulate facial expressions. In the first procedure, deformable face geometry is of paramount importance, as it helps represent all facial expressions in a realistic manner. This can be achieved using different devices, such as RBG-D cameras or 3D face scanning [269]. The second procedure involves parameterising specific nodes on the face mesh to simplify modelling the facial expressions or generating the speech movements, such as coarticulation effects. The final procedure regards how much the nodes should be displaced to generate a particular facial expression or mouth motion. Various modelling techniques to produce facial animations have been published; these can be divided into two categories: geometrical techniques and physical techniques.

2.3.1 3D Face Reconstruction

Input Modalities

In this method, 3D face models are reconstructed by capturing the face using optical sensors and then representing the captured data in a 3D domain based on the illumination data. Previously, a multi-view setup was used to capture the performance of real faces, where the subject is surrounded by pairwise stereo cameras. The 3D face is reconstructed by exploring the face geometry captured by each stereo pair via triangulation and then aggregating all the geometries obtained by each camera in a consistent manner. Marker data [7, 25, 122] or high illumination [249, 254, 148] can be used during the recording to aid reconstruction. However, these techniques require expensive materials for building and operating; therefore, research tends to use lightweight cameras instead. Lightweight and low-cost cameras, such as RGB and RGB-D cameras, are typically used.

RGB cameras are equipped with complementary metal oxide semiconductors or charge coupled device (CCD) sensors to capture the three channels of an image and encode them separately into red, green and blue. Modern cameras use three CCD sensors to capture each channel signal and use a Bayer filter to arrange them in a square array. The 3D face can be reconstructed using physical dimensions of each colour channel, such as geometry, illumination and surface reflectance. Because this type of camera is easily accessible, it has been widely used by researchers for offline [99, 100, 129] or online [40, 41, 166] 3D face reconstructions and motion capture. However, 3D face reconstruction from monocular data is challenging due to the complexity of the image deformation process, which involves representing multiple physical dimensions in one colour measurement. Therefore, researches tend to use straightforward techniques for image deformation, such as image-based 3D face reconstruction, and apply data-driven priors.

RGB-D cameras are provided by passive or active depth sensors that capture both color and depth data. This solves a problem of depth ambiguity in the monocular reconstruction techniques, since a coarse geometry can be estimated. Accordingly, these cameras have been used for 3D face reconstruction and tracking approaches [162, 192, 253].

Most passive depth cameras are based on a stereo camera setup, where a 3D point is reconstructed via triangulation between two views of the point on a calibrated stereo setup. Matching between pixels in the two views is challenging. Therefore, other features of epipolar geometry, such as colour and edges, can be used. However, if these features cannot be detected, the reconstruction process will fail.

To solve this problem, a projector can be placed instead of one of the stereo system cameras. Another problem can be released using projectors is that geometric features and structures can not be reconstructed as long as they are smaller than the projected pattern. These issues can be tackled by using cameras that have both passive and active sensors. Generally, these cameras have poor depth data and a low signal-to-noise ratio compared to data captured by 3D scanning devices. In addition, undesired pixels that appear between the background and foreground at depth discontinuities cannot be modelled and therefore complicate the process of face reconstruction. Thus, most 3D face reconstruction approaches rely on image formation or 3D scanning devices.

Image Formation Models

This method involves the reconstruction of a 3D face model from an image in an inverse rendering through a sequence of mathematical processes. To define a face's geometry, a mapping from 2D space to 3D space (R^2 to R^3) is required, where the range is $x(u) = (x(u), y(u), z(u))$, $x(u) \subset R^3$, and the domain is $u = (u, v) \subset R^2$ which is called UV space. The face geometry can be represented in a 3D space as triangle or quad meshes. A triangle mesh $M = (V, F)$ consists of a set of points $V = p = (x, y, z)$ and a set of triangles (face) $F = (p_i, p_j, p_k)$ where $p_i, p_j, p_k \in V$. A given image can be treated as a geometry surface that consist of pixels that define the distance between a known reference frame and a visible point in the scene. The image can be represented as $I = (x, y, f(x, y))$ [222].

The mapping process between the 3D geometry and a 2D image is achieved through projection camera models, such as orthographic projection [28, 97] , Weak Perspective projection [38, 218, 246] or Full Perspective projection [100, 129, 230, 256]. Some material properties describe light reflections on the skin should be modeled. Modeling the light reflections on faces is challenging since faces reflect a different amount of brightness and a specific amount of diffuse, due to various factors such as oily skin or sweat [150]. To tackle these issues, the face can be represented as a simple appearance model that ignores subsurface effects [36, 134]. Moreover, some models, such as Spherical Harmonics [185] and Environment Maps [6], can be used to compute incident illumination on faces.

3D scanning

Another method that is frequently used to construct 3D face models is laser scanning devices. There are several devices available to capture faces [33]. The 3DMD device, designed for medical imaging, uses four cameras to capture the face and produce high-quality coloured textures. The FR1 device captures faces in real time using a single camera and a projected infrared pattern and provides non-coloured images. The FR2 device captures faces using three cameras and an LCD stripe pattern and provides colour texture map images. Konica Minolta 910 captures faces using a single camera and laser stripe and generates 3D data via triangulation. Polhemus captures faces using a single camera that is fixed on a handheld wand and a laser strip for triangulating the 3D data.

3D scan devices provide a 3D surface from a single viewpoint. 3D scans are registered to compose 3D models that represent a 3D surface from various viewpoints. 3D models can be rigid models that describe the texture and geometry of an object or morphable models [32, 80] that adjust the texture and shape of an object by morphing between the scanned data. However, the main stage of constructing 3D models from the 3D scanned data is the establishment of correspondences among the vertices of the scanned data to make the data consistent so that all meshes have the same number of vertices, the same triangulation and the same anatomical structure of each vertex. For example, if the index i of a vertex in a mesh corresponds to the nose root, it is essential that the vertex of the nose root in every single mesh in the data is represented with the same index number.

Multiple approaches to aligning the data have been published: the rigid alignment approach uses affine transformations, such as a least squares linear system or the Iterative

Closest Point (ICP) [23, 212], to align two different meshes; the non-rigid alignment approach involves deforming the 3D models using interpolation techniques, such as Thin Plate Splines [127] or motion segmentation techniques such as optical flow [32]. Although these approaches easily state the correspondence problem, it is still challenging to achieve an accurate and robust alignment for highly inconstant face meshes. For example, some smooth regions of the face, such as the forehead or cheek, have a very detailed anatomical meaning, which complicates measurements of the correspondence and thus affects the resulting animation. Once the vertex correspondences are established among the 3D data, a 3D morphable model can be constructed.

The most widely used approach to building the 3D morphable model is the approach presented by Blanz and Vetter [32]. Their approach is based on learning a low-dimensional face subspace from high-resolution 3D scanned data. Several models based on this approach have been built using different input modalities [43, 144], different number of 3D scans [34, 35, 125, 153], different facial expressions either collected from one person [30] or several persons [12]. Blanz and Vetter's model is composed of 200 scans of real faces with neutral poses. It consists of the geometry and the illumination textures of the real faces. The non-rigid alignment approach was applied to align the 3D face data. A principle component analysis (PCA) was applied to the geometry and the illumination components separately, resulting in two models: one for the shape and the other for the texture. The principle components of a dataset and their corresponding standard deviations are computed using the PCA within a multi-variate Gaussian distribution framework. Using this approach, new faces with different skin reflectance can be synthesised. 3D morphable models enable the construction of a high-quality representation of 3D faces from insufficient sources of data, such as 2D images or noisy 3D scanned data. They also provide a mechanism for encoding any 3D face in a low dimensional space, which gives a compact representation that simplifies the analysis of 3D faces.

2.3.2 Parameterising Face Models

The purpose of parameterising face models is to facilitate the representation of facial expressions using a small number of vertices. Early work was presented by Parke [195] involved using parameters to personalise the face model (conformation parameters) and modify the model to produce sets of emotional or physical facial expressions, such as smiles and blinks (expression parameters). There are some properties which should be taken into account to define facial parameters that are capable of representing changes in facial expressions:

- Complete: parameters should represent all possible facial expressions.
- Independence: parameters should work independently such that the outcome of each parameter does not affect the outcomes of the other parameters
- Minimal: a small number of parameters should be used to represent facial expressions accurately and to be easily interpreted.
- Intuitive: each parameter should be labelled based on its function (e.g. blink or jaw rotation).
- Physically Plausible: all parameters should represent the observable expression, in which unrealistic facial expressions cannot be presented when the parameters are combined.

Different sets of parameters for facial expression have been introduced [69, 138, 141, 145, 151, 159, 195, 251] in an attempt to include these properties. Two standard sets of parameters exist: the Facial Action Coding Scheme [81] and Moving Pictures Experts Group-4 (MPEG-4) [51]. The Facial Action Coding Scheme is a set of parameters that describes the movements of the facial muscles, the tongue and the jaw based on the facial anatomy analysis. Forty-four basic action units (AUs) are included in this set. Complex compound facial expressions can be recreated with a combination of these units. The AUs are independent, but it is not guaranteed that the animation of the face's mesh will also be independent. Although this is potentially problematic, this set was used to model the facial expressions [95, 141].

MPEG-4 facial parameterisation was derived from this set. MPEG-4 is an international standard of ISO/IEC for describing and representing facial motions and speakers' gestures during speech to achieve an adequate animation [233]. Two types of parameters are defined in the standard: facial definition parameters for identifying the face size, shape and texture; and facial animation parameters for defining facial deformation and expressions. MPEG-4 uses a standard parameter, which makes it reliable and widely used by researchers for facial animation [45, 242]. Another method that is based on a statistical technique for parameterisation relies on extracting parameters that have the most variations in the dataset by applying PCA to the data [32, 60, 108, 207].

2.3.3 Modelling Techniques

Facial animation modelling techniques can be classified into two categories: geometric techniques and physical techniques. Geometric techniques involve the deformation of a facial surface by manipulating a geometric control structure. The most popular approach is interpolation, also known as morph targets or blend shapes, in which the animation is achieved by applying an interpolation function to a set of face poses (key-frames) created by the animator or software to generate the middle face pose or key-frame over time [194]. These values can be controlled using parametric curves, such as Bezier curves [197]. However, this technique fails to generate more complex facial expressions, which leads to inadequate and insufficient animation. To remedy this, Joshi et al. [137] localised the morph targets, for example, to generate an eye blink pose so that only the region around the eye would be manipulated.

The physical techniques involve modelling the function and structure of the face. These techniques accurately provide details of the facial movements, but they are extremely difficult to implement due to the complex structure of the facial anatomy, which contains muscle, skin, fatty tissue and bone. Facial expressions are produced by contractions of the muscles, which create stretching, wrinkles and creases. Hence, the elastic nature of the face must be considered when modelling facial expressions. These techniques can be divided into two areas: tension networks that involve treating the face mesh as a network of masses and springs [159, 204]; and finite element models that involve modelling the skin's elastic properties [54, 151, 152].

2.4 Visual Speech Animation

Over the last decades, facial animation and visual speech synthesis have received a lot of attention by researchers [175], because of the development of necessary computing power to achieve appropriate mouth animations. Combining a high quality of a synthetic visual speech signal with a synthetic or original auditory speech signal improves the intelligibility of speech in noise [5, 193], especially for hard of hearing people and foreign language learners. In fact, hard of hearing people rely on audiovisual speech to enhance speech perception, because it provides additional visual information since they perceive audio signals in a distorted way. Foreign language learners find difficulty in perceiving or producing new phonemes that are not in their native language. Therefore, achieving accurate visualisation of the speech articulators can help to improve their pronunciation which leads to a better perception of their acoustic signals. In addition to that visual speech animation plays a vital role in many different applications of human interaction such as animated story narration and virtual avatar.

Prediction of appropriate articulations can be achieved by providing synthesis engines with audiovisual speech signals such as set of synthesis rules (e.g classifying phonemes into visemes) or a collection of original articulators' movements. Recently, researchers have focused attention on applying deep learning approaches to generate speech animation. Generally, based on that the audiovisual speech synthesis approaches can be classified into three main categories which are viseme driven approaches, data driven approaches and deep learning based approaches. The next three sections review these approaches in more detail.

2.4.1 Viseme Driven Approaches

Viseme driven approaches involve segmenting audio speech signals into phonemes, which are then classified into visual units called visemes (see Section 2.1.4). These approaches are often based on classifying many phonemes into one viseme, in which phonemes that have the same visual appearance are mapped to the same viseme (see Section 2.1.4). Recently, Taylor et al [236] proposed a method based on a deep learning to automatically classify many phonemes into many visemes. Viseme parameters are then interpolated using dominance functions [57, 171] or with co-articulation rules incorporated [57, 79, 200].

The previous approaches are based on determining the weight of the target phoneme against the neighboring segments and their influence on the corresponding control parameters (Cohen and Massaro model) [57]. However, this model can not deliver realistic animation results especially for labiodental and bilabial consonant. To tackle these issues, Cosi et.al [63] modified this model by adding a shape function and a temporal resistance function to model more speech features such as speech rate. Although this model enhanced animating the speech articulators for labiodental and bilabial consonants, they still behave unrealistically. Moreover, this model requires a manual control for the optimization process to achieve the method convergence. Goff and Benoit [157] analysed trajectories of 8 parameters measured from a French talker to calculate the model parameters. Nevertheless, an analysis of a larger corpus of lip shapes is required to refine the dominance functions. King and Parent [146] modified the dominance functions by representing visemes as dynamic shapes (curves) instead of key-frames. However, their model fails to consider coarticulation effects on the tongue and teeth over time.

The rule-based models [21, 199, 200] take into account only visemes that have an impact on the neighbouring ones (backward and forward coarticulations (see Section 2.1.5)). Xu et.al [259] produced speech animation by blending curves of each pair of diphone (adjoin two phones in an utterance) coarticulation. Recently, Edwards et.al [79] categorised visual speech production rules into constraints, conventions and habits. Each category describes a set of linguistic rules that are followed to achieve the final animation correctly. Charalambous et.al [52] animated their model by varying a viseme weight using dynamic linguistic rules and emotional features extracted from audio signals. However, these models typically fail to fully take co-articulation effects into account, thus leading to unrealistic lip motions. For this reason, data-driven approaches are more favored as they are based on animating faces according to captured data from real speakers, which guarantee considering coarticulation effects. The next section covers data-driven approaches for visual speech animation.

2.4.2 Data Driven Approaches

Data driven approaches involve capturing motion data from a real speaker to produce a synthesized talking head [16, 82, 165, 190] or to reanimate faces in images and videos [30, 70, 258, 263]. The original facial features can be captured using marker based techniques or markerless techniques. Marker based techniques involve capturing the original facial features by tracking the markers that are placed on the talker's face. These markers are either placed using colours [82] or fluorescent markers [110]. Markerless techniques involve tracking the facial features in 3D using motion capture systems such as the VICON system [3], in 2D frames by using image processing techniques such as snakes [237], or using facial modelling techniques such as active appearance model [60]. Recent works have established high quality of facial features' tracking results [247], and have improved to commercial softwares such as Faceware Technologies [2].

The captured data is either organised based on phonetic information (sample-based approaches) [39, 43, 61, 107, 143, 155, 236], or processed using statistical models to control the facial motion that is learned from the training data (learning-based approaches) [101, 142, 202, 260, 261]. Data driven approaches allow to estimate visual speech motion occurring in actual speech, since they are based on processing captured motion data from real speakers. However, in the sample based approaches, the quality of the animated visual speech is based on organising the phonetic information units correctly, where incorrect usage of a single unit in a sentence significantly affects the perceived quality of the entire sequence. Bregler et al [39] presented a method for constructing a new sequence of visemes by concatenating triphones (a sequence of three successive phones) from a given video sequence. This method does not involve dynamic factors in speech, because it models the effect of coarticulation with the segmented triphone from video, instead of using ad hoc co-articulation models. However, this approach cannot be considered a generative approach, because faces need to be trained before applying the coarticulation. Cao at al [43] extended the combination of the triphones approach to longer phoneme segments. They proposed a greedy graph search algorithm that examines a set of continuous motion segments which match the phonemes in the dataset. Kshirsagar and Thalmann [155] proposed an approach for synthesising a speech motion using visyllables segments instead of phoneme segments. Taylor at al. [236] presented a method for generating a dynamic continues visual speech animation by connecting parameters of Active Appearance Model(AAM) based on a given phonetic input sequence. Learning-based approaches are based on training data that is collected from real speakers. Such data is difficult to predict how much is required and sufficient to generate the desired results. The lack of this information has provided the motivation to investigate the impact of using different amounts of data during training on the resulting animation, as will be presented in Chapter 4.

2.4.3 Deep Learning Based Approaches

A recent line of visual speech animation research has focused on deep learning approaches because of their efficiency in learning representations progressively from raw features, which can improve the accuracy level dramatically over using hand crafted features [154].

Previously, researchers tend to use Hidden Markov Models (HMMs) rather than neural networks due to their insufficiency for processing speech into intuitive states. However, with the recent development in deep learning, neural networks have become more popular in modern models. The broad use of these methods is mostly aimed to perform a sequence to sequence prediction [229]. A typical example of this is Fan et.al's model [88], which uses bidirectional long short-term memory (BLSTM) [120] (an architecture of recurrent neural network (RNN)) to model audiovisual stereo data for animating a realistic 3D head. They used the deep BLSTM neural network for learning a regression model by minimising the sum of square error between a phoneme sequence extracted from audio signals and a shape features sequence extracted from the lower part of the face images. New faces of the talking head were then rendered by predicting the shape features from the text of any input audio signals (natural or synthesized speech). Following similar deep architecture, Suwajanakorn et.al [231] synthesized photorealistic videos. Inspired by these approaches, Zhou et.al [266] proposed a deep neural network architecture based on three stages of LSTMs: the first stage predicts phonemes' sequence from audio, the second stage predicts landmarks of the lower part of the face from video frames, and the last stage uses the outputs of the previous stages to predict speech motion curves and JALI parameters to drive a JALI or a face-rig. However, approaches of this kind carry with them various well-known limitations, such as they are subject dependent and require a large amount of data for training to be adapted to new faces.

Taylor et.al [235] followed the same deep learning approach but using a sliding window predictor which allows to include coarticulation effects and context neighborhoods, where their results showed how the sliding window architecture significantly outperformed the LSTM for generating realistic visual speech animation. However, this approach does not provide fully automatic speech animation, since the animation technique is based on feeding the system with visual speech signals extracted from a character for the learning process and then retargeting to a face rig to predict the final animation, which can lead to a potential imperfect animation. Consequently, the initial computational step needs to be adjusted for each character. A possible solution is to train the model using videos of multiple speakers who have different facial characteristics and then select the most matched characteristics to the face rig at the prediction stage. Again, a comprehensive dataset is required to train the model which is highly cost and time consuming. This issue may refer to the importance of investigating the impact of differences in facial features between real faces and 3D faces on the resulting animation, which has provided the motivation to investigate this at a detailed level, as will be presented in Chapter 5.

Convolutional neural networks (CNN) have been used by Karras et.al [142] to animate a 3D face mesh by audio signals only. They learned a deep neural network system to map audio waveforms to vertices of a 3D face mesh model. Their model is based on audio signals without a transcript, where the model was divided into sub networks that model two acoustic features which are formants (resonance frequencies of the vocal tract) and an excitation signal (characteristics of a talker's sound such as pitch and timbre). One major drawback of this approach is that the proposed model performs sufficiently as long as the input audio signal is within the range of the training dataset, and also the resulting animation is well synced with the audio as long as the tempo of the input audio signal is not too fast. Chung et.al [55] proposed a model based on CNN that uses features of audio signals extracted using Mel-frequency cepstral coefficients (MFCCs) and a still frame image to generate speaker independent videos. The resulting video frames are blurry due to using an L_1 loss at the pixel level, which makes the deblurring stage is required. Recently, Liu et.al [163] proposed a framework that uses CNN architecture to perform a sequence to sequence prediction that maps audio and text to facial features extracted from pixels and landmarks for synthesizing talking faces. However, this framework is based on an offline speech recognition software for processing audio and text, which impedes applying end to end training. Also, different postures of faces can not be converged well in the encoder that was used to extract the facial features, which makes this approach restricted with standard faces only.

Although, deep learning approaches have revolutionised visual speech animation, synthesizing talking faces with lip sync is still a challenging task due to several issues. First, the efficiency of deep learning approaches is based heavily on the training data, where using more data enhances the results; however, feature extraction of audiovisual data for training with a low level of noise is still challenging. Second, the complex structure of the human face [203] makes representing a face sufficiently using current linear based approaches hard [235]. Finally, synchronising lip movements with audio signals is a demand in visual speech animation [264], hence, exploring a deep learning approach that takes the lip sync and coarticulation effects into account is still an open issue.

2.5 Evaluation Approaches for Visual Speech Animation Quality

The evaluation of visual speech can be categorised into two approaches: objective approaches involve algorithmic metrics, and subjective approaches involve human participants. The quality of a synthesised speech model can be evaluated objectively by comparing the trajectories of the motion captured from an animated face against the trajectories of the motion captured from a real face using different methods, such as a dynamic time warping [215], a root mean square error (RMSE) [252], or the peak signal to noise ratio (PNSR) [143] to measure the similarity. The subjective evaluation is the most frequently used method to judge the quality of synthetic visual speech signals using different approaches, which include a subjective assessment [39, 62, 87], comparing between synthetic signal and ground truth signal visually [58, 239], comparing between different versions of synthetic signals visually [236, 238], testing the user's ability for perceiving an uttered sentence in a noisy environment with the aid of the synthetic signal (intelligibility test) [62, 191], and choice testing, where users are asked to determine whether a presented animation is real or synthetic [24, 86]. Each of these tests evaluates how the resulting lip movements are synchronised to speech, in which good results indicate a high-quality synthetic signal and poor results indicate a weak synthetic signal.

Generally, the subjective evaluation can be categorised into two tests. The first test is the intelligibility test, which examines the quality of lip synchronisation of a talking head by presenting videos of the synthetic signal versus the ground truth signal in a noisy environment. The second test is the naturalness test, which examines the smoothness of the animated lip movement compared to a human lip motion and the likelihood that those sounds would be produced by asking the participants to rate the quality of the lip motions or to compare between different synthetic signals or between the ground truth signal and the synthetic one. The first method provides an overall impression of the quality of the 3D talking head lip-sync, but there remains a lack of clarity concerning which viseme or speech segments are synchronised accurately and which are synchronised poorly. The first procedure of the second method provides user information about the overall quality of the 3D talking head lip-sync but leaves no space for comparison with other systems. It also does not provide any information about the strengths and weaknesses of the synthetic signal (i.e. which viseme is poorly synthesised). The second procedure provides more information about the weaknesses and strengths and a more quantitative measure of the overall effectiveness of the synthesised signal. However, it does not provide enough information about the quality of synthetic signal perception [66].

2.6 Summary

Visual speech animation can be achieved via an extensive understanding of speech production and perception, and facial animation and speech synthesis techniques. Understanding the mechanism of producing the speech by a speaker and receiving the uttered signal by a listener provides a good background for modeling and controlling vocal tract articulators, in addition to evaluating the quality of such models. Knowledge of facial animation techniques helps to understand the basic principles of building speech synthesis systems, where this includes reconstruction of 3D face models, parameterisation of the models, and modelling techniques.

The 3D face model reconstruction is the first stage for animating faces that depends on using cameras, 2D image deformation, or 3D face scans. Reconstructing a 3D face model using laser scan devices has proven to be effective in providing sufficient 3DMMs from a low-dimensional face subspace presented to detect facial expressions or lip motions, and reanimate faces in 2D video frames, but they need time to be gathered and processed. Instead, in Chapter 3 of this thesis, 3D head poses that are generated using commercial software, are used to create a 3DMM for detecting lip motions of a real speaker in 2D videos. Parameterising 3D face models is a mediation stage between the construction stage and the modelling stage. This stage is essential especially for visual speech animation due to coarticulation effects that influence the speech articulators in different ways (i.e. width and height of the lips). Such an issue provides the motivation to investigate the functionality of each set of facial features landmarks in the animation process. This investigation is presented in Chapter 3 as well.

Realism in visual speech animation is still a challenging task because people can detect any slight defect in audiovisual synchronisation. Several approaches of visual speech animation have been proposed which are based on phonetic information (viseme driven approaches), or motion data captured from real speakers (data driven approaches). Data-driven approaches are favored by researchers due to providing speech animation according to the lip motions of a real speaker, including coarticulation effects. However, the quality of the resulting animation depends on the amount of training dataset which is difficult to predict how much is required. A recent line of research promotes the use of deep learning approaches in speech animation, as evidence has been found for improved visual speech animation when audio and visual signals were introduced in the training dataset. Again, these approaches depend heavily on the amount of data for the training process which is difficult to extract and processed. Furthermore, for subject independent approaches, a potential imperfect animation may be generated due to mismatch between facial features of the real speaker and a face rig. For these reasons, an investigation of using different amounts of data during constructing a neutral 3D head pose and training the 3DMM on the resulting animation is introduced in Chapter 4. Followed by another investigation that studies the impact of differences in facial features between real faces and 3D faces on the final animation (Chapter 5).

Chapter 3

Mapping Process

3.1 Introduction

This chapter presents the mapping process between 2D video frames of a real speaker and the corresponding 3D head through a 3D morphable model. The aim of this mapping is to animate the lips of 3D talking head according to lip motion of a real speaker in 2D video frames. In this chapter, the stages of this process are presented. The first stage is constructing the 3D morphable model which includes collecting 3D head data and morphing between these 3D heads (see the red dotted box in Figure 3.1). The presented 3DMM based on synthetic 3D head poses that was generated using a commercial software, which establishes vertex correspondences among the generated head models. This functionality simplifies processing these models, therefore generating a large number of data becomes customisable and controllable. This stage will be presented in Section 3.2.

As a preparation for the second stage, a review of accuracy level for a range of facial features tracking systems is presented in Section 3.3. Then the second stage which is mapping 2D video frames of a real speaker to the corresponding 3D head model is presented in Section 3.4. This includes tracking the facial landmarks in 2D videos, estimating the head pose and fitting the head pose shape (see the green dotted box in

Figure 3.1). Furthermore, an overview of the 2D dataset that will be used for creating, animating and evaluating the 3D talking head will be presented in Section 3.5.

Finally, a study that investigates the functionality of each set of the facial features landmarks in the mapping process is presented in Section 3.6. The aim of this study is to determine the required facial features set that will be used to achieve the desired 3D lip motions. Then the resulting 3D lip motions will be evaluated from both front-view and side-view. These stages are presented in Chapter 4. Another study that involves investigating the impact of mapping between non-similar faces on the resulting 3D animation is presented in Chapter 5.

3.2 Constructing the 3D Morphable Model

A 3D Morphable head Model (3DMM) is a statistical model of head shape, built from a set of synthetic 3D head poses. Principal Component Analysis (PCA) [136, 255] is only applied to the shape data to construct a model that spans a subspace of head pose shape learned from synthetic 3D head poses. Morphing between these head poses in PCA space can be used to transfer head poses from one pose to a different pose, or generate new head poses. This section presents the stages of constructing the model and defines the components of a Morphable Model used throughout this thesis.

3.2.1 Collecting Head Data

A 3D Morphable head Model (3DMM) requires a set of head poses for training. These are often generated by taking scans of real people. Instead, FaceGen software [1] is used to produce synthetic head poses that are derived from hundreds of high-resolution 3D scans of human faces. The produced head models have correspondences of vertices, which makes processing these models uncomplicated. A lack of this functionality is considered one of the major obstacles for analysing 3D scans of real faces. Although many techniques have been considered [212, 216, 265], it is still challenging, in addition to the consumed time for collecting such data. This software is based on a statistical

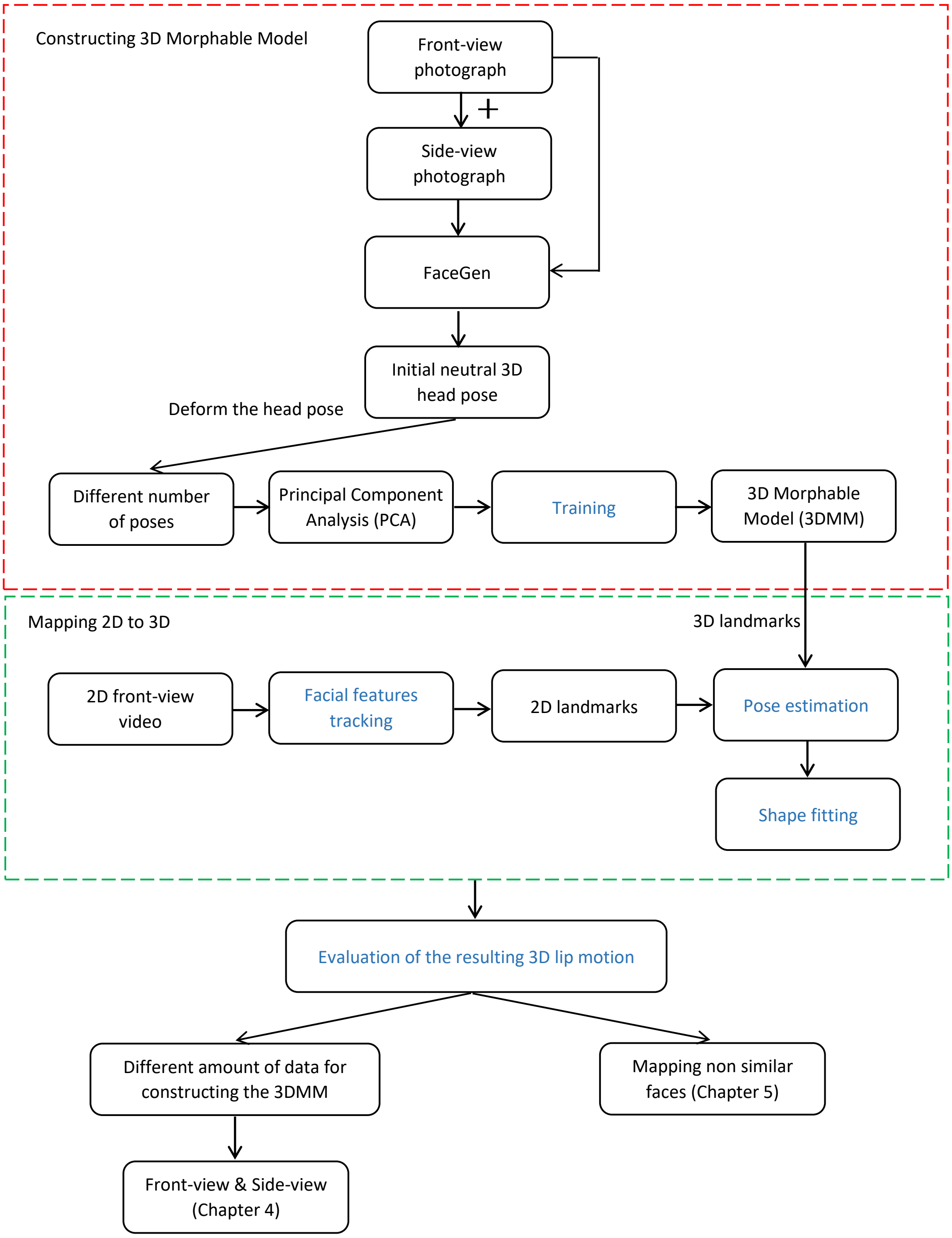

Constructing 3D Morphable Model
Front-view photograph
+
Side-view photograph
FaceGen
Initial neutral 3D head pose
Deform the head pose
Different number of poses
Principal Component Analysis (PCA)
Training
3D Morphable Model (3DMM)
Mapping 2D to 3D
3D landmarks
2D front-view video
Facial features tracking
2D landmarks
Pose estimation
Shape fitting
Evaluation of the resulting 3D lip motion
Different amount of data for constructing the 3DMM
Mapping non similar faces (Chapter 5)
Front-view & Side-view (Chapter 4)

Figure 3.1 Schematic view of the proposed methodology to animate a 3D talking head.

technique called Principal Component Analysis (PCA) [136, 255] (more details in the next section 3.2.2) to convert a subspace of shapes or textures into a set of values of linearly uncorrelated variables (principal components (PCs)) by using an orthogonal transformation. Morphing between 3D head shapes (i.e. longer thinner head and shorter wider head) or textures (i.e. skin tone) can be generated by adding or subtracting PCs to or from the mean. FaceGen software provides a high resolution of 3D head mesh as shown in Figure 3.2, which contains a mouth mesh (inner mouth, tongue and teeth) that can be exported separately. The number of vertices of the 3D head is 5968, 5850 vertices for the head and 118 vertices for the mouth.

FaceGen was chosen based on the number of 3D head poses that contain a wide range of mouth visemes, and also the provided high resolution 3D head mesh. An initial neutral head pose can be generated using photographs of a real person, either a front-view only photograph or front and side views as shown in Figure 3.3. Next some facial points can be placed on the front-view photograph only or front- and side-view photographs. The software can then be used to deform the face into a range of poses as explained in the previous paragraph. The software includes 16 default viseme poses (shown in Figure 3.4), which are parameterised so that different intensities of each viseme can be generated, i.e. different amounts of openness for an open-mouthed viseme. Figure 3.5 shows an example of different intensities of viseme ah. This functionality is used for generating the 3D datasets. Each head pose created using FaceGen automatically has vertex correspondence, something which is more complex to achieve with scanned data. FaceGen also generates tongue and teeth poses, but they are excluded since this thesis is concentrating on lip shape.

3.2.2 3D Morphable Model

Principle Component Analysis (PCA) is a statistical technique that is widely used to identify and express patterns of data to highlight their similarities and differences. Furthermore, it is used to analyse and compress datasets that contain a large number of interdependent variables by transforming them into uncorrelated variables which called principal components (PCs) (eigenvectors of the covariance matrix) that hold the maximal variation in the dataset [136, 255].

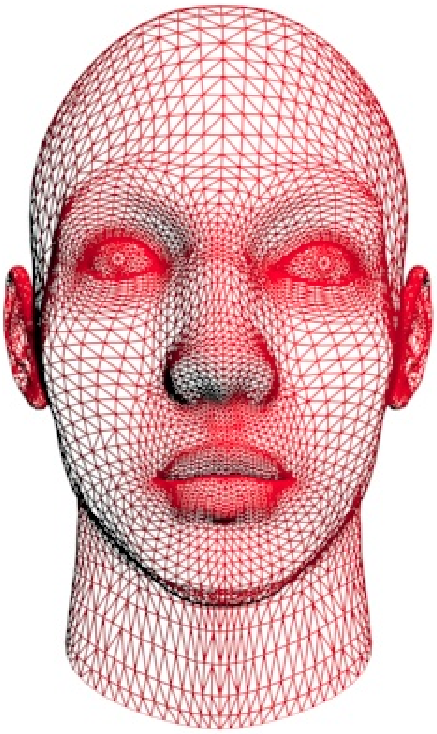
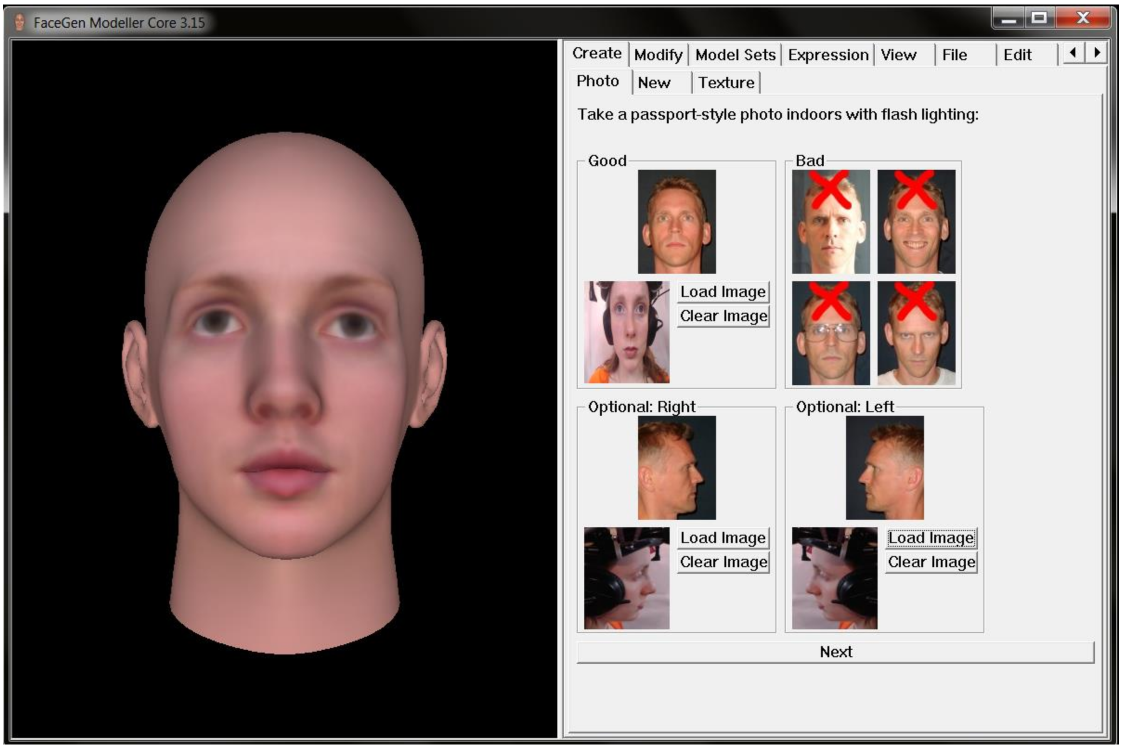


Figure 3.2 Wireframe view of a neutral 3D head pose, generated using FaceGen software.

Given a set of head poses, Principal Component Analysis (PCA) can be applied to the vertices to generate a 3DMM. Only shape needs to be considered, since every head pose shares the same texture. The geometry of the head is represented by a shape vector $S = (X_1, Y_1, Z_1, \ldots, X_n, Y_n, Z_n)^\top$, containing the X, Y, Z coordinates of the vertices (5850 vertices), where n is the number of FaceGen poses used to build the 3DMM. The 3DMM consists of a PCA model of the shape, which is represented as:

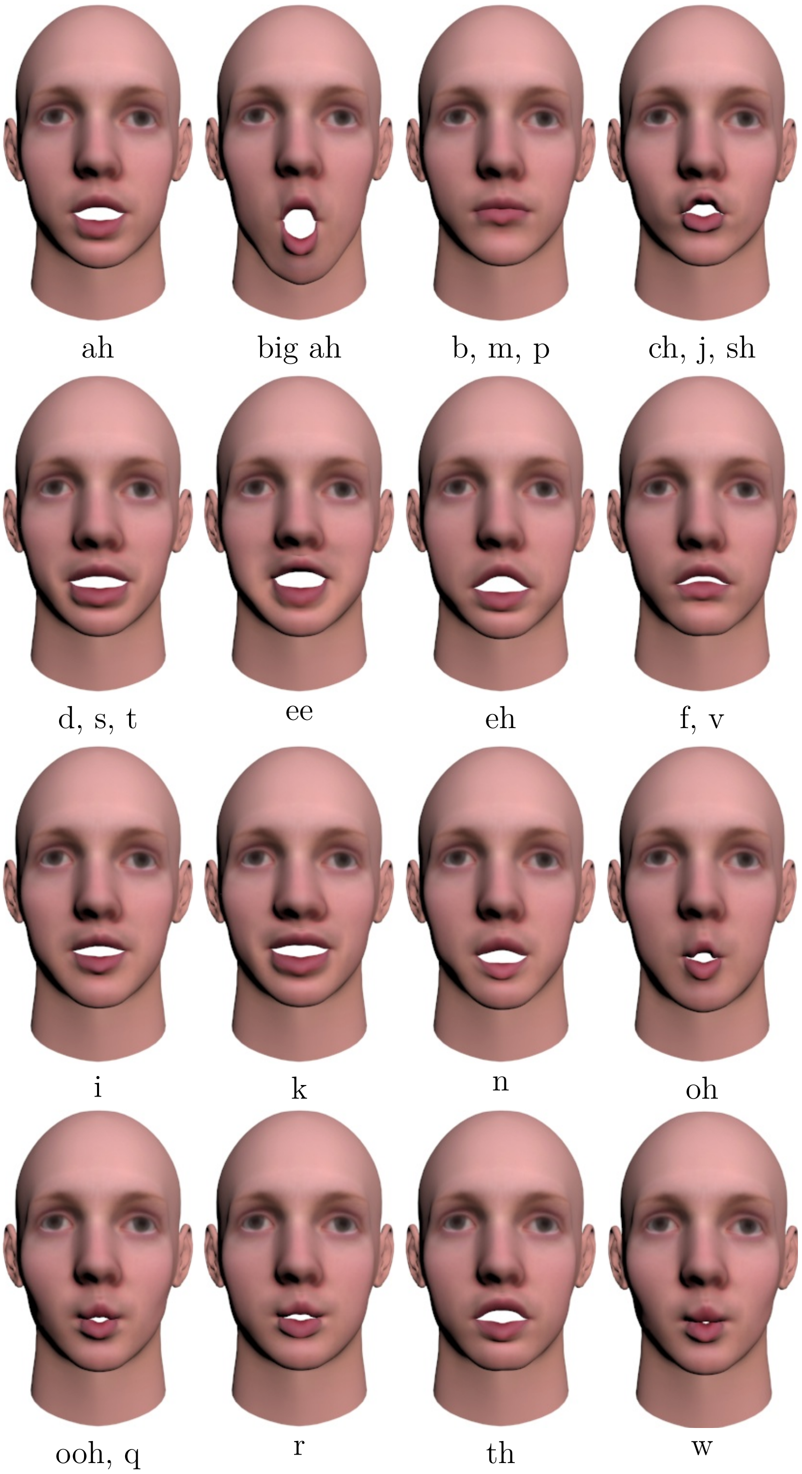
$$M := \{\overline{F}, \sigma, V\} \tag{3.1}$$

where $\overline{F} \in R^{3N}$ is the mean vector of the example meshes (mean pose) with N being the number of mesh vertices, and $\sigma \in R^{n-1}$ denotes the standard deviation, where $V = [v_1, \ldots, v_{n-1}] \in R^{3N \times n-1}$ is a set of principal components in the model.



Take a passport-style photo indoors with flash lighting:

Figure 3.3 Generating initial 3D head pose using FaceGen software.



ah

big ah

b, m, p

ch, j, sh

d, s, t

ee

eh

f, v

i

k

n

oh

ooh, q

r

th

w

Figure 3.4 Viseme poses generated using FaceGen software.

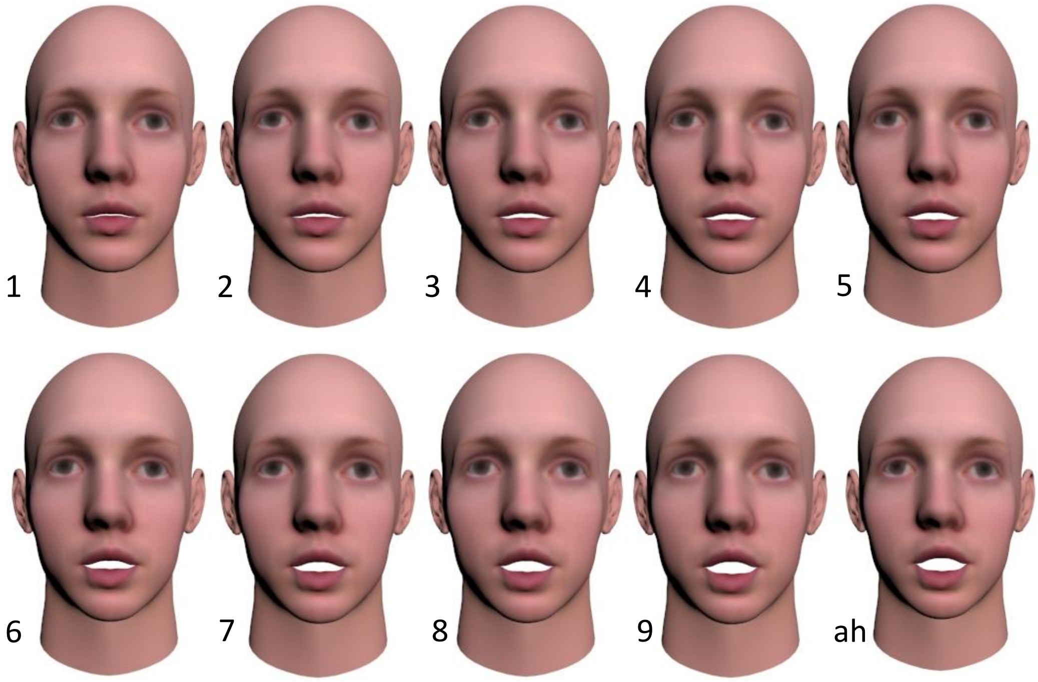


Figure 3.5 An example of different intensities of viseme ah, generated using FaceGen software.

A new pose can be generated as follows:

$$S = \overline{F} + \sum_{i=1}^{K} \alpha_i \sigma_i v_i \tag{3.2}$$

where $K \leq n - 1$ is the number of principal components and $\alpha_i \in R^K$ is the shape coefficient [125].

For analysing the variations in the built head model, the directions of largest variance in the PCA space can be visualised by Equation 3.2, where α_i is set to a specific value and all other parameters are set to zero. The resulting head mesh S can then be rendered. Figure 3.6 shows the mean of the head model along with the top three principal components of shape variation. From this figure, it can be observed that the principal modes of variation capture trends of lip shape deformation due to viseme intensities.

3.3 Facial Features Tracking

The first stage in driving 3D visual speech using 2D information is to track the visual speech information in 2D videos. This can be done using facial landmarks related to speech production. Since this thesis is concerned with animating 3D lips based on 2D information, developing an automatic facial features tracking approach is beyond the scope of this research. Consequently, a number of facial features tracking systems were tested to choose an appropriate tracker. The tested systems were: Faceware Analyser [2, 211], Face Plus Plus [89] and random cascaded-regression copse (R-CR-C) [91]. Table 3.1 and Figure 3.7 show the differences between the tested systems in terms of the number of the facial features landmarks.

In order to validate each tracking system, video frames of a female speaker (ID: S17) and a male speaker (ID: S48) from the Audiovisual Lombard Grid Speech corpus [10] were used. For each real speaker, frames from the front view video files were chosen to be mapped to each corresponding 3D synthetic head. The frames contain different phonemes of different visemes. For example, rounded viseme (phoneme /ao/), open

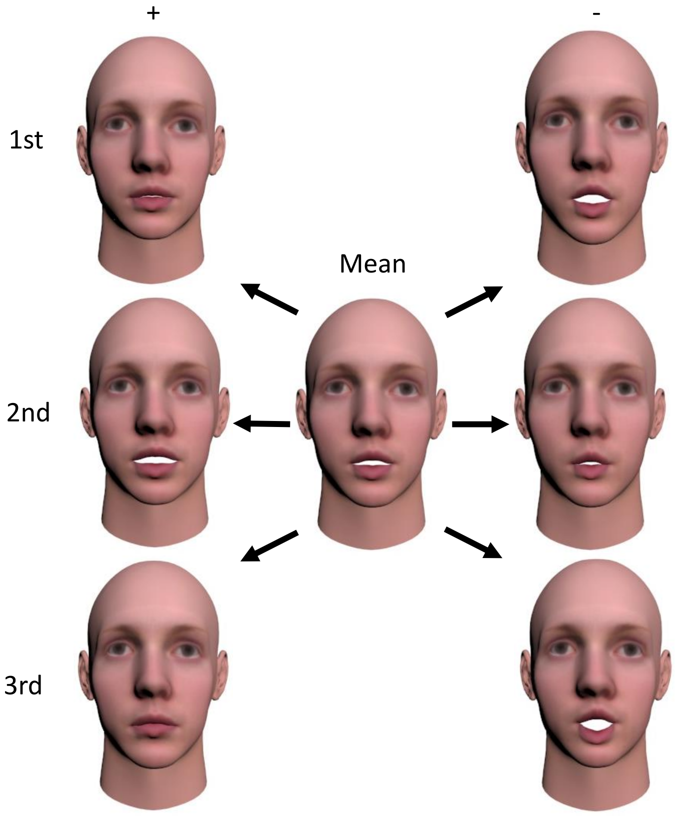
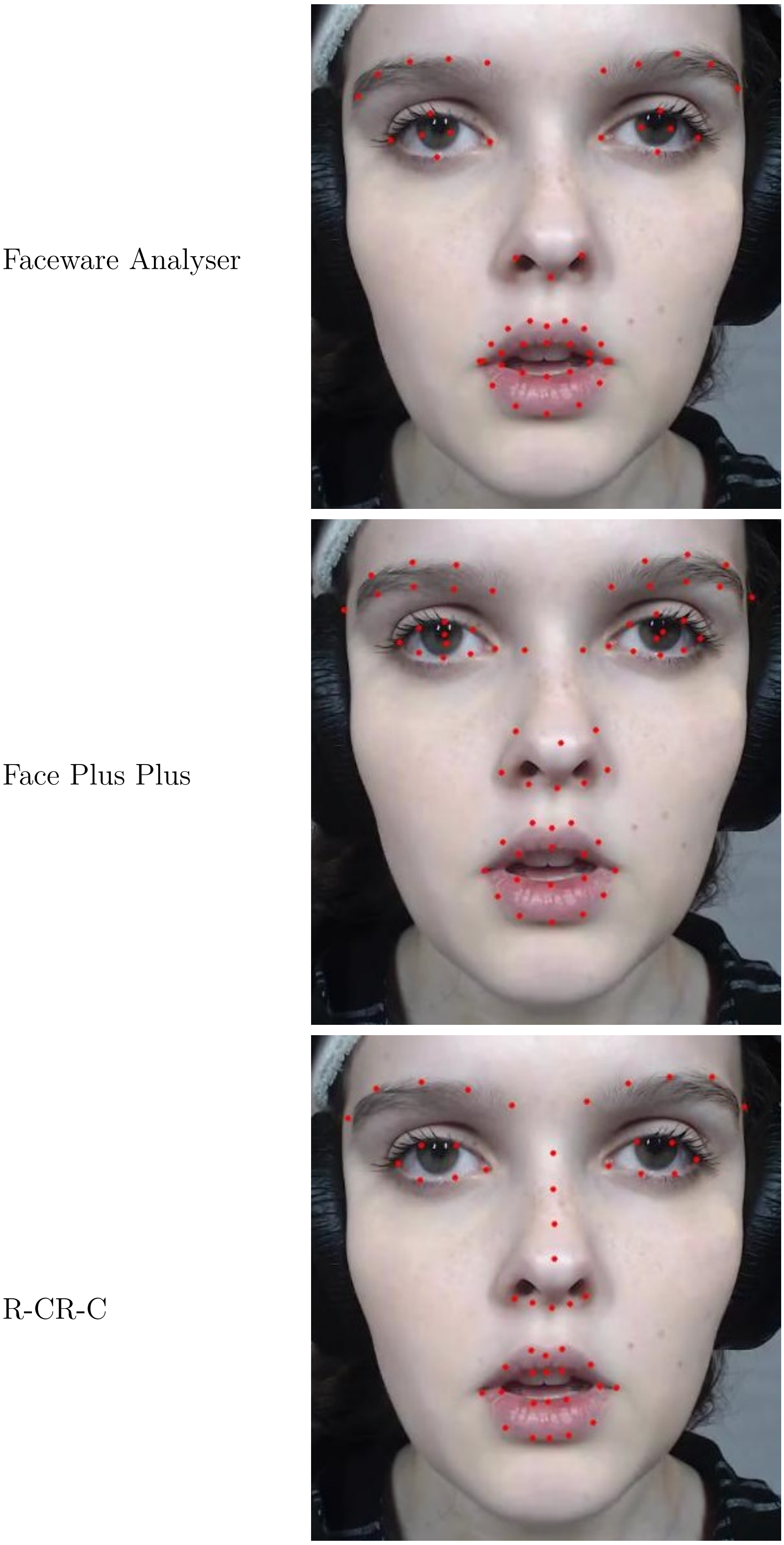


Figure 3.6 Visualisations of the mean head pose and the first three principal components of head shape for the 3DMM that contains 161 head poses. Each visualised as additions and subtractions from the mean.



Faceware Analyser

Face Plus Plus

R-CR-C

Figure 3.7 A comparison of facial features landmarks tracking systems. The facial landmark points labelled on a real speaker for each tracking system.

Tracking system	Landmarks description

Faceware Analyaser	Facial features: 51;
	Mouth: Inner contour: 12; Outer contour: 14;
	Nose: 3;
	Eyes: 12;
	Eyebrows:10

Face Plus Plus	Facial features: 64;
	Mouth: Inner contour: 6; Outer contour: 12;
	Nose: 10;
	Eyes: 20;
	Eyebrows:16

R-CR-C	Facial features: 51;
	Mouth: Inner contour: 8; Outer contour: 12;
	Nose: 9;
	Eyes: 12;
	Eyebrows:10

Table 3.1 Description of landmarks for each tracking system.

spread viseme (phoneme /ay/), bilabial viseme (phoneme /b/), spread viseme (phoneme /ih/), alveolar viseme (phoneme /s/), and protruding-rounded (phoneme /uw/). Each 3DMM was trained using 161 poses, and front- and side-view photographs were used to generate the initial neutral 3D head pose in FaceGen software. The resulting 3D head animation was then compared to the original ground-truth 2D video frames.

For the comparison, the facial features in the ground-truth 2D video frames and the front-view (2D) of the corresponding 3D animation were tracked using Faceware Analyser software [2]. This software provides processing video files as a batch, which speeds the evaluation process. The tracked facial features were used to calculate two geometric articulatory measurements. The first was a width measurement defined by the horizontal distance between the right and left inner corners of the lips. The second was a height measurement defined by the distance between the top and the bottom middle of the inner mouth contour. The Euclidean distance between the midpoint of the inner corners of the eyes and the nose tip's point was used to normalise the landmarks, in order to correct the distance between the camera and the real speaker or the synthetic 3D head. The maximum and minimum mouth measurements of all visual articulatory features for each real speaker and the corresponding 3D head in the video frames were used for normalisation, giving a [0-1] scale. Given the width and height values for each frame of animation, for the 2D video frames of a real speaker and the corresponding 3D animation, the root mean square error (RMSE) over each phoneme frames was used to evaluate the effectiveness of each tracking system.

Figure 3.8 shows an example of consecutive frames of the phoneme/ih/ for a real speaker (ID: S48) and the corresponding 3D head that was fitted to using Faceware Analyser landmarks set (second row), Face Plus Plus landmark set (third row), and R-CR-C landmarks set (forth row). This Figure shows how using Faceware Analyser and Face Plus Plus landmark sets in the mapping process gives poor performance. Due to including the first set, a small number of nose landmarks which lead to stretching the nose mesh area, and the second set a large number of eyebrows, eyes, and nose landmarks which restrict the movement of the upper part of the face.

Table 3.2 shows the RMSE results averaged over each phoneme frames for width and height of the mouth aperture of the real speakers and their corresponding 3D heads for each tracking system. The 3D head models that were fitted to their corresponding real speakers using R-CR-C tracking system give the lowest RMSE scores for width and height for the two real speakers. R-CR-C was selected based on the number of landmarks that encode the mouth and the nose bridge, which are essential for producing adequate 3D lip motions as is presented in Section 3.6, in addition to the level of accuracy that is provided by the system for different face pose angles.

3.4 Mapping 2D to 3D

To generate the 3D animation, 2D video of a speaker needs to be mapped to the 3DMM. An approach to fit the 3DMM to 2D video frames of a real speaker can be applied to perform this function. Mapping 2D to 3D is an active research area, either based on illumination or depth data [74, 100, 101] or tracked markers on real faces [25, 123]. For the algorithms based on illumination or depth data, the appearance and ranges of

Phoneme	Faceware				Face Plus Plus				R-CR-R			
	S17		S48		S17		S48		S17		S48	
	W	H	W	H	W	H	W	H	W	H	W	H
/ao/	0.234	0.095	0.164	0.441	0.248	0.120	0.685	0.341	**0.162**	**0.036**	**0.062**	**0.025**
/ay/	0.529	0.205	0.175	0.051	0.190	0.217	0.041	0.046	**0.048**	**0.082**	**0.039**	**0.028**
/b/	0.153	0.157	0.382	0.104	0.210	0.192	0.305	0.257	**0.098**	**0.138**	**0.149**	**0.025**
/ih/	0.232	0.292	0.367	0.125	0.338	0.032	0.244	0.063	**0.036**	**0.025**	**0.156**	**0.029**
/s/	0.078	0.066	0.190	0.205	0.218	0.081	0.388	0.363	**0.066**	**0.047**	**0.047**	**0.106**
/uw/	0.241	0.212	0.216	0.107	0.264	0.165	0.548	0.350	**0.185**	**0.077**	**0.107**	**0.084**

Table 3.2 The RMS error averaged over frames of different phonemes for width (W) and height (H) of the mouth of the real speakers and their corresponding 3D heads that were mapped to each other using different facial features tracking systems. Values in bold means the decreased RMS error.

lip motions are hard to detect, thus it requires user-specific training or labeling (see [74, 100, 101, 129]). As a result, two different approaches of reconstructing 3D faces from 2D images with the aid of landmarks were tested in order to select a suitable approach that provides a proper detection of lip motion during speech. The first approach is presented by Bas et al [18], which is based on fitting a 3DMM to 2D images using edges and landmarks. The second approach is presented by Huber et al [126], which is based on only landmarks for the fitting process.

In order to validate each approach, the procedure presented in the previous (Section 3.3) was followed. Figure 3.9 shows an example of consecutive frames of the phoneme /ih/ for a real speaker (ID: S17) and the corresponding 3D head that were fitted to using Bas et al.'s method (second row) and Huber et al.'s method (third row). This Figure shows how the 3D head model fails to detect the real speakers' mouth shape, when Bas et al.'s method was used for the mapping process. The 3D lips are slightly opened, due to fit the 3D head to the edges of the real speaker, which extends the 3D mesh, resulting in restricting forming the lips. Figure 3.10 illustrates how the 3D head model gave a poor performance when it was fitted to the real speaker (ID: S48) using Bas et al.'s method. This gives the 3D lips a considerably opened mouth shape during the utterance of the phoneme /uw/.

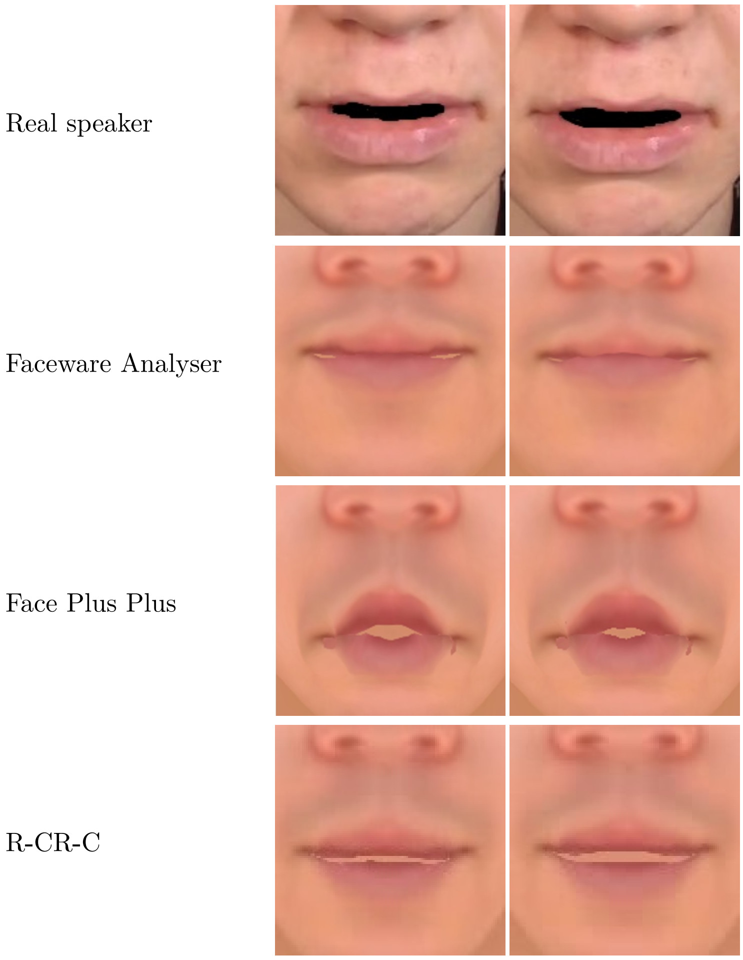


Figure 3.8 Consecutive frames of the phoneme /ih/ during utterance of the word "bin" for a real speaker (ID: S48) (first row), and the corresponding 3D heads that are animated using Faceware landmark set (second row), Face Plus Plus landmark set (third row) and R-CR-R landmark set (forth row).

Table 3.3 shows the RMSE results averaged over each phoneme frames for width and height of the mouth aperture of the real speakers and their corresponding 3D heads for the two tested approaches. The 3D head models that were fitted to their corresponding real speakers using Huber et al.'s method give the lowest RMSE scores for width and height. Therefore, Huber et al.'s method was chosen based on the level of accuracy for lip motion detection during speech provided by this approach. Furthermore, it does not require any aid from the user as in Bas et al.'s method that requires manual adjustments for landmarks and edges weights.

Phoneme	Bas et al.'s Method				Huber et al.'s Method			
	S17		S48		S17		S48	
	W	H	W	H	W	H	W	H
/ao/	0.468	0.600	0.711	0.240	**0.162**	**0.036**	**0.062**	**0.025**
/ay/	0.183	0.244	0.047	0.874	**0.048**	**0.082**	**0.039**	**0.028**
/b/	0.417	0.203	0.410	0.276	**0.098**	**0.138**	**0.149**	**0.025**
/ih/	0.201	0.338	0.500	0.626	**0.036**	**0.025**	**0.156**	**0.029**
/s/	0.332	0.301	0.484	0.494	**0.066**	**0.047**	**0.047**	**0.106**
/uw/	0.188	0.541	0.622	0.194	**0.185**	**0.077**	**0.107**	**0.084**

Table 3.3 The RMS error averaged over frames of different phonemes for width (W) and height (H) of the mouth of the real speakers and their corresponding 3D heads that were mapped to each other using two different approaches. Values in bold means the decreased RMS error.

To generate the 3D animation, the facial features of a real speaker in a front-view 2D video frame need to be tracked, and the corresponding 3D landmarks need to labelled on the corresponding 3D head. Given the tracked 2D landmarks and the corresponding 3D landmarks, a pose of the face is estimated and fitted to the real speaker's mouth shape. The following sections explain the 2D facial landmarks tracking process and how the pose of the 3DMM is estimated and fitted to the mouth shape of a real speaker using the camera matrix method presented by Huber et al [126].

3.4.1 2D Facial Landmarks Tracking

In order to track the facial features of a real speaker in a video, the random cascaded-regression copse (R-CR-C) approach presented by Feng et al [91] is used, which regresses a set of extracted facial feature landmarks from the input image $f(I,\theta)$ to fit a predictive shape model $\delta\theta$ to the true shape. Given a set of labelled 2D images as a training set $T = \{I_1, \ldots, I_N\}$, random sub-sampling is applied on T to generate W subsets $W = \{T_1, \ldots, T_W\}$. Then a single CR thread is trained using each subset defining a copse as follows:

$$U = \{R_1, R_2, \ldots, R_W\}, \tag{3.3}$$

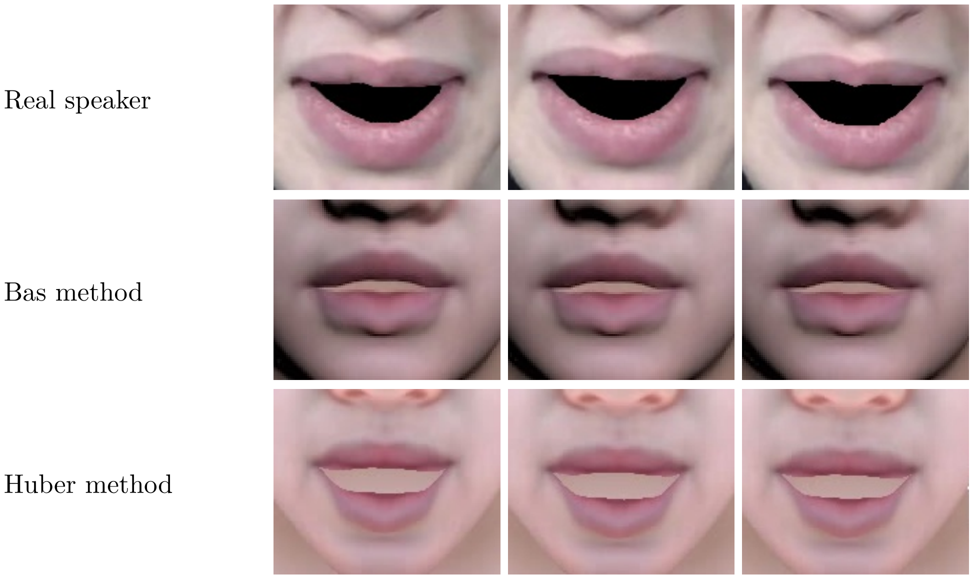


Figure 3.9 Consecutive frames of the phoneme /ih/ during utterance of the word "bin" for a real speaker (ID: S17) (first row), and the corresponding 3D heads that are animated using Bas et al's method (second row) and Huber et al's method (third row).

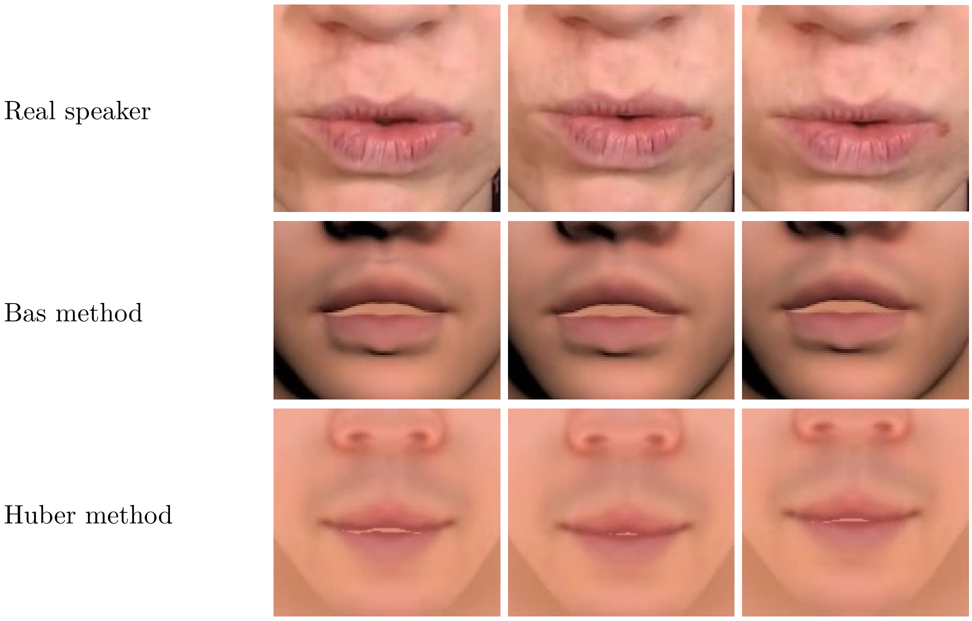


Figure 3.10 Consecutive frames of the phoneme /uw/ during utterance of the word "soon" for a real speaker (ID: S48) (first row), and the corresponding 3D heads that are animated using Bas et al's method (second row) and Huber et al's method (third row).

each CR thread R is formed by D weak regressors for each subset W as follows:

$$R_w = \{r_{w,1}, r_{w,2}, \ldots, r_{w,D}\} \tag{3.4}$$

where $r_{w,d} = \{A_{w,d}, b_{w,d}\}$ $(d = 1, \ldots, D)$, A_d presents the projection matrix and b_d presents the offset of the dth regressor. A series of linear regressors R_d are used to learn this mapping from a training dataset, where

$$R_d : \delta\theta = A_d f(I, \theta) + b_d \tag{3.5}$$

Based on that, when a video is run, a learned landmark detection model using the Ibug-Helen test set [214] will detect and track the facial features of the real speaker.

3.4.2 Pose Estimation

Given 51 2D landmarks and the corresponding 3D landmarks a pose of the face is estimated using the Gold Standard Algorithm [112]. It computes a least squares approximation of the camera matrix that is used to reconstruct the 3D shape from given 2D-3D point pairs [125]. Figure 3.11 shows the facial landmarks labelled on a video frame of a real speaker (left) and on the corresponding 3D head model (right) that correspond to a set of Ibug[1] facial landmarks (the contour landmarks were excluded). Firstly, the labelled 2D landmarks in the video frame $x_i \in R^3$ and the corresponding 3D head model landmarks $X_i \in R^4$ are presented in homogeneous coordinates, then they are normalised using similarity transforms that transform the centroid of the 2D and 3D landmarks to the origin, making the the Root Mean Square distance from their origin $\sqrt{2}$ for the 2D landmarks and $\sqrt{3}$ for the 3D landmarks as presented in the following:

$$\tilde{x_i} = Tx_i, T \in R^{3\times3} \text{ and } \tilde{X_i} = UX_i, U \in R^{4\times4}$$

[1] https://ibug.doc.ic.ac.uk/resources/facial-point-annotations

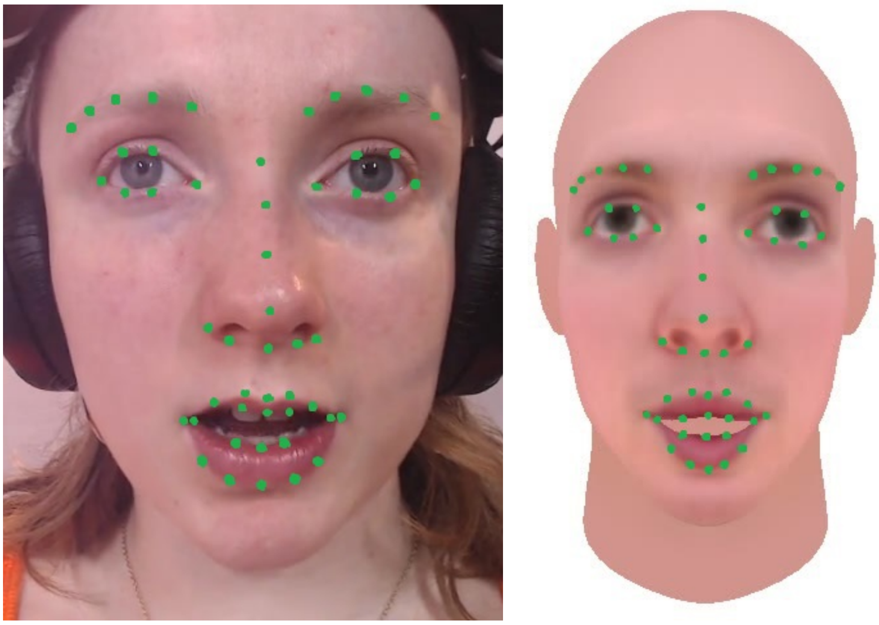


Figure 3.11 The facial landmark points labelled on a real speaker (left) and the corresponding 3D head model (right).

Using the Gold Standard Algorithm [112], a normalised camera matrix can be computed $\tilde{C} \in R^{3\times 4}$ as follows:

$$\begin{bmatrix} \tilde{X}_i^T & 0^T \\ 0^T & \tilde{X}_i^T \end{bmatrix} \begin{bmatrix} \tilde{C}_1^T \\ \tilde{C}_2^T \end{bmatrix} = \begin{bmatrix} \tilde{x}_i \\ \tilde{y}_i \end{bmatrix} \tag{3.6}$$

where $\tilde{C}_1^T$ and $\tilde{C}_2^T$ are the first and the second rows of $\tilde{C}$ while the third row is $[0, 0, 0, 1]$, and $\tilde{x}_i$ and $\tilde{y}_i$ present the coordinates of $\tilde{x}_i$. Then the final camera matrix that minimises$\sum_i \|x_i - CX_i\|^2$ is denormalised as follows:

$$C = T^{-1}\tilde{C}U \tag{3.7}$$

3.4.3 Shape Fitting

The most likely vector of PCA shape coefficients, α, is found by minimising the following cost function:

$$E = \sum_{i=1}^{3L} \frac{(y_{3D,i} - y_{2D,i})^2}{2\sigma_{2D}^2} + \|\alpha\|_2^2 \tag{3.8}$$

where L is the number of landmarks, $y_{2D,i}$ is the 2D landmarks represented in homogeneous coordinates, σ_{2D}^2 is an ad hoc variance of these landmarks, and $y_{3D,i}$ is the projected 3D landmarks to a 2D plane using the camera matrix [125]. In more detail, $y_{3D,i} = P_i(\hat{V}_h\alpha+\overline{v})$, where P_i is the i-th row of matrix P which includes copies of the matrix C on its diagonal, and $\hat{V}_h$ is a matrix of a modified PCA basis that contains the rows that correspond to the landmarks that the shape is mapped to. In addition to inserting a row of zeros after each third row of the matrix V, and then multiplying the basis vectors by the square root of their corresponding eigenvalue. With this, the cost function presented in Equation 3.8 can be represented as a standard linear least squares form. Using such an uncomplicated linear system makes iteration of the pose estimation, and the shape fitting stages for refining the estimates run fast. For refining the face pose, the pose estimate can employ the shape estimate instead of using the mean face. The shape estimate can employ the refined camera matrix to enhance fitting the shape.

3.5 A 2D Dataset for Creating, Animating and Evaluating the 3DMM

In order to personalise the 3DMM, either front-view photograph or front- and side-view photographs of a real speaker are required, which can be supplied to FaceGen software for creating the initial neutral 3D head pose (see the red dotted box in Figure 3.12). Furthermore, 2D front-view videos of a real speaker are required to animate the 3DMM (see the green dotted box in Figure 3.12). Besides, the resulting 3D lip motions should be evaluated; front- and side-view 2D videos of the corresponding real speaker can be used to compare against (see the brown dotted box in Figure 3.12). The width and height of the animated 3D lips can be assessed using the front-view videos (see the grey dotted box in Figure 3.12), while the side-view videos can be used to assess the 3D lip protrusion (see the yellow dotted box in Figure 3.12). Thus, a corpus that contains both front- and side-view 2D videos that include rich phonetic features of real speakers is required.

The Audiovisual Lombard Grid Speech corpus [10] will be used in this thesis. The corpus is provided by the University of Sheffield to support joint computational behavioural studies in speech perception. The form of the corpus was based on the Grid corpus [59], which is inspired by the coordinate response measure (CRM) corpus [183]. The sentence of the CRM corpus is in the form: "READY" <call sign> GO TO <colour> <digit>. The CRM corpus was formed by eight call signs, four colours and eight digits, resulting in 2048 sentences that were uttered by eight speakers. The Grid corpus altered the CRM sentence structure with richer phonetic features, by including four commands, four colours, four preposition, 25 letters, ten digits and four adverbs. Table 3.4 presents the structure of the GRID corpus sentences. Each sentence contains a 6 word sequence and is formed as follows: "<command> <color> <preposition> <letter> <digit> <adverb>". An example sentence is: "bin blue at A 3 please". Consequently, this corpus contains greater variety and richer high-level semantic details due to including the filler words (command, preposition, and adverb), which are not static.

The corpus consists of both front- and side-view videos of 54 speakers (30 female and 24 male) uttering sentences from the GRID corpus [59]. Each speaker utters 100 sentences, 50 in a Lombard condition and 50 in plain conditions. Figure 3.13 shows the helmet that was used to record the front- and side-view videos for each real speaker. Audio files are sampled at 48 kHz. Front video files are sampled at 24 fps (frames per second) with the frame sizes of 720 × 480, while side video files are sampled at 30 fps with a frame size of 864 × 480. Figure 3.14 shows example frames of recorded videos from front- and side-view cameras. Only front-view videos of plain sentences are used for mapping between the 3DMM and the corresponding real speaker, while both front- and side-view videos are used for evaluating performance of the 3DMM.

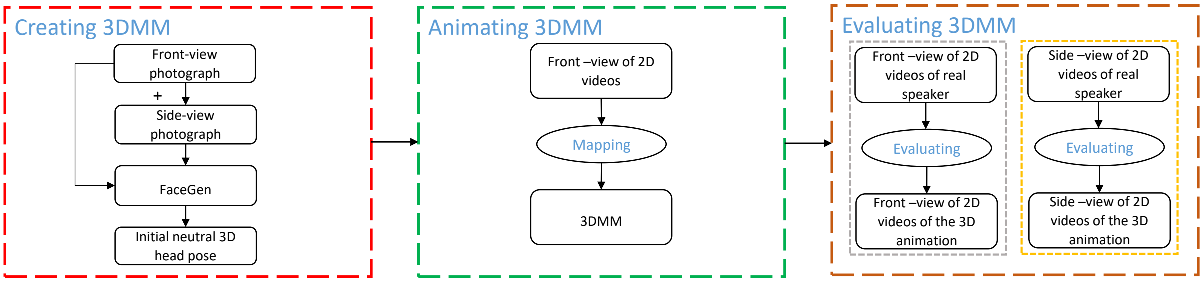

Creating 3DMM
Front-view photograph
+
Side-view photograph
FaceGen
Initial neutral 3D head pose
Animating 3DMM
Front –view of 2D videos
Mapping
3DMM
Evaluating 3DMM
Front –view of 2D videos of real speaker
Evaluating
Front –view of 2D videos of the 3D animation
Side –view of 2D videos of real speaker
Evaluating
Side –view of 2D videos of the 3D animation

Figure 3.12 Flow diagram of using 2D dataset for creating, animating and evaluating the 3DMM.

command	color	preposition	letter	digit	adverb
bin	blue	at	A-Z	1-9,	again
lay	green	by	excluding	zero	now
place	red	in	W		please
set	white	with			soon

Table 3.4 The structure of the GRID sentences [59].

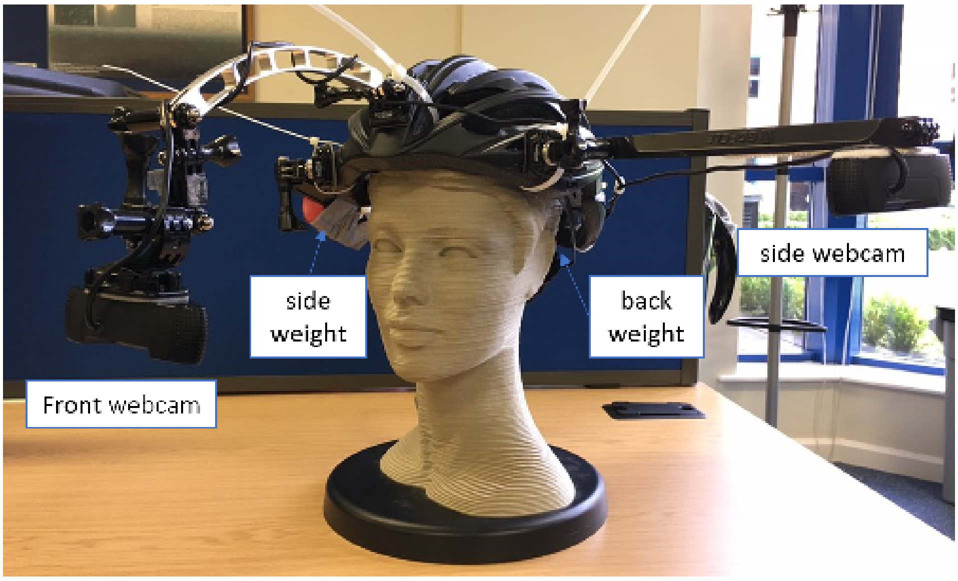


Figure 3.13 The used helmet for recording the front- and side-view videos of real speakers.

A pool of 27 speakers (12 male and 15 female) of the Audiovisual Lombard Grid Speech corpus was selected to validate the performance of the 3D head models. This pool contains videos of real speakers who their faces are not obscured by glasses or facial hair,

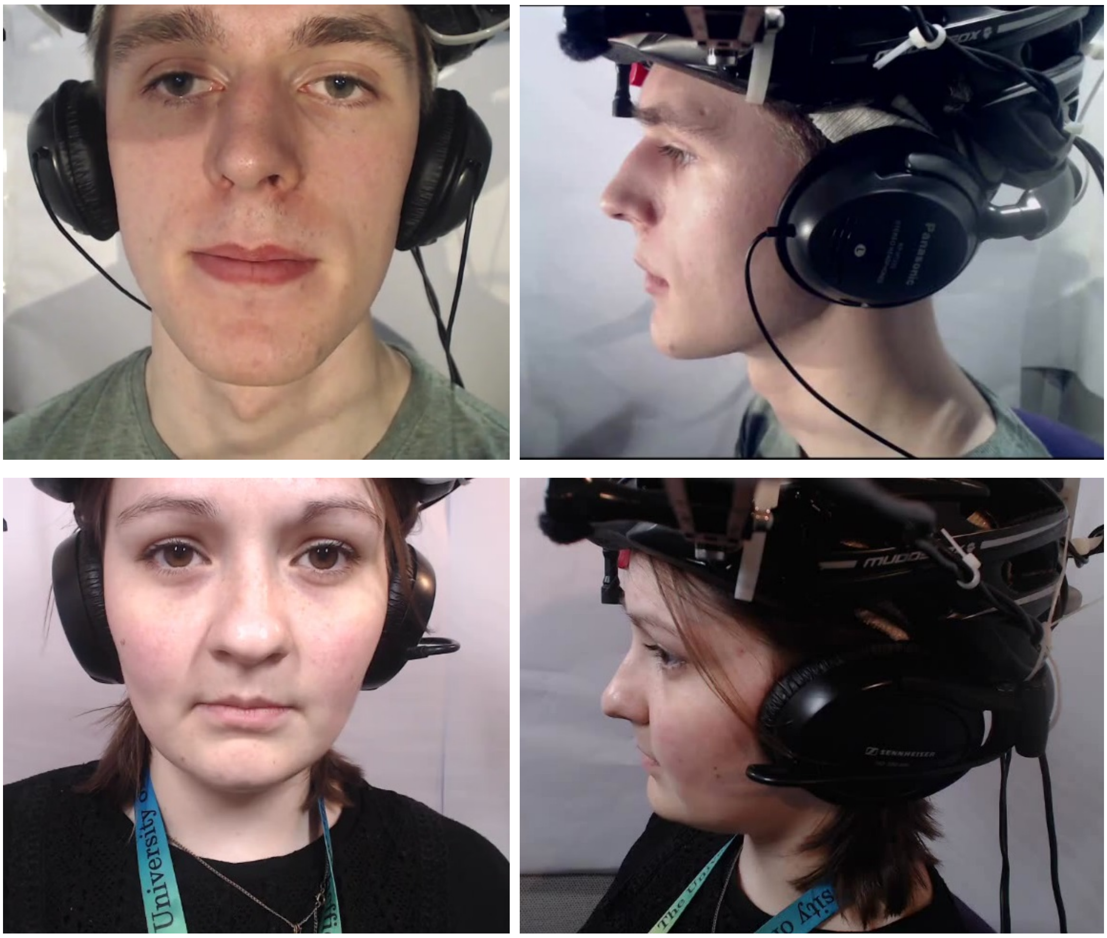


Figure 3.14 Selected examples of front- and side-view frames of the dataset.

heads are not tilted downward, and bottom chin points are visible. The purpose of this step was to facilitate the facial landmark annotation process in the Facegen software tool (see Section 3.2.1). This is considered the main limitation of this thesis, where choosing an inappropriate photograph of a real speaker leads to an insufficient neutral 3D head pose, consequently resulting in undesired 3D lip motions.

There is a variety of facial appearances between individuals due to the basic differences in facial features. For example, person A has thicker lips than person B, and person C has a longer nose than person D. Therefore, facial features of real speakers of the selected pool were classified to investigate different mouth shapes of the real speakers on the resulting 3D lip motions.

To analyse the facial features of the real speakers, a method that presented by Roelfose et al [209] was applied. They used morphometrical methods to classify facial features of South African males photos to investigate common and rare features in this community. For each speaker, video frames were investigated to select the appropriate frame, where the face has a neutral pose shape. Faceware Analyser software was used to process the video frames to extract 12 facial landmarks that are shown in Figure 3.15. Then 13 measurements were taken using the Euclidean distance between the extracted landmarks to calculate a total of 12 indices (see Figure 3.16). Each index was calculated by dividing the smaller measurement by the larger measurement and multiplying the quotient by 100. Ranges of each index then were used to classify the features into different morphological categories (low, middle and high). In order to create the classes for each index, the distributional properties of the data were investigated using box-whisker plots as presented in Table 3.5.

Based on the facial features classification of the real speakers, six real speakers (four female (IDs: S15, S17, S24, and S32) and two male (IDs: S20 and S48)) were selected. In order to investigate the performance of the 3DMM using different sets of facial landmarks (see Section 3.6) and using different amounts of data during creating and training the 3DMM (see Chapter 4). Since this thesis looks into animating lips of a 3D head using 2D videos of a real speaker, classes of two indices relate to mouth features, which are

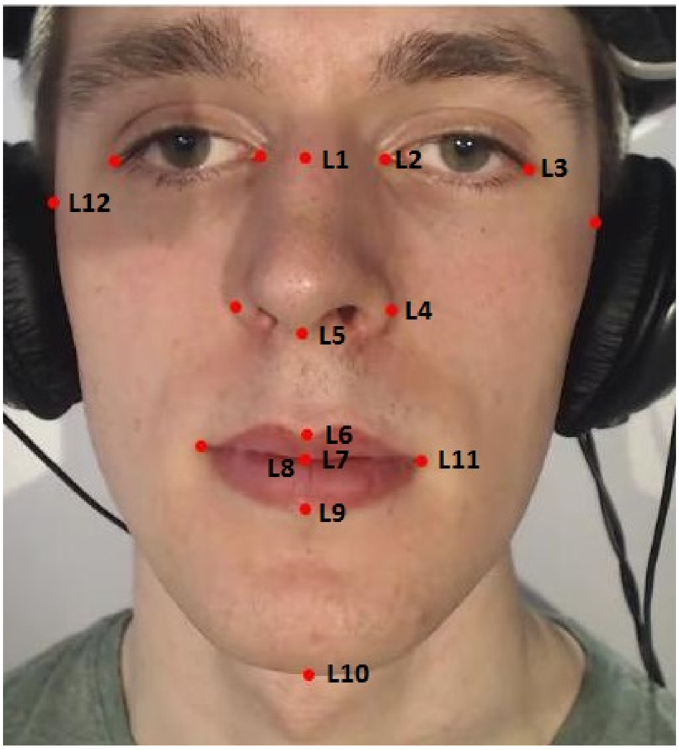


Figure 3.15 Used biometric landmarks of the face (L1 = nasion, L2 = endocanthion, L3 = exocanthion, L4 = alare, L5 = subnasale, L6 = labiale superius, L7 = stomion1, L8= stomion2 L9 = labiale inferius, L10 = gnathion, L11 = cheilion, L12 = zygion)

vertical mouth height (index 7) and mouth width (index 10) were considered to select the six speakers. For example, some of the selected speakers have low (ID: S48), middle (ID: S32) and high (ID: S17) vertical mouth height, and some of them have low (ID: S32), middle (IDs: S15, S24, and S48) and high (ID: S20) mouth width.

For each real speaker, front-view videos of four plain sentences were selected to be mapped to the corresponding 3D head. The selected sentences contain four commands, four prepositions, four different letters and digits, and four adverbs (see Table 3.4) to give phonetic variation in the pool. Although all 25 letters and ten digits were not included in the four selected sentences, which may reduce the chance of presenting the included phonetic features, the filler words (commands, prepositions, and adverbs) provide a variety of the English phonemes.

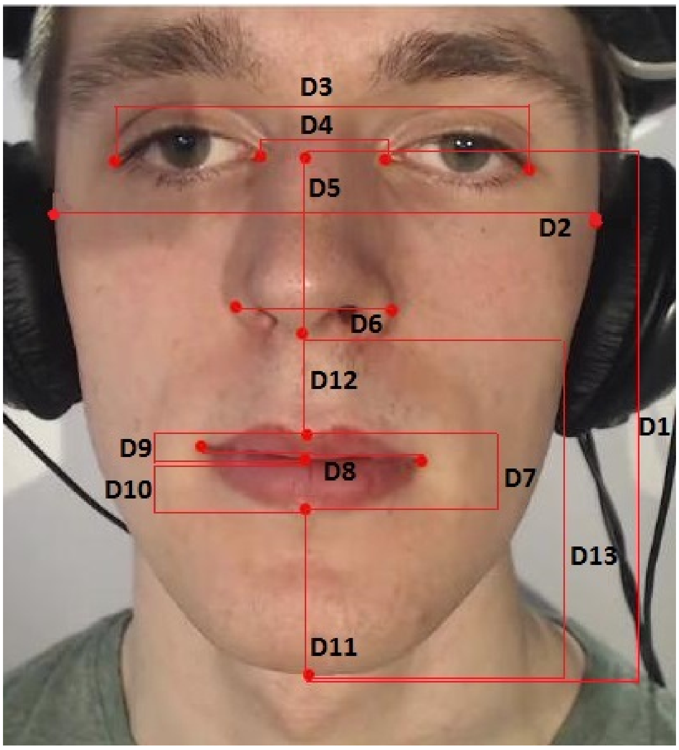


Figure 3.16 Measurements taken from each frame.

3.6 Investigating Facial Feature Landmarks in the Mapping Process

Estimation of a 3D face pose from a 2D image can be achieved by minimising the difference between the 2D landmarks identified on the real face and the corresponding 3D landmarks labelled on the 3D face. Different numbers of feature points have been used for 3D facial expression recognition, for example 74 points [42], 75 points [41], or 83 points [232]. Therefore, it is worth investigating the functionality of each set (lips, nose, eyes, eyebrows, and contours of the face) of the facial features landmarks in the mapping process for 3D lip motion detection. The questions that arise in this context are:

1. What is the role of each feature set of the facial features in the mapping process?

2. Is it possible to animate the lips of a 3D head using the lip's landmarks only?

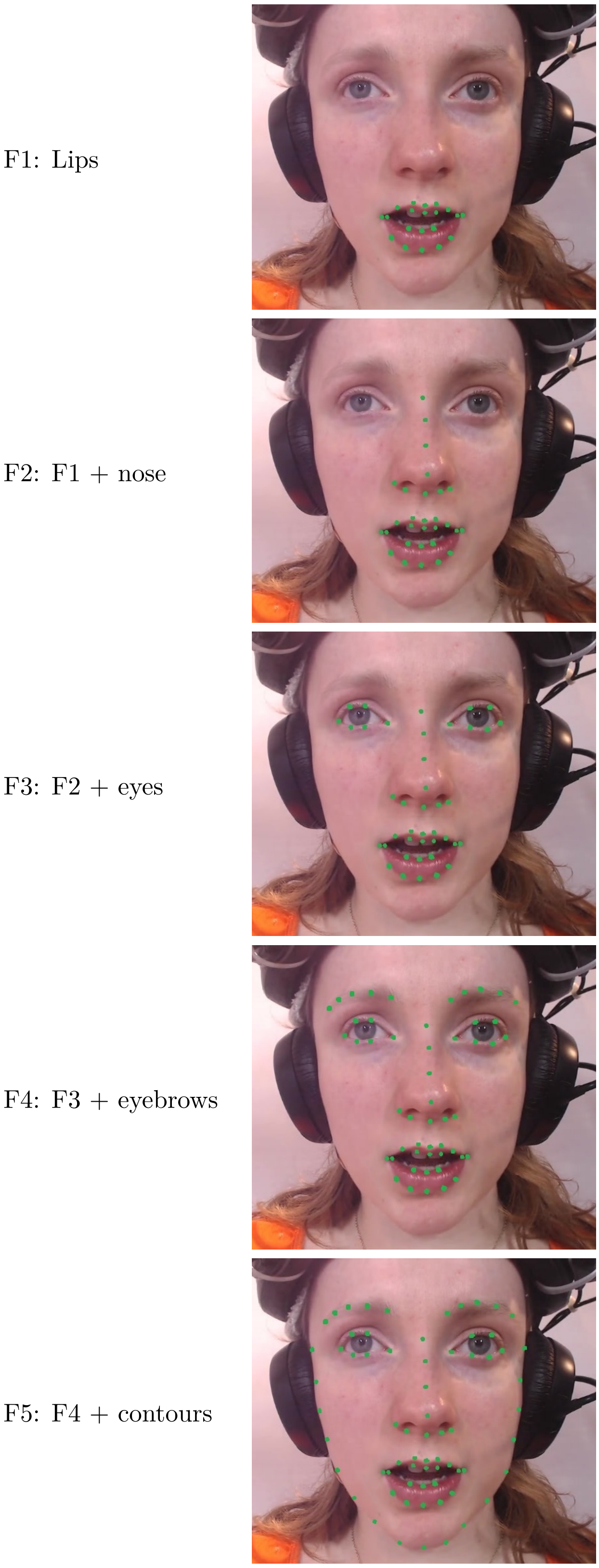
Index	Calculation	Low	Middle	High
Facial	(100 * (D1/D2))	Short, wide <83.68	83.68−89.52	Long, narrow >89.52
Intercanthal	(100 * (D4/D3))	Close <33.59	33.59−35.50	Far apart >35.50
Nasal	(100 * (D6/D5))	Narrow <73.55	73.55−95.97	Wide >95.97
Nasofacial	(100 * (D5/D1)	Short <42.23	42.23−47.49	Long >47.49
Nose-face width	(100 * (D6/D2))	Narrow <28.91	28.91−32.66	Wide >32.66
Lip area	(100 * (D7/D8))	Thin <29.79	29.79−35.53	Thick >35.53
Vertical mouth height	(100 * (D7/D1))	Low, thin <13.58	13.58−16.22	High, thick >16.22
Upper lip thickness	(100 * (D9/D1))	Thin <3.93	3.93−5.61	Thick >5.61
Lower lip thickness	(100 * (D10/D1))	Thin <9.23	9.23−11.44	Thick >11.44
Mouth width	(100 * (D8/D3))	Narrow <49.18	49.18−54.80	Wide >54.80
Chin size	(100 * (D11/D1))	Short <22.79	22.79−27.18	Long >27.18
Nose-upper-lips	(100 * (D12/D13))	Narrow <25.88	25.88−29.08	Wide >29.08

Table 3.5 Metrical features (indices) of the audio-visual lombard grid speech corpus's speakers.

In this experiment front-view videos of real speakers were mapped to their corresponding 3D heads using different sets of facial features landmarks. Five sets were used to map the real faces to the corresponding 3D heads. Figure 3.17 shows the sets.

3.6.1 Evaluation

In order to validate the performance of the animated 3D lip motion, videos of four female speakers (IDs: S15, S17, S24 and S32) and two male speakers (IDs: S20 and S48) from the Audiovisual Lombard Grid Speech corpus [10] were used (see Section 3.5). For each real speaker, four plain sentences from the front view video files were chosen



F1: Lips

F2: F1 + nose

F3: F2 + eyes

F4: F3 + eyebrows

F5: F4 + contours

Figure 3.17 Sets of landmarks that used to map real faces to the corresponding 3D heads.

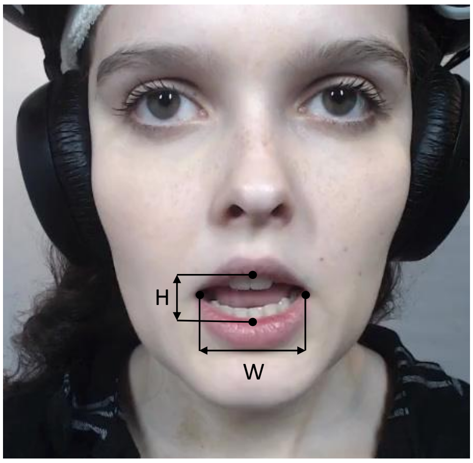


Figure 3.18 Definition of the critical landmarks to measure the width (W) and height (H) of mouth aperture.

to be mapped to each corresponding 3D synthetic head that is generated using front and side-view photographs and trained using 161 poses. The chosen sentences contain different words (e.g different verbs (bin, lay, place and set) and letters (a, b, etc.)), in order to contain the maximum number of English phonemes. The resulting 3D head animation was then compared to the original ground-truth 2D videos. This was done for each set of landmarks shown in Figure 3.17.

For the comparison, Faceware Analyser software [2] was used to track the facial features in the ground-truth 2D video and the front-view (2D) of the corresponding 3D animation. Two geometric articulatory measurements, as shown in Figure 3.18, were calculated from the extracted facial features, following the previous literature [4]. These are:

- Width of mouth aperture measurement (W), which is defined by the horizontal distance between the right and left inner corners of the lips.
- Height of mouth aperture measurement (H), which is defined by the distance between the top and the bottom middle of the inner mouth contour.

In order to correct the distance between the camera and the real speaker or the talking 3D head, all the landmarks were normalised by using the Euclidean distance between the midpoint of the inner corners of the eyes and the nose tip's point following a study in [9]. These points are considered reference of a measure in imaging systems, since they are not affected by the articulations or the facial expressions [48]. All visual articulatory features for the real speakers and their corresponding 3D heads were normalised by their corresponding maximum and minimum mouth measurements in the videos. This gives all the articulatory measurements on a [0-1] scale. Given the width and height values for each frame of animation, for both the 2D video for a real speaker and the corresponding 3D animation, the root mean square error (RMSE) over a sentence was used to evaluate the role of each set of the facial features in the mapping process following a study in [252]. More discussion about evaluating the quality of visual speech is presented in Section 2.5.

3.6.2 Results and Discussion

Figure 3.19 shows an example of consecutive frames of a real speaker (ID: S 24) and the corresponding 3D head that was fitted to the real speaker's face using different sets of landmarks, during utterance of the phoneme /b/ of the word "bin" from sentence "bin white in N 3 now". This Figure shows that the performance of the animated 3D lips improves when F3 set was used for the mapping process, and further improves when F4 set was used which contains the eyebrows landmarks, due to keeping the spatial distance between the facial features of the 3D head during the animation process. In case of using the set F2, the lips were completely distorted, because vertices of the cheeks' mesh were extended. Using the set F3 that contains the eyes landmarks restricted the cheeks, thus the lip motions were enhanced. Also, it can be noticed that the 3DMM gives poor estimation to the real speaker's pose when F5 set was used. This is because of using the contour landmarks that slightly shrinks and restricts the mesh of the 3D face, resulting in a partially opened mouth shape. Figure A.1 shows another example of consecutive frames of a real speaker (ID: S 48) and the corresponding 3D head that was fitted to the real speaker's face using different sets of landmarks, during utterance of the phoneme /b/ of the word "bin" from sentence "bin white with V 7 soon". This Figure shows how the 3D mouth shape matched the real speaker when the set F3 was used for the animation

Speaker ID	F1: Lips		F2: F1+Nose		F3: F2+Eyes		F4: F3+eyebrows		F5: F4+Contour	
	W	H	W	H	W	H	W	H	W	H
S15	0.518	0.140	0.296	0.183	0.146	0.117	**0.131**	**0.087**	0.153	0.099
S17	0.252	0.192	0.116	0.118	0.095	0.119	**0.092**	**0.095**	0.129	0.106
S20	0.366	0.347	0.187	0.201	0.278	0.167	**0.244**	**0.155**	0.245	0.192
S24	0.303	0.158	0.287	0.137	0.276	0.136	**0.219**	0.123	0.254	**0.105**
S32	0.339	0.128	0.247	0.156	0.165	0.071	**0.111**	**0.056**	0.157	0.087
S48	0.378	0.111	0.198	0.125	**0.125**	**0.068**	0.149	0.071	0.180	0.091

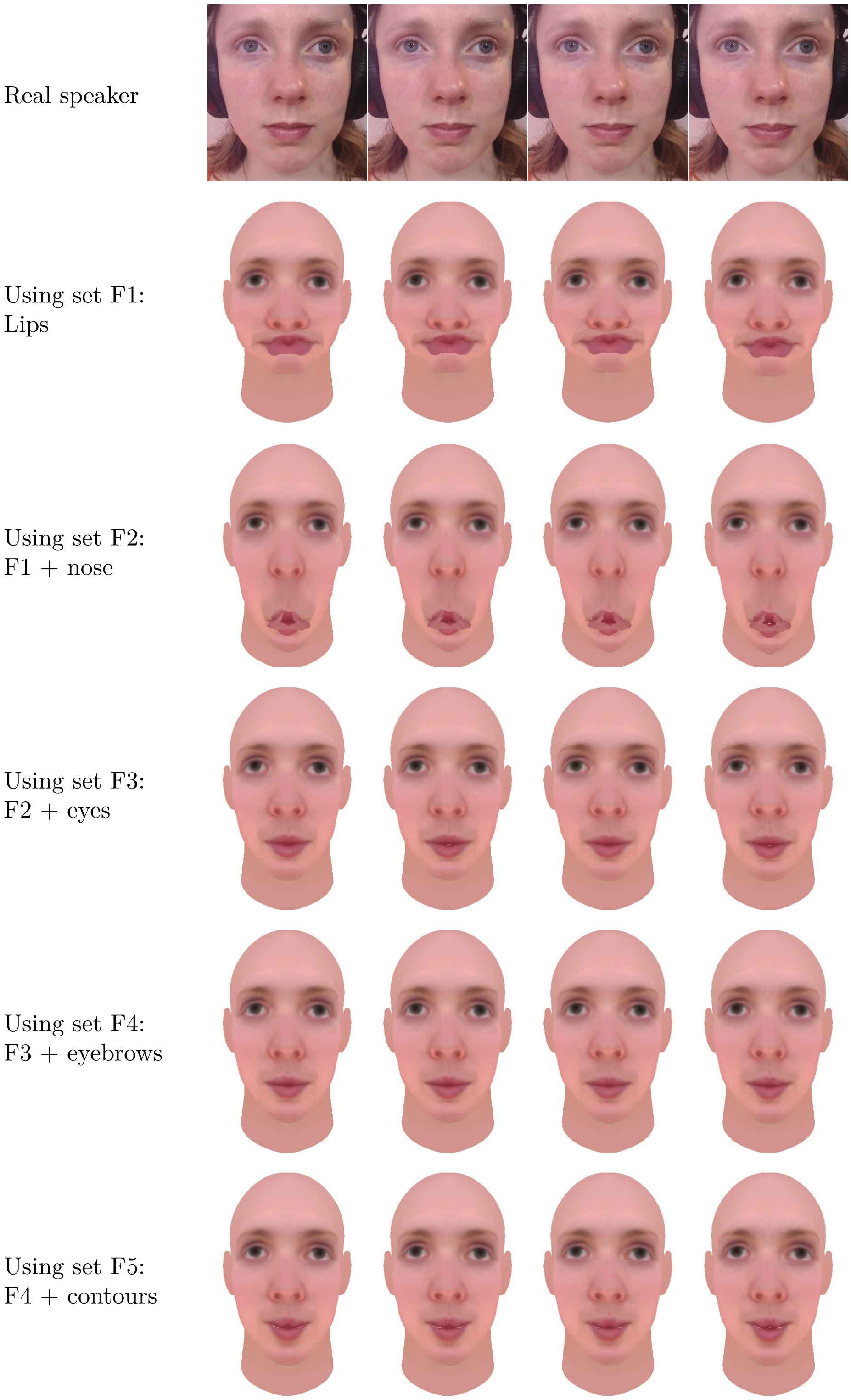
Table 3.6 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers and their corresponding 3D heads that were mapped to each other using different sets of facial features landmarks. Values in bold means the decreased RMS error. Width and height error=±0.001.

process. However, the texture was distorted around the mouth area as illustrated in Figure A.2, that confirms the importance of using the eyebrows landmarks.

Table 3.6 shows the RMSE results averaged over 4 sentences for width and height of the mouth aperture of the real speakers and their corresponding 3D heads. The 3D head models that were fitted to their corresponding real speakers using F4 set give the lowest RMSE scores for height for four out of six of the speakers and width for five out of six of the speakers. For the width, a t-test suggests a significant difference in RMSE results for the 3D heads that were fitted to the real speakers' faces using F4 set versus F1 set (p=0.0059) and F5 set (p=0.0065), this is perhaps due to stretching the mesh of the lip when F1 set was used and shrinking when F5 set was used. There is no significant difference for F2 set (p=0.1061) and F3 set (p=0.1300). Also, there is a significant difference for height between the 3D heads that are fitted to the real faces using F4 set and F1 set (p=0.0193), F2 set (p=0.0130) and F3 set (p=0.0220), although there is no significant difference for F5 set (p=0.1074). Due to slight changes in lip closure when contour landmarks are used. Based on these results, the F4 set will be used to animate the 3D heads in the rest of this thesis.

3.7 Summary

This chapter presented how the 3D Morphable Model was constructed using synthetic 3D head poses that were generated using a commercial software (FaceGen), which provides vertices correspondences for the generated models, solving a problem of the variability in vertices numbers of the scanned 3D real faces. First, the synthetic 3D head poses dataset was introduced which was used to train the 3D Morphable Model, that was generated by applying PCA to the vertices of the head poses. The stages for mapping between 2D video frames of a real speaker and the 3D head model were then presented. This mapping process was achieved following the method presented by Huber et al [125] that automatically fits 3D faces to 2D images using markers on both 2D and 3D faces. Furthermore, the dataset that are used to create the initial 3D head pose, animate the 3D heads and evaluate the resulting 3D lip motion was reviewed. In addition to, a study was presented that investigates the role of each set of the facial features in the mapping process. The results confirmed that using facial features landmarks, including eyebrows, eyes, nose, and lips gives the best performance of the 3D lip motions, while including the contour landmarks impedes achieving the desired lip motions due to restriction of the face mesh. In the next chapter, the implementation and evaluation of performance of the animated 3D lips using the facial features landmarks (i.e. F4 set presented in Section 3.6) will be presented. Also, experiments with different data sets being used to build different 3DMMs for a real speaker will be presented.



Real speaker

Using set F1:
Lips

Using set F2:
F1 + nose

Using set F3:
F2 + eyes

Using set F4:
F3 + eyebrows

Using set F5:
F4 + contours

Figure 3.19 Consecutive frames of the phoneme /b/ during utterance of the word bin from sentence "bin white in N 3 now" for a real speaker (ID: S 24) and the corresponding 3D head animated using each set of facial features landmarks.

Chapter 4

3D Visual Speech Animation Using 2D Videos

4.1 Introduction

The intelligibility of speech communication is increased when visual signals such as facial expressions and lip motions are combined with speech [158]. It is therefore paramount that such visual signals are well integrated with an original, or synthetic, auditory speech signal to produce a realistic talking virtual character. This chapter presents an approach for visual speech animation that involves driving the lip motion of a synthetic 3D head, in accordance with tracked lip motion in front-view 2D videos of a real speaker. This is achieved through a mapping between corresponding landmarks identified in the 2D videos, and a 3DMM built using 3D synthetic head poses (see Section 3.2). The 3D synthetic head poses are generated using commercial software (FaceGen [1]) to train the model as explained in Chapter 3. The 3DMM is fitted to front-view 2D video of a real speaker (from [10]) using the method in Huber et al.'s work [125], which uses a 3DMM to reconstruct 3D faces from images by minimising the difference between 2D and the corresponding 3D landmarks. Using geometric features such as landmarks to estimate shape is independent of illumination effects and limited resolution. This makes detection of lip motions in 2D videos accessible and does not require manual labelling or training as in the mapping algorithms based on depth (i.e. RGB_D) data [74, 100, 101, 129]. Given a shape estimate, a linear method can be derived to fit the estimated shape to the real face pose in a manner which is repetitive and fast.

3DMMs [32] have been widely used for different applications such as 3D visual speech animation [67, 186], face recognition [31, 102], controlling 3D avatars for gaming [49], visual dubbing [98], and face reanimation [30, 240, 205]. For these kinds of applications, a large number of publicly available 3D scanned datasets that include either real neutral face poses only [34, 198] or neutral poses and different expressions [102, 262] were used. Such models and approaches, however, do not consider how using different amounts of data in different stages of constructing a 3DMM influences the final animation results. For example, how scanning the face from different views (i.e. front and side-views) and using different numbers of face scans would affect the resulting animation. For this reason, a series of experiments were conducted to investigate this. The experiments address two main questions:

1. Would using both front- and side-view photographs, rather than just a front-view photograph, in the construction of the initial 3D head pose produce better animation results?

2. Would using different intensities of the same viseme shape (e.g. different amounts of mouth openness for the same viseme) when constructing the 3DMM produce better animation results?

In this chapter, four different 3DMMs are generated for each real speaker using either front-view photograph only or front- and side-view photographs to create the initial 3D head pose, and using different numbers of 3D head poses to train each 3DMM (see the red dotted box in Figure 4.1). Then 2D front-view videos of the real speaker are mapped to each corresponding 3DMM (see the green dotted box in Figure 4.1). Finally, the front-view videos of a real speaker are used to evaluate the performance of each corresponding 3DMM (see the left diagram in the brown dotted box in Figure 4.1). Additionally, the lip protrusion of each 3DMM will be evaluated using the side-view videos in this chapter as well (see the right diagram in the brown dotted box in Figure 4.1). The experiments examine to what extent the side-view photograph contributes to give a closer 3D head shape to the real speaker, which enhances the resulting 3D lip motions. Furthermore, they investigate the effects of using different intensities of each viseme on width and height of the mouth aperture, which provide a variety of mouth shapes during speech.

This chapter is organised as follows: Section 4.2 will present the data sets used to build different 3DMMs for a speaker. Section 4.3 will present the front-view evaluation of the 3D lip motions. This includes the evaluation of the final synthetic 3D animation results for each 3DMM, in comparison with ground-truth data (the front-view videos of a real speaker [10]) (Section 4.3.1), and the discussion of the results (Section 4.3.2). Section 4.4 will provide the side-view evaluation of the resulting animation, including the evaluation of the final synthetic 3D animation results for each 3DMM, in comparison with ground-truth data (the side-view videos of a real speaker [10]) and the discussion of the results. Finally, Section 4.5 summaries this chapter.

4.2 The Data Sets of 3DMM

Four data sets were used to build different 3DMMs for a speaker. Table 4.1 summarises the data sets, and the red dotted box in Figure 4.1 explains the stages of preparing these data sets. The differentiating factors are whether a front-view photo only or front- and side-view photos are used in constructing the neutral head pose and whether 17 (16 visemes and a neutral pose) or 161 poses (10 intensity variations of 16 visemes and a neutral pose) are used for a 3DMM.

Figure 4.2 shows the front and side photographs of a real speaker (ID: S32) and the corresponding 3D heads that were generated using a front-view photograph only (left), and front- and side-view photographs (right). More figures are presented in Appendix B (Figures B.1 - B.3) that show how the 3D face is flattened when only the front-view photograph is used to generate the initial neutral 3D head pose, and how the lips are

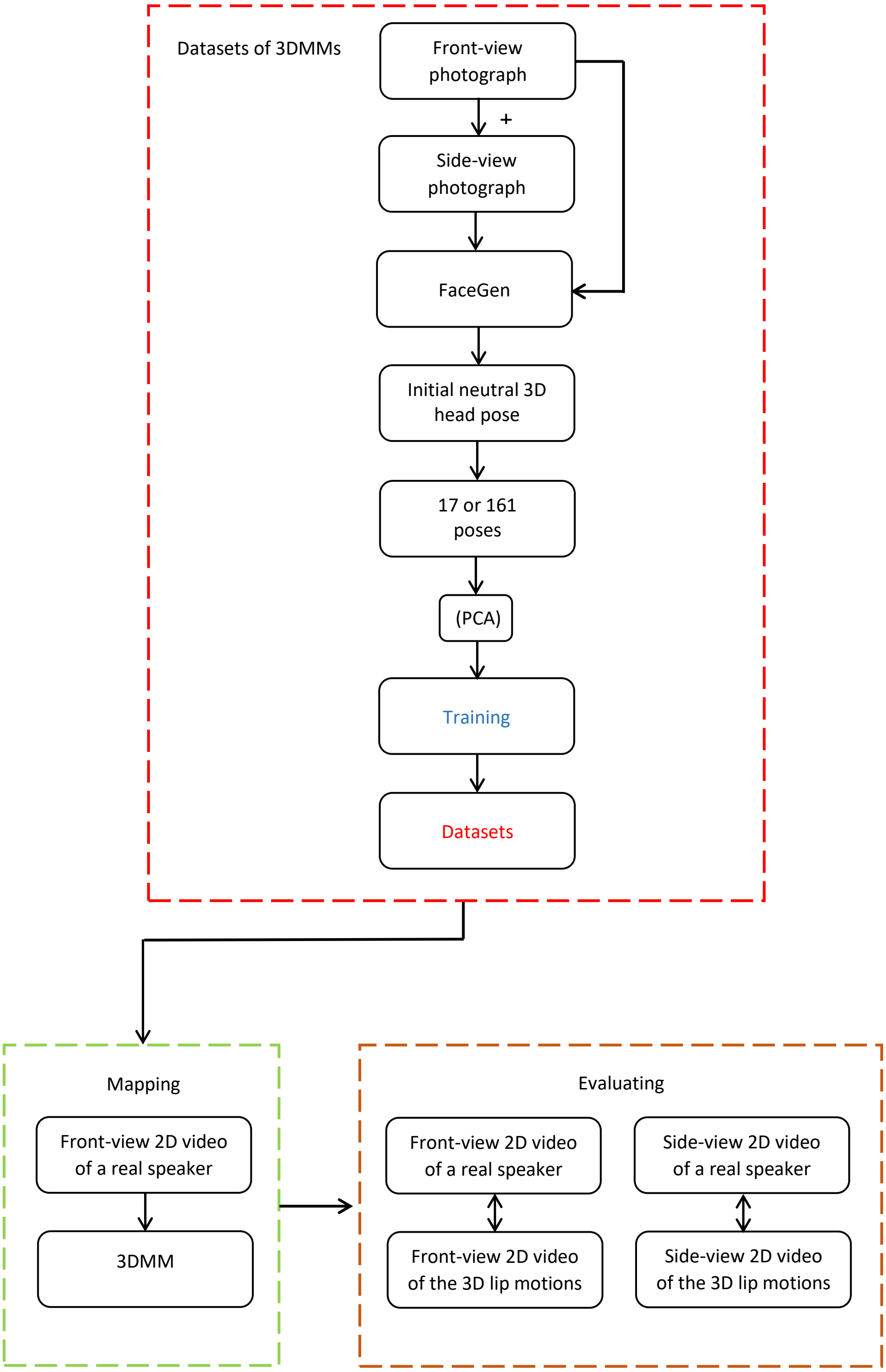

Datasets of 3DMMs
Front-view photograph
+
Side-view photograph
FaceGen
Initial neutral 3D head pose
17 or 161 poses
(PCA)
Training
Datasets
Mapping
Front-view 2D video of a real speaker
3DMM
Evaluating
Front-view 2D video of a real speaker
Side-view 2D video of a real speaker
Front-view 2D video of the 3D lip motions
Side-view 2D video of the 3D lip motions

Figure 4.1 Schematic view of animating and evaluating 3D lips of each 3DMM.

more protruded when both the front and side-view photographs are used, giving the 3D head a natural shape which is closer to the real speaker. Each of the data sets was used in producing a 3DMM, which was subsequently used in the process described in Chapter 3.

Number of poses	Front-view	Front- & side-views
17 poses	Dataset 1	Dataset 3
161 poses	Dataset 2	Dataset 4

Table 4.1 17 or 161 poses in combination with front-view only or front- and side-views photos are used to prepare four data sets.

4.3 Front-view Evaluation of 3D Lip Motion

In this section, the front-view videos are used to evaluate the performance of different 3DMMs that created for each real speaker. Either front-view photograph only or front- and side-view photographs are used to create the initial neutral 3D head pose, and 17 poses or 161 poses are used to train the model. The aim of this section is to investigate whether using a side-view photograph with the front-view photograph to create the initial neutral 3D heads pose, and training the 3DMM with different numbers of viseme intensities enhances the resulting 3D lip motions.

4.3.1 Evaluation

The evaluation process presented in Section 3.6.1 was followed to evaluate the performance of the animated 3D lip motion of each 3DMM. This includes using videos of four female and two male speakers from the Audiovisual Lombard Grid Speech corpus [10], choosing four plain sentences from the front-view videos for each speaker to be mapped to each corresponding 3D head, and comparing 2D videos of the resulting 3D animation to the original ground-truth 2D videos. This was done for each of the 3DMMs built for the 4 data sets summarised in Table 4.1. Faceware Analyser software [2] was used to track the facial features of both the real speakers and the corresponding 3D animation in 2D videos. Two articulatory measurements (see Figure 3.18) that represent the width and height of the mouth aperture were calculated. Then the extracted features were

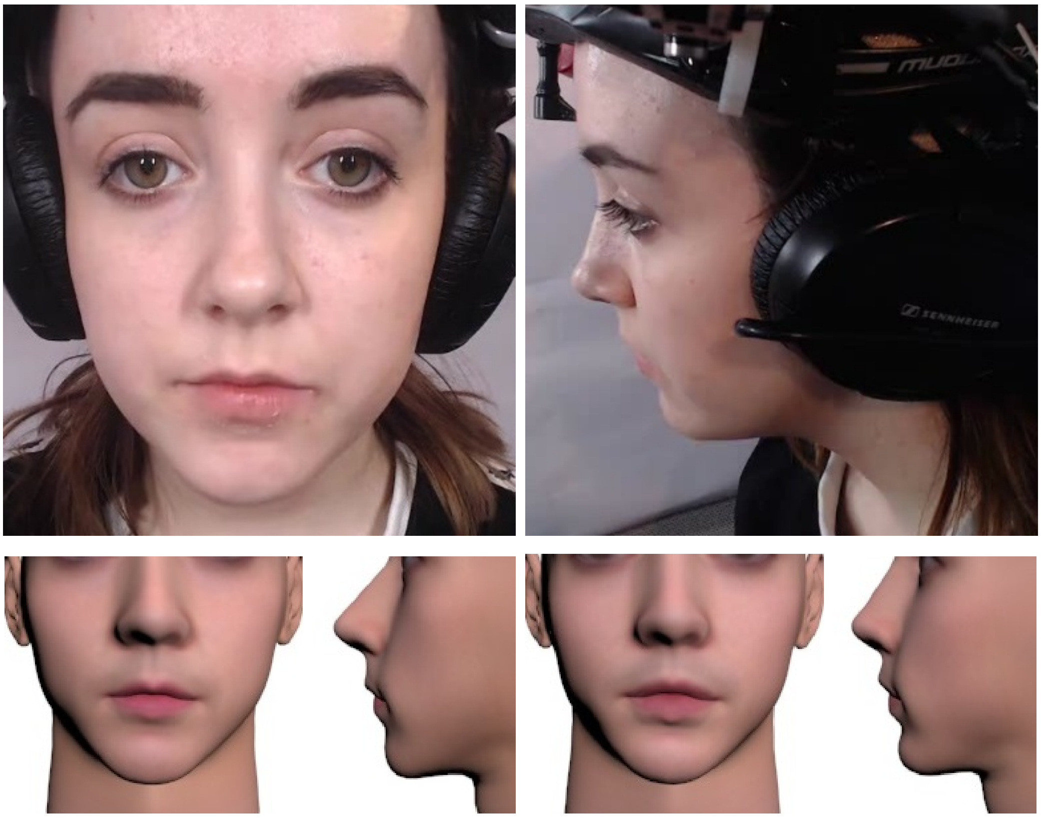


Figure 4.2 First row: Front (left) and side (right) photographs of a real speaker (ID: S32); Second row: front and side view of the corresponding 3D heads generated using front photograph only (left) and front and side photographs (right) – the lips are more protruded in the image on the right.

normalised to correct the distance between the camera and the speaker, and scaled to give the articulatory measurements on a [0-1] scale as explained in Section 3.6.1. Given the width and height values for each frame of animation, for the 2D videos of a real speaker and the corresponding 3D animation, the root mean square error (RMSE) over a sentence was used to evaluate the effectiveness of each 3DMM.

4.3.2 Results and Discussion

This section presents results of mapping front-view 2D videos of each real speaker to each corresponding 3DMM. The 3DMMs were created using front-view photograph only or front- and side-view photographs to generate the initial neutral 3D head pose; and then 17 or 161 3D head poses were used to train the 3DMMs as explained in Section 4.2. Figure 4.3 shows an example of consecutive frames of the phoneme /w/ during utterance of the letter "Y" from the sentence "place green in Y zero again" for a real speaker (ID: S17) and the corresponding 3D head for each data set. This figure shows that the performance of the animated 3D lips improves when front- and side-view photos are used to generate the initial neutral 3D head pose in FaceGen (see the last two rows in Figure 4.3), where the mouth aperture becomes smaller. This gives the 3D lips a closer shape to the corresponding real speaker. For example, the thickness of the lips in Figure B.1 shows the front and side photographs of the same real speaker and the corresponding 3D heads that were generated using the front-view photograph only (left) and the front- and side-view photographs (right). Also, the performance of the 3D lips further improves and becomes closer to the real speaker when a larger number of 3D head poses (i.e. different viseme intensities) are used to train the 3DMM.

Figure B.4 illustrates how the 3D head model fails to detect the phoneme /b/ during utterance of the word "bin" from the sentence "bin white in N 3 now" for a real speaker (ID: S 24) when less information was used to generate the 3D head model. The lips are partially opened when the front-view photograph only was used to generate the initial neutral 3D head pose (i.e. Dataset 1 and Dataset 2). When the front- and side-view photographs were used, the lips become closer (e.g. Dataset 3) and the performance is enhanced when a larger number of poses were used to train the 3D head model (i.e. Dataset 4). Figure B.5 illustrates how the 3D lips become more protruding during uttering the phoneme /uw/ of the word "soon" from the sentence "bin white with V 7 soon" when front- and side-view photographs were used (i.e. Dataset 3), and have closer shape to the real speaker when a larger number of poses were used to train the model (i.e. Dataset 4).

Figure 4.4 shows the trajectories of the width and the height parameters of the mouth aperture for the real speaker (ID: S17) and the corresponding 3D heads whilst uttering the sentence "place green in Y zero again". Whilst all the trajectories generated using the animation pipeline generally follow the real speaker's trajectory, the trajectories of the 3D heads that contain 161 poses and which are generated using front- and side-view photos (i.e. Dataset 4) are much closer to the ground truth trajectory. Thus, using different intensities of viseme data in the construction of the 3DMM, as well as one extra photograph in the construction of the 3D head, improves the performance of the resulting 3D lip motions.

For the width parameter (W), the trajectories of the 3D head that is generated using the front-view photo only and trained with 161 poses (i.e. Dataset 2) is closer to the ground truth's trajectory especially for spread phonemes such as /ey/, /iy/ and /s/, due to using larger number of poses for training the model which gives the model more variations of spread mouth shapes. For the height parameter (H), the model that is generated using front- and side-view photos and trained with 17 poses (i.e. Dataset 3) performs properly for the protruding rounded phonemes such as /w/ and /ow/, due to using side-view photo for creating the initial neutral 3D head pose which gives the lips more protruding shape as shown in Figure B.1. This is confirmed by Table 4.2 that shows the RMSE results for each phoneme for the two corresponding 3D heads of the real speaker (ID: S17). Also, the same behaviour can be observed in Figure B.6 that shows the trajectories of the width and the height parameters of the mouth aperture for a real speaker (ID: S32) and the corresponding 3D heads whilst uttering the sentence "place blue at Y 4 now".

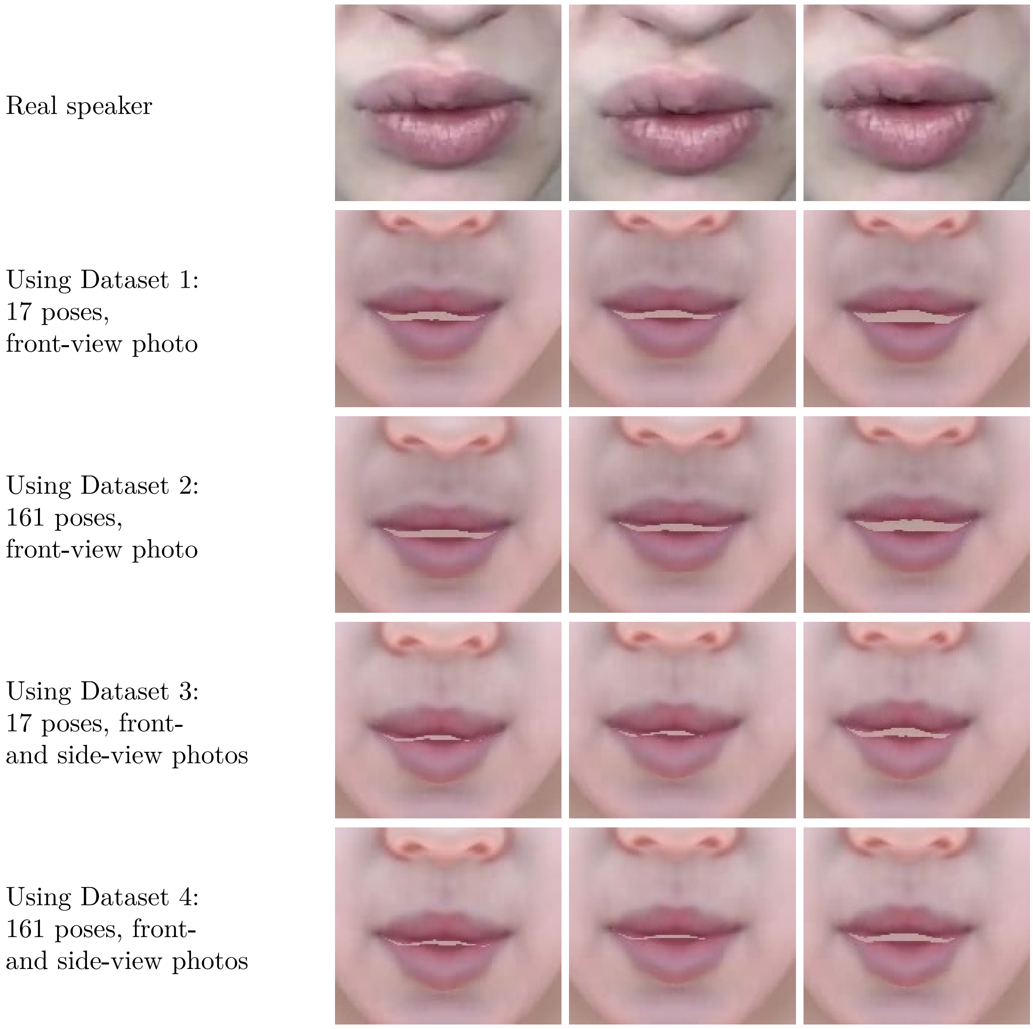

Real speaker
Using Dataset 1: 17 poses, front-view photo
Using Dataset 2: 161 poses, front-view photo
Using Dataset 3: 17 poses, front- and side-view photos
Using Dataset 4: 161 poses, front- and side-view photos

Figure 4.3 Consecutive frames of the phoneme /w/ during utterance of the letter y from sentence "place green in Y zero again" for a real speaker (ID: S17) and the corresponding 3D head for each data set.

Table 4.3 shows the RMSE results averaged over 4 sentences for width and height of the mouth aperture of the real speakers and their corresponding 3D heads. The 3DMMs that contain 161 poses and are generated using front-view photo only (i.e. Dataset 2) give closer results to the 3DMMs that are generated using Dataset 4 for most of the speakers. Due to including larger number of poses that give more variations in mouth width which makes the model closer to the real speaker. This can be clearly observed from Figure 4.5 that shows the error bars of the width parameter (W) of the mouth aperture of the

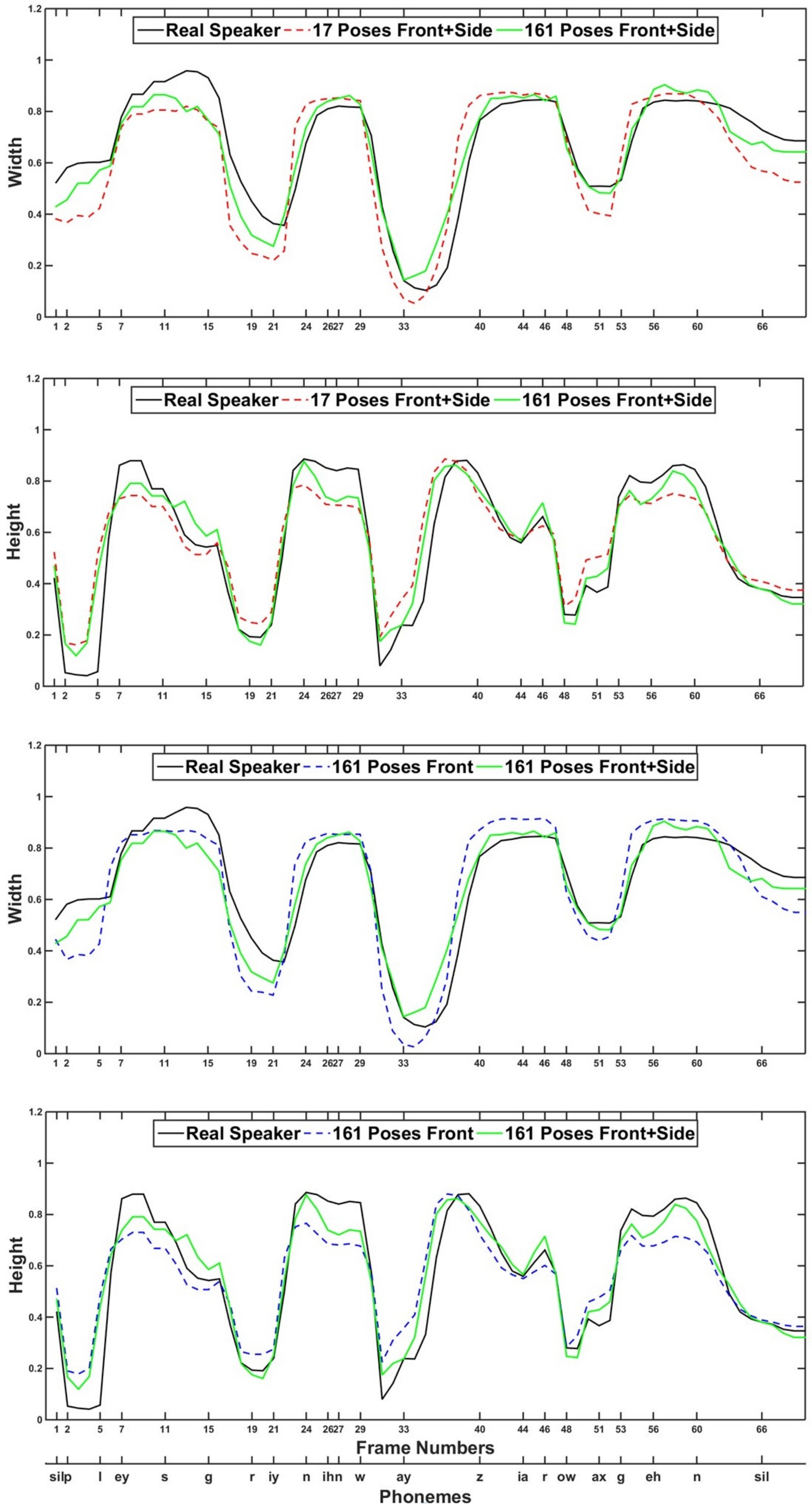


Figure 4.4 Width and height of mouth trajectories of 2D frames of the real speaker (ID:S17) and the corresponding 3D heads. Top two compare height and width between 17 and 161 poses (both with front- and side-view photos), while the bottom two compare height and width between front- view photo only and front- and side-view photos (both with 161 poses).

ID	Phoneme	Front photo		Front + side photos	
		161 poses		17 poses	
		W	H	W	H
S17	ey	**0.040**	0.120	0.100	**0.097**
	iy	**0.145**	0.052	0.149	**0.051**
	s	**0.082**	0.075	0.137	**0.049**
	ow	0.051	0.059	**0.048**	**0.050**
	w	**0.122**	0.140	0.128	**0.114**

Table 4.2 The RMS error averaged over frames of spread phonemes (/ey/, /iy/, and /s/) and protruding rounded phonemes (/w/ and /ow/) for width (W) and height (H) of the mouth of a real speaker (ID: S17) and the corresponding 3D heads created using Dataset 2 and Dataset 3 during utterance of sentence "place green in Y zero again". Values in bold indicate the lowest RMS errors for width and height in each row.

corresponding 3D heads of each real speaker. Uncertainty in the reported measurement is calculated by adding and subtracted 2 pixels to and from all the articulatory features for both the real speakers and their corresponding 3D heads.

The 3DMMs that are generated using front- and side-view photos and trained with 17 poses (i.e. Dataset 3) give better results than the 3DMMs that are generated using front-view photo only and trained with 17 poses (i.e. Dataset 1) for most of the speakers, which emphasizes on the importance of using the side-view photo to construct the initial neutral 3D head pose in enhancing the final animation.

This interpretation is similar to the height parameter (H), where the 3DMMs that contain 17 poses and which are generated using both front- and side-view photos (i.e. Dataset 3) give closer results for most of the speaker. This is due to using the side-view photos to create the initial neutral 3D head pose which gives the lips more thickness shape that is closer to the real speaker (see Figure 4.3), and subsequently affects the closure of the lips (see Figure B.4). This is noticeable from Figure 4.6 that shows the error bars of the height parameter (H) of the mouth aperture of the corresponding 3D heads of each real speaker.

For the 3D heads that contain 161 poses, a t-test suggests a significant difference in RMSE results for the 3D heads that use front- and side-view photos versus front-view

ID	Front photo				Front+side photos			
	17 poses		161 poses		17 poses		161 poses	
	W	H	W	H	W	H	W	H
S15	0.152	0.120	0.154	0.117	**0.129**	0.102	0.131	**0.087**
S17	0.121	0.137	0.115	0.128	0.120	0.109	**0.092**	**0.095**
S20	0.239	0.166	0.247	0.158	**0.229**	0.156	0.244	**0.155**
S24	0.287	0.141	0.223	0.151	0.260	0.142	**0.219**	**0.123**
S32	0.117	0.067	0.115	0.075	0.210	0.067	**0.111**	**0.056**
S48	0.199	0.086	0.175	0.080	0.203	0.075	**0.149**	**0.071**

Table 4.3 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers and their corresponding 3D heads. Values in bold indicate the lowest RMS errors for width and height in each row. Width and height error=±0.001.

photos only (p=0.0292 for width and p=0.0009 for height). Also, there is a significant difference for height between the 3D heads containing 161 poses and 17 poses that are generated using front- and side-view photos (p=0.0135), although there is no significant difference for the width (p=0.0967). This is due to limited changes in this parameter over time during uttering some phonemes, in comparison with the height parameter (H) that detects the closure of the lips.

In summary, the evaluation produced two sets of results: using both a front- and side-view photos in the construction of the neutral 3D head pose improves the results in comparison to just using the front-view photo. In addition, increasing the number of 3D head poses (different viseme intensities) to train the 3DMM further improves the performance of the 3D lip motions. An example video of the resulting 3D lip motions available at: https://www.youtube.com/watch?v=PzBxbDugvmA. These findings support the usefulness of using both front- and side-view photos to construct the neutral pose of the 3D head, and different intensities of each viseme to train the 3DMM. In the next section, further evaluation work of the resulting animation will be presented that makes use of side-view videos.

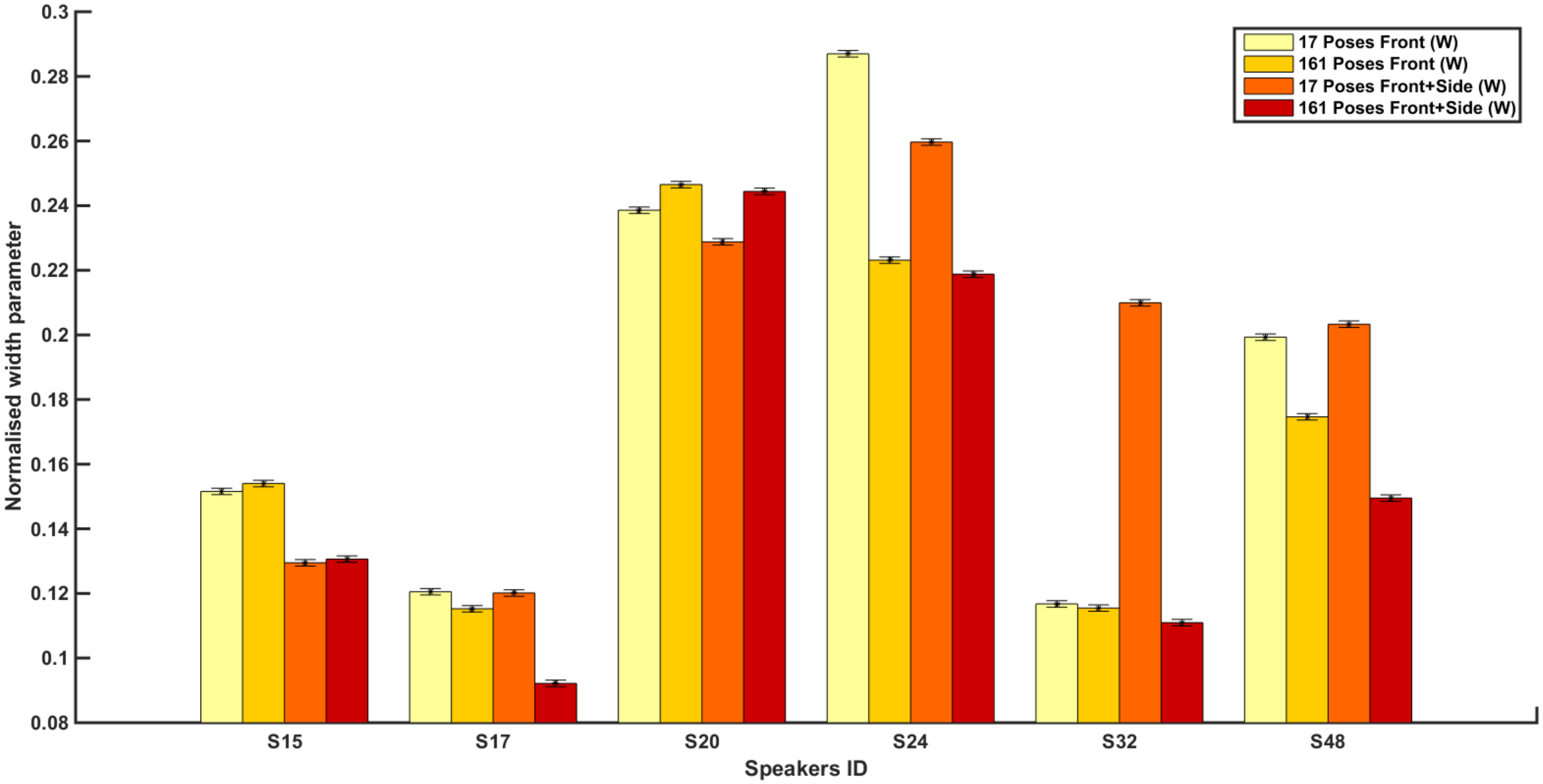


Figure 4.5 Error bars of the width of the mouth aperture parameter (W) of the corresponding 3D heads of each real speaker.

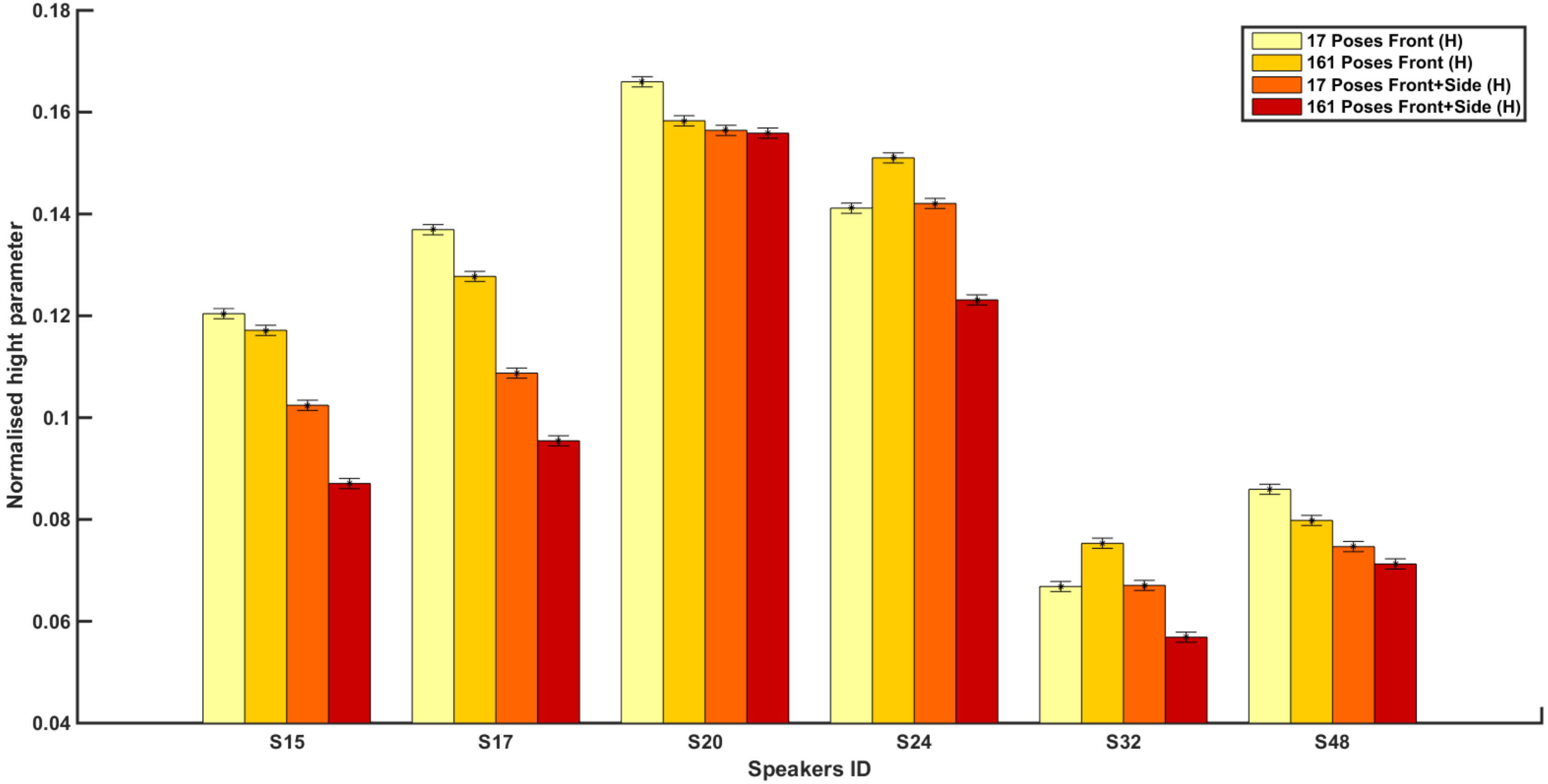


Figure 4.6 Error bars of the height of the mouth aperture parameter (H) of of the corresponding 3D heads of each real speaker.

4.4 Side-View Evaluation of 3D Lip Motion

In the second language learning (L_2) field, four skills of students should be improved which are writing, reading, speaking, and listening [29, 124, 160]. The fundamental factor of improving listening and speaking is pronunciation, due to its role in understanding and being understood, also increasing self-confidence and promoting social interactions of students in the surrounded environment [47, 94, 184]. There is a diversity in the phonetic system of languages, which makes learners are required to learn new movements of the speech articulators and do lots of effort to achieve sufficient pronunciation. A typical example, Turkish learners often encounter difficulties in producing rounded vowels, due to their inability to make the essential muscular effort for producing such sounds sufficiently [119]. Another example is Japanese learners who are not able to pronounce rounded sound vowels that require lip protrusion, due to lacking distinction between rounded and unrounded vowels in their daily conversation [76].

3D talking heads can provide aid for learning pronunciation, where they can be presented as virtual tutors on the computer screens in an engaging manner of the learning process, starting from reading, to the pronunciation, ending with conversation practice such as [174]. However, much investigation and accurate evaluation are needed before presenting a new synthetic speech tool that can provide an unquestionable and efficient aid to the pronunciation teachers, where any lack of consistency between the audio signal and the visual signal leads to an ambiguous animated signal, resulting in incorrect learning. This provided the motivation to investigate the impact of using different amounts of data during constructing the 3DMM on the quality of lip protrusion of the animated 3D heads.

The quality of 3D visual speech animation can be assessed objectively by re-synthesizing a set of sentences according to ground truth speech signals. For each sentence, quality measurement is defined by computing the similarity between the ground-truth and the synthesized speech signal. This can be achieved using the location of specific feature points in the face to calculate different geometric articulatory features for example width and height of the mouth [53] or the mouth aperture [208] and chin height [109]. Such approaches, however, do not consider measuring the quality of lip protrusion of the synthesised signal. Therefore, in this section front-view videos are used to generate the 3D lip motion, and side-view videos are used to evaluate the resulting 3D lip motion (i.e. lip protrusion). Different amounts of data in the mapping process are investigated. To facilitate this, a data set that contains both front- and side-view videos of speakers [10] is used.

Unlike the previous section, where the results are evaluated only from a front viewpoint, the results are evaluated from side views in this section. The aim is to investigate whether or not adequate lip protrusion effects are produced when using only front-view videos in the mapping process for the 3DMM.

4.4.1 Evaluation

In order to validate the lip protrusion of our 3D talking heads, videos of four female speakers (IDs: S15, S17, S24 and S32) and two male speakers (IDs: S20 and S48) from the Audiovisual Lombard Grid Speech corpus [10] were used. The front-view videos were used for the mapping process and the side-view videos were used in the evaluation process. For each real speaker, four plain sentences from the front-view video files were chosen to be mapped to each corresponding 3D synthetic head (built using FaceGen, as explained in Chapter 3). The selected sentences contain various words (e.g different verbs, colours, letters, etc) in which the maximum number of English phonemes are included. A side view of the resulting 3D animation was then compared to the original ground-truth 2D side-view videos to measure lip protrusion. This was done for each of the 3DMMs built for the 4 data sets summarised in Table 4.1 that is presented in Section 4.2. Figure 4.7 shows a schematic view of the stages of evaluating the resulting 3D motions from the side-view.

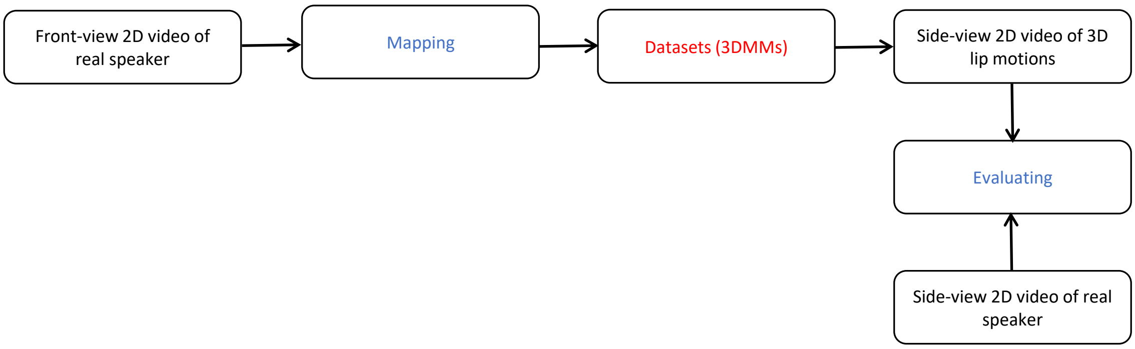

Front-view 2D video of real speaker
Mapping
Datasets (3DMMs)
Side-view 2D video of 3D lip motions
Evaluating
Side-view 2D video of real speaker

Figure 4.7 Schematic view of the stages of the evaluation process of 3D lip motion from the side-view.

Extraction of Side-View Features

For each video, ground-truth 2D side-view video or the side-view (2D) of the corresponding 3D animation, a reference plane parallel to the frontal plane is determined manually (see [4, 75, 149]). This is illustrated as a black line in Figure 4.8. From this a crop area for the mouth is calculated according to the dimensions specified by the red dotted square shown in Figure 4.8. The lip profile is then segmented from the background (by mapping to black and white and thresholding (see Figure 4.9)) and a lip protrusion value is calculated, defined by the horizontal distance between the foremost point of the upper lip and the previous calculated reference plane, shown by the green line in Figure 4.8. Given the upper lip protrusion value for each frame for both the 2D video for a real speaker and the corresponding 3D animation, the root mean square error (RMSE) over a sentence was used to evaluate the effectiveness of each 3DMM (see Section 4.4.2).

Normalisation of the Extracted Features

Following the same procedure that is presented in Chapter 3 (see Section 3.6.1), in order to correct the distance between the camera and the real speaker or the talking 3D head, all the landmarks were normalised by using the Euclidean distance between the root of the nose and the nose tip's point, since these were not affected by the articulations

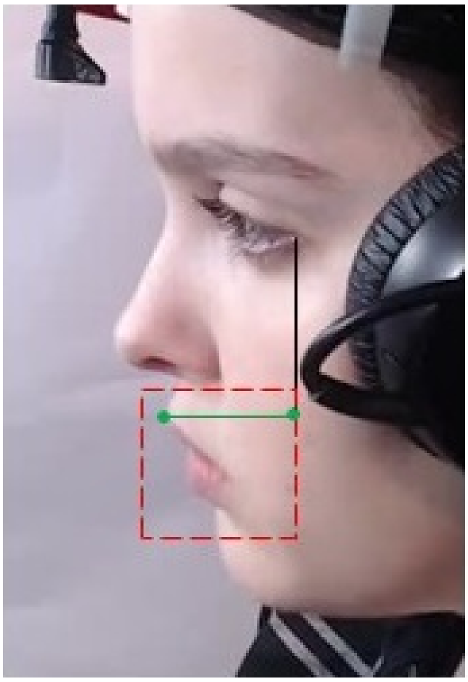


Figure 4.8 Definition of the critical measurements to determine the lip protrusion parameter in the side-view of a real speaker. The black vertical line represents a reference plane, the green horizontal line represents the measured distance for the lip protrusion, and the red dotted square represents the cropped mouth area.

[48]. The nose landmarks for both the real speakers and their corresponding 3D heads are detected following a similar procedure to detecting the upper lip protrusion. All visual articulatory features for the real speakers and their corresponding 3D heads were normalised by their corresponding maximum and minimum lip protrusion measurement in the videos, giving a [0-1] scale. Also, since the camera system used in the recording process was positioned differently on each real speaker's head (see Figures 3.13 and 3.14), the virtual camera used in calculations for the side-view videos of the animated 3D heads was assumed to be directed at the mouth and the vertical angle was manually adjusted to give a visual approximation to the real camera positioning for each mapped speaker. Figure 4.10 shows an example of consecutive frames for the side-view 2D video of a real speaker (ID: S 32), and different angles of the virtual camera for the corresponding 3D head. The virtual camera angle was adjusted vertically but not horizontally.

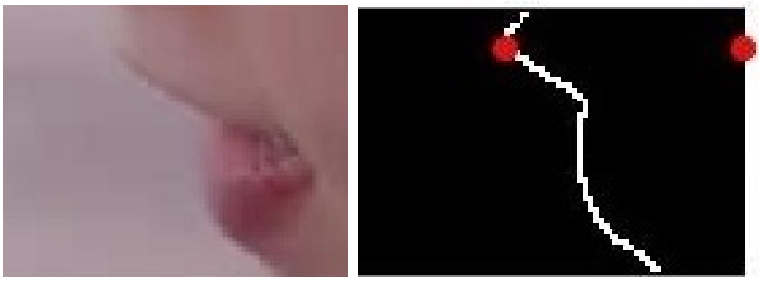


Figure 4.9 Cropped mouth area from side view frame of a real speaker (left), and the corresponding binary image and calculated contours with highlighted points for the lip protrusion parameter calculation (right).

4.4.2 Results and Discussion

Figure 4.11 shows an example of consecutive frames of the phoneme /w/ during utterance of the word "white" from the sentence "bin white at A 9 now" for a real speaker (ID: S17) and the corresponding 3D head for each data set used in building the 3DMM. This Figure shows how the lips are flattened when the front-view photo only was used to generate the initial neutral head pose in FaceGen, and 17 poses of the 3D head were used to train the 3DMM (i.e. Dataset 1). The performance of the animated lips improves (i.e. is more protruded) when front- and side-view photos are used to generate the initial neutral 3D head pose and further improves when a larger number of 3D head poses (i.e. different viseme intensities) are used for training the 3DMM. Figure B.1 shows how using front- and side-view photos contribute to enhance the protrusion of the lips. Figure 4.12 shows how the lips become more protruding during utterance of the phoneme /w/ of the word "white" from the sentence "bin white in N 3 now" for a real speaker (ID: S24), when further information is added during generating the initial neutral 3D head pose and training the 3DMM (i.e. Dataset 4).

Figure 4.13 shows the trajectories of the upper lip protrusion parameter (UL) for the side-view video of a real speaker (ID: S17) and the corresponding 3D head for the utterance "bin white at A 9 now". Generally, the trajectories of the 3D head that contains 161 poses and which are generated using front- and side-view photos (i.e. Dataset 4)

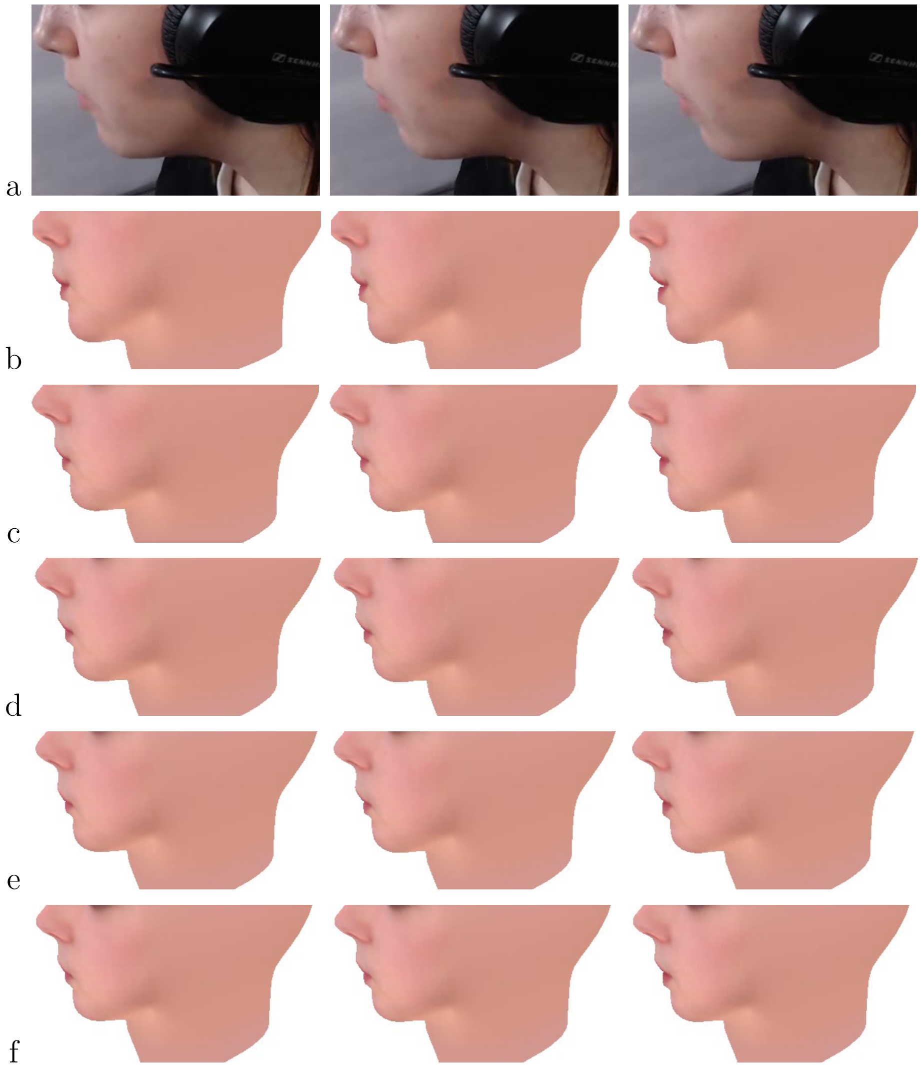


Figure 4.10 Consecutive frames for the side-view of a real speaker (ID: S 32) (a); and the corresponding 3D head (the virtual camera positioned at the centre 0° (b), 12°(c), 15°(d), 17°(e), and 21°(f) respectively).

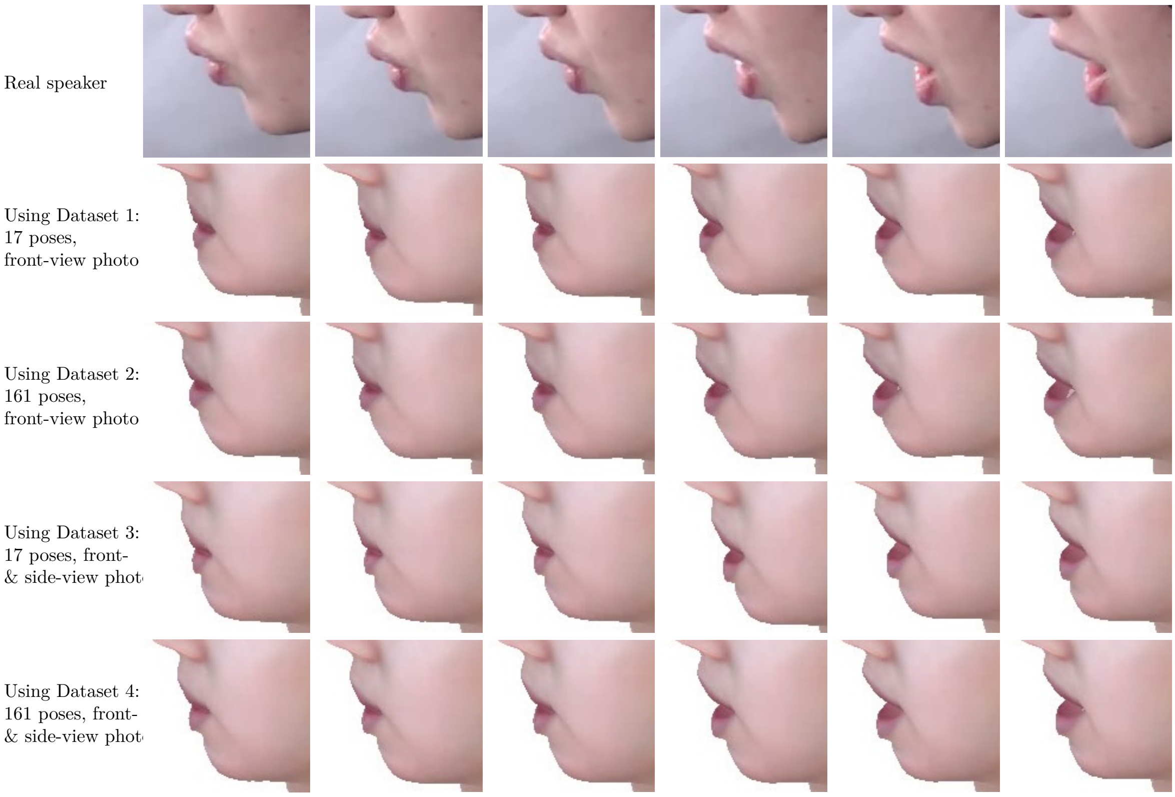

Real speaker
Using Dataset 1: 17 poses, front-view photo
Using Dataset 2: 161 poses, front-view photo
Using Dataset 3: 17 poses, front- & side-view phot
Using Dataset 4: 161 poses, front- & side-view phot

Figure 4.11 Consecutive frames of the phoneme /w/ for the utterance of the word "white" from the sentence "bin white at A 9 now" for a real speaker (ID: S17) and the 3D head produced for each data set.

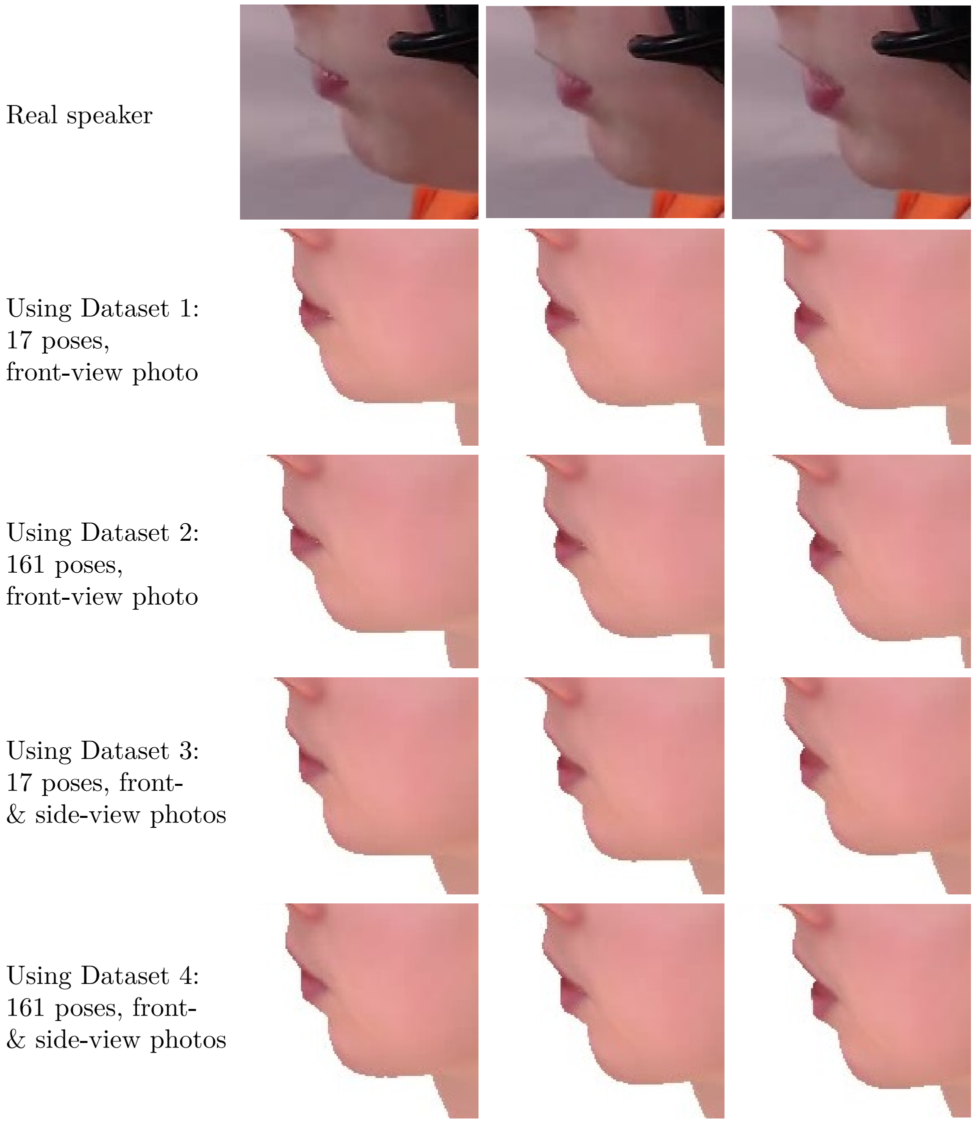

Real speaker
Using Dataset 1: 17 poses, front-view photo
Using Dataset 2: 161 poses, front-view photo
Using Dataset 3: 17 poses, front- & side-view photos
Using Dataset 4: 161 poses, front- & side-view photos

Figure 4.12 Consecutive frames of the phoneme /w/ for the utterance of the word "white" from the sentence "bin white in N 3 now" for a real speaker (ID: S24) and the 3D head produced for each data set.

are much closer to the ground truth trajectories. The trajectories of the 3D head that contains 161 poses and which are generated using a front-view photograph (i.e. Dataset 2) show how the model performs properly for the bilabial phoneme /b/ and the rounding phonemes /w/ and /aw/. This is due to including a larger number of viseme intensities. The trajectories of the 3D head that contains 17 poses and which are generated using front- and side-view photos (i.e. Dataset 3) are more adequate for the open lips phonemes /ay/, /ae/, /ey/ and /ay/. This confirms the importance of using the side-view photograph to create the initial 3D head pose that gives the 3D head a more protruding lips shape. Thus, using side-view photo to construct the initial 3D head pose, as well as a larger number of poses to train the 3DMM, enhances the performance of the resulting 3D lip motions.

Table 4.4 shows the RMSE results averaged over 4 sentences for the upper lip protrusion parameter of the real speakers and their corresponding 3D heads. The 3DMMs that contain 161 poses and which are generated using both front- and side-view photographs (Dataset 4) give the lowest RMSE scores for all the speakers. For some speakers, the 3DMMs that contain 161 poses and which are generated using only front-view photo (Dataset 2) give lower RMSE scores, in comparison with the 3DMMs that contain 17 poses and which are generated using both front- and side-view photographs (Dataset 3). Therefore, using larger number of poses for training the 3DMMs improves the resulting 3D lip motions, in addition to using side-view photo in the construction of the initial neutral 3D head pose in FaceGen software. This can be observed from Figure 4.14 that shows error bars of the upper lip protrusion parameter (UL) of the corresponding 3D heads of each real speaker. Uncertainty in the lip protrusion measurement is calculated by adding and subtracted 2 pixels to and from all the features that were used to determine this measurement.

For the 3D heads that contain 161 poses, a t-test suggests a significant difference in RMSE results for the 3D heads that use front- and side-view photos (Dataset 4) versus front-view photos only (Dataset 2) (p=0.0094) for the upper lip protrusion. Also, there

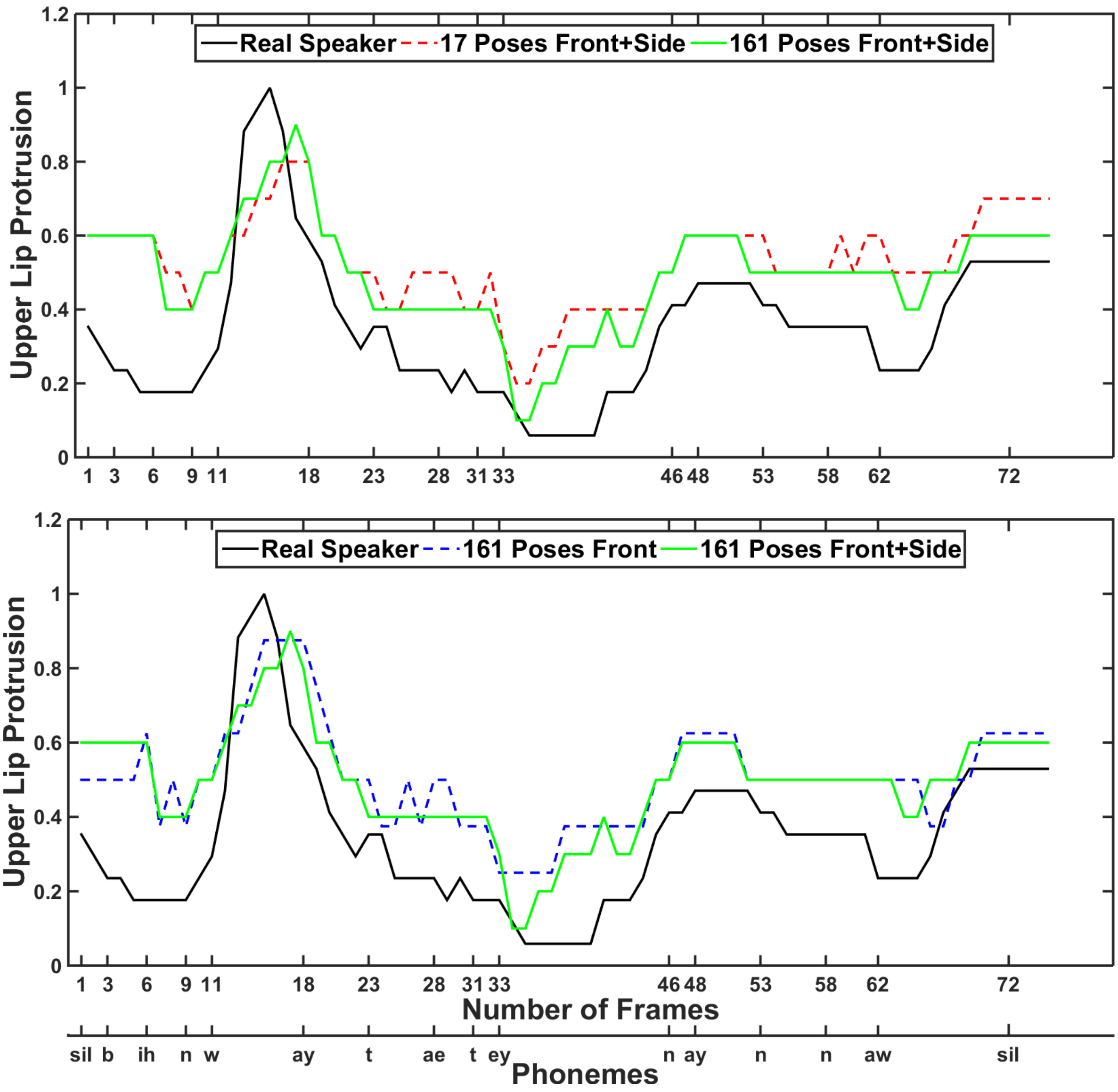


Figure 4.13 Upper lip protrusion trajectories of 2D frames of the side-view video of a real speaker (ID:S17) and the corresponding 3D head, whilst uttering the sentence "bin white at A 9 now". Top: comparison of upper lip protrusion for models constructed with 17 and 161 poses (both with front- and side-view photos) against the real speaker; Bottom: comparison of upper lip protrusion for models constructed using front-view photo only and front- and side-view photos (both with 161 poses) against the real speaker.

is a significant difference between the 3D heads containing 17 and 161 poses that are generated using front- and side-view photos (Datasets 3 and 4, respectively) (p=0.0114).

ID	Front photo		Front+side photos	
	17 poses	161 poses	17 poses	161 poses
	UL	UL	UL	UL
S15	0.318	0.291	0.278	**0.236**
S17	0.262	0.217	0.219	**0.157**
S20	0.248	0.256	0.256	**0.246**
S24	0.312	0.307	0.244	**0.224**
S32	0.303	0.338	0.349	**0.266**
S48	0.291	0.275	0.297	**0.258**

Table 4.4 The RMS error averaged over 4 sentences for upper lip (UL) protrusion of the real speakers and their corresponding 3D heads. Values in bold indicate lowest RMSE. UL error=±0.001.

The side-view evaluation of the 3D lip motions for each 3DMM confirmed how using the side-view photograph to create the initial 3D head pose and larger number of 3D head poses to train the 3DMMs enhanced the performance of the resulting 3D lip motions. Further experiments can be conducted to improve the results, which involve using different horizontal and vertical angles for the virtual camera to predict the angle of the camera system that was used to record the real speakers' data. Moreover, the lower lip protrusion of the 3DMMs can be analysed to investigate the impact of using different amounts of data during creation of the 3DMMs on the lower lip as well.

4.5 Summary

This chapter has presented a 3D talking head based on fitting a 3DMM, created using synthetic data, to 2D video frames of a real speaker. Different amount of data were used to create different 3DMMs for each real speaker. For each 3DMM, two sets of photographs (front-view photograph only or front- and side-view photographs) were used for generating the initial neutral 3D head pose. Then different numbers of 3D head poses (17 or 161 poses) were used for training. The evaluation of the effectiveness of each

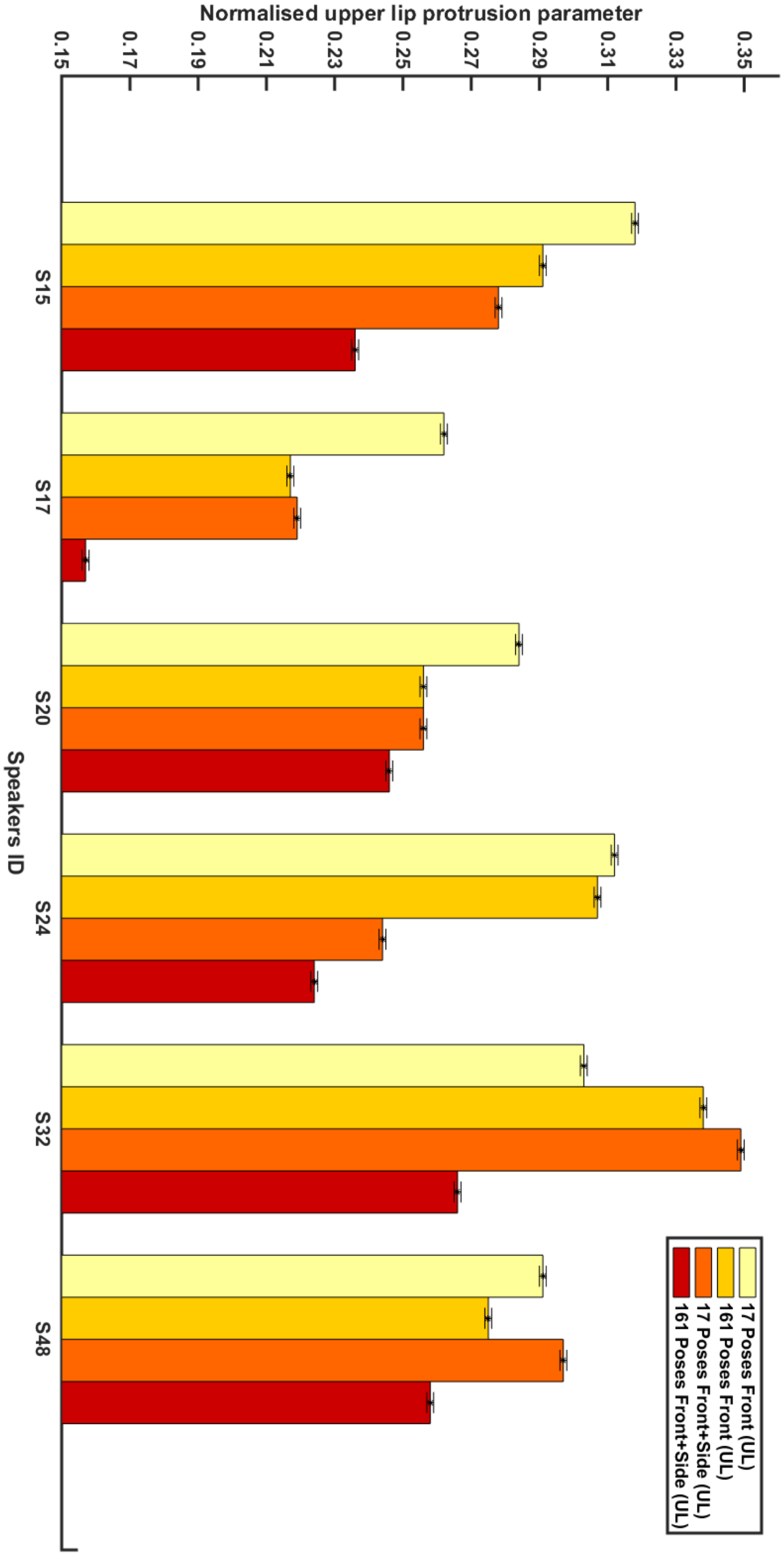


Figure 4.14 Error bars of the upper lip protrusion parameter (UL) of the corresponding 3D heads of each real speaker.

3DMM from both front and side-views was presented. The performance of the animated 3D lip motion was evaluated using ground truth data to compare against.

For the front view evaluation, two articulatory measurements were extracted and calculated from front-view 2D videos of the real speakers and the 3D lip motions: the width and height of the mouth aperture. One articulatory measurement which is the upper lip protrusion was extracted and calculated from side-view 2D videos of the real speakers and their corresponding 3D heads for the side-view evaluation. The RMSE was used over each sentence to calculate the difference between the articulatory measurements of the real speaker and the 3D animation. In comparison with the ground truth videos, the results of the front and side-view evaluations show that using both front- and side-view photos in the construction of the neutral pose of the 3D head improves the results in comparison to just using a front-view photo. In addition, increasing the number of 3D head poses (using different viseme intensities) to train the 3DMM further improves the performance of the 3D lip motions from both front and side-view. These results confirm the importance of using both front- and side-view photos for constructing the neutral pose of the 3D head, and different intensities of each viseme for training the 3DMM. The next chapter will investigate the impact of the spatial relations of real speakers' facial features on the resulting 3D lip motion.

Chapter 5

Mapping Non Similar Faces

5.1 Introduction

In Chapter 4, a data-driven approach to animating a 3D talking head using the tracked 2D lip motions of a corresponding real speaker was presented. This technique illustrates how using different amounts of data during the creation of a 3DMM that corresponds to a real speaker affects the performance of 3D lip motions. This chapter will investigate the impact of mapping between a real speaker and a non-corresponding 3DMM on the resulting 3D lip motions. Realistic 3D lip motions are used for many applications, such as movies and games where there is a convergence between customers and 3D animation presented through historical or believable characters that promote emotional, immersive content. For this kind of application, mapping between non-similar faces (i.e. mapping between videos of a real speaker and a non-corresponding 3DMM) has been chosen as a case study to determine the extent to which real speakers match 3DMMs in facial features to achieve sufficient 3D lip motions. The results of this study provide a greater understanding of the impact of similarities and differences in facial features between real speakers and 3DMMs on the resulting 3D lip motions. The results could therefore define criteria of facial feature classification that enable the animation of lips on a 3D head using videos of several real speakers.

To define these criteria, a classification of the facial features of real speakers is necessary. Thus, a facial metrical analysis of speakers of the Audio-Visual Lombard Grid Speech Corpus [10] was applied. The facial features of each speaker were classified into three classes: low, middle and high. Two dimensions of mouth features were investigated in this thesis: vertical mouth height and mouth width. Based on this, the mapping between real speakers and non-corresponding 3DMMs was investigated. 2D videos of a real speaker were mapped to a 3DMM that corresponded to a different speaker who was classified under the same or a different class.

This chapter is organised as follows: facial features classifications of real speakers is presented in Section 5.2. History of facial features classifications is reviewed in Section 5.2.1. Metrical analyses of facial features of the audio-visual Lombard Grid Speech corpus's speakers is presented in Section 5.2.2. Measurements used to describe indices of the facial features are presented in Section 5.2.3. The evaluation of this process is presented in Section 5.3, where Section 5.4 presents the objective test results and Section 5.5 presents the subjective test results. Finally, Section 5.6 concludes this chapter.

5.2 Facial Features Classification

5.2.1 Background

The most commonly used method of identifying individuals is facial photographs. Several methods can be used to analyse facial morphology in two photographs to be compared (face mapping [56]): superimposition, morphological characteristics, anthropometrical measurements and morphometrics (combination of morphology and measurements).

- **Superimposition:** In this method, a known image is placed on top of another to compare the two [15]. The outlines of the face on the two images are traced to be fitted over each other. For example, the face outline of an individual is placed over a suspect's face. More reference points on the face can be used. Additionally, a combination of mixers, monitors and video cameras can be used to fade two photographs into each other.

- **Morphological Characteristics:** In this method, facial features are classified morphologically into relevant categories. Then, a match between two or more photographs can be detected by comparing the various categories. Penry and Ryan [201] suggested dividing the face into different morphological regions and then classifying each region into categories or classes. In this study, faces were investigated by examining the morphological characteristics and then categorising them into proper classes. The face outlines were categorised into three classes: angular, rounded and mixed. The facial features were classified by dividing the face into sections, as described below:

  1. The head is divided into four equal horizontal portions: top of the head to the hairline, hairline to cranium, cranium to bottom of the nose and bottom of the nose to end of the chin.
  2. The face is divided into three equal horizontal portions: hairline to cranium, cranium to bottom of the nose and bottom of the nose to end of the chin.
  3. The area between the bottom of the nose and the end of the chin is horizontally divided into three equal portions.

  Based on these portions, facial features can be classified. For example, the ears are one-third of the length of the face [201]. In the case of a thin face, if the ears are more than one-third, they are classified as large. Following the same procedure, the rest of the facial features can be classified.

- **Anthropometric Measurements:** In this method, several measurements between facial landmarks are taken. Then, the facial features are classified using indices to avoid using an absolute size that can be altered when enlarging the photographs. Hrdlicka [220] used the indices to classify whole body parts; he was inspired by Martin and Saller [170], who used the indices to calculate the proportions of facial features in 1914. Hrdlicka [220] included some indices that describe faces and skulls, such as the cephalic index (cranial breadth/cranial length *100) and the total facial index (menton-nasion height/diameter bizygomatic maximum *100). However, Hrdlicka [220] conducted his studies on living subjects rather than photographs. Farkas [90] compared the measurements taken from a live subject and those taken from photographs to validate their reliabilities.

- **Morphometrical Methods:** In this method, the measurements and morphology of the face are combined to generate a reliable analytic procedure for facial identification. Iscan [130] classified facial features into different morphological classes using measurements from photographs (this procedure is called photoanthropometry). He used standard facial landmarks visible in photographs so that several measurements between landmarks could be taken and indices could be calculated. Different morphological classes can be identified from these indices to calculate the proportions of the face.

5.2.2 Chosen Approach

To analyse the facial features of the Audio-Visual Lombard Grid Speech Corpus's speakers, a method presented by Roelfose et al. [209] was followed. They used morphometrical methods to classify the facial features of South African males in photos to investigate common and rare features in this community. This method is based on both measurements and morphology of the face, which provide a reliable procedure for facial features classification. Furthermore, it is based on indices measurements which solve the problem of using an absolute size that can be changed when the photos are enlarged.

Because lip shapes are affected by facial movements such as smiling and crying, the lips must be assessed when the subject has a neutral face shape: the eyes are open, the lips make gentle contact and the jaw, neck and facial muscles should not be stretched or contracted [44]. Thus, videos of the Audio-Visual Lombard Grid Speech Corpus were investigated for each speaker (27 speakers) to select the appropriate frame (all video frames were chosen when the speakers were silent). The selected frames were processed using Faceware Analyser software to obtain x and y coordinates for each landmark. Figure 3.15 shows the landmarks that were utilised to take the facial measurements. A description of the used landmarks is presented below, where L1, L2, etc. refer to landmark 1, landmark 2, etc.

L1 Nasion (n): This landmark is placed on the midpoint between the inner corners of the eyes.

L2 Endocanthion (en): This landmark is placed on the inner corner of the eye.

L3 Exocanthion (ex): This landmark is placed on the outer corner of the eye.

L4 Alare (al): This landmark is placed on the border of the nostril wing of the nose.

L5 Subnasale (sn): This landmark is placed on the lower border of the nasal septum.

L6 Labiale superius (ls): This landmark is placed on the midpoint of the outer contour of the upper lip.

L7 Stomion1 (sto1): This landmark is placed on the midpoint of the inner contour of the upper lip.

L8 Stomion2 (sto2): This landmark is placed on the midpoint of the inner contour of the lower lip.

L9 Labiale inferius (li): This landmark is placed on the midpoint of the outer contour of the lower lip.

L10 Gnathion (gn): This landmark is placed on the bottom of the chin.

L11 Cheilion (ch): This landmark is placed on the outer corner of the lips.

L12 Zygion (zy): This landmark is placed on the zygomatic arch.

To calculate the indices, several measurements were taken from the frames of the speakers. As shown in Figure 3.16, 13 measurements were taken from each frame using the Euclidean distance between the predetermined facial landmarks (Figure 3.15). A brief description of each measurement is presented below, where D1, D2, etc. refer to distance 1, distance 2, etc.

D1 Gnathion to nasion `(L10_L1):` This measurement is used to determine the height of the face. It is measured from the midpoint between the inner corners of the eyes to the lower visible point of the chin.

D2 Zygion to zygion `(L12_L12):` This measurement determines the width of the face below the level of the eyes.

D3 Exocanthion to exocanthion `(L3_L3):` This measurement assesses the distance between the outer corners of the eyes.

D4 Endocanthion to endocanthion `(L2_L2):` This measurement assesses the distance between the inner corners of the eyes.

D5 Nasion to subnasale `(L1_L5):` This measurement determines the length of the nose from the middle of the nasal root to the lower border of the nasal septum.

D6 Alare to alare `(L4_L4):` This dimension measures the width of the nose between the borders of the nostril wings of the nose.

D7 Labiale superius to labiale inferius `(L6_L9):` This measurement determines the height of the lips. It assesses the distance between the midpoint of the outer contour of the upper lip and that of the lower lip.

D8 Cheilion to cheilion `(L11_L11):`This measurement determines the width of the mouth. It assesses the distance between the outer corners of the mouth.

D9 Labiale superius to stomion1 `(L6_L7):` This measurement determines the thickness of the upper lip. It assesses the distance between the midpoints of the outer and inner contours of the upper lip.

D10 Labiale superius to stomion2 (L8_L9): This measurement determines the thickness of the lower lip. It assesses the distance between the midpoints of the inner and outer contours of the lower lip.

D11 Labiale inferius to gnathion (L9_L10): This measurement determines the vertical height of the chin. It assesses the distance between the midpoint of the outer contour of the lower lip and the lowest midpoint on the chin.

D12 Subnasale to Labiale superius (L5_L6): This measurement assesses the distance between the lower border of the nasal septum and the midpoint of the outer contour of the upper lip.

D13 Subnasale to gnathion (L5_L10): This measurement assesses the distance between the lower border of the nasal septum and the lowest midpoint on the chin.

5.2.3 Basic Statistics and Indices for Each Speaker

The described measurements in the previous section were used to calculate 12 indices. Each index was computed by dividing the smaller measurement by the larger measurement and multiplying the quotient by 100. The reason for using the indices was to nullify the effect of absolute size. This means that any difference in the size of the face on the frame will not affect the outcome of the results. For each index, the mean, standard deviation and ranges were computed (Table 5.1).

The ranges of each index were used to categorise the features into different morphological classes. The classes of each index were created by investigating the distributional properties of the data with box-whisker plots. Outliers were defined as any value more than 1.5 away from the top or bottom of the box (interquartile range). The classes were determined by the outliers. For example, for the index of the mouth width, the mean was 52.18, and the standard deviation was 4.02 (Table 5.1). The outliers for this index were 49.18 and 54.80. Using these values, three ranges were determined, with the lowest comprising values less than 49.18 (thus covering the range between 44.57 and 49.17), the middle class comprising values between 49.18 and 54.80 and the third class comprising values greater than 54.80 (thus covering the range between 54.81 and 58.98). Values less than 49.18 constituted a narrow mouth width, between 49.18 and 54.80 middle, and values greater than 54.80 constituted a wide mouth width in relation to the distance between the outer corners of the eyes. Figure 5.1 shows the classification of the 12 indices for each speaker in the corpus with 80% of the calculated confidence intervals for each class of each index, where yellow, orange and red circles represent low, middle and high classes, respectively.

Index	Mean	Standard Deviation	Min	Max
I1- Facial	86.24	4.96	77.106	97.51
I2- Intercanthal	34.69	2.56	29.93	40.55
I3- Nasal	80.75	9.05	67.24	95.97
I4- Nasofacial	44.58	4.14	34.20	52.16
I5- Nose-face width	30.78	2.50	25.10	36.54
I6- Lip area	32.87	5.11	23.13	42.60
I7- Vertical mouth height	15.22	2.50	11.94	21.20
I8- Upper lip thickness	4.85	1.46	2.48	8.32
I9- Lower lip thickness	10.31	1.33	8.31	13.27
I10- Mouth width	52.13	4.02	44.57	58.98
I11- Chin size	24.97	3.30	17.87	32.06
I12- Nose-Upper-Lips	27.56	2.88	18.98	32.68

Table 5.1 Basic descriptive statistics for the indices.

5.3 Mapping between Non Similar Faces

In this experiment, the effect of differences and similarities in the facial features between real speakers and 3DMMs on the resulting 3D lip motions was investigated objectively and subjectively. 2D video frames of a real speaker were mapped to a corresponding 3DMM and a 3DMM that corresponded to a different real speaker. The mapping between non-similar faces was based on the classes of two of the 12 indices: index 7 (vertical mouth height) and index 10 (mouth width). As proven in the previous chapters, the best performance of 3D lip motions can be achieved when front- and side-view photographs are used to create the initial 3D head pose, and 161 poses are used to train the 3DMM.

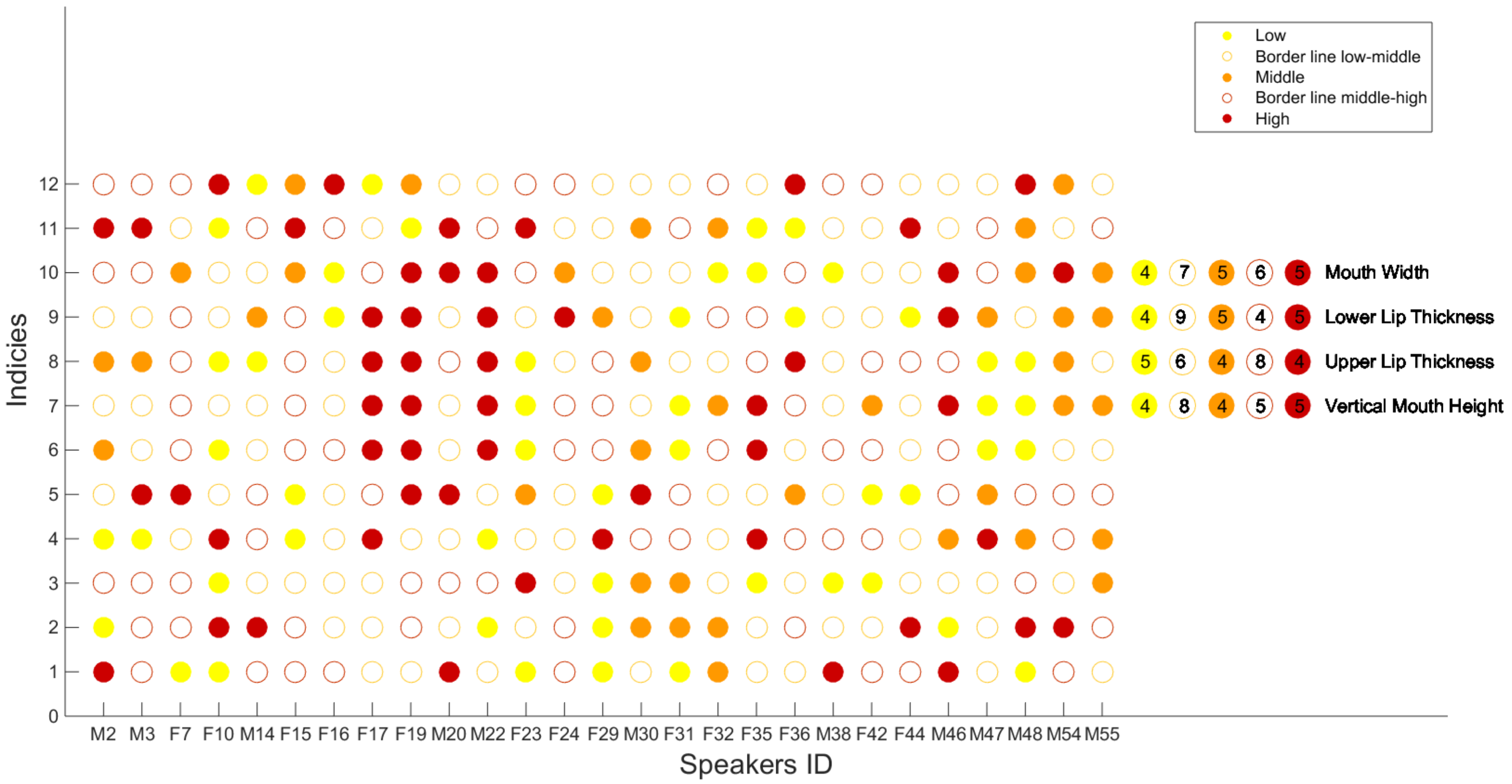


Figure 5.1 Classification of indices for each speaker of the audio-visual Lombard grid speech corpus (80% confidence level for each class of each index). The x axis shows the speaker's ID (where M refers to male speaker and F refers to female speaker) and the y axis shows the indices' number.

Therefore, the same procedure was followed to generate a 3DMM for each speaker classified under one of the three classes of indices 7 and 10.

5.4 Objective Evaluation

For each speaker, four plain sentences from the front-view video files were chosen to be mapped to the corresponding 3D head and the non-corresponding 3D heads. For example, referring to index 7 in Figure 5.1, a real speaker (ID: S17) was classified into the high class. 2D videos of this speaker were mapped to the corresponding 3D head and the 3D heads that corresponded to other speakers in the same class (high class) (i.e. speaker IDs: S19, S22, S35 and S46), low class (i.e. speaker IDs: S23, S31, S47 and S48) and middle class (i.e. speaker IDs: S32, S42, S54 and S55). An example of this process is shown in Figure 5.2. Next, 2D videos of the resulting 3D lip motions of each head were compared with the original ground-truth 2D videos. For comparison, the procedure presented in Section 3.6 was followed. This included tracking the facial features in the ground-truth 2D video and the front-view (2D) of the 3D animation, normalising and scaling the extracted features and calculating two geometric articulatory measurements: width (W) and height (H) of the mouth aperture.

Results and Discussion

Index 7 (Vertical Mouth Height)

Tables 5.2 shows the RMSE results averaged over four sentences for the width and height of the mouth aperture of real speakers in the low class of index 7 and their corresponding, non-corresponding middle and non-corresponding high 3D heads. From this table, it is clear that the RMSE results varied when 2D videos of real speakers were mapped to the non-corresponding low 3D heads or the non-corresponding middle 3D heads . When the 2D videos were mapped to the non-corresponding high 3D heads, the corresponding 3D heads gave the lowest RMSE scores for height for all speakers and for width for two

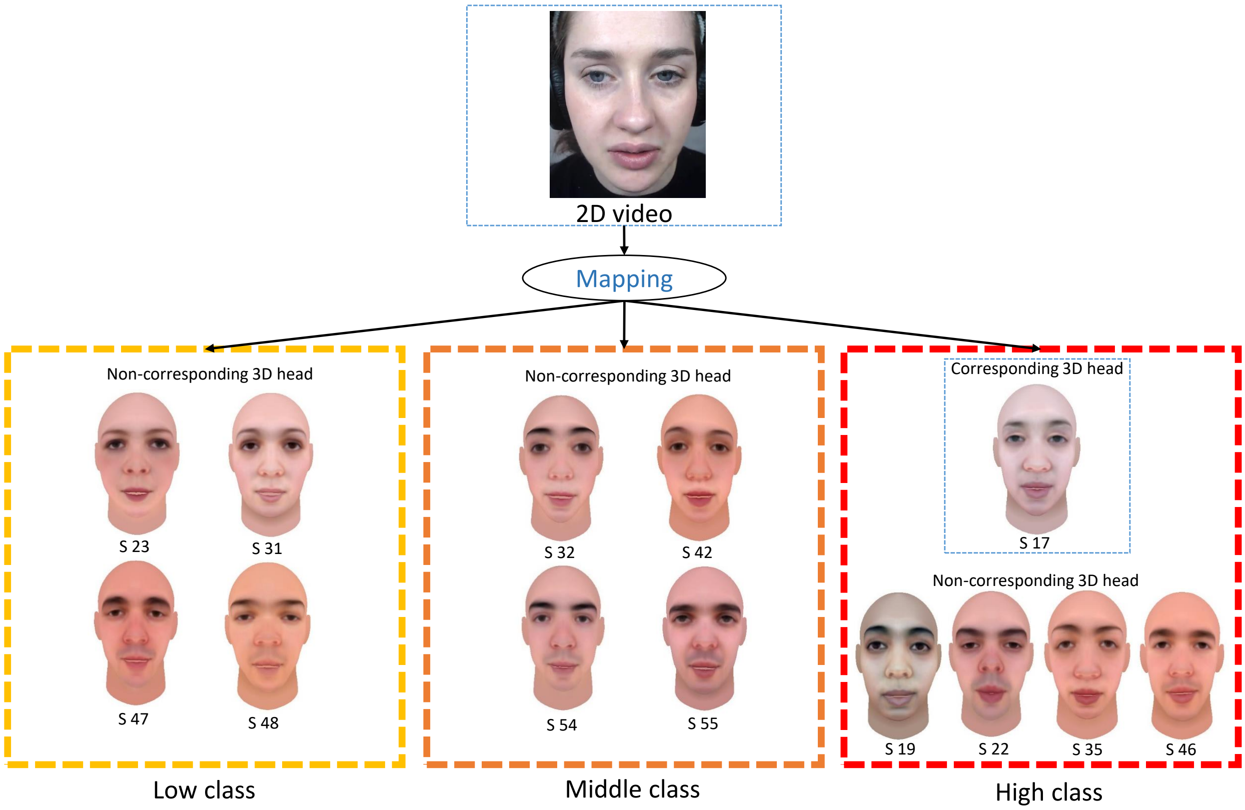


Figure 5.2 An example of the mapping process between 2D video frames of a real speaker (ID: S17) who classified under the high class of index 7, the corresponding 3D head, and the non-corresponding 3D heads.

out of four speakers who were classified in the low to middle class (ID: S31) and the middle class (IDs: S48) of index 10 (mouth width). The corresponding 3D heads of the real speakers (IDs: S23 and S47) failed to give the lowest score for width because their corresponding real speakers were classified under the middle to high class of index 10, which is very close to most of the non-corresponding high 3D heads.

For the corresponding low 3D head of each real speaker and the non-corresponding low 3D heads, a t-test suggested no significant difference in RMSE results for width and height. Additionally, no significant difference was found between the corresponding low 3D heads and the non-corresponding middle 3D heads for all speakers for height and three out of four speakers for width. The significant difference in width given by the corresponding low 3D head of a real speaker (ID: S31) was due to its low mouth width.

A significant difference was found between three out of four of the corresponding low 3D heads and the non-corresponding high 3D heads for height. The corresponding low 3D head of a real speaker (ID: S48) suggested no significant difference. This may be due to the large distance between the nose tip and the upper lip (index 12), which reduces the height of the mouth aperture. For width, three out of four of the corresponding 3D heads showed a significant difference. The corresponding low 3D head of a real speaker (ID: S23) suggested no significant difference; this may because it was classified in the middle to high class of indices 10 and 12.

Figure 5.3 provides an example of consecutive frames of the phoneme /ih/ during the utterance of the word "in" from the sentence "bin white in O 7 now" for a real speaker (ID: S47) who was classified in the low class of index 7, the corresponding 3D head, the non-corresponding middle 3D heads and the non-corresponding high 3D heads. This figure illustrates how the non-corresponding high 3D heads failed to detect the uttered phoneme and that the mouth was completely closed due to lip thickness, while the corresponding 3D head and the non-corresponding middle 3D heads gave the closest mouth shapes to the real speaker. These findings and the clear visual discrimination in the 3D lip motions presented by each 3D head suggest that an appropriate animation can be achieved by mapping between 2D videos of real speakers and the corresponding 3D head or non-corresponding middle 3D heads but not the non-corresponding high 3D heads.

Figure 5.4 shows the trajectories of the width and the height parameters of the mouth aperture for the real speaker (ID: S31) classified under the low class of index 7, the corresponding 3D head, the non-corresponding middle 3D head (ID: S32) and the non-corresponding high 3D head (ID: S19), whilst uttering the sentence "set white at D zero please". Whilst all the trajectories generated using the animation pipeline generally follow the real speaker's trajectory, the trajectories of the corresponding 3D head and the non-corresponding middle 3D head closer to the ground truth trajectories. For the width, the trajectory of the non-corresponding high 3D head shows a marked rise for bilabial phoneme /p/, which confirms that the lips stretch due to sharp touch between the upper and lower lips caused by the lip thickness. Also, for the height, the trajectory of this 3D head shows steep drops for dental phonemes such as /s/, /t/ and /d/, which confirms that the lips are semi-closed during uttering these phonemes.

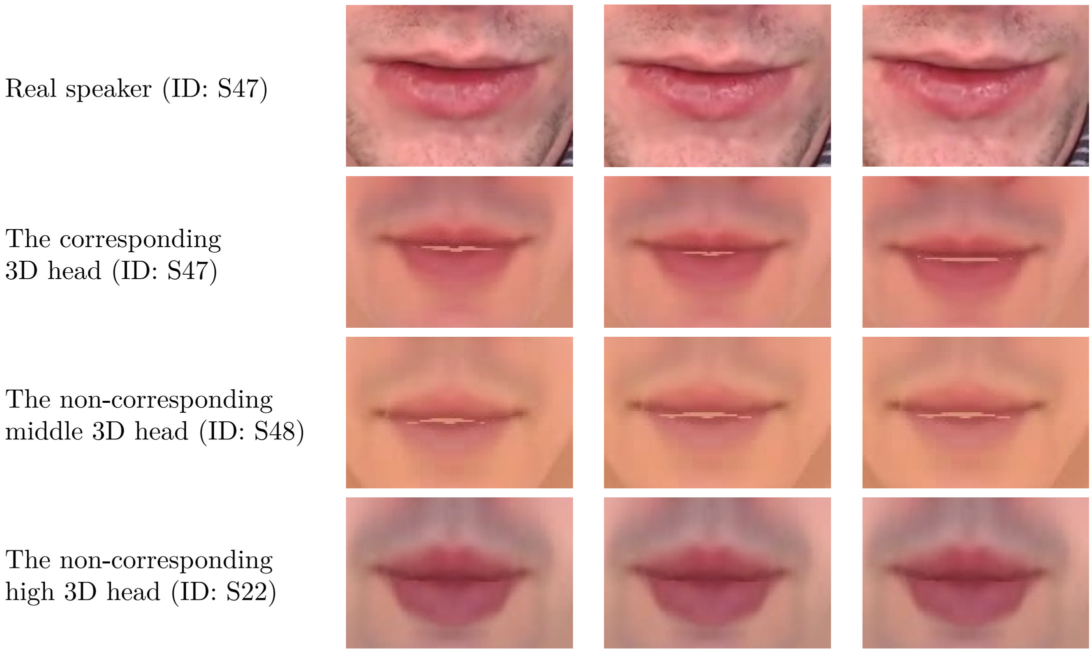

Real speaker (ID: S47)
The corresponding 3D head (ID: S47)
The non-corresponding middle 3D head (ID: S48)
The non-corresponding high 3D head (ID: S22)

Figure 5.3 Consecutive frames of the phoneme /ih/ during utterance of the word "in" from sentence "bin white in O 7 now" for a real speaker (ID: S47) who is classified under the low class of index 7, the corresponding 3D head, the non-corresponding middle and the non-corresponding high 3D heads.

Table 5.3 shows the RMSE results averaged over four sentences for the width and height of the mouth aperture of real speakers classified in the middle class of index 7 (vertical mouth height), their corresponding 3D heads, the non-corresponding low 3D heads and the non-corresponding high 3D heads. This table shows variations in the RMSE results for width due to variations in the mouth width of real speakers. When 2D videos of the real speakers were mapped to their corresponding 3D heads and the non-corresponding middle 3D heads, t-test results suggested no significant difference in RMSE results for height for all speakers and for three out of four speakers for width. The corresponding 3D head of a real speaker (ID: S54) suggested a significant difference for the width; this may be due to a large mouth width (index 10). For the non-corresponding

2D video	Corresponding 3D head (low)		Non-corresponding 3D head (low)										T-test (P value)	
			S 23		S 31		S 47		S 48					
	W	H	W	H	W	H	W	H	W	H			W	H
S 23	0.270	0.131			0.232	**0.114**	0.253	0.137	**0.188**	0.145			0.1399	0.9195
S 31	0.193	0.163	0.208	0.201			**0.189**	0.229	0.219	**0.125**			0.2946	0.5523
S 47	0.218	0.099	0.212	**0.081**	0.227	0.149			**0.140**	0.082			0.4501	0.8448
S 48	0.149	**0.071**	**0.131**	0.101	0.157	0.083	0.188	0.073					0.6168	0.2153
			Non-corresponding 3D head (middle)											
			S 32		S 42		S 54		S 55					
	W	H	W	H	W	H	W	H	W	H			W	H
S 23	0.270	**0.131**	**0.241**	0.135	0.256	0.168	0.265	0.185	0.277	0.152			0.2696	0.0734
S 31	**0.193**	0.163	0.229	**0.159**	0.281	0.206	0.245	0.218	0.248	0.213			0.0132	0.0767
S 47	0.218	0.099	0.250	**0.091**	0.234	0.131	**0.170**	0.103	0.173	0.098			0.6233	0.4975
S 48	0.149	**0.071**	0.157	**0.071**	0.187	0.078	**0.139**	0.104	0.166	0.072			0.2762	0.2772
			Non-corresponding 3D head (high)											
			S 17		S 19		S 22		S 35		S 46			
	W	H	W	H	W	H	W	H	W	H	W	H	W	H
S 23	0.270	**0.131**	**0.209**	0.193	0.234	0.202	0.278	0.211	0.244	0.208	0.262	0.179	0.1059	0.0003
S 31	**0.193**	**0.163**	0.321	0.185	0.278	0.223	0.246	0.286	0.243	0.216	0.251	0.213	0.0070	0.0209
S 47	0.218	**0.099**	0.198	0.145	0.204	0.137	**0.149**	0.191	0.188	0.143	0.168	0.194	0.0226	0.0073
S 48	**0.149**	**0.071**	0.198	0.083	0.184	0.084	0.164	0.145	0.157	0.101	0.156	0.096	0.0508	0.0532

Table 5.2 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers classified under the low class of index 7 (vertical mouth height), their corresponding 3D heads and the non-corresponding 3D heads. Values in bold means the decreased RMS error. The last column shows p value of the t-test results between each corresponding 3D head and the non-corresponding 3D heads for width and height.

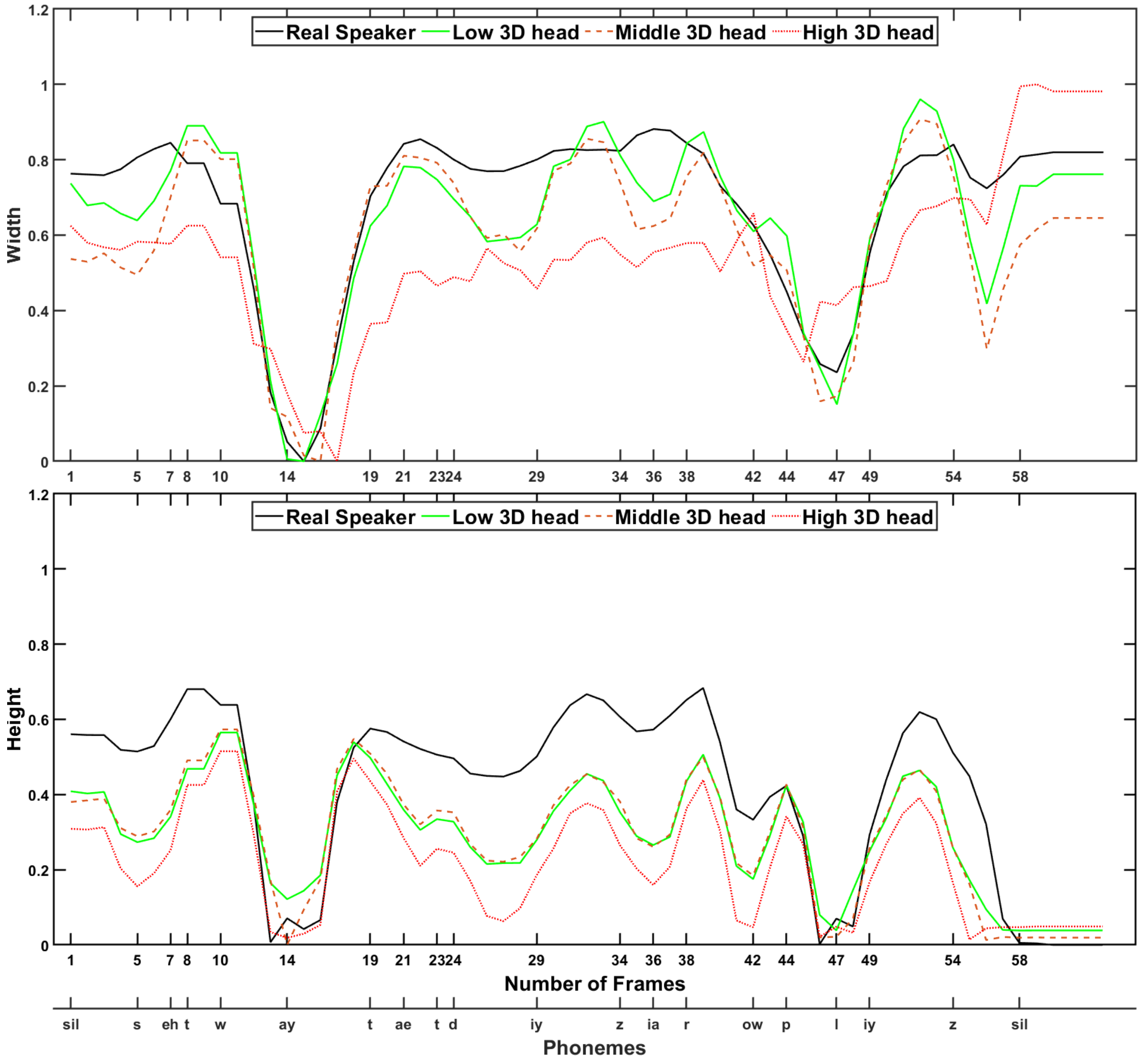


Figure 5.4 Width and height of mouth trajectories of 2D frames of the real speaker (ID: S 31) classified under the low class of index 7, the corresponding 3D head, the non-corresponding middle 3D head (ID: S 32) and the non-corresponding high 3D head (ID: S 19), whilst uttering the sentence "set white at D zero please".

high 3D heads, t-test results showed a significant difference in the RMSE scores for width for the corresponding 3D head of a real speaker (ID: S32); this is probably due to its small mouth width.

For the height, t-test results showed a significant difference for two of the corresponding middle 3D heads; this may because their corresponding real speakers (IDs: S32 and S42) were classified in the low class and the low to middle class of index 10, respectively. This makes the mouth of the non-corresponding high 3D heads shrink to fit the real speakers' mouths; thus, the lips are not closed or opened adequately. Figure 5.5 confirms these findings by showing an example of consecutive frames of the phoneme /b/ during the utterance of the word "bin" from the phrase "bin white at U three again" for a real speaker (ID: S32) classified in the middle class of index 7, the corresponding 3D head, the non-corresponding low 3D heads and the non-corresponding high 3D heads. These findings may confirm that the resulting 3D lip motions become sufficient and adequate when 2D videos of the real speakers classified in the middle class of index 7 are mapped to the corresponding 3D head, the non-corresponding middle 3D heads and the non-corresponding low 3D heads and when they are mapped to the non-corresponding high 3D heads that relate to real speakers who have a similar mouth width.

Table 5.4 shows the RMSE results averaged over four sentences for the width and height of the mouth aperture of real speakers classified in the high class of index 7 (vertical mouth height), their corresponding 3D heads, the non-corresponding low 3D heads and the non-corresponding middle 3D heads. From this table, it can be observed that the corresponding 3D head of a real speaker (ID: S19) gave the lowest RMSE scores for width for real speakers (IDs: S22 and S46); this may due to similarities in mouth width. Additionally, the non-corresponding low 3D heads (IDs: S23, S47 and S48) gave the lowest scores because their corresponding real speakers were classified in the middle class and the middle to high class of index 10, and most of the real speakers (IDs: S17, S19, S22 and S46) were classified in the middle to high class or the high class. This was confirmed by t-test results that showed no significant difference for width between the corresponding 3D heads and the non-corresponding low 3D heads.

2D video	Corresponding 3D head (middle)		Non-corresponding 3D head (middle)										T-test (P value)	
			S 32		S 42		S 54		S 55					
	W	H	W	H	W	H	W	H	W	H			W	H
S 32	0.111	**0.056**			0.122	0.089	**0.106**	0.112	0.174	0.059			0.3790	0.1837
S 42	**0.198**	0.083	0.261	**0.078**			0.248	0.138	0.220	0.088			0.0653	0.4274
S 54	0.281	0.132	**0.255**	0.152	0.263	**0.130**			0.256	0.149			0.0118	0.2324
S 55	**0.113**	0.107	0.154	**0.105**	0.126	0.118	0.134	0.117					0.0953	0.2687
			Non-corresponding 3D head (low)											
			S 23		S 31		S 47		S 48					
	W	H	W	H	W	H	W	H	W	H			W	H
S 32	**0.111**	0.056	0.119	0.068	0.123	**0.047**	0.144	0.075	0.163	0.089			0.0820	0.2141
S 42	0.198	0.083	**0.185**	**0.075**	0.224	0.125	0.255	0.092	0.208	0.130			0.2669	0.1870
S 54	0.281	**0.132**	**0.249**	0.151	0.265	0.180	**0.249**	0.158	0.273	0.133			0.0351	0.0943
S 55	0.113	0.107	0.119	**0.096**	0.119	0.111	**0.112**	0.108	0.122	0.104			0.0997	0.5385
			Non-corresponding 3D head (high)											
			S 17		S 19		S 22		S 35		S 46			
	W	H	W	H	W	H	W	H	W	H	W	H	W	H
S 32	**0.111**	**0.056**	0.185	0.132	0.145	0.102	0.135	0.193	0.176	0.145	0.161	0.123	0.0060	0.0055
S 42	0.198	**0.083**	0.205	0.140	0.209	0.090	0.200	0.171	0.229	0.156	**0.172**	0.098	0.6150	0.0394
S 54	0.281	0.132	0.226	0.148	**0.209**	0.147	0.241	**0.129**	0.267	0.136	0.131	0.194	0.0130	0.1931
S 55	0.113	0.107	0.102	**0.103**	**0.097**	0.111	0.116	0.114	0.099	0.127	0.116	0.110	0.0832	0.2022

Table 5.3 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers classified under the middle class of index 7 (vertical mouth height), their corresponding 3D heads and the non-corresponding 3D heads. Values in bold means the decreased RMS error. The last column shows p value of the t-test results between each corresponding 3D head and the non-corresponding 3D heads for width and height.

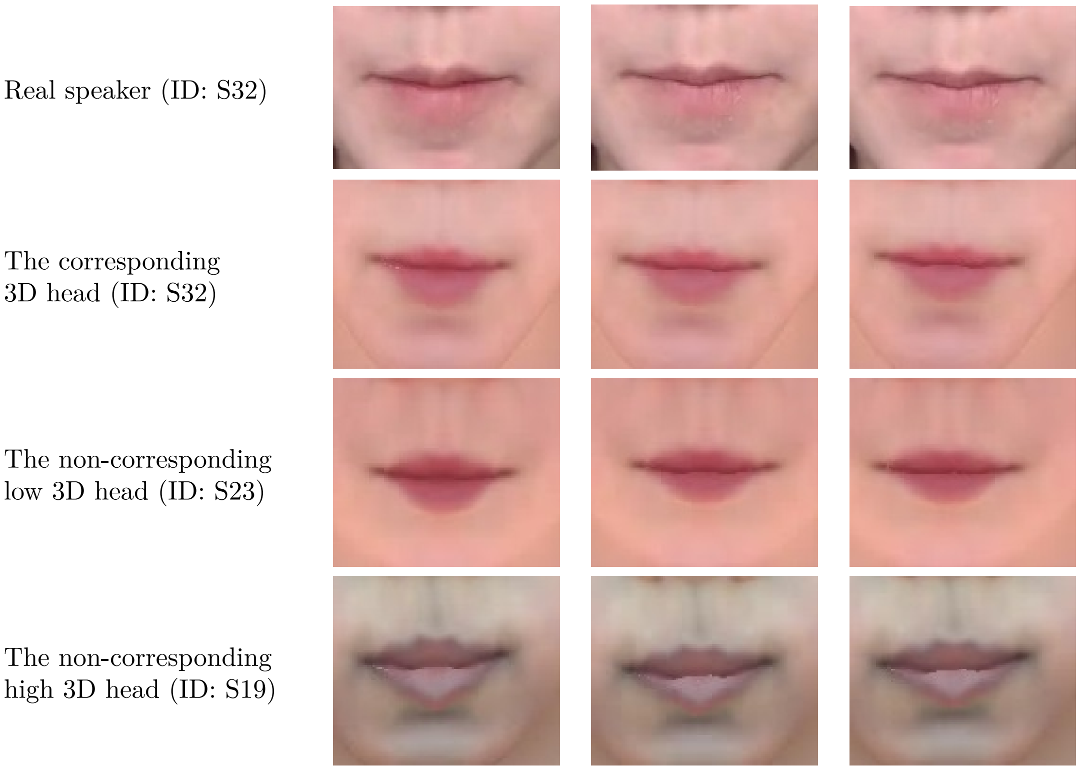

Real speaker (ID: S32)
The corresponding 3D head (ID: S32)
The non-corresponding low 3D head (ID: S23)
The non-corresponding high 3D head (ID: S19)

Figure 5.5 Consecutive frames of the phoneme /b/ during utterance of the word "bin" from sentence "bin white at U three again" for a real speaker (ID: S32) who classified under the middle class of index 7, the corresponding 3D head, the non-corresponding low 3D head and the non-corresponding high 3D head.

For the height, the corresponding high 3D heads gave the lowest score for most of the speakers from different classes. T-test results showed no significant difference in the RMSE scores for height between four out of five of the corresponding high 3D heads and the non-corresponding middle 3D heads. However, there was a significant difference between three out of five of the corresponding high 3D heads and the non-corresponding low 3D heads. Based on these findings, it can be concluded that mapping between 2D videos of real speakers classified in the high class of index 7 and the non-corresponding low 3D heads cannot achieve any reasonable 3D lip animations. While it is possible to achieve reasonable 3D lip motion using 2D videos of real speakers who have a middle vertical mouth height, to animate 3D heads corresponding to real speakers who have a high vertical mouth height, they must be similar in other facial features, such as lower or upper lip thickness (indices 8 and 9, respectively), mouth width (index 10) or the distance between the nose and the upper lip (index 12).

This is indicated by Figure 5.6 , which shows how the lips of the non-corresponding 3D heads fail to give the mouth shape of the phoneme /b/. This figure also shows how the non-corresponding middle 3D head (ID: S54) gives a semi-opened mouth shape due to its high mouth width, which is similar to that of the real speaker (ID: S22). However, it fails to deliver the correct mouth shape due to its middle upper and lower lip thickness. Figure 5.7 shows an example of consecutive frames of the phoneme /p/ during utterance of the word "please" from the phrase "lay white with A 5 please" for a real speaker (ID: S46) classified in the high class of index 7, the corresponding 3D head, the non-corresponding middle 3D heads and the non-corresponding low 3D heads. This figure illustrates how the non-corresponding middle 3D (ID: S54) heads gave a mouth shape more similar to that of the real speaker due to similarities in mouth width (index 10) and middle lower lip thickness. The non-corresponding low 3D heads (ID: S48) gave a more accurate mouth shape (semi-closed mouth shape) because of the large distance between the nose and the upper lip (index 12), the middle mouth width (index 10) and the middle upper and lower lip thickness (indices 8 and 9), while the non-corresponding low 3D heads (ID: S31) failed to give the correct mouth shape due to the low to middle mouth width and upper and lower lip thickness. The distortion in the texture around the mouth's corners of the non-corresponding low 3D heads (IDs: S48 and S31) was due to differences in mouth width between the real speaker and the 3D heads.

Figure 5.8 shows the trajectories of the width and the height parameters of the mouth aperture for the real speaker (ID: S22) classified under the high class of index 7, the corresponding 3D head, the non-corresponding middle 3D head (ID: S55) and the non-corresponding low 3D head (ID: S48), whilst uttering the sentence "set white with S 1 now". The trajectories of the corresponding 3D head closer to the ground truth trajectories for both width and height. For the width, what can be clearly seen in this figure is the steady decline of the trajectories of the non-corresponding low 3D head. For the height, the trajectories of the non-corresponding low 3D head show a marked increase for the rounding lips phonemes such as /w/ and /aw/, alveolar phonemes such as /th/, and dental phonemes such as /s/, /t/, and /n/, which confirms that the lips are widely opened during uttering these phonemes that require semi-opened mouth shape.

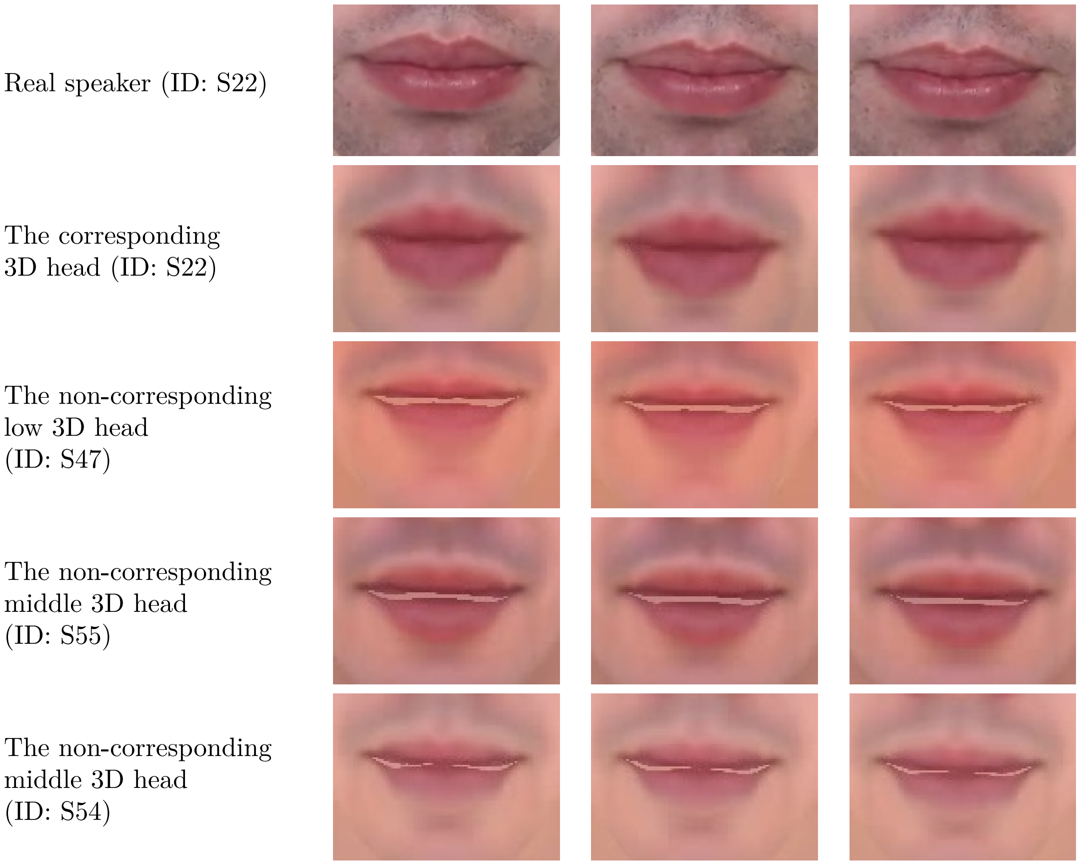

Real speaker (ID: S22)
The corresponding 3D head (ID: S22)
The non-corresponding low 3D head (ID: S47)
The non-corresponding middle 3D head (ID: S55)
The non-corresponding middle 3D head (ID: S54)

Figure 5.6 Consecutive frames of the phoneme /b/ during utterance of the word "bin" from sentence "bin green by Q zero again" for a real speaker (ID: S22) who classified under the high class of index 7 (first row), the corresponding 3D head (second row), the non-corresponding low 3D head (third row) and the non-corresponding middle 3D head (last row).

Index 10 (Mouth Width)

Table 5.5 shows the RMSE results averaged over four sentences for width and height of the mouth aperture of real speakers classified in the low class of index 10 (mouth width), their corresponding 3D heads, the non-corresponding middle 3D heads and the non-corresponding high 3D heads. From this table, it can be observed that the 3D head

2D video	Corresponding 3D head (high)		Non-corresponding 3D head (high)										T-test (P value)	
			S 17		S 19		S 22		S 35		S 46			
	W	H	W	H	W	H	W	H	W	H			W	H
S 17	**0.092**	**0.095**			0.101	0.130	0.112	0.113	0.107	0.111	0.111	0.110	0.0080	0.0210
S 19	**0.215**	0.122	0.259	0.132			0.222	0.134	0.250	**0.120**	0.247	0.128	0.0337	0.1266
S 22	0.172	**0.137**	0.136	0.150	**0.125**	0.184			0.127	0.157	0.149	0.156	0.0062	0.0469
S 35	0.311	**0.149**	0.270	0.181	0.294	0.152	**0.261**	0.186			0.499	0.226	0.7465	0.0919
S 46	0.231	0.118	0.198	0.142	**0.166**	0.136	0.227	**0.070**	0.244	0.108			0.2847	0.8234
			Non-corresponding 3D head (low)											
			S 23		S 31		S 47		S 48					
	W	H	W	H	W	H	W	H	W	H			W	H
S 17	0.092	**0.095**	**0.087**	0.110	0.097	0.116	0.112	0.128	0.111	0.110			0.2021	0.0158
S 19	0.215	0.122	0.240	0.122	0.224	0.125	0.254	**0.079**	**0.214**	0.130			0.1337	0.5459
S 22	0.172	**0.137**	0.126	0.149	**0.111**	0.160	0.139	0.142	0.179	0.153			0.1070	0.0338
S 35	0.311	0.149	0.315	**0.128**	0.320	0.115	0.319	0.149	**0.302**	0.156			0.5214	0.2934
S 46	0.231	**0.118**	0.267	0.148	0.237	0.166	0.267	0.137	**0.215**	0.138			0.3082	0.0224
			Non-corresponding 3D head (middle)											
			S 32		S 42		S 54		S 55					
	W	H	W	H	W	H	W	H	W	H			W	H
S 17	**0.092**	**0.095**	0.120	0.112	0.112	0.126	0.107	0.122	0.109	0.113			0.0060	0.0065
S 19	**0.215**	**0.122**	0.262	0.125	0.224	0.127	0.238	0.140	0.281	0.131			0.0642	0.0783
S 22	0.172	**0.137**	0.300	0.233	**0.132**	0.149	0.147	0.156	0.168	0.157			0.7269	0.1609
S 35	0.311	0.149	0.325	**0.130**	0.325	0.153	**0.293**	0.156	0.313	0.141			0.7177	0.5501
S 46	**0.231**	0.118	0.269	0.191	0.245	0.114	0.255	**0.104**	0.252	0.144			0.0171	0.3760

Table 5.4 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers classified under the high class of index 7 (vertical mouth height), their corresponding 3D heads and the non-corresponding 3D heads. Values in bold means the decreased RMS error. The last column shows p value of the t-test results between each corresponding 3D head and the non-corresponding 3D heads for width and height.

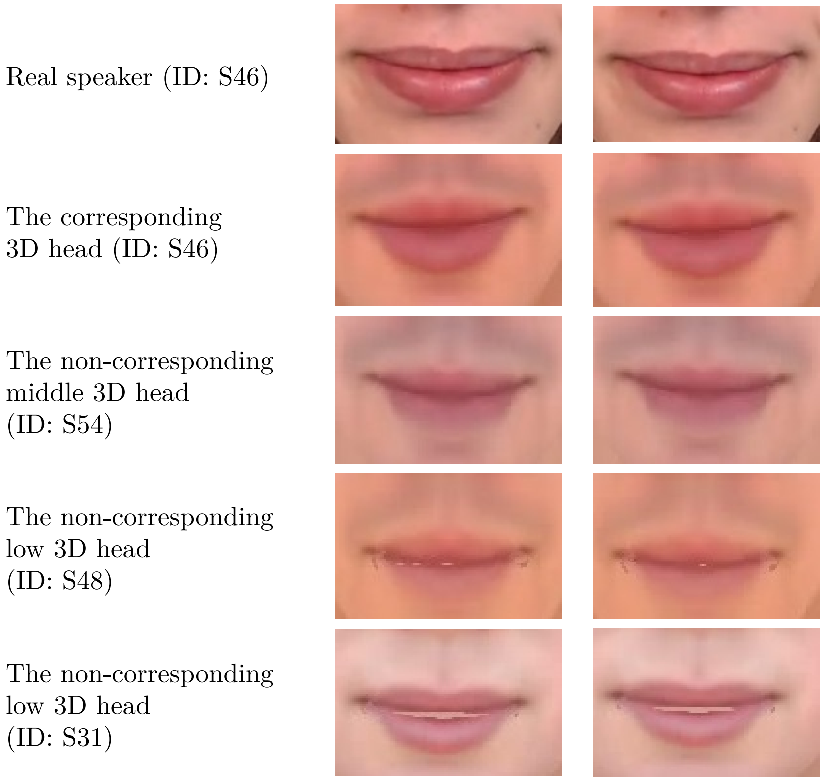

Real speaker (ID: S46)
The corresponding 3D head (ID: S46)
The non-corresponding middle 3D head (ID: S54)
The non-corresponding low 3D head (ID: S48)
The non-corresponding low 3D head (ID: S31)

Figure 5.7 Consecutive frames of the phoneme /p/ during utterance of the word "please" from sentence "lay white with A 5 please" for a real speaker (ID: S46) who classified under the high class of index 7 (first row), the corresponding 3D head (second row), the non-corresponding middle 3D head (third row) and the non-corresponding low 3D heads (last two rows).

that corresponded to a real speaker (ID: S32) gave the lowest RMSE score for three out of four speakers for width and for all speakers for height, when it was mapped to 2D videos of real speakers who classified under the low class. This may be due to its middle vertical mouth height (i.e. index 7). This explains why it failed to give the lowest score for width for a real speaker (ID: S35) with a high vertical mouth height.

The RMSE results varied when 2D videos of the real speakers were mapped to the non-corresponding middle 3D heads and the non-corresponding high 3D heads. For mapping between 2D videos of real speakers and the non-corresponding middle 3D heads, the 3D head of a real speaker (ID: S32) gave the lowest score for both width and height because its middle vertical mouth height was similar to or the same as most of

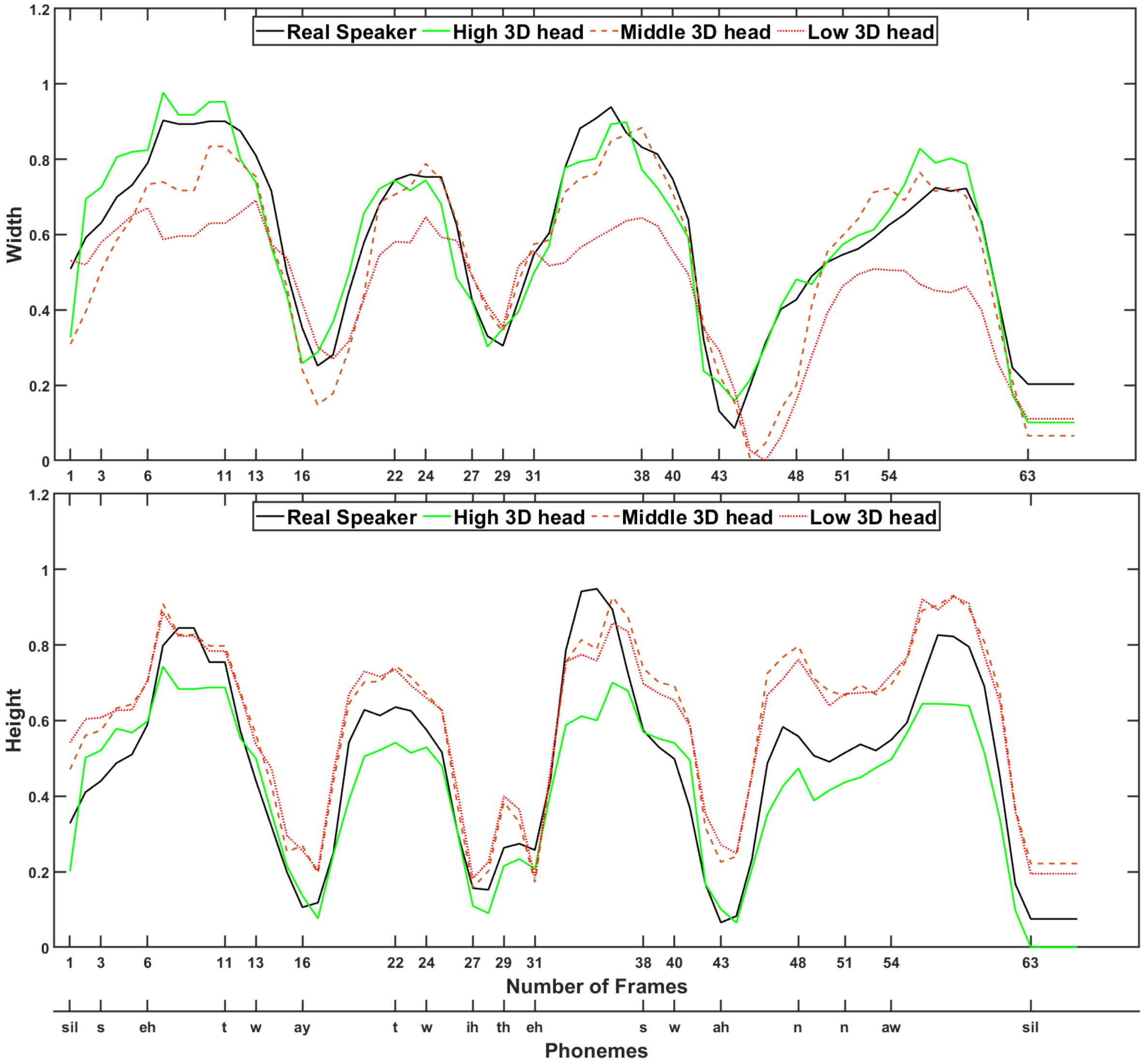


Figure 5.8 Width and height of mouth trajectories of 2D frames of the real speaker (ID: S 22) classified under the high class of index 7, the corresponding 3D head, the non-corresponding middle 3D head (ID: S 55) and the non-corresponding low 3D head (ID: S 48), whilst uttering the sentence "set white with S 1 now".

the non-corresponding middle 3D heads with middle to high (IDs: S7, S15 and S24) or middle (ID: S55) vertical mouth heights. What is striking in this table is that the corresponding 3D head of a real speaker (ID: S32) gave the lowest RMSE score for both width and height, when the 2D videos were mapped to the non-corresponding middle 3D heads. Additionally, some of the non-corresponding high 3D heads gave the lowest scores for width for real speakers with similar vertical mouth heights. For example, the non-corresponding high 3D head (ID: S20) gave the lowest score for a real speaker (ID: S16) with the same vertical mouth height (i.e. low to middle).

Figure 5.9 gives an example of consecutive frames for the real speaker (ID: S16), the corresponding 3D head and the non-corresponding high 3D head (ID: S20) during utterance of the phoneme /ih/ of the word "bin" from the phrase "bin white with M 2 soon". Also, the non-corresponding high 3D head (ID: S22) gave the lowest score for a real speaker (ID: S35); this may because they both had high vertical mouth heights. However, a t-test suggested a significant difference in the RMSE results between the corresponding 3D head of a real speaker (ID: S32) and the non-corresponding high 3D heads for both width and height. There was also a significant difference between the corresponding 3D head of a real speaker (ID: S38) and the non-corresponding high 3D heads for width.

These finding confirm that 2D videos of real speakers who have middle or wide mouth widths can be used to animate 3D heads that correspond to real speakers who have narrow mouth widths, as long as they have similar vertical mouth heights, lip thicknesses and distances between the nose and the upper lip. For example, Figure 5.10 gives an example of consecutive frames of a real speaker (ID: S38) classified in the low class of index 10, the corresponding 3D heads and the non-corresponding 3D heads. This Figure reveals that the non-corresponding high 3D head (ID: S54) produced a mouth shape more similar to the real speaker than the non-corresponding middle 3D head (ID: S55) due to closer classes of indices 7, 8, 9 and 12 of the corresponding real speakers (see Figure 5.1), while the non-corresponding high 3D head (ID: S22) failed to give a more accurate shape because of its lip thickness.

2D video	Corresponding 3D head (low)		Non-corresponding 3D head (low)										T-test (P value)	
			S 16		S 32		S 35		S 38					
	W	H	W	H	W	H	W	H	W	H			W	H
S 16	0.173	**0.097**			**0.162**	**0.097**	0.176	0.108	0.191	0.149			0.7290	0.3157
S 32	**0.111**	**0.056**	0.135	0.118			0.176	0.145	0.140	0.138			0.0930	0.0107
S 35	0.311	0.149	**0.258**	0.143	0.325	**0.130**			0.306	0.167			0.5385	0.8495
S 38	0.296	0.138	0.273	0.147	**0.204**	**0.111**	0.214	0.121					0.0928	0.3905
			Non-corresponding 3D head (middle)											
			S 7		S 15		S 24		S 48		S 55			
	W	H	W	H	W	H	W	H	W	H	W	H	W	H
S 16	0.173	0.097	**0.127**	0.120	0.137	0.100	0.154	**0.093**	0.212	0.096	0.181	0.103	0.5231	0.3162
S 32	**0.111**	**0.056**	0.117	0.198	0.133	0.085	0.152	0.103	0.163	0.089	0.175	0.059	0.0235	0.1006
S 35	0.311	0.149	0.277	0.200	**0.274**	**0.140**	0.334	0.145	0.302	0.156	0.313	0.141	0.3838	0.5471
S 38	0.296	0.138	0.300	0.183	0.287	**0.112**	0.247	0.117	**0.174**	0.119	0.212	0.114	0.0901	0.5430
			Non-corresponding 3D head (high)											
			S 19		S 20		S 22		S 46		S 54			
	W	H	W	H	W	H	W	H	W	H	W	H	W	H
S 16	0.173	0.097	0.179	0.095	**0.126**	0.127	0.145	0.103	0.181	**0.094**	0.213	0.116	0.7957	0.1912
S 32	0.111	**0.056**	0.145	0.102	0.134	0.072	0.135	0.193	0.161	0.123	**0.106**	0.112	0.0484	0.0325
S 35	0.311	0.149	0.294	0.152	0.288	**0.132**	**0.261**	0.186	0.499	0.226	0.293	0.156	0.7312	0.2610
S 38	0.296	0.138	0.269	0.111	0.254	0.111	0.254	**0.105**	0.229	0.145	**0.199**	0.123	0.0111	0.0561

Table 5.5 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers classified under the low class of index 10 (mouth width), their corresponding 3D heads and the non-corresponding 3D heads. Values in bold means the decreased RMS error. The last column shows p value of the t-test results between each corresponding 3D head and the non-corresponding 3D heads for width and height.

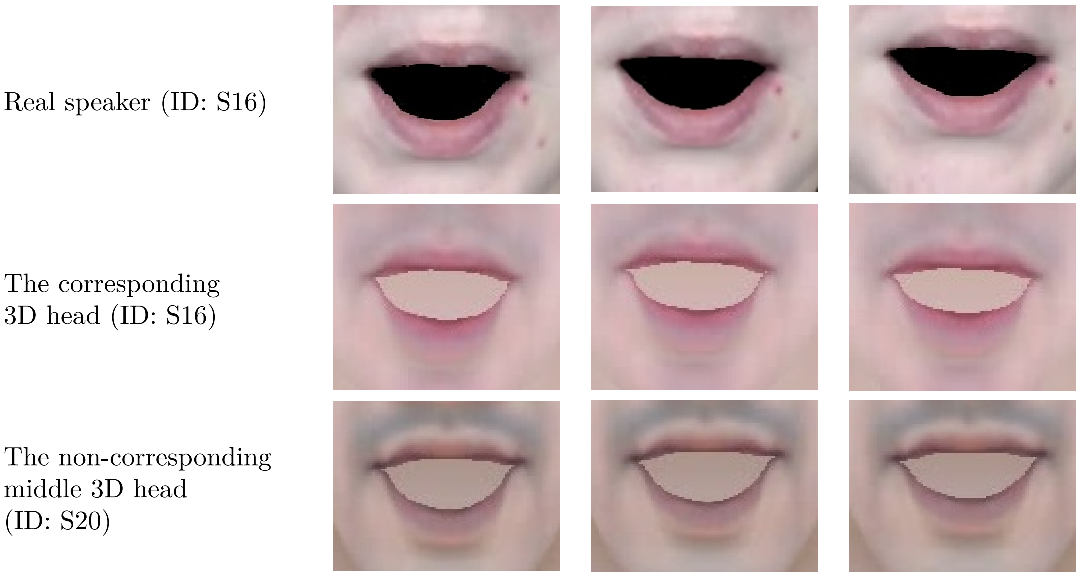

Real speaker (ID: S16)
The corresponding 3D head (ID: S16)
The non-corresponding middle 3D head (ID: S20)

Figure 5.9 Consecutive frames of the phoneme /ih/ during utterance of the word "bin" from sentence "bin white with M 2 soon" for a real speaker (ID: S16) who classified under the low class of index 10 (mouth width), the corresponding 3D head (second row), the non-corresponding high 3D head (third row).

Table 5.6 shows the RMSE results averaged over four sentences for the width and height of the mouth aperture of real speakers classified in the middle class of index 10, their corresponding 3D heads, the non-corresponding low 3D heads and the non-corresponding high 3D heads. From this table, it can be observed that three of the corresponding heads (IDs: S7, S15 and S55) gave the lowest scores for width, when the 2D videos of real speakers were mapped to the non-corresponding low 3D heads and the non-corresponding middle 3D heads. However, t-test results showed a significant difference in RMSE results between the corresponding 3D head of a real speaker (ID: S15) versus all the non-corresponding 3D heads for height and width. Also, there was a significant difference in the RMSE results for width between the corresponding 3D head of a real speaker (ID: S7) and the non-corresponding low 3D heads and between the corresponding 3D head of a real speaker (ID: S24) and the non-corresponding high 3D heads.

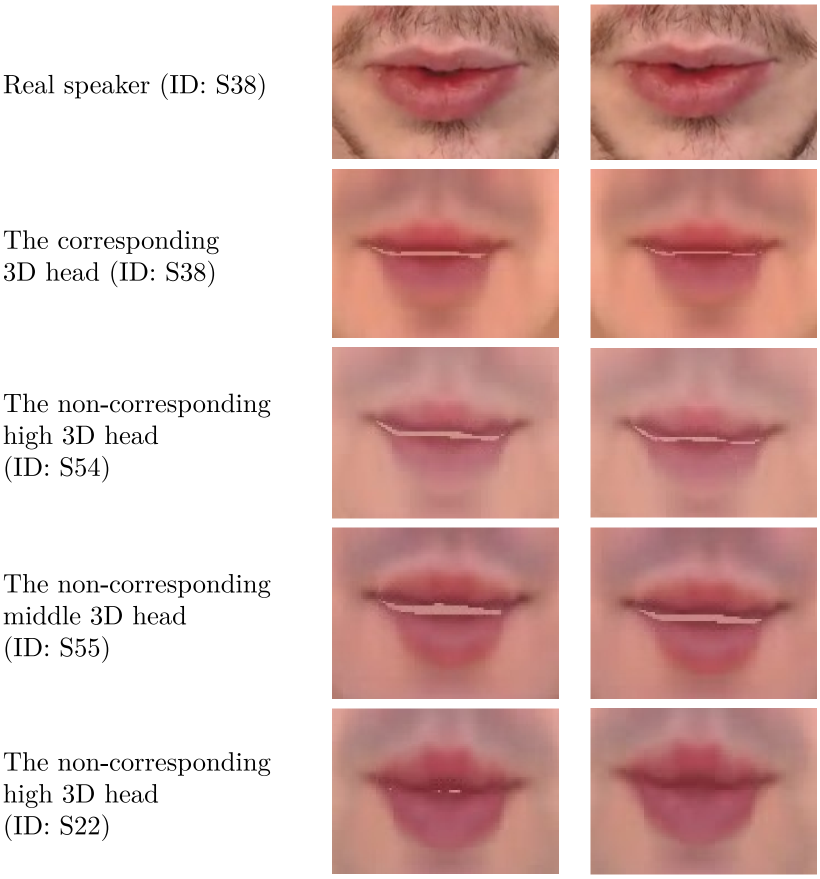

Real speaker (ID: S38)
The corresponding 3D head (ID: S38)
The non-corresponding high 3D head (ID: S54)
The non-corresponding middle 3D head (ID: S55)
The non-corresponding high 3D head (ID: S22)

Figure 5.10 Consecutive frames of the phoneme /uw/ during utterance of the word "two" from sentence "bin white in I 2 soon" for a real speaker (ID: S38) who classified under the low class of index 10 (mouth width), the corresponding 3D head (second row), the non-corresponding middle 3D head (third row) and the non-corresponding high 3D heads (last two rows).

These findings may prove that 2D videos of real speakers who have middle mouth width can be used to animate 3D heads that correspond to real speakers that have narrow, middle or wide mouth widths, as long as they have similar lip thicknesses or distances between the nose and the upper lip. Figure 5.11 shows an example of consecutive frames of the phoneme /th/ from the word "three" during uttering the phrase "bin white in N 3 now" by a real speaker (ID: S24) classified in the middle class of index 10, the corresponding 3D heads, the non-corresponding low 3D heads and the non-corresponding high 3D heads. This figure shows how the 3D heads gave the correct mouth shape regardless of the mouth width. The mouth aperture of the non-corresponding middle

3D head (ID: S32) (third row) is slightly wide compared to the real speaker due to its middle vertical mouth height (index 7).

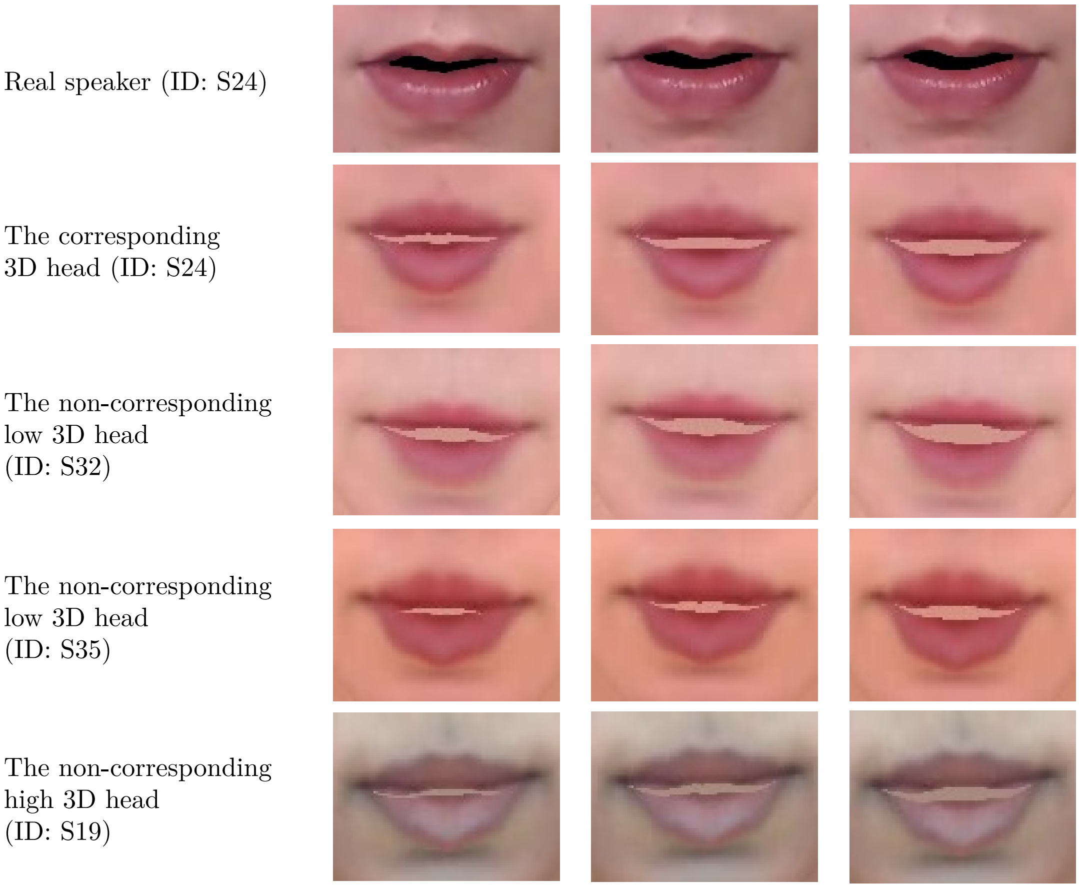

Real speaker (ID: S24)
The corresponding 3D head (ID: S24)
The non-corresponding low 3D head (ID: S32)
The non-corresponding low 3D head (ID: S35)
The non-corresponding high 3D head (ID: S19)

Figure 5.11 Consecutive frames of the phoneme /th/ during utterance of the word "three" from sentence "bin white in N 3 now" for a real speaker (ID: S24) who classified under the middle class of index 10 (mouth width), the corresponding 3D head (second row), the non-corresponding low 3D head (third and forth rows) and the non-corresponding high 3D heads (last row).

Table 5.7 shows the RMSE results averaged over four sentences for the width and height of the mouth aperture of real speakers classified in the high class of index 10 (mouth width), their corresponding 3D heads, the non-corresponding low 3D heads and the non-corresponding middle 3D heads. From this table, it can be noticed that the corresponding 3D head (ID: S19) gave the lowest RMSE scores for width for most of the speakers, when it was mapped to 2D videos of real speakers who are classified under the

2D video	Corresponding 3D head (middle)		Non-corresponding 3D head (middle)										T-test (P value)	
			S 7		S 15		S 24		S 48		S 55			
	W	H	W	H	W	H	W	H	W	H	W	H	W	H
S 7	**0.273**	0.128			0.318	0.140	0.279	**0.124**	0.296	0.185	0.302	0.148	0.6312	0.1985
S 15	**0.131**	**0.087**	0.158	0.191			0.186	0.132	0.164	0.151	0.179	0.134	0.0081	0.0177
S 24	0.219	0.123	**0.172**	0.142	0.237	0.122			0.245	0.142	0.260	**0.118**	0.6584	0.3001
S 48	0.149	**0.071**	0.158	0.188	**0.133**	0.081	0.144	0.095			0.166	0.072	0.8754	0.2506
S 55	0.143	0.167	0.118	0.116	0.119	**0.103**	0.122	0.104	**0.113**	0.107			0.1241	0.3809
			Non-corresponding 3D head (low)											
			S 16		S 32		S 35		S 38					
	W	H	W	H	W	H	W	H	W	H			W	H
S 7	**0.273**	**0.128**	0.310	0.138	0.295	0.149	0.298	0.131	0.293	0.129			0.0064	0.1481
S 15	**0.131**	**0.087**	0.179	0.139	0.183	0.130	0.175	0.139	0.174	0.136			0.0002	0.0002
S 24	0.219	0.123	**0.217**	0.116	0.255	0.124	0.275	0.117	0.228	**0.112**			0.1558	0.1046
S 48	0.149	**0.071**	0.122	0.090	0.157	**0.071**	0.157	0.101	**0.113**	0.127			0.4464	0.1105
S 55	**0.113**	0.107	0.145	0.109	0.154	**0.105**	0.116	0.114	0.120	0.116			0.1122	0.2056
			Non-corresponding 3D head (high)											
			S 19		S 20		S 22		S 46		S 54			
	W	H	W	H	W	H	W	H	W	H	W	H	W	H
S 7	0.273	0.128	**0.180**	0.122	0.257	0.137	0.256	**0.121**	0.274	**0.121**	0.325	0.126	0.5645	0.4410
S 15	**0.131**	**0.087**	0.146	0.136	0.173	0.134	0.164	0.163	0.166	0.136	0.170	0.149	0.0022	0.0005
S 24	**0.219**	0.123	0.244	0.133	0.251	0.129	0.255	0.133	0.258	0.120	0.228	**0.114**	0.0062	0.5007
S 48	0.149	0.071	0.184	0.078	0.156	**0.068**	0.164	0.145	0.156	0.096	**0.139**	0.104	0.2129	0.1107
S 55	0.113	0.107	**0.097**	0.111	0.114	**0.099**	0.099	0.127	0.116	0.110	0.134	0.117	0.8886	0.2747

Table 5.6 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers classified under the middle class of index 10 (mouth width), their corresponding 3D heads and the non-corresponding 3D heads. Values in bold means the decreased RMS error. The last column shows p value of the t-test results between each corresponding 3D head and the non-corresponding 3D heads for width and height.

high class. This may due to the middle distance between the nose tip and the upper lip and the high vertical mouth height. This also explains the significant difference suggested by the t-test result for height. Also, it can be noticed that the non-corresponding low 3D head (ID: S16) gave the lowest RMSE score for width for most of the speakers. This is probably due to the large distance between the nose tip and the upper lip. Another notable finding shown in this table is that the non-corresponding middle 3D head (ID: S7) gave the lowest RMSE score for most of the speakers. This may be due to the middle to high vertical mouth height (index 7), upper lip thickness (index 8), lower lip thickness (index 9) and distance between the nose tip and the upper lip (index 12). The t-test results suggested no significant difference for width and height between four out of five of the corresponding 3D heads, the non-corresponding high 3D heads and the non-corresponding middle 3D heads, while there is no significant difference between the non-corresponding low 3D heads, the corresponding 3D heads for height and three out of five of the corresponding 3D heads for width.

Such findings suggest that animating the wide mouths of 3D heads can be achieved using 2D videos of real speakers who have narrow or middle mouth widths as long as they have similar lip thicknesses, shapes and distances between the nose and the upper lip. Figure 5.12 shows an example of consecutive frames of a real speaker (ID: S54) classified in the high class of index 10, the corresponding 3D head, the non-corresponding low 3D head and the non-corresponding middle 3D head during utterance of the letter "B" from the phrase "lay white by B 8 again". This figure shows how all the 3D heads gave the correct mouth shape for the phoneme /b/, including the non-corresponding low 3D head.

Summary

This section summaries the objective test results for index 7 (vertical mouth height) and index 10 (mouth width). The evaluation produced these sets of results:

- 3D heads correspond to real speakers who have low vertical mouth height can be animated using only 2D videos of real speakers who have similar or middle vertical mouth height.

2D video	Corresponding 3D head (high)		Non-corresponding 3D head (high)										T-test (P value)	
			S 19		S 20		S 22		S 46		S 54			
	W	H	W	H	W	H	W	H	W	H	W	H	W	H
S 19	**0.215**	**0.122**			0.218	0.129	0.222	0.134	0.247	0.128	0.238	0.140	0.0967	0.0297
S 20	**0.239**	0.166	0.240	**0.149**			0.275	0.208	0.333	0.162	0.252	0.152	0.1933	0.9064
S 22	0.172	0.137	0.125	0.184	**0.117**	**0.134**			0.149	0.156	0.147	0.156	0.0182	0.1391
S 46	0.231	0.118	**0.166**	0.136	0.216	0.172	0.227	**0.070**			0.255	0.104	0.4785	0.9160
S 54	0.281	0.132	**0.209**	0.147	0.242	0.166	0.240	**0.129**	0.250	0.131			0.0148	0.2814
			Non-corresponding 3D head (low)											
			S 16		S 32		S 35		S 38					
	W	H	W	H	W	H	W	H	W	H			W	H
S 19	**0.215**	0.122	0.268	0.130	0.262	0.125	0.250	**0.120**	0.242	0.151			0.0062	0.2575
S 20	0.239	0.166	**0.223**	**0.155**	0.326	0.157	0.327	0.173	0.281	0.156			0.1332	0.2708
S 22	0.172	**0.137**	**0.124**	0.138	0.300	0.233	0.127	0.157	0.147	0.142			0.9564	0.2634
S 46	**0.231**	0.118	0.256	0.105	0.269	0.191	0.244	0.108	0.248	**0.099**			0.0244	0.7461
S 54	0.281	**0.132**	**0.255**	0.135	**0.255**	0.152	0.267	0.136	0.293	0.133			0.2289	0.2081
			Non-corresponding 3D head (middle)											
			S 7		S 15		S 24		S 48		S 55			
	W	H	W	H	W	H	W	H	W	H	W	H	W	H
S 19	0.215	**0.122**	**0.209**	0.169	0.244	0.130	0.254	**0.122**	0.214	0.130	0.281	0.131	0.1284	0.1582
S 20	0.239	0.166	**0.208**	0.350	0.274	0.158	0.308	0.171	0.294	**0.148**	0.309	0.163	0.1023	0.4491
S 22	0.172	0.137	0.140	**0.116**	**0.110**	0.135	0.129	0.146	0.179	0.153	0.168	**0.157**	0.1011	0.5825
S 46	0.231	0.118	**0.203**	**0.116**	0.255	0.149	0.261	0.122	0.215	0.138	0.252	0.144	0.6259	0.0679
S 54	0.281	**0.132**	**0.240**	0.156	0.273	0.155	0.254	0.142	0.273	0.133	0.256	0.149	0.0255	0.0252

Table 5.7 The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers classified under the high class of index 10 (mouth width), their corresponding 3D heads and the non-corresponding 3D heads. Values in bold means the decreased RMS error. The last column shows p value of the t-test results between each corresponding 3D head and the non-corresponding 3D heads for width and height.

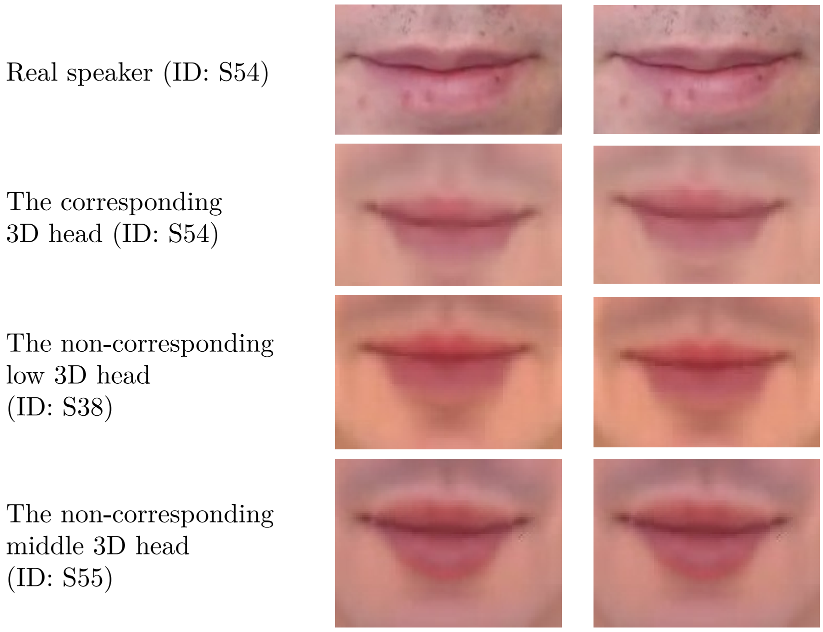

Real speaker (ID: S54)
The corresponding 3D head (ID: S54)
The non-corresponding low 3D head (ID: S38)
The non-corresponding middle 3D head (ID: S55)

Figure 5.12 Consecutive frames of the phoneme /b/ during utterance of the letter "B" from sentence "lay white by B 8 again" for a real speaker (ID: S54) who classified under the high class of index 10 (mouth width), the corresponding 3D head (second row), the non-corresponding low 3D head (third row) and the non-corresponding middle 3D heads (last row).

- 3D heads correspond to real speakers who have middle vertical mouth height can be animated using 2D videos of real speakers who have similar or low vertical mouth height. Also, they can be animated using 2D videos of real speakers who have high vertical mouth height as long as they have a similar mouth width.

- 3D heads of real speakers who have high vertical mouth height can be animated using 2D videos of real speakers who have similar vertical mouth height. Also, they can be animated using 2D videos of real speakers who have middle vertical mouth height as long as they are similar in other facial features such as upper and lower lip thickness, the distance between the nose and the upper lip, and mouth width.

- 3D heads of real speakers who have different mouth widths can be animated using 2D videos of real speakers who have similar or different mouth widths, as long as they are identical in other facial features such as vertical mouth height, lip thickness, and distance between the nose and the upper lip.

These findings confirm that the vertical mouth height has a significant impact on the mapping process, while other features should be considered with the mouth width indicating its decreased effect on the mapping process.

5.5 Subjective Evaluation

Subjective test is conducted to compare the naturalness of animations generated by mapping between 2D videos of real speakers and 3D heads with either the corresponding or non-corresponding 3D heads defined in Section 5.4. This test investigates the impact of similarities and differences in the vertical mouth height and the mouth width features between real speakers and 3DMMs on the resulting 3D lip motions. The baseline animation modalities that will be used to guide the evaluation comparison are the animation of the corresponding 3D head and the non-corresponding 3D heads. For example, 2D video of a real speaker from the Audio-Visual Lombard Grid corpus who is classified under one class of index 7 or index 10 will be used to animate the corresponding 3D head and the non-corresponding 3D heads that relate to other speakers who are classified under the other two classes. Videos of the animated 3D heads will be synchronised with clean audio signal of the real speaker. The clean audio is used to enable the subjects to judge the extent to which the animated lip movement is as smooth as a real speaker's and how likely it was that the movement would produce those sounds. The evaluation will address two main points:

- The impact of differences between the real speakers and the 3D heads in the vertical mouth height and the mouth width on the resulting 3D lip animation.

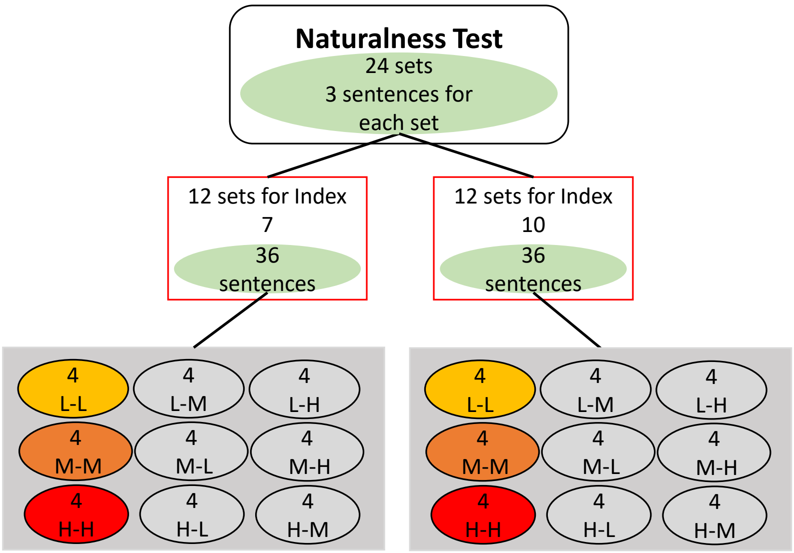
- The ranges of differences and similarities in the vertical mouth height and the mouth width between real speakers and 3D heads that provide sufficient 3D lip motions.

Stimuli

Seventy-two 2D videos of 3D talking heads were played in this test. Three separate animations were presented side by side for each set (24 sets in total presented in a random order for each participant). Twelve sets showed the resulting animation for mapping between non-similar faces based on the classes of index 7, while the other 12 sets showed the resulting animation for mapping between non-similar faces based on the classes of index 10. For each set, three 2D videos of 3D animations were presented side by side. All three heads were animated using 2D videos of one real speaker classified in one of the three classes. One animation corresponded to that real speaker, and the other two corresponded to different real speakers classified in the other two classes. Figure 5.13 illustrates the structure of the stimuli. The subjects used a play button to repeat each sentence and watched each video three times. After each set of videos, the subjects were asked to choose which 3D talking head had the most natural lip movements and which had the least. The selection scores of a subject for the best and the worst choices were used to evaluate the impact of the vertical mouth height and the mouth width on the resulting animation. Figure 5.14 shows the graphical user interface that was used to present 2D videos of the animation to the subjects.

Subjects

Two groups of participants with normal hearing and vision were recruited from the Department of Computer Science, University of Sheffield, and tested individually in an acoustically isolated booth with visual signals presented on a computer screen combined with acoustic signals presented binaurally through headphones. The first group consisted of native English speakers N_e (12 native English speakers), and the second group consisted of non-native English speakers NN_e (15 non-native English speakers from different Arabic countries [i.e. Bahrain, Egypt, Iraq, Libya and Saudi Arabia]; IELTS score ϵ [5.5,9]). Where N and NN denote native English speakers and non-native English speakers, respectively, and the subscript e denotes English. The non-native



Stimulus	Description	Stimulus	Description
L-L	Low 2D speaker to corresponding low 3D head	M-H	Middle 2D speaker to high 3D head
L-M	Low 2D speaker to middle 3D head	H-H	High 2D speaker to corresponding high 3D head
L-H	Low 2D speaker to high 3D head	H-L	High 2D speaker to low 3D head
M-M	Middle 2D speaker to corresponding middle 3D head	H-M	High 2D speaker to middle 3D head
M-L	Middle 2D speaker to low 3D head		

Figure 5.13 Structure of stimuli.

English speakers were recruited to participate in this experiment to investigate how the synthesised signals would be perceived and evaluated by the non-native speakers; since such animation has been utilised for customer services and entertainment such as films and games [257], and also proved to be effective in using for pronunciation training systems [11, 72, 73, 83, 84, 172, 248]. This study was ethically approved via the University of Sheffield's ethics review procedure (application number: 024196).

Results

Figure 5.15 shows the most and least natural choice rates given by the two groups for the corresponding 3D heads for index 7 (a) and index 10 (b). Generally, the differences in vertical mouth height (index 7) between real speakers and 3D head models have a greater

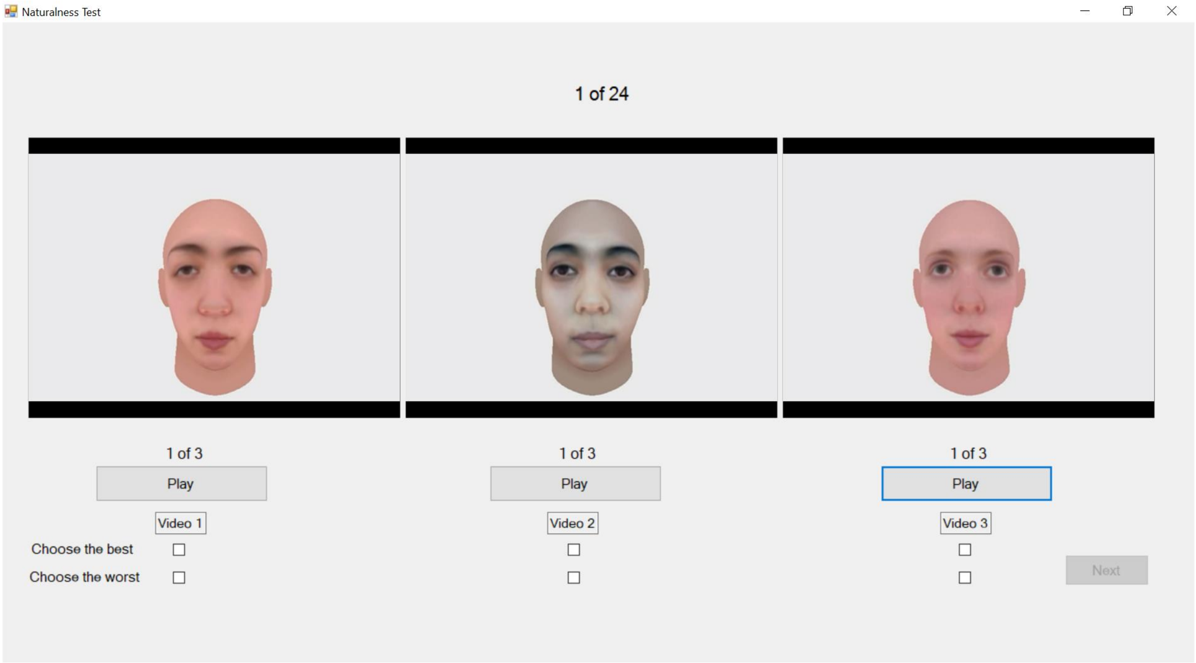


Figure 5.14 Screenshot of the graphical user interface for the naturalness test.

influence on the resulting animation, where the two groups voted for the corresponding 3D heads as most natural. The two groups were able to distinguish the corresponding 3D heads from the non-corresponding 3D heads with a modest, slightly better than the chance level of accuracy: the N_e group chose 39.58% of the corresponding 3D heads stimuli as the most natural lip motions, whereas the NN_e group chose 37.22% of the corresponding 3D heads stimuli. This is confirmed by a t-test result that showed no significant difference between the two groups for selecting the corresponding 3D heads as the best choice (p=0.6514).

For the mouth width (index 10), the N_e group outperformed the NN_e group for voting for the corresponding 3D heads as most natural. This may because the stimuli were presented in their language, suggesting that the differences in mouth width have less impact on the 3D lip motions, which made the non-native group (NN_e) were not able to select the corresponding 3D heads as the best choice. However, no significant difference was found between the two groups for voting for the corresponding 3D heads as most natural (p=0.1482).

Figure 5.16 shows the most and least natural choice rates for the corresponding 3D heads and the non-corresponding 3D heads for each class of index 7 for the two groups. The N_e group found the corresponding 3D heads of each class to be most natural, while the NN_e voted only for the corresponding high 3D heads. The two groups were able to distinctively choose the corresponding high 3D heads as having the most natural lip motions. This is confirmed by t-test result that showed no significant difference between the two groups for choosing the corresponding high 3D heads as having most natural lip motions (p=0.0981). However, for the high class, t-test results showed no significant difference between the corresponding 3D heads and the non-corresponding low 3D heads for the N_e group (p=0.3225) and the NN_e group (p=0.1643). Also, no significant difference was found between the corresponding 3D heads and the non-corresponding middle 3D heads for the N_e group (p=0.2229) and the NN_e group (p=0.0611). The two groups found the animation generated by mapping 2D videos of a real speaker classified in the low class to the non-corresponding high 3D heads to be the least natural. A t-test result showed no significant difference between the two groups for voting for the non-corresponding high 3D heads as least natural (p=0.7882). This confirms the objective test results provided in Table 5.2. However, no significant difference was found between the non-corresponding high 3D heads and the corresponding low 3D heads for the two groups (p=0.1362 for N_e and p=0.1442 for NN_e) or between the non-corresponding high 3D heads and the corresponding middle 3D heads for the two groups (p=0.1891 for N_e and p=0.1781 for NN_e).

Figure 5.17 shows the most and least natural choice rates for the corresponding 3D heads and the non-corresponding 3D heads for each class of index 10 for the two groups. For the high class, t-test results showed a significant difference between the corresponding 3D heads and the non-corresponding middle 3D heads for the N_e group (p=0.0280). The N_e group found the corresponding 3D heads of the middle and the high classes to be the most natural. For the low class, the two groups found the non-corresponding high 3D heads to have the least natural lip motions. This was confirmed by t-test result that showed no significant difference between the two groups for selecting the non-corresponding high 3D head as having the least natural lip motions (p=0.7575). A significant difference was suggested by the t-test results between the corresponding 3D heads and the non-corresponding high 3D heads for the two groups (p=0.0448 for the N_e group and p=0.0046 for the NN_e group). Also, a significant difference was found between the non-corresponding high 3D heads and the non-corresponding middle 3D heads for the two groups (p=0.0012 for the N_e group and p=0.0009 for the NN_e group). This may prove that reasonable 3D lip motions cannot be achieved when 2D videos of real speakers with narrow mouth widths are mapped to 3D heads that relate to real speakers with wide mouth widths.

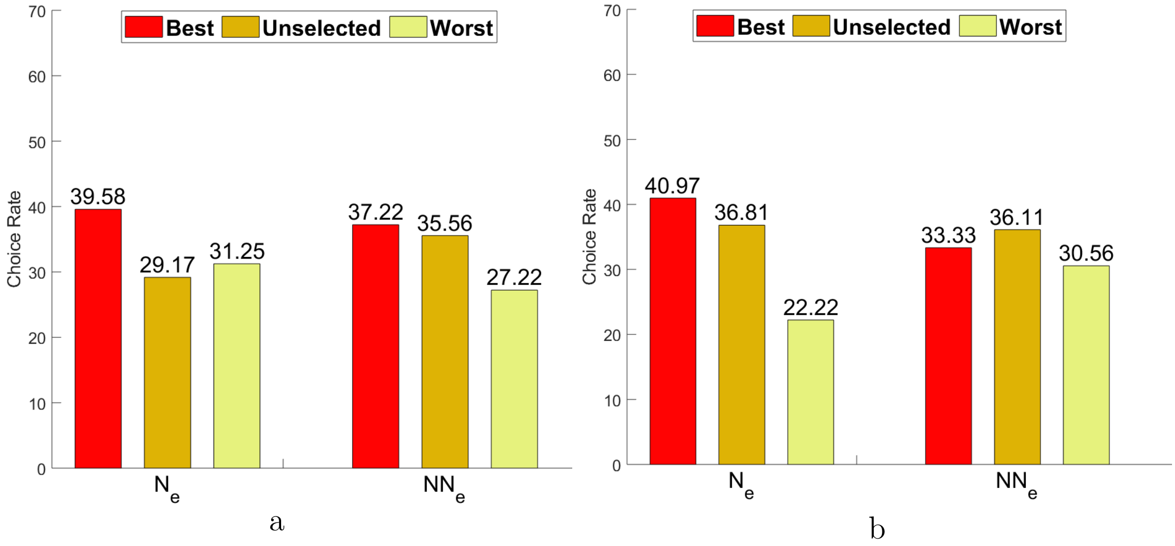


Figure 5.15 Results for the best, unselected and worst rates for the corresponding 3D heads for the N_e and NN_e groups: (a) Rates for index 7; (b) Rates for index 10.

Discussion

This study presented a two-fold aim: investigating the impact of similarities and differences in the facial features between real speakers and 3DMMs on the resulting 3D lip motions, and defining the ranges of differences in facial features between real speakers and 3D heads that allow adequate 3D lip motions to be achieved. The main observations and discussion points are listed below.

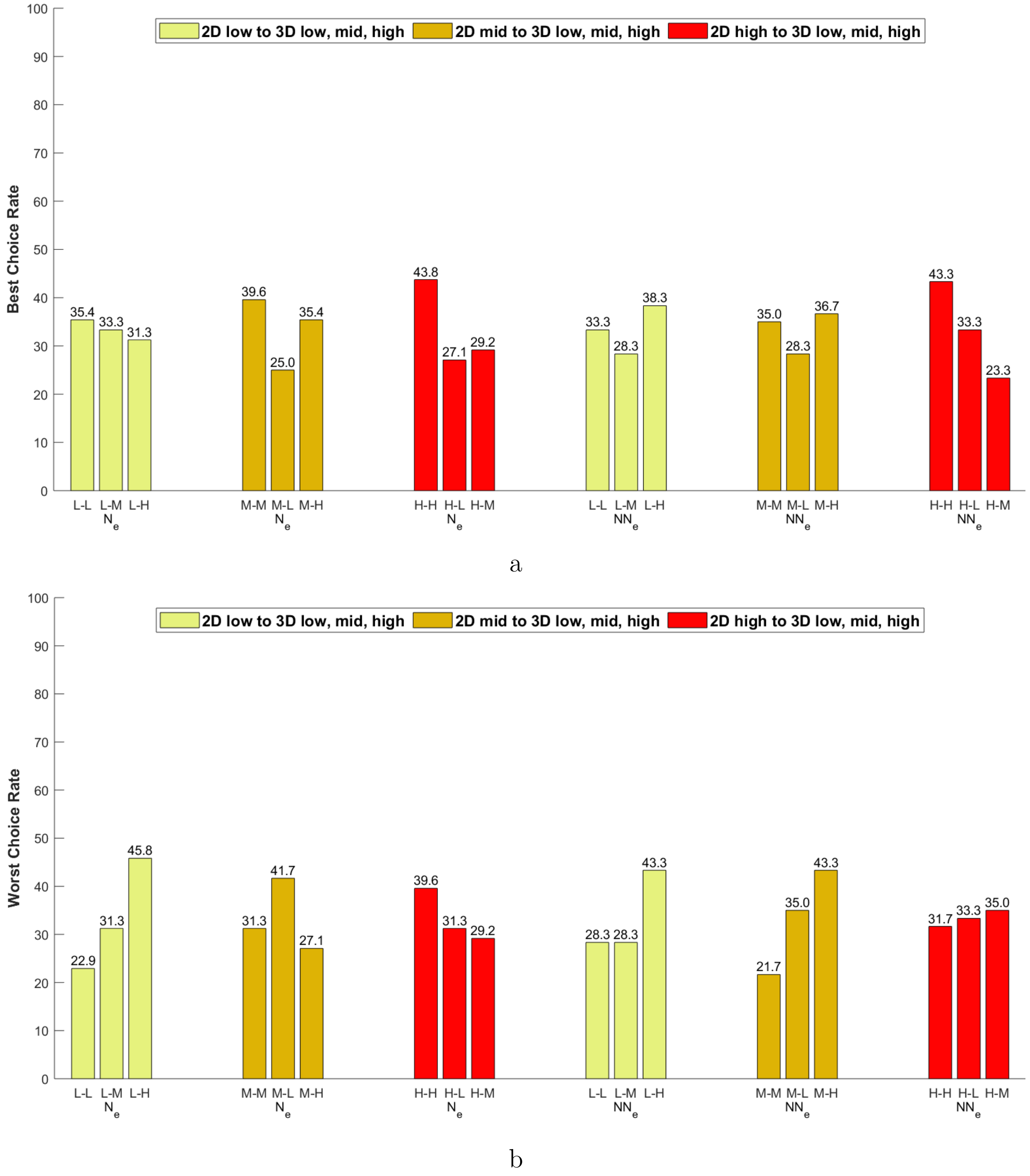
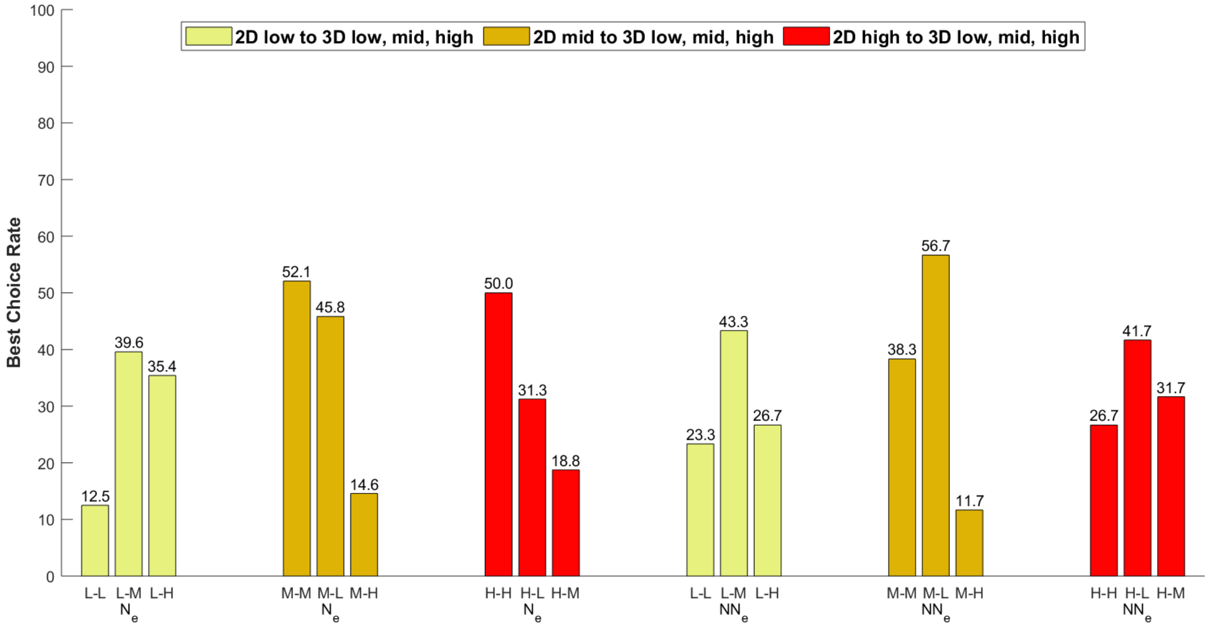
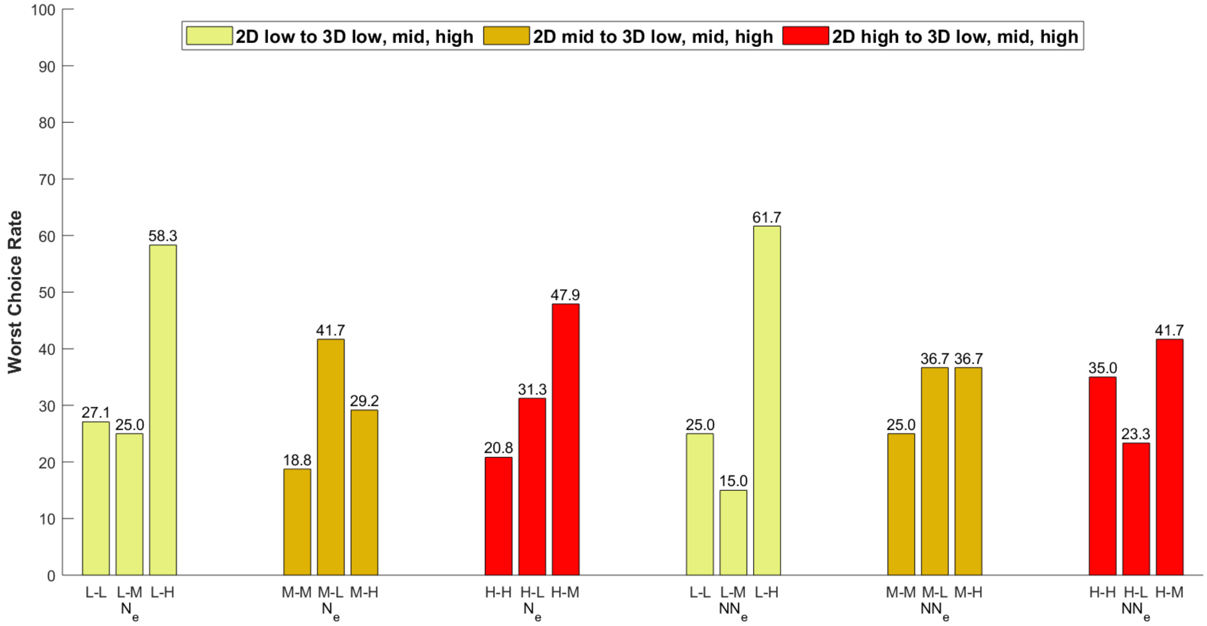


Figure 5.16 Results for the best and worst choice rates for the corresponding 3D heads and the non-corresponding 3D heads for each class of index 7 for the N_e and NN_e groups: (a) Rates of the best answer ; (b) Rates of the worst answer.



a



b

Figure 5.17 Results for the best and worst choice rates for the corresponding 3D heads and the non-corresponding 3D heads for each class of index 10 for the N_e and NN_e groups: (a) Rates of the best answer ; (b) Rates of the worst answer.

- Native English-speaking participants were able to distinguish between the corresponding and non-corresponding 3D heads slightly better than non-native English-speaking participants for the two tested indices (indices 7 and 10). The non-native English-speaking participants were able to select the corresponding 3D heads as most natural for index 7 only. This may be because the non-native participants had lower linguistic competence than the natives, which made them were not able to spot any slight inaccuracy in visual speech animation. This suggests that the differences between real speakers and 3DMMs in the vertical mouth height (index 7) have a greater influence on the resulting animation than the mouth width (index 10).

- The two groups were able to distinguish between the corresponding 3D heads and the non-corresponding low 3D heads for the high class of index 7, where the subjects chose the lip motions of the corresponding high 3D heads as the most natural. However, the results for the least natural choice for this class were contrary for the NN_e group and convergent for the N_e group. This indicates that selecting one of three choices is more difficult than choosing between two options. This could also apply to the most natural choice answers for the low class, where the results were close for each 3D head alternate, although the subjects were able to select the non-corresponding high 3D heads as the least natural. This indicates that the difference between the corresponding 3D low heads and the non-corresponding high 3D heads is distinguishable, which was also confirmed by the objective test results (see Section 5.4).

- For the low class of index 10, the two groups were able to distinctly choose the non-corresponding high 3D heads as the least natural, which confirms that the difference is significant and distinguishable between these classes for this index. These findings are not comparable with the objective test results due to an unbalanced number of male and female speakers in each class of the tested indices. Consequently, presenting all possible methods of mapping to the subjects was restricted by this factor, as it is not reasonable to display a 3D head of a real male speaker combined with a female audio signal to the participants. This suggests that in order to accurately investigate the effects of differences and similarities in the facial features between real speakers and 3DMMs on the resulting 3D visual speech animation, a large amount of data is essential. However, the most natural choice answers for this class indicate confusion between the corresponding 3D heads and the non-corresponding 3D heads. For example, the NN_e group chose the non-corresponding low 3D heads as the most natural for the high class. The performance of the NN_e group for the high class of index 10 is chaotic for the most and least natural choice answers; this may be because the variations in this index (mouth width) are not noticeable to non-native participants in comparison to index 7 (vertical mouth height), which has a greater effect on lip closure.

5.6 Summary

In this chapter, an investigation of the effects of differences and similarities in facial features between real speakers and 3DMMs on the mapping process was presented. The facial features of real speakers were represented in 12 indices, and each index was classified into three classes: low, middle and high. In this thesis, two indices representing vertical mouth height (index 7) and mouth width (index 10) were investigated separately by mapping between real speakers from different classes to their corresponding 3D heads and 3D heads that corresponded to different speakers in the same class or different classes.

The resulting 3D lip motions were evaluated objectively and subjectively. The results of the objective test suggest that for index 7, the mapping between real speakers with low vertical mouth height and the 3D heads that correspond to real speakers with high vertical mouth height, or vice versa, leads to unpleasant 3D lip motions. For index 10, the results varied between the classes, which confirms that mouth width does not have significant effects on the mapping process, whilst other facial features should be considered, such as lip thickness. The subjective evaluation results suggest that native

English-speaking participants are able to distinguish between the corresponding and the non-corresponding 3D heads slightly better than non-native speakers. For the two tested indices, the two groups of participants chose the non-corresponding high 3D heads as having the most unnatural lip motions when they were mapped to real speakers classified in low classes. For index 7, the two groups selected the corresponding high 3D heads as having the most natural lip motions. This is not the case with index 10, where only the native-speaking participants were able to select the corresponding high 3D heads as having the most natural lip motions. This may confirm that mouth width does not have considerable effects on the resulting 3D lip motions due to limited changes in this feature during speech in comparison with index 7, which affects lip closure.

Chapter 6

Conclusions

This thesis has investigated driving the 3D lips of a 3D head using tracked lip motions from 2D videos of a real speaker with the aid of a 3DMM. Mapping between the tracked landmarks in the 2D videos and the corresponding 3D landmarks was achieved by implementing a method that presented by Huber et al. [125] that reconstructs 3D faces from images and videos using a 3DMM. The 3DMM used in this thesis was built using synthetic 3D head poses generated with the commercial software FaceGen [1], which provides two functionalities that are essential to tracking and analysing lip motions in a detailed manner: the software can be supplied with a front-view photograph or front- and side-view photographs to create the initial 3D head pose to personalise the 3DMM, and the generated 3D heads have similar vertex correspondences that facilitate the generation of a large number of 3D head poses that are used to train the 3DMM.

It is important to determine which set of facial features can be used in the mapping process to achieve adequate 3D lip motions. In Chapter 3, different sets of facial features were used to map 2D videos of a real speaker to the corresponding 3DMM. The test results showed that using feature points that represent the lips, nose, eyes and eyebrows for the mapping process produced the desired 3D lip motions. Including the contour landmarks impedes adequate 3D lip motions by restricting the face mesh, which affects the lip movements. These findings confirm that adequate lip motions can be achieved using all sets of facial features. Each set plays a vital role in which a combination set of them controls the movement of smooth regions of the face's mesh, such as the cheek and the forehead that would affect the resulting lip motions.

It was also important to determine how using different amounts of data while constructing the 3DMM can influence the resulting 3D lip motions. In Chapter 4, four 3DMMs were created for each real speaker using different amounts of data. Either the front-view photograph only or the front- and side-view photographs were used to create the initial 3D head pose, and the 3DMM was trained using either 17 poses (neutral 3D head pose and 16 viseme) or 161 poses (neutral 3D head pose and 10 variations of each viseme). Each 3DMM was fitted to the front-view videos of the corresponding real speaker; then, the performance of each 3DMM was evaluated in comparison with the real speaker's videos. Two articulatory measurements were extracted and calculated from 2D videos of the real speakers and the 3D lip motions, which are the width and height of the mouth aperture. The RMSE was used to measure the difference between the articulatory measurements of the real speaker and the 3D animation. The test results of the objective evaluation confirmed that using the front- and side-view photographs to create the initial 3D head pose and training the 3DMM with 161 poses enhanced the resulting 3D lip motions. These results indicate how using a side-view photograph for creating the initial neutral head pose gives the 3D head closer shape to the corresponding real speaker and how this affects the resulting 3D lip motions. In addition, training the model with a larger number of head poses gives the 3D head model an ability to accurately detect the lip movements of the real speaker, resulting in smooth 3D lip motions.

Also, Chapter 4 presented a side-view evaluation of the performance of each 3DMM compared to the real speaker's videos. The upper lip protrusion was measured, extracted and calculated from side-view 2D videos of the real speakers and the 3D lip motions. The RMSE was used to measure the difference between the articulatory measurement of the real speaker and the 3D animation. The results also confirmed that using the front- and side-view photographs to create the initial 3D head pose and 161 poses to train the 3DMM improved the performance of the 3DMM. These findings prove using extra data during the construction of the model achieve adequate 3D lip protrusion when only front-view 2D videos are used for the mapping process.

Chapter 5 investigated the impact of differences and similarities in facial features between the real speakers and the 3DMMs on the 3D lip animation by mapping 2D videos of a real speaker to corresponding and non-corresponding 3D heads. The facial features of the Lombard speech grid corpus' speakers [10] were analysed and represented into 12 indices, and each index was classified into three classes: low, middle and high. Two indices related to the mouth features were investigated: vertical mouth height and mouth width. The mapping between non-similar faces was achieved with the three classes of each index. The resulting 3D lip motions were tested objectively and subjectively. In regard to the vertical mouth height index, the results showed that mapping between real speakers classified in the low class and 3D heads of speakers classified in the high class, or vice versa, produced undesirable 3D lip motions. In regard to the mouth width index, there were variations in the results of mapping between the classes, suggesting a reduced effect on the mapping process. Two groups of subjects participated in this study: native English speakers and non-native English speakers (native Arabic speakers). For the two tested indices, the non-corresponding high 3D heads were chosen by the two groups as having the worst lip motions when they were fitted to 2D videos of real speakers classified in the low class. For vertical mouth height index, the corresponding high 3D heads were chosen by the two groups as having the most natural lip motions. For the mouth width index, only the native-speaking subjects chose the corresponding high 3D heads as having the most natural lip motions, confirming a reduced effect on the resulting 3D lip motions due to limited variations in this index during speech.

6.1 Original Contributions

- **Animating 3D lips using 2D videos with the aid of a 3DMM:** A 3D talking head was animated, according to tracked lip motions of a real speaker in 2D videos. This is achieved through a mediation of a 3DMM. Most previous works related to reconstructing 3D faces from images by utilising illumination or depth data, or required 3D scans of real faces for personalising the generated 3D models and detect lip motions accurately. Here, the 3DMM was constructed using synthetic 3D head poses that generated using photos of a real speaker. Then the mapping between the 2D landmarks in front-view videos of the real speaker and the corresponding 3D landmarks labelled on the 3DMM was accomplished following a method presented by Huber et al [125] that reconstructs 3D faces from images using a 3DMM.

- **Identifying a set of facial features landmarks for achieving desired 3D lip motions:** The effect of each set of facial features landmarks on the resulting animation was examined. This investigation illustrated the functionality of each set of the facial landmarks in the mapping process, where the 3D heads were fitted to the corresponding real speakers using different sets of facial features. The objective test results revealed that utilising a set that contains eyebrows, eyes, nose, and lips landmarks gives more accurate and smoother lip motions, while including the contour landmarks produces insufficient animation.

- **Identifying the required amount of data to construct and train 3DMMs for producing efficient 3D lip motions:** Presenting a study that investigated the impact of using different amount of data for creating the initial 3D head pose, and training the 3DMM on the 3D lip motions. The study investigated whether using front-view photograph or front- and side-view photographs for creating the initial 3D head pose would improve the resulting 3D lip motions. Also, training the 3DMM with a larger number of poses would further enhance the results. In comparison with front-view 2D videos of the corresponding real speakers, the results showed that the performance of the 3D lip motion enhanced when front- and side-view photographs were used to create the initial 3D head pose, and when the 3DMM was trained with a larger number of poses.

- **Evaluation of 3D lip motions from the side-view:** Producing a study that evaluated the 3D lips' protrusion of the 3DMMs that were created using different amount of data for generating the initial 3D head pose and training the model. Side-view 2D videos of the corresponding real speakers were used to compare against to investigate the accuracy level of the 3D lips protrusion when front-view videos are used for the mapping process. The quantitative evaluation confirmed that using further information during constructing and training the 3DMM enhances the final animation.

- **First study of investigating the influence of differences in facial features between real faces and 3D faces on the final animation:** Introducing a study that investigated whether the 3D lip motions are influenced when front-view 2D videos of a real speaker are mapped to a non-corresponding 3DMM. The mapping between non-similar faces was implemented based on differences and similarities of two features of the mouth, which are vertical mouth height and mouth width. The objective and subjective evaluation results confirmed that a poor animation can be achieved when 2D videos of real speakers who have low vertical mouth height are mapped to 3D heads that correspond to real speakers who have high vertical mouth height, or vice versa. Suggesting a significant impact of differences in this index on the resulting 3D lip motions.

6.2 Limitations and Future Work

The main practical limitation is that the initial 3D head pose of each real speaker is created based on placing facial points on a well-positioned face in a good quality of front-view photo or front- and side-view photos. This enables the personalizing of the generated 3DMM, but any misleading in the facial landmarks annotation process due to choosing inappropriate photographs introduces a potential source of errors. Care must be taken when posing the initial neutral 3D head setup for the deforming face shapes to preserve the efficiency of the final animation. Fortunately, this is a parameterising step that only needs to be achieved once per speaker. Moving forward, one possible direction for future work is to use image processing techniques or computer vision techniques to enhance the quality of the photographs or the real speakers' head position.

By using front-view 2D videos of a real speaker, a robust model of speech animation that mimics real speech motions is achieved. It is currently valuable to add facial expressions. An interesting future direction would be investigating other facial feature motions during speech, such as eyebrow movements and facial expressions, to make the head appears more realistic. A training set that includes facial expression poses could be used to animate the upper part of the face. It is also worth exploring whether training the model with these poses enhances the resulting 3D lip motions or whether the face must be segmented into sub-regions to morph each region independently.

The side-view evaluation of the 3D lip motions presented in Chapter 4 requires further investigation. An exploration of how using more vertical and horizontal angles for the virtual camera to predict the angle of the used camera for recording the real speakers would enhance the results. Furthermore, protrusion of the lower lip of the 3D heads can be evaluated as well.

Another study could extend the mapping between the non-similar faces (conducted in Chapter 5) by including all indices of facial features to examine the impact of each feature on the resulting 3D lip motions. Furthermore, using a large amount of data that includes more speakers would expand the pool of the resulting 3D motions, and further studies should examine the impact of differences and similarities between real speakers and 3DMMs subjectively by presenting all the cases of mapping between non-similar faces.

Another direction for future work would be adding a tongue and teeth to the 3D talking head to determine whether they can be animated using observed data from 2D videos and to investigate the impact of using a large number of poses to train the 3D head model. Furthermore, the impact of differences and similarities in facial features between real speakers and the animated 3D heads (Chapter 5) on the resulting 3D mouth motions can be tested subjectively by testing the intelligibility of the 3D talking heads.

Bibliography

[1] Facegen modeller, 1998. https://facegen.com/. Accessed: 2018-12-01.

[2] Faceware technologies, 2014. http://facewaretech.com/. Accessed: 2018-12-01.

[3] Vicon systems, 2013. http://vicon.com/. Accessed: 2018-12-01.

[4] Abry, C. and Boë, L.-J. (1986). " laws" for lips. *Speech communication.*

[5] Agelfors, E., Beskow, J., Karlsson, I., Kewley, J., Salvi, G., and Thomas, N. (2006). User evaluation of the synface talking head telephone. In *International Conference on Computers for Handicapped Persons*, pages 579–586. Springer.

[6] Alexander, O., Fyffe, G., Busch, J., Yu, X., Ichikari, R., Jones, A., Debevec, P., Jimenez, J., Danvoye, E., Antionazzi, B., et al. (2013). Digital ira: Creating a real-time photoreal digital actor. In *ACM SIGGRAPH 2013 Posters*, page 1. ACM.

[7] Alexander, O., Rogers, M., Lambeth, W., Chiang, J.-Y., Ma, W.-C., Wang, C.-C., and Debevec, P. (2010). The digital emily project: Achieving a photorealistic digital actor. *IEEE Computer Graphics and Applications*, 30(4):20–31.

[8] Algadhy, R., Gotoh, Y., and Maddock, S. (2016). Analysis of visemes in the grid corpus. *UK Speech 2016 Conference at the University of Sheffield.*

[9] Alghamdi, N., Maddock, S., Barker, J., and Brown, G. J. (2017). The impact of automatic exaggeration of the visual articulatory features of a talker on the intelligibility of spectrally distorted speech. *Speech Communication*, 95:127–136.

[10] Alghamdi, N., Maddock, S., Marxer, R., Barker, J., and Brown, G. J. (2018). A corpus of audio-visual lombard speech with frontal and profile views. *The Journal of the Acoustical Society of America*, 143(6):EL523–EL529.

[11] Ali, A. Z. M., Segaran, K., and Hoe, T. W. (2015). Effects of verbal components in 3d talking-head on pronunciation learning among non-native speakers. *Journal of Educational Technology & Society*, 18(2):313–322.

[12] Amberg, B., Knothe, R., and Vetter, T. (2008). Expression invariant 3d face recognition with a morphable model. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE.

[13] Aschenberner, B. and Weiss, C. (2005). Phoneme-viseme mapping for german video-realistic audio-visual-speech-synthesis.

[14] Auer Jr, E. T. and Bernstein, L. E. (1997). Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *The Journal of the Acoustical Society of America*, 102(6):3704–3710.

[15] Aulsebrook, W., İşcan, M., Slabbert, J., and Becker, P. (1995). Superimposition and reconstruction in forensic facial identification: a survey. *Forensic Science International*, 75(2-3):101–120.

[16] Badin, P., Elisei, F., Bailly, G., and Tarabalka, Y. (2008). An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data. In *International Conference on Articulated Motion and Deformable Objects*, pages 132–143. Springer.

[17] Barker, L. J. (2003). Computer-assisted vocabulary acquisition: The cslu vocabulary tutor in oral-deaf education. *Journal of Deaf Studies and Deaf Education*, 8(2):187–198.

[18] Bas, A., Smith, W. A., Bolkart, T., and Wuhrer, S. (2016). Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences. In *Asian Conference on Computer Vision*, pages 377–391. Springer.

[19] Beeler, T., Hahn, F., Bradley, D., Bickel, B., Beardsley, P., Gotsman, C., Sumner, R. W., and Gross, M. (2011). High-quality passive facial performance capture using anchor frames. In *ACM Transactions on Graphics (TOG)*, volume 30, page 75. ACM.

[20] Berkovitz, B. K., Holland, G. R., and Moxham, B. J. (2017). *Oral Anatomy, Histology and Embryology E-Book*. Elsevier Health Sciences.

[21] Beskow, J. (1995). Rule-based visual speech synthesis. In *Fourth European Conference on Speech Communication and Technology*.

[22] Beskow, J. (2004). Trainable articulatory control models for visual speech synthesis. *International Journal of Speech Technology*, 7(4):335–349.

[23] Besl, P. J. and McKay, N. D. (1992). Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics.

[24] Bettinger, F., Cootes, T. F., and Taylor, C. J. (2002). Modelling facial behaviours. In *BMVC*, volume 2, pages 797–806.

[25] Bickel, B., Botsch, M., Angst, R., Matusik, W., Otaduy, M., Pfister, H., and Gross, M. (2007). Multi-scale capture of facial geometry and motion. In *ACM transactions on graphics (TOG)*, volume 26, page 33. ACM.

[26] Binnie, C. A., Montgomery, A. A., and Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of speech and hearing research*, 17(4):619–630.

[27] Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one*, 8(4):e60603.

[28] Black, M. J. and Yacoob, Y. (1995). Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proceedings of IEEE international conference on computer vision*, pages 374–381. IEEE.

[29] Blake, R. (2016). Technology and the four skills. *Language Learning & Technology*, 20(2):129–142.

[30] Blanz, V., Basso, C., Poggio, T., and Vetter, T. (2003). Reanimating faces in images and video. In *Computer graphics forum*, volume 22, pages 641–650. Wiley Online Library.

[31] Blanz, V., Romdhani, S., and Vetter, T. (2002). Face identification across different poses and illuminations with a 3d morphable model. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 202–207. IEEE.

[32] Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co.

[33] Boehnen, C. and Flynn, P. (2005). Accuracy of 3d scanning technologies in a face scanning scenario. In *Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM'05)*, pages 310–317. IEEE.

[34] Booth, J., Roussos, A., Ponniah, A., Dunaway, D., and Zafeiriou, S. (2018). Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2-4):233–254.

[35] Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., and Dunaway, D. (2016). A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5543–5552.

[36] Borshukov, G. and Lewis, J. P. (2005). Realistic human face rendering for the matrix reloaded. In *ACM Siggraph 2005 Courses*, page 13. ACM.

[37] Bouaziz, S., Wang, Y., and Pauly, M. (2013). Online modeling for realtime facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):40.

[38] Brand, M. and Bhotika, R. (2001). Flexible flow for 3d nonrigid tracking and shape recovery. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE.

[39] Bregler, C., Covell, M., and Slaney, M. (1997). Video rewrite: driving visual speech with audio. In *Siggraph*, volume 97, pages 353–360.

[40] Cao, C., Bradley, D., Zhou, K., and Beeler, T. (2015). Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (ToG)*, 34(4):46.

[41] Cao, C., Weng, Y., Lin, S., and Zhou, K. (2013a). 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):41.

[42] Cao, C., Weng, Y., Zhou, S., Tong, Y., and Zhou, K. (2013b). Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425.

[43] Cao, Y., Faloutsos, P., Kohler, E., and Pighin, F. (2004). Real-time speech motion synthesis from recorded motions. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 345–353. Eurographics Association.

[44] Carey, J. C., Cohen, M. M., Curry, C. J., Devriendt, K., Holmes, L. B., and Verloes, A. (2009). Elements of morphology: standard terminology for the lips, mouth, and oral region. *American Journal of Medical Genetics Part A*, 149(1):77–92.

[45] Carretero, M. R., You, L., Xiao, Z., and Zhang, J. J. (2019). Hybrid integration of euclidean and geodesic distance-based rbf interpolation for facial animation. *ICGST Journal of Graphics, Vision and Image Processing*, 19(2):1–7.

[46] Cassell, J., Sullivan, J., Churchill, E., and Prevost, S. (2000). *Embodied conversational agents*. MIT press.

[47] Celce-Murcia, M. and McIntosh, L. (1991). Teaching english as a second or foreign language.

[48] Çeliktutan, O., Ulukaya, S., and Sankur, B. (2013). A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, 2013(1):13.

[49] Chai, J.-x., Xiao, J., and Hodgins, J. (2003). Vision-based control of 3d facial animation. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 193–206. Eurographics Association.

[50] Chang, F.-J., Tran, A. T., Hassner, T., Masi, I., Nevatia, R., and Medioni, G. (2018). Expnet: Landmark-free, deep, 3d facial expressions. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 122–129. IEEE.

[51] Chang, S.-F., Sikora, T., and Purl, A. (2001). Overview of the mpeg-7 standard. *IEEE Transactions on circuits and systems for video technology*, 11(6):688–695.

[52] Charalambous, C., Yumak, Z., and van der Stappen, A. F. (2019). Audio-driven emotional speech animation for interactive virtual characters. *Computer Animation and Virtual Worlds*, 30(3-4):e1892.

[53] Chen, T. (2001). Audiovisual speech processing. *IEEE Signal Processing Magazine*, 18(1):9–21.

[54] Choe, B., Lee, H., and Ko, H.-S. (2001). Performance-driven muscle-based facial animation. *The Journal of Visualization and Computer Animation*, 12(2):67–79.

[55] Chung, J. S., Jamaludin, A., and Zisserman, A. (2017). You said that? *arXiv preprint arXiv:1705.02966*.

[56] Clement, J. and Ranson, D. (1998). *Craniofacial Identification in Forensic Medicine*. Taylor & Francis.

[57] Cohen, M. M. and Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech. In *Models and techniques in computer animation*, pages 139–156. Springer.

[58] Cohen, M. M., Massaro, D. W., and Clark, R. (2002). Training a talking head. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, page 499. IEEE Computer Society.

[59] Cooke, M., Barker, J., Cunningham, S., and Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424.

[60] Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685.

[61] Cosatto, E. and Graf, H. P. (1998). Sample-based synthesis of photo-realistic talking heads. In *Proceedings Computer Animation'98 (Cat. No. 98EX169)*, pages 103–110. IEEE.

[62] Cosatto, E. and Graf, H. P. (2000). Photo-realistic talking-heads from image samples. *IEEE Transactions on multimedia*, 2(3):152–163.

[63] Cosi, P., Caldognetto, E. M., Perin, G., and Zmarich, C. (2002a). Labial coarticulation modeling for realistic facial animation. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 505–510. IEEE.

[64] Cosi, P., Cohen, M. M., and Massaro, D. W. (2002b). Baldini: Baldi speaks italian! In *Seventh International Conference on Spoken Language Processing*.

[65] Cosi, P., Fusaro, A., and Tisato, G. (2003). Lucia a new italian talking-head based on a modified cohen-massaro's labial coarticulation model. In *Eighth European Conference on Speech Communication and Technology*.

[66] Cosker, D., Paddock, S., Marshall, D., Rosin, P. L., and Rushton, S. (2005). Toward perceptually realistic talking heads: Models, methods, and mcgurk. *ACM Transactions on Applied Perception (TAP)*, 2(3):270–285.

[67] Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., and Black, M. J. (2019). Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10101–10111.

[68] De Menezes, M., Rosati, R., Baga, I., Mapelli, A., and Sforza, C. (2011). Three-dimensional analysis of labial morphology: effect of sex and age. *International journal of oral and maxillofacial surgery*, 40(8):856–861.

[69] DeCarlo, D., Metaxas, D., and Stone, M. (1998). An anthropometric face model using variational techniques. In *SIGGRAPH*, volume 98, pages 67–74.

[70] Deng, Z., Chiang, P.-Y., Fox, P., and Neumann, U. (2006a). Animating blendshape faces by cross-mapping motion capture data. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, pages 43–48. ACM.

[71] Deng, Z., Neumann, U., Lewis, J. P., Kim, T.-Y., Bulut, M., and Narayanan, S. (2006b). Expressive facial animation synthesis by learning speech coarticulation and expression spaces. *IEEE transactions on visualization and computer graphics*, 12(6):1523–1534.

[72] Dey, P., Maddock, S., and Nicolson, R. (2010a). A talking head for speech tutoring. In *Proceedings of the SSPNET 2nd International Symposium on Facial Analysis and Animation*, pages 14–14.

[73] Dey, P., Maddock, S. C., and Nicolson, R. (2010b). Evaluation of a viseme-driven talking head. In *TPCG*, pages 139–142.

[74] Dinev, D., Beeler, T., Bradley, D., Bächer, M., Xu, H., and Kavan, L. (2018). User-guided lip correction for facial performance capture. In *Computer Graphics Forum*, volume 37, pages 93–101. Wiley Online Library.

[75] Dohen, M., Loevenbruck, H., and Hill, H. (2009). Recognizing prosody from the lips: Is it possible to extract prosodic focus from lip features? In *Visual speech recognition: Lip segmentation and mapping*, pages 416–438. IGI Global.

[76] Duan, R., Zhang, J., Cao, W., and Xie, Y. (2014). A preliminary study on asr-based detection of chinese mispronunciation by japanese learners. In *Fifteenth annual conference of the international speech communication association.*

[77] Eberhardt, S. P., Bernstein, L. E., Demorest, M. E., and Goldstein Jr, M. H. (1990). Speechreading sentences with single-channel vibrotactile presentation of voice fundamental frequency. *The Journal of the Acoustical Society of America*, 88(3):1274–1285.

[78] Edge, J. D. (2004). *Techniques for the Synthesis of Visual Speech.* PhD thesis, Department of Computer Science, University of Sheffield.

[79] Edwards, P., Landreth, C., Fiume, E., and Singh, K. (2016). Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics (TOG)*, 35(4):127.

[80] Egger, B., Smith, W. A., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., et al. (2019). 3d morphable face models–past, present and future. *arXiv preprint arXiv:1909.01815.*

[81] Ekman, P. and Friesen, W. V. (1978). Facial action coding system. *Consulting Psychologists Press, Palo Alto, CA.*

[82] Elisei, F., Odisio, M., Bailly, G., and Badin, P. (2001). Creating and controlling video-realistic talking heads. In *AVSP*, pages 90–97. Citeseer.

[83] Engwall, O. (2012). Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. *Computer Assisted Language Learning*, 25(1):37–64.

[84] Engwall, O. and Bälter, O. (2007). Pronunciation feedback from real and virtual language teachers. *Computer Assisted Language Learning*, 20(3):235–262.

[85] Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of speech and hearing research*, 12(2):423–425.

[86] Ezzat, T., Geiger, G., and Poggio, T. (2002). *Trainable videorealistic speech animation*, volume 21. ACM.

[87] Ezzat, T. and Poggio, T. (1998). Miketalk: A talking facial display based on morphing visemes. In *Proceedings Computer Animation'98 (Cat. No. 98EX169)*, pages 96–102. IEEE.

[88] Fan, B., Wang, L., Soong, F. K., and Xie, L. (2015). Photo-real talking head with deep bidirectional lstm. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4884–4888. IEEE.

[89] Fan, H., Cao, Z., Jiang, Y., Yin, Q., and Doudou, C. (2014). Learning deep face representation. *arXiv preprint arXiv:1403.2802*.

[90] Farkas, L. G., Bryson, W., and Klotz, J. (1980). Is photogrammetry of the face reliable?. *Plastic and Reconstructive surgery*, 66(3):346–355.

[91] Feng, Z.-H., Huber, P., Kittler, J., Christmas, W., and Wu, X.-J. (2015). Random cascaded-regression copse for robust facial landmark detection. *IEEE Signal Processing Letters*, 22(1):76–80.

[92] Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11(4):796–804.

[Flemming] Flemming, E. 24.910 Topics in Linguistic Theory: Laboratory Phonology, Spring 2007. (Massachusetts Institute of Technology: MIT OpenCourseWare. http://ocw.mit.edu/. [Online; accessed 4-Novmber-2015]. License: Creative Commons BY-NC-SA.

[94] Florez, M. C. (1998). *Improving adult ESL learners' pronunciation skills*. ERIC, National Clearinghouse for ESL Literacy Education.

[95] Frydrych, M., Kätsyri, J., Dobsík, M., and Sams, M. (2003). Toolkit for animation of finnish talking head. In *AVSP 2003-International Conference on Audio-Visual Speech Processing*.

[96] Fyffe, G., Jones, A., Alexander, O., Ichikari, R., and Debevec, P. (2014). Driving high-resolution facial scans with video performance capture. *ACM Transactions on Graphics (TOG)*, 34(1):8.

[97] Garg, R., Roussos, A., and Agapito, L. (2013). A variational approach to video registration with subspace constraints. *International journal of computer vision*, 104(3):286–314.

[98] Garrido, P., Valgaerts, L., Sarmadi, H., Steiner, I., Varanasi, K., Perez, P., and Theobalt, C. (2015). Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In *Computer graphics forum*, volume 34, pages 193–204. Wiley Online Library.

[99] Garrido, P., Valgaerts, L., Wu, C., and Theobalt, C. (2013). Reconstructing detailed dynamic face geometry from monocular video. *ACM Trans. Graph.*, 32(6):158–1.

[100] Garrido, P., Zollhöfer, M., Casas, D., Valgaerts, L., Varanasi, K., Pérez, P., and Theobalt, C. (2016a). Reconstruction of personalized 3d face rigs from monocular video. *ACM Transactions on Graphics (TOG)*, 35(3):28.

[101] Garrido, P., Zollhöfer, M., Wu, C., Bradley, D., Pérez, P., Beeler, T., and Theobalt, C. (2016b). Corrective 3d reconstruction of lips from monocular video. *ACM Trans. Graph.*, 35(6):219–1.

[102] Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., and Vetter, T. (2018). Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE.

[103] Ghosh, A., Fyffe, G., Tunwattanapong, B., Busch, J., Yu, X., and Debevec, P. (2011). Multiview face capture using polarized spherical gradient illumination. In *ACM Transactions on Graphics (TOG)*, volume 30, page 129. ACM.

[104] Gibelli, D., Codari, M., Rosati, R., Dolci, C., Tartaglia, G. M., Cattaneo, C., and Sforza, C. (2015). A quantitative analysis of lip aesthetics: the influence of gender and aging. *Aesthetic plastic surgery*, 39(5):771–776.

[105] Gick, B., Wilson, I., and Derrick, D. (2012). *Articulatory phonetics.* John Wiley & Sons.

[106] Giegerich, H. J. (1992). *English phonology: An introduction.* Cambridge University Press.

[107] Graf, H. P., Cosatto, E., and Ezzat, T. (2000). Face analysis for the synthesis of photo-realistic talking heads. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 189–194. IEEE.

[108] Guiard-Marigny, T., Tsingos, N., Adjoudani, A., Benoit, C., and Cani, M.-P. (1996). 3d models of the lips for realistic speech animation. In *Computer Animation*, pages 80–89. IEEE Computer Society.

[109] Gutierrez-Osuna, R., Kakumanu, P. K., Esposito, A., Garcia, O. N., Bojórquez, A., Castillo, J. L., and Rudomín, I. (2005). Speech-driven facial animation with realistic dynamics. *IEEE transactions on multimedia*, 7(1):33–42.

[110] Hallgren, A. and Lyberg, B. (1998). Visual speech synthesis with concatenative speech. In *AVSP'98 International Conference on Auditory-Visual Speech Processing.*

[111] Hardcastle, W. J. (1976). *Physiology of speech production: an introduction for speech scientists.* Academic Press.

[112] Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision.* Cambridge university press.

[113] Hashim, H. A. and AlBarakati, S. (2003). Cephalometric soft tissue profile analysis between two different ethnic groups: a comparative study. *J Contemp Dent Pract*, 4(2):60–73.

[114] Hazan, V., Kim, J., and Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication*, 52(11-12):996–1009.

[115] Hazan, V., Sennema, A., Iba, M., and Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by japanese learners of english. *Speech communication*, 47(3):360–378.

[116] Hennekam, R. C., Cormier-Daire, V., Hall, J. G., Méhes, K., Patton, M., and Stevenson, R. E. (2009). Elements of morphology: standard terminology for the nose and philtrum. *American Journal of Medical Genetics Part A*, 149(1):61–76.

[117] Hewlett, N. and Beck, J. M. (2013). *An introduction to the science of phonetics*. Routledge.

[118] Hilder, S., Theobald, B.-J., and Harvey, R. (2010). In pursuit of visemes. In *Auditory-Visual Speech Processing 2010*.

[119] Hişmanoğlu, M. (2007). The [ɔ:] and [ʊ] contrast as a fossilized pronunciation error of turkish learners of english and solutions to the problem. *Journal of Language and Linguistic Studies*, 3(1).

[120] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[121] Honda, K., Kurita, T., Kakita, Y., and Maeda, S. (1995). Physiology of the lips and modelingof lip gestures. *Journal of Phonetics*, 23(1-2):243–254.

[122] Huang, H., Chai, J., Tong, X., and Wu, H.-T. (2011). Leveraging motion capture and 3d scanning for high-fidelity facial performance acquisition. In *ACM Transactions on Graphics (TOG)*, volume 30, page 74. ACM.

[123] Huang, X., Zhang, S., Wang, Y., Metaxas, D., and Samaras, D. (2004). A hierarchical framework for high resolution facial expression tracking. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 22–22. IEEE.

[124] Hubbard, P. (2009). *Computer assisted language learning: Critical concepts in linguistics*. Routledge.

[125] Huber, P., Hu, G., Tena, R., Mortazavian, P., Koppen, P., Christmas, W. J., Ratsch, M., and Kittler, J. (2016). A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*.

[126] Huber, P., Kopp, P., Christmas, W., Rätsch, M., and Kittler, J. (2017). Real-time 3d face fitting and texture fusion on in-the-wild videos. *IEEE Signal Processing Letters*, 24(4):437–441.

[127] Hutton, T. J., Buxton, B. F., Hammond, P., and Potts, H. W. (2003). Estimating average growth trajectories in shape-space using kernel smoothing. *IEEE transactions on medical imaging*, 22(6):747–753.

[128] Hwang, H.-S., Kim, W.-S., and McNamara Jr, J. A. (2002). Ethnic differences in the soft tissue profile of korean and european-american adults with normal occlusions and well-balanced faces. *The Angle Orthodontist*, 72(1):72–80.

[129] Ichim, A. E., Bouaziz, S., and Pauly, M. (2015). Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)*, 34(4):45.

[130] Iscan, M. (1993). Introduction to techniques for photographic comparison; potential and problems. *Forensic analysis of the skull*.

[131] Jackson, P. L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review*.

[132] Jain, V. and Zhang, H. (2006). Robust 3d shape correspondence in the spectral domain. In *IEEE International Conference on Shape Modeling and Applications 2006 (SMI'06)*, pages 19–19. IEEE.

[133] Jeffers, J. and Barley, M. (1980). *Speechreading (lipreading)*. Charles C. Thomas Publisher.

[134] Jensen, H. W., Marschner, S. R., Levoy, M., and Hanrahan, P. (2001). A practical model for subsurface light transport. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 511–518. ACM.

[135] Jiang, L., Zhang, J., Deng, B., Li, H., and Liu, L. (2018). 3d face reconstruction with geometry details from a single image. *IEEE Transactions on Image Processing*, 27(10):4756–4770.

[136] Jolliffe, I. (2011). *Principal component analysis*. Springer.

[137] Joshi, P., Tien, W. C., Desbrun, M., and Pighin, F. (2006). Learning controls for blend shape based realistic facial animation. In *ACM Siggraph 2006 Courses*, page 17. ACM.

[138] Kähler, K., Haber, J., and Seidel, H.-P. (2001). Geometry-based muscle modeling for facial animation. In *Graphics interface*, volume 2001, pages 37–46.

[139] Kaiser, A. R., Kirk, K. I., Lachs, L., and Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research*.

[140] Kalberer, G. A. and Van Gool, L. (2001). Face animation based on observed 3d speech dynamics. In *Proceedings Computer Animation 2001. Fourteenth Conference on Computer Animation (Cat. No. 01TH8596)*, pages 20–251. IEEE.

[141] Kalra, P., Mangili, A., Thalmann, N. M., and Thalmann, D. (1992). Simulation of facial muscle actions based on rational free form deformations. In *Computer Graphics Forum*, volume 11, pages 59–69. Wiley Online Library.

[142] Karras, T., Aila, T., Laine, S., Herva, A., and Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):94.

[143] Kawai, M., Iwao, T., Maejima, A., and Morishima, S. (2015). Automatic generation of photorealistic 3d inner mouth animation only from frontal images. *Journal of information processing*, 23(5):693–703.

[144] Kemelmacher-Shlizerman, I. (2013). Internet based morphable model. In *Proceedings of the IEEE international conference on computer vision*, pages 3256–3263.

[145] King, S. A. (2001). *A facial model and animation techniques for animated speech*. PhD thesis, The Ohio State University.

[146] King, S. A. and Parent, R. E. (2005). Creating speech-synchronized animation. *IEEE Transactions on visualization and computer graphics*, 11(3):341–352.

[147] King, S. A., Parent, R. E., and Olsafsky, B. (2000). An anatomically-based 3d parametric lip model to support facial animation and synchronized speech. In *Proc. Deform 2000*, pages 7–9. Citeseer.

[148] Klaudiny, M. and Hilton, A. (2012). High-detail 3d capture and non-sequential alignment of facial performance. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 17–24. IEEE.

[149] Klause, F., Stone, S., and Birkholz, P. (2017). A head-mounted camera system for the measurement of lip protrusion and opening during speech production. *28.Konferenz Elektronische Sprachsignalverarbeitung, Saarbrücken.*

[150] Klehm, O., Rousselle, F., Papas, M., Bradley, D., Hery, C., Bickel, B., Jarosz, W., and Beeler, T. (2015). Recent advances in facial appearance capture. In *Computer Graphics Forum*, volume 34, pages 709–733. Wiley Online Library.

[151] Koch, R. M., Gross, M. H., and Bosshard, A. A. (1998). Emotion editing using finite elements. In *Computer Graphics Forum*, volume 17, pages 295–302. Wiley Online Library.

[152] Koch, R. M., Gross, M. H., Carls, F. R., von Büren, D. F., Fankhauser, G., and Parish, Y. I. (1996). Simulating facial surgery using finite element models. *Technischer Bericht/Eidgenössische Technische Hochschule, Departement Informatik*, 246.

[153] Koppen, P., Feng, Z.-H., Kittler, J., Awais, M., Christmas, W., Wu, X.-J., and Yin, H.-F. (2018). Gaussian mixture 3d morphable face model. *Pattern Recognition*, 74:617–628.

[154] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

[155] Kshirsagar, S. and Magnenat-Thalmann, N. (2003). Visyllable based speech animation. In *Computer Graphics Forum*, volume 22, pages 631–639. Wiley Online Library.

[156] Ladefoged, P. (2005). *Vowels and consonants.* Blackwell Oxford, UK.

[157] Le Goff, B. and Benoît, C. (1996). A text-to-audiovisual-speech synthesizer for french. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 4, pages 2163–2166. IEEE.

[158] Le Goff, B., Guiard-Marigny, T., Cohen, M. M., and Benoît, C. (1994). Real-time analysis-synthesis and intelligibility of talking faces. In *SSW*, pages 53–56.

[159] Lee, Y., Terzopoulos, D., and Waters, K. (1995). Realistic modeling for facial animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 55–62. ACM.

[160] Levy, M. and Stockwell, G. (2013). *CALL dimensions: Options and issues in computer-assisted language learning.* Routledge.

[161] Lewis, J. P., Anjyo, K., Rhee, T., Zhang, M., Pighin, F. H., and Deng, Z. (2014). Practice and theory of blendshape facial models. *Eurographics (State of the Art Reports)*, 1(8):2.

[162] Li, H., Yu, J., Ye, Y., and Bregler, C. (2013). Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph.*, 32(4):42–1.

[163] Liu, N., Zhou, T., Ji, Y., Zhao, Z., and Wan, L. (2020). Synthesizing talking faces from text and audio: An autoencoder and sequence-to-sequence convolutional neural network. *Pattern Recognition*, 102:107231.

[164] Löfqvist, A. (1990). Speech as audible gestures. In *Speech production and speech modelling*, pages 289–322. Springer.

[165] Ma, J., Cole, R., Pellom, B., Ward, W., and Wise, B. (2004). Accurate automatic visible speech synthesis of arbitrary 3d models based on concatenation of diviseme motion capture data. *Computer Animation and Virtual Worlds*, 15(5):485–500.

[166] Ma, L. and Deng, Z. (2019). Real-time hierarchical facial performance capture. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, page 11. ACM.

[167] Ma, W.-C., Hawkins, T., Peers, P., Chabert, C.-F., Weiss, M., and Debevec, P. (2007). Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 183–194. Eurographics Association.

[168] MacKinnon, P. C. and Morris, J. F. (2005). *Oxford textbook of functional anatomy. Volume 3. Head and neck.* Oxford University Press.

[169] MacLeod, A. and Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British journal of audiology*, 21(2):131–141.

[170] Martin, R. and Saller, K. (1957). Lehrbuch der anthropologie [textbook of anthropology]. *Stuttgart: Fischer.*

[171] Massaro, D., Cohen, M., Tabain, M., Beskow, J., and Clark, R. (2012). 12 animated speech: research progress and applications.

[172] Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*, volume 1. Mit Press.

[173] Massaro, D. W. (2004). Symbiotic value of an embodied agent in language learning. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, pages 10–pp. IEEE.

[174] Massaro, D. W. (2005). The psychology and technology of talking heads: Applications in language learning. In *Advances in Natural Multimodal Dialogue Systems*, pages 183–214. Springer.

[175] Mattheyses, W. and Verhelst, W. (2015). Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66:182–217.

[176] Mattys, S. L., Bernstein, L. E., and Auer, E. T. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception & Psychophysics*, 64(4):667–679.

[177] McConville, J. T. (1976). Human variability and respirator sizing.

[178] McGrath, M. (1985). An examination of cues for visual and audio-visual speech perception using natural and computer-generated faces. *Ph. D Thesis, Univ. of Nottingham.*

[179] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746.

[180] Middelweerd, M. and Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *The Journal of the Acoustical Society of America*, 82(6):2145–2147.

[181] Montgomery, A. A. and Jackson, P. L. (1983). Physical characteristics of the lips underlying vowel lipreading performance. *The Journal of the Acoustical Society of America*, 73(6):2134–2144.

[182] Moore, R. K. (2015). Articulatory phonetics. University Lecture, The University of Sheffield.

[183] Moore, T. J. (1981). Voice communication jamming research. In *Advisory Group for Aerospace Research and Development Conference Proceedings*, number 311. Citeseer.

[184] Morley, J. (1991). The pronunciation component in teaching english to speakers of other languages. *TESOL quarterly*, 25(3):481–520.

[185] Müller, C. (2006). *Spherical harmonics*, volume 17. Springer.

[186] Musti, U., Ouni, S., Zhou, Z., and Pietikäinen, M. (2014). 3d visual speech animation from image sequences. In *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, page 47. ACM.

[187] Öhman, S. E. (1967). Numerical model of coarticulation. *The Journal of the Acoustical Society of America*, 41(2):310–320.

[188] Olson, I. R., Gatenby, J. C., and Gore, J. C. (2002). A comparison of bound and unbound audio–visual information processing in the human cerebral cortex. *Cognitive Brain Research*, 14(1):129–138.

[189] Ostermann, J. and Weissenfeld, A. (2004). Talking faces-technologies and applications. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 826–833. IEEE.

[190] Ouni, S. and Gris, G. (2018). Dynamic lip animation from a limited number of control points: Towards an effective audiovisual spoken communication. *Speech Communication*, 96:49–57.

[191] Ouni, S., Massaro, D. W., Cohen, M. M., Young, K., and Jesse, A. (2003). Internationalization of a talking head. In *Proc. of 15th International Congress of Phonetic Sciences, Barcelona, Spain*, pages 286–318.

[192] Pala, P., Seidenari, L., Berretti, S., and Del Bimbo, A. (2019). Enhanced skeleton and face 3d data for person re-identification from depth cameras. *Computers & Graphics*, 79:69–80.

[193] Pandzic, I. S., Ostermann, J., and Millen, D. (1999). User evaluation: Synthetic talking faces for interactive services. *The visual computer*, 15(7-8):330–340.

[194] Parke, F. I. (1972). Computer generated animation of faces. In *Proceedings of the ACM annual conference-Volume 1*, pages 451–457. ACM.

[195] Parke, F. I. (1974). A parametric model for human faces. Technical report, UTAH UNIV SALT LAKE CITY DEPT OF COMPUTER SCIENCE.

[196] Parke, F. I. and Waters, K. (1996). *Computer facial animation*. CRC press.

[197] Parke, F. I. and Waters, K. (2008). *Computer facial animation*. AK Peters/CRC Press.

[198] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. (2009). A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. Ieee.

[199] Pelachaud, C. (1991). *Communication and coarticulation in facial animation*. University of Pennsylvania. Moore School of Electrical Engineering ....

[200] Pelachaud, C., Badler, N. I., and Steedman, M. (1996). Generating facial expressions for speech. *Cognitive science*, 20(1):1–46.

[201] Penry, J. and Ryan, I. (1971). *Looking at faces and remembering them: A guide to facial identification*. Elek.

[202] Pham, H. X., Wang, Y., and Pavlovic, V. (2018). End-to-end learning for 3d facial animation from speech. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 361–365. ACM.

[203] Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., and Salesin, D. H. (2006). Synthesizing realistic facial expressions from photographs. In *ACM SIGGRAPH 2006 Courses*, pages 19–es.

[204] Platt, S. M. and Badler, N. I. (1981). Animating facial expressions. In *ACM SIGGRAPH computer graphics*, volume 15, pages 245–252. ACM.

[205] Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., and Moreno-Noguer, F. (2018). Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833.

[206] Rabiner, L. R., Juang, B.-H., and Rutledge, J. C. (1993). *Fundamentals of speech recognition*, volume 14. PTR Prentice Hall Englewood Cliffs.

[207] Revéret, L., Bailly, G., and Badin, P. (2000). Mother: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *Int. Conference of Spoken Language Processing, ICSLP'2000.*

[208] Revéret, L. and Benoît, C. (1998). A new 3d lip model for analysis and synthesis of lip motion in speech production.

[209] Roelofse, M. M., Steyn, M., and Becker, P. J. (2008). Photo identification: facial metrical and morphological features in south african males. *Forensic Science International*, 177(2-3):168–175.

[210] Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., and Abrams, H. B. (2006). Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics*, 27(3):465.

[211] Rogers, M., Walker, K., Williams, T. G., Gorce, M. D., and Tosas, M. (2015). Building systems for tracking facial features across individuals and groups. US Patent 9,111,134.

[212] Rosato, M., Chen, X., and Yin, L. (2008). Automatic registration of vertex correspondences for 3d facial expression analysis. In *2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems*, pages 1–7. IEEE.

[213] Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, 17(6):405–409.

[214] Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18.

[215] Salvador, S. and Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580.

[216] Sandbach, G., Zafeiriou, S., Pantic, M., and Yin, L. (2012). Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697.

[217] Schwartz, J.-L., Berthommier, F., and Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, 93(2):B69–B78.

[218] Shi, F., Wu, H.-T., Tong, X., and Chai, J. (2014). Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Transactions on Graphics (TOG)*, 33(6):222.

[219] Skipper, J. I., Small, S. L., Nusbaum, H. C., and van Wassenhove, V. (2007). Hearing Lips and Seeing Voices: How Cortical Areas Supporting Speech Production Mediate Audiovisual Speech Perception. *Cerebral Cortex*, 17(10):2387–2399.

[220] Stewart, T. D. (1947). *Hrdlička's practical anthropometry*. Wistar Institute fo Anatomy and Biology.

[221] Stropahl, M. and Debener, S. (2017). Auditory cross-modal reorganization in cochlear implant users indicates audio-visual integration. *NeuroImage: Clinical*, 16:514–523.

[222] Stylianou, G. and Lanitis, A. (2009). Image based 3d face reconstruction: a survey. *International Journal of Image and Graphics*, 9(02):217–250.

[223] Sumby, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215.

[224] Summerfield, Q. (1976). Some preliminaries to comprehensive account of audio-visual speech perception. *Hearing by eye: the psychology of lipreading*, 3:746–748.

[225] Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica*, 36(4-5):314–331.

[226] Summerfield, Q. (1989). Lips, teeth, and the benefits of lipreading. *Handbook of research on face processing*, pages 223–233.

[227] Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 335(1273):71–78.

[228] Sumner, R. W. and Popović, J. (2004). Deformation transfer for triangle meshes. *ACM Transactions on graphics (TOG)*, 23(3):399–405.

[229] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[230] Suwajanakorn, S., Kemelmacher-Shlizerman, I., and Seitz, S. M. (2014). Total moving face reconstruction. In *European Conference on Computer Vision*, pages 796–812. Springer.

[231] Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):1–13.

[232] Tang, H. and Huang, T. S. (2008). 3d facial expression recognition based on automatically selected features. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 1–8. IEEE.

[233] Tao, H., Chen, H. H., Wu, W., and Huang, T. S. (1999). Compression of mpeg-4 facial animation parameters for transmission of talking heads. *IEEE Transactions on circuits and systems for video technology*, 9(2):264–276.

[234] Tatham, M. (1969). The control of muscles in speech. *Occas. Pap*, (3).

[235] Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A. G., Hodgins, J., and Matthews, I. (2017). A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):1–11.

[236] Taylor, S. L., Mahler, M., Theobald, B.-J., and Matthews, I. (2012). Dynamic units of visual speech. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 275–284. Eurographics Association.

[237] Terzopoulos, D. and Waters, K. (1993). Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579.

[238] Thangthai, A., Milner, B., and Taylor, S. (2016). Visual speech synthesis using dynamic visemes, contextual features and dnns.

[239] Theobald, B.-J., Bangham, J. A., Matthews, I. A., Glauert, J. R., and Cawley, G. C. (2003). 2.5 d visual speech synthesis using appearance models. In *BMVC*, pages 1–10.

[240] Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., and Theobalt, C. (2015). Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6):183–1.

[241] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395.

[242] Tolba, R. M., Al-Arif, T., and El Horbaty, E.-S. M. (2018). Realistic facial animation review: Based on facial action coding system. *Egyptian Computer Science Journal*, 42(1).

[243] Trujillo, F. (2001). Speech production process. University Lecture, University of Granada.

[244] Tyler, R. S., Fryauf-Bertschy, H., Kelsay, D. M., Gantz, B. J., WOODWORTH, G. P., Parkinson, A., et al. (1997). Speech perception by prelingually deaf children using cochlear implants. *Otolaryngology-Head and Neck Surgery*, 117(3):180–187.

[245] Uysal, T., Baysal, A., Yagci, A., Sigler, L. M., and McNamara Jr, J. A. (2012). Ethnic differences in the soft tissue profiles of turkish and european–american young adults with normal occlusions and well-balanced faces. *The European Journal of Orthodontics*, 34(3):296–301.

[246] Vlasic, D., Brand, M., Pfister, H., and Popović, J. (2005). Face transfer with multilinear models. In *ACM transactions on graphics (TOG)*, volume 24, pages 426–433. ACM.

[247] Walder, C., Breidt, M., Bülthoff, H., Schölkopf, B., and Curio, C. (2009). Markerless 3d face tracking. In *Joint Pattern Recognition Symposium*, pages 41–50. Springer.

[248] Wang, L., Chen, H., Li, S., and Meng, H. M. (2012). Phoneme-level articulatory animation in pronunciation training. *Speech Communication*, 54(7):845–856.

[249] Wang, Y., Huang, X., Lee, C.-S., Zhang, S., Li, Z., Samaras, D., Metaxas, D., Elgammal, A., and Huang, P. (2004). High resolution acquisition, learning and transfer of dynamic 3-d facial expressions. In *Computer Graphics Forum*, volume 23, pages 677–686. Wiley Online Library.

[250] Wängler, H. (1972). Physiologische phonologie. *Marburg: N. G. Elwert Verlag.*

[251] Waters, K. (1987). A muscle model for animation three-dimensional facial expression. *Acm siggraph computer graphics*, 21(4):17–24.

[252] Wei, L. and Deng, Z. (2014). A practical model for live speech-driven lip-sync. *IEEE computer graphics and applications*, 35(2):70–78.

[253] Weise, T., Bouaziz, S., Li, H., and Pauly, M. (2011). Realtime performance-based facial animation. In *ACM transactions on graphics (TOG)*, volume 30, page 77. ACM.

[254] Wilson, C. A., Ghosh, A., Peers, P., Chiang, J.-Y., Busch, J., and Debevec, P. (2010). Temporal upsampling of performance geometry using photometric alignment. *ACM Transactions on Graphics (TOG)*, 29(2):17.

[255] Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

[256] Wu, C., Bradley, D., Gross, M., and Beeler, T. (2016). An anatomically-constrained local deformation model for monocular face capture. *ACM Transactions on Graphics (TOG)*, 35(4):115.

[257] Xie, L., Wang, L., and Yang, S. (2016). Visual speech animation. *Handbook of Human Motion, Springer (Ed.). Springer*, pages 1–30.

[258] Xie, W., Shen, L., Yang, M., and Jiang, J. (2018). Facial expression synthesis with direction field preservation based mesh deformation and lighting fitting based wrinkle mapping. *Multimedia Tools and Applications*, 77(6):7565–7593.

[259] Xu, Y., Feng, A. W., Marsella, S., and Shapiro, A. (2013). A practical and configurable lip sync method for games. In *Proceedings of Motion on Games*, pages 131–140. ACM.

[260] Yu, J., Jiang, C., Li, R., Luo, C.-W., and Wang, Z.-F. (2016). Real-time 3d facial animation: From appearance to internal articulators. *IEEE Transactions on Circuits and Systems for Video Technology.*

[261] Yu, L., Yu, J., and Ling, Q. (2019). Deep neural network based 3d articulatory movement prediction using both text and audio inputs. In *International Conference on Multimedia Modeling*, pages 68–79. Springer.

[262] Zhang, L., Snavely, N., Curless, B., and Seitz, S. M. (2008). Spacetime faces: High-resolution capture for˜ modeling and animation. In *Data-Driven 3D Facial Animation*, pages 248–276. Springer.

[263] Zhao, Y., Oveneke, M. C., Jiang, D., and Sahli, H. (2018). A video prediction approach for animating single face image. *Multimedia Tools and Applications*, pages 1–22.

[264] Zhou, H., Liu, Y., Liu, Z., Luo, P., and Wang, X. (2019a). Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9299–9306.

[265] Zhou, W., Zhao, C., Lu, L., and Zhao, Q. (2019b). Dense correspondence of 3d facial point clouds via neural network fitting. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3731–3735. IEEE.

[266] Zhou, Y., Xu, Z., Landreth, C., Kalogerakis, E., Maji, S., and Singh, K. (2018). Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)*, 37(4):1–10.

[267] Zhuang, Z., Guan, J., Hsiao, H., and Bradtmiller, B. (2004). Evaluating the representativeness of the lanl respirator fit test panels for the current us civilian workers. *Journal of the International Society for Respiratory Protection*, 21:83–93.

[268] Zhuang, Z., Landsittel, D., Benson, S., Roberge, R., and Shaffer, R. (2010). Facial anthropometric differences among gender, ethnicity, and age groups. *Annals of occupational hygiene*, 54(4):391–402.

[269] Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., and Theobalt, C. (2018). State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer Graphics Forum*, volume 37, pages 523–550. Wiley Online Library.

Appendix A

Mapping 2D to 3D

Investigating Functionality of Facial Feature Landmarks in the Mapping Process

Consecutive Frames of the 3D Lip Motions

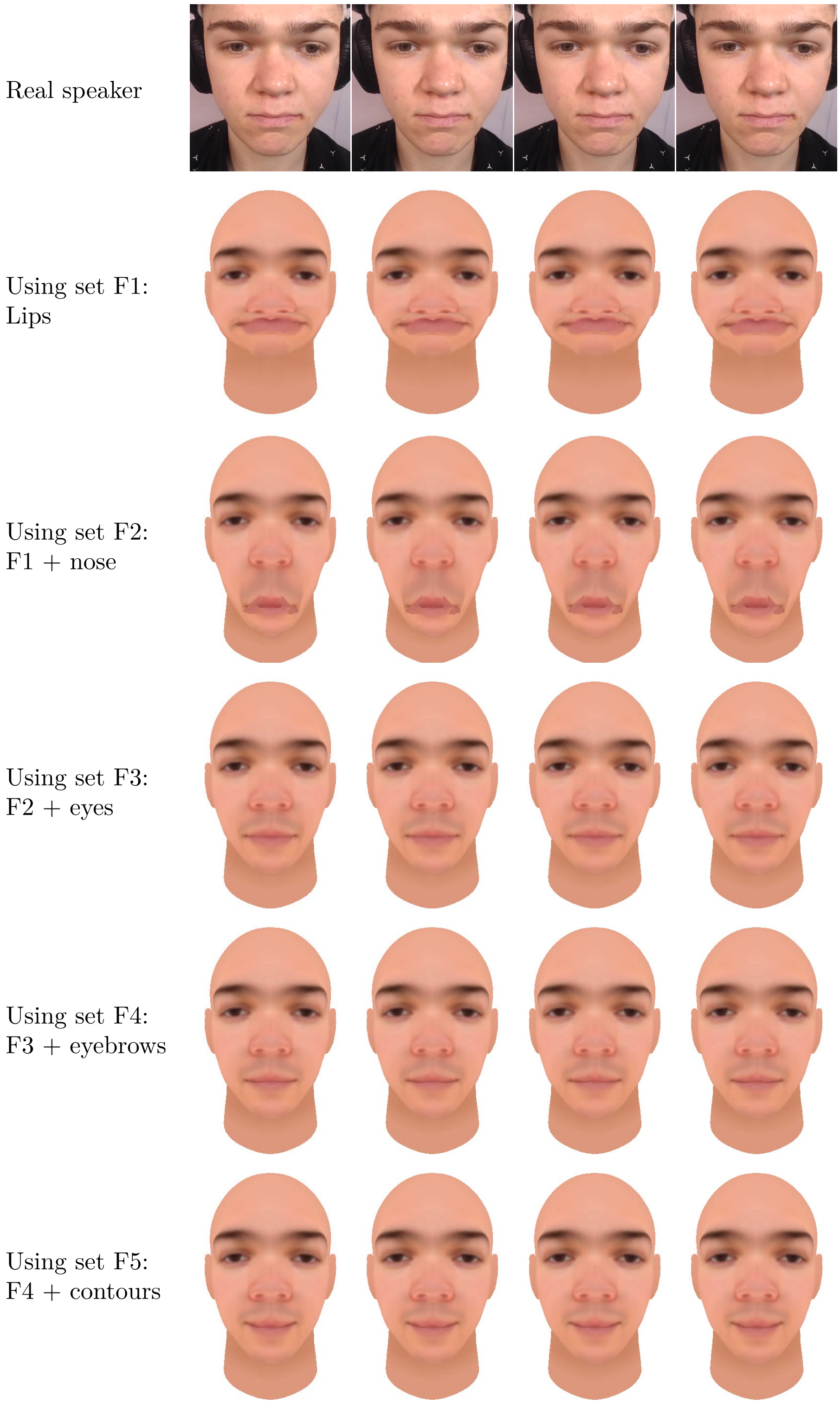


Figure A.1 Consecutive frames of the phoneme /b/ during utterance of the word bin from sentence "bin white with V 7 soon" for a real speaker (ID: S 48) and the corresponding 3D head animated using each set of facial features landmarks.

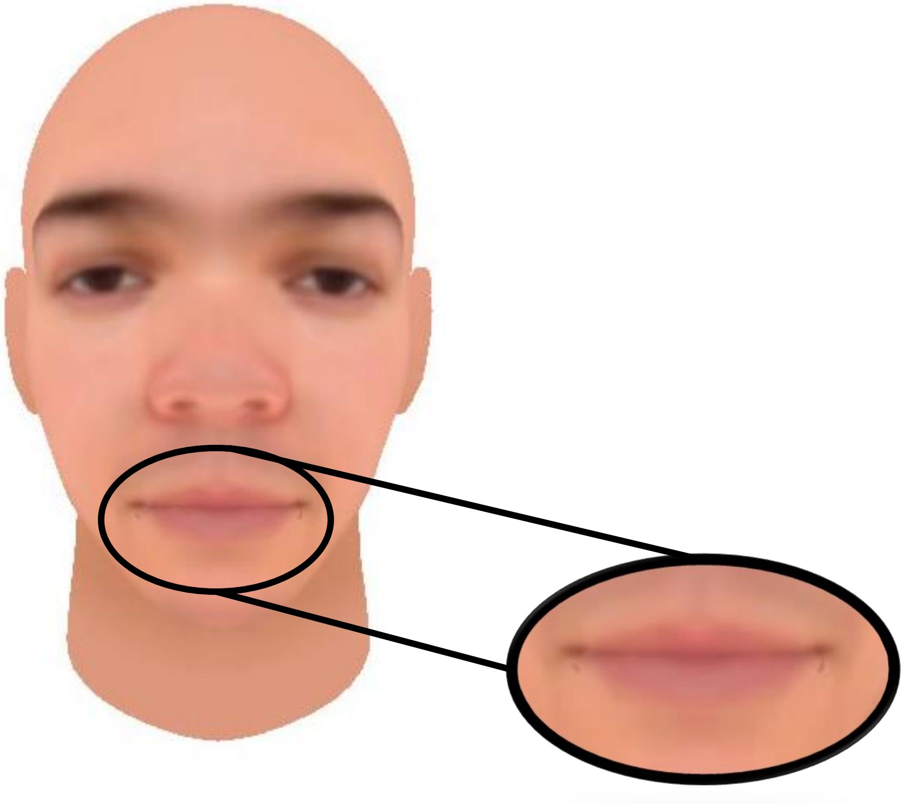


Figure A.2 An example of texture distortion around the mouth area, due to using the set F3 for animating the 3D lips.

Appendix B

3D Visual Speech Animation Using 2D Videos

This appendix provides more example figures of the Front and side-view of initial natural 3D head poses, and frames and trajectories of the 3D lip motion (Chapter 4).

Front and Side-view of Initial Natural 3D Head Poses

This section shows more examples of the front and side-view photographs of real speakers (IDs: S17, S48 and S24) and their corresponding 3D heads that generated using a front-view photograph only (left), and front- and side-view photographs (right).

Consecutive Frames of the 3D Lip Motions

The following figures show more examples of consecutive frames of the real speakers and the corresponding 3D heads that generated using the four datasets that presented in section 4.2.

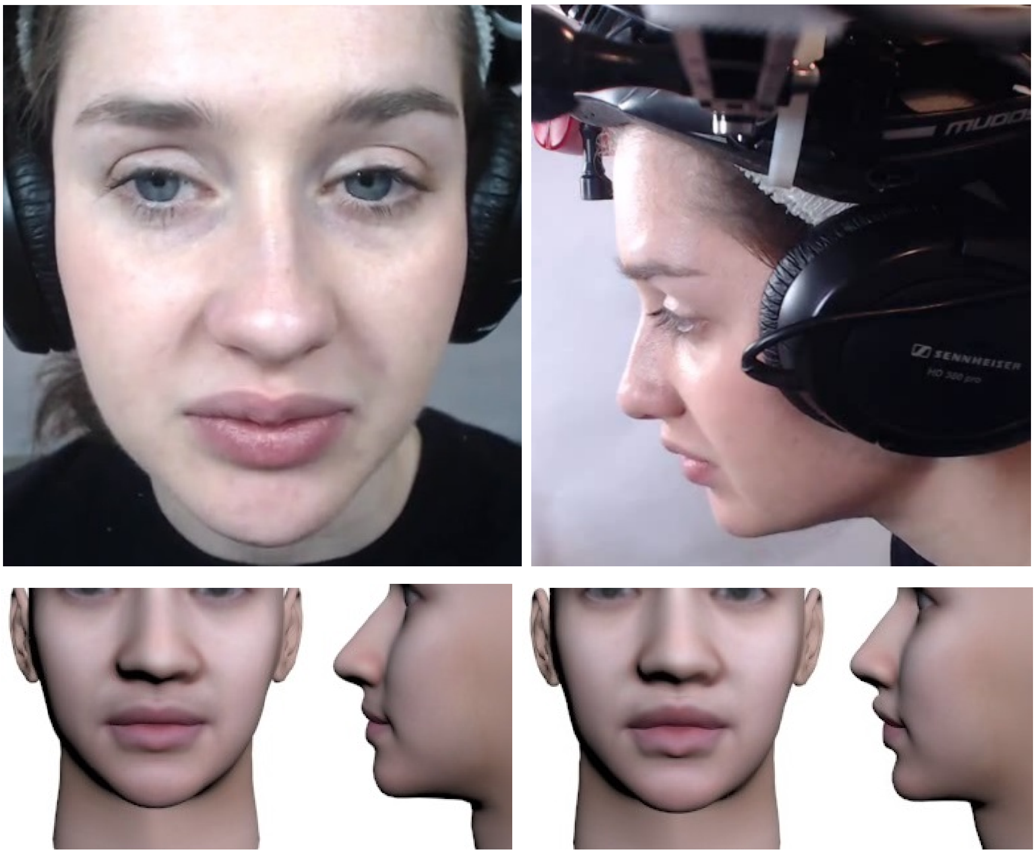


Figure B.1 First row: Front (left) and side (right) photographs of a real speaker (ID: S17); Second row: front and side view of the corresponding 3D heads generated using front photograph only (left) and front and side photographs (right) – the lips are more protruded in the image on the right.

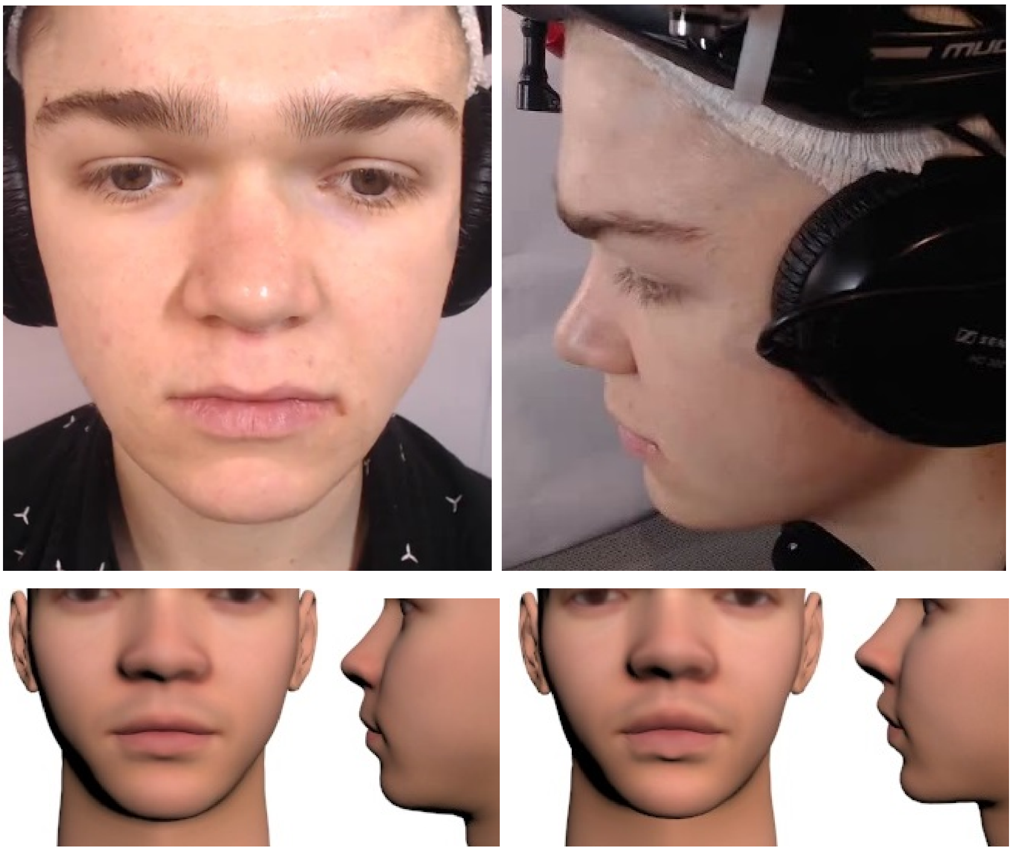


Figure B.2 First row: Front (left) and side (right) photographs of a real speaker (ID: S48); Second row: front and side view of the corresponding 3D heads generated using front photograph only (left) and front and side photographs (right) – the lips are more protruded in the image on the right.

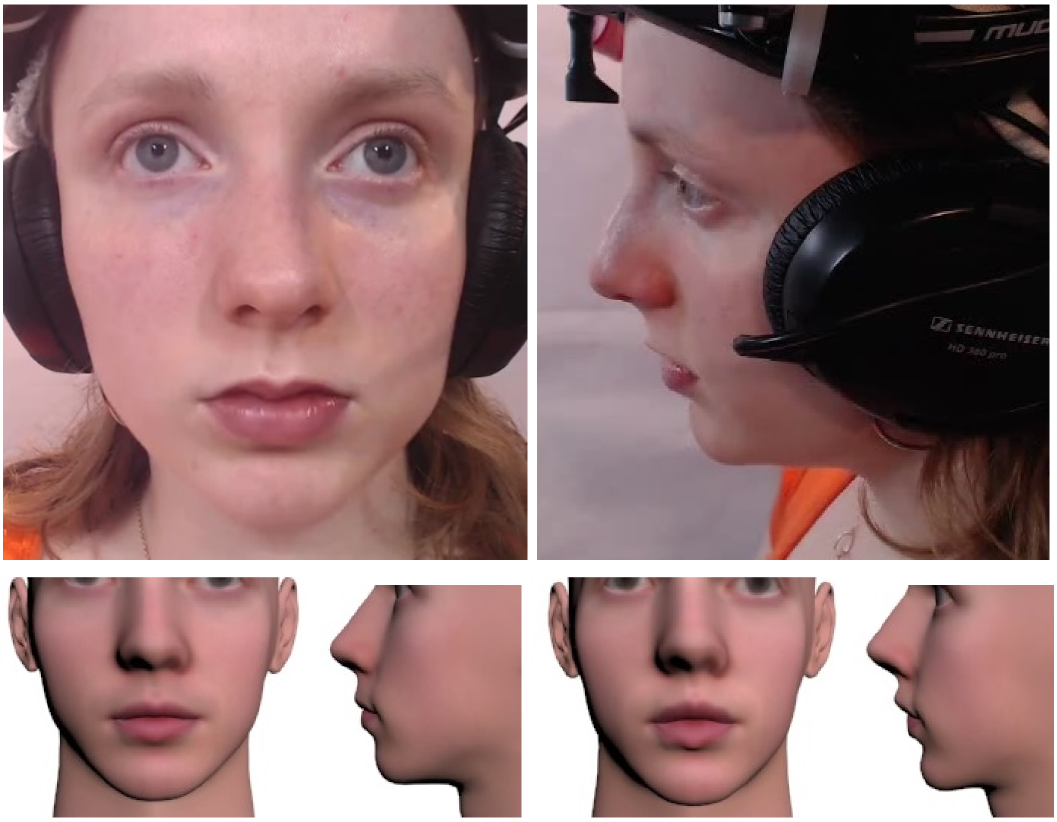


Figure B.3 First row: Front (left) and side (right) photographs of a real speaker (ID: S24); Second row: front and side view of the corresponding 3D heads generated using front photograph only (left) and front and side photographs (right) – the lips are more protruded in the image on the right.

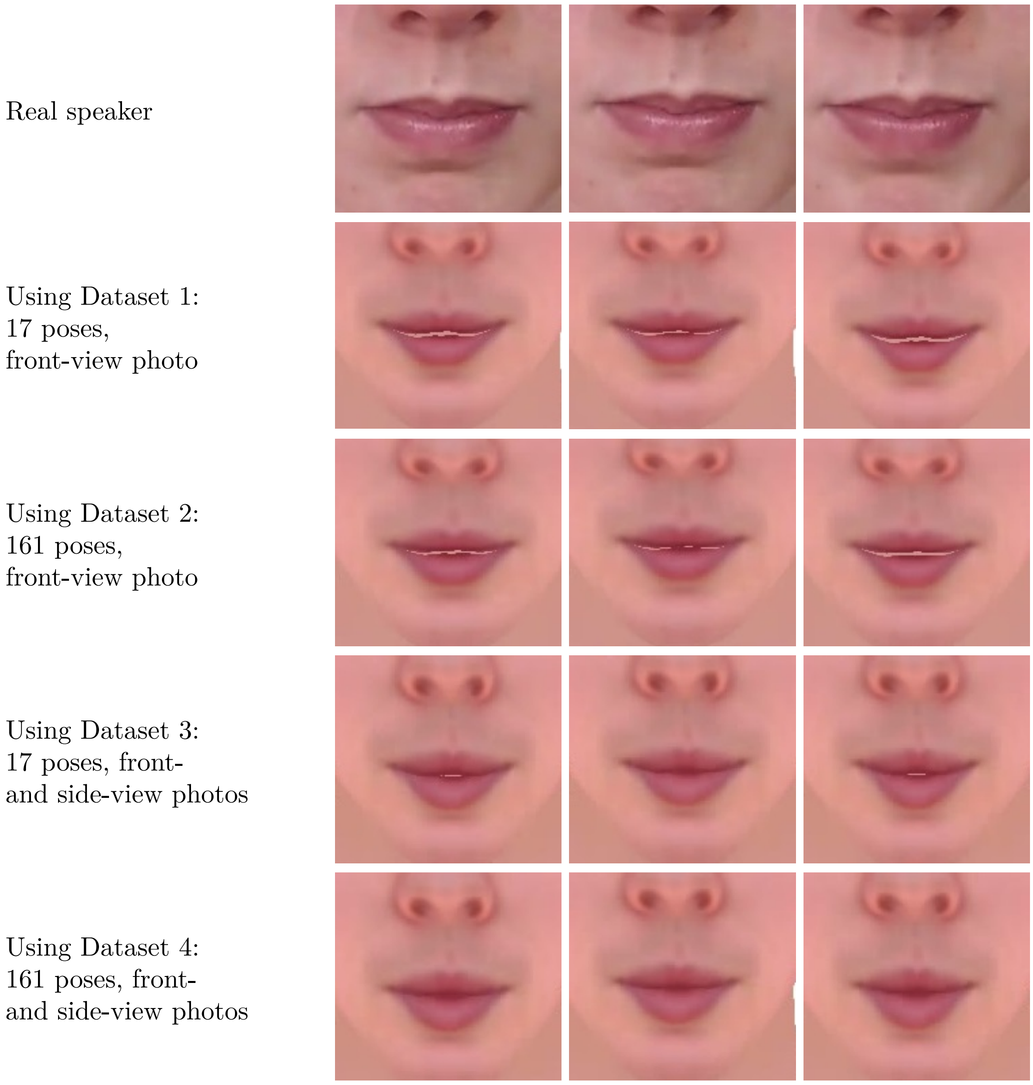
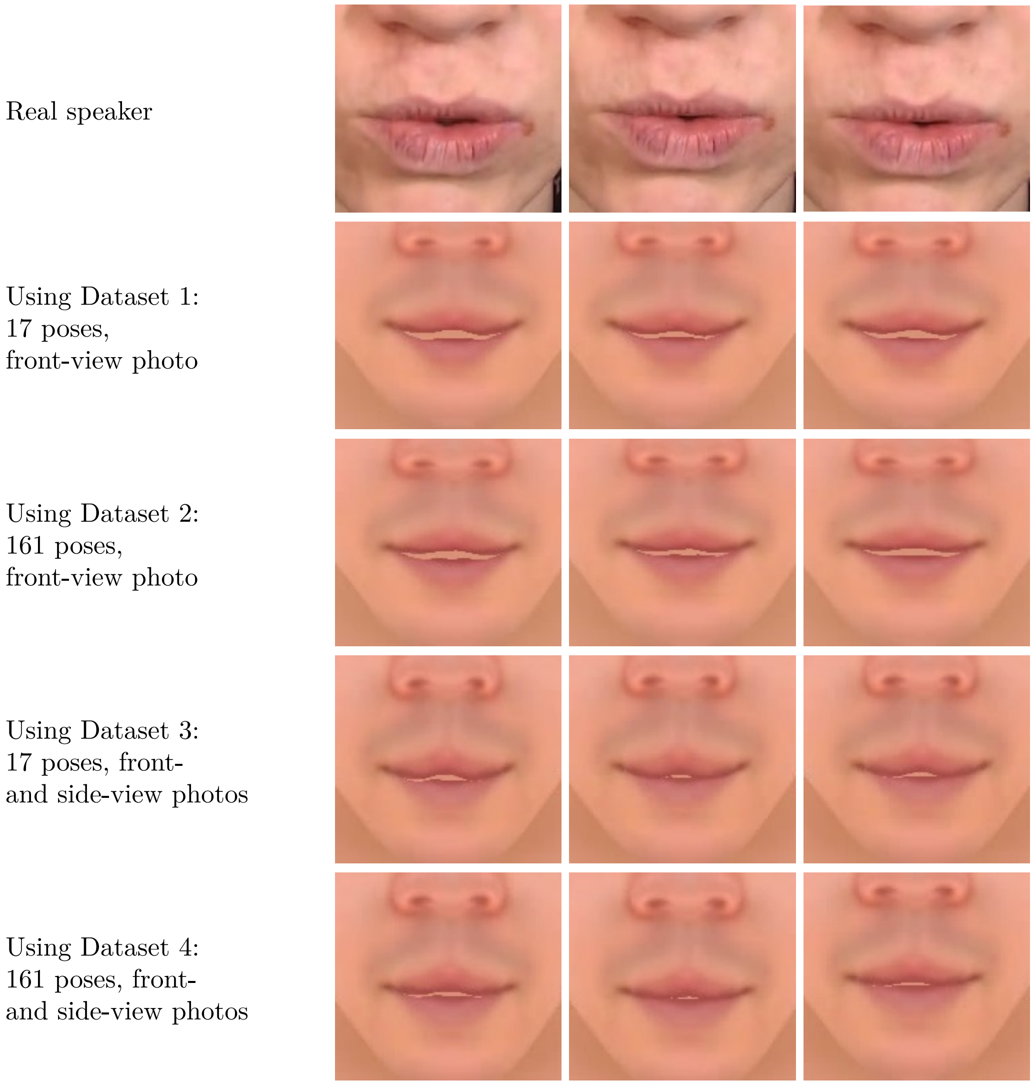


Figure B.4 Consecutive frames of the phoneme /b/ during utterance of the word bin from the sentence "bin white in N 3 now" for a real speaker (ID: S24) and the corresponding 3D head for each data set.

3D lip Motions Trajectories

The figure below shows width and height of mouth trajectories of 2D frames of the real speaker (ID:S32) and the corresponding 3D heads, during utterance of the sentence "place blue at Y 4 now". Top two compare height and width between 17 and 161 poses



Real speaker

Using Dataset 1:
17 poses,
front-view photo

Using Dataset 2:
161 poses,
front-view photo

Using Dataset 3:
17 poses, front-
and side-view photos

Using Dataset 4:
161 poses, front-
and side-view photos

Figure B.5 Consecutive frames of the phoneme /uw/ during utterance of the word soon from the sentence "bin white with V 7 soon" for a real speaker (ID: S48) and the corresponding 3D head for each data set.

(both with front- and side-view photos), while the bottom two compare height and width between front- view photo only and front- and side-view photos (both with 161 poses).

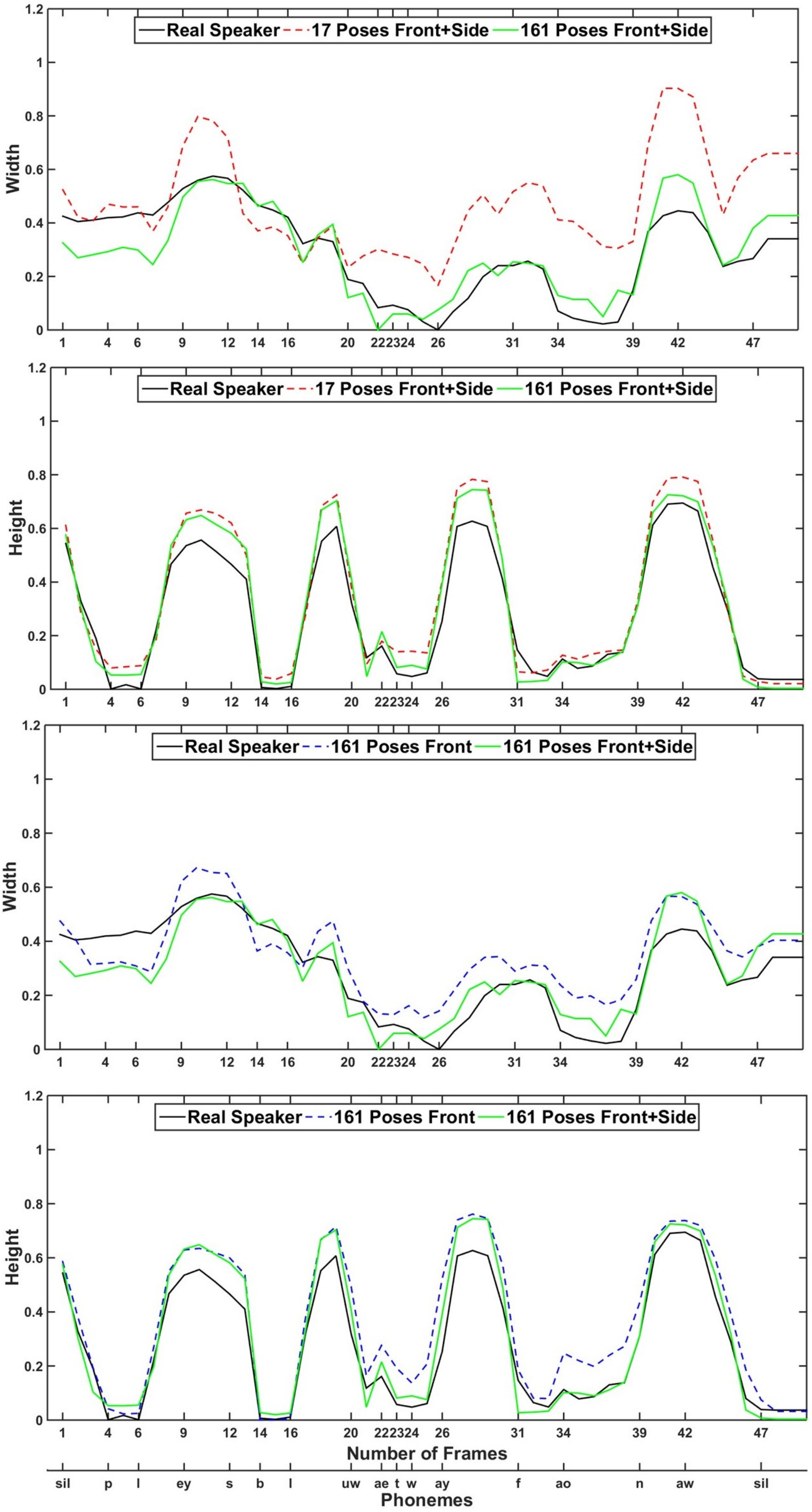


Figure B.6 Width and height of mouth trajectories of 2D frames of the real speaker (ID:S32) and the corresponding 3D heads. Top two compare height and width between 17 and 161 poses (both with front- and side-view photos), while the bottom two compare height and width between front- view photo only and front- and side-view photos (both with 161 poses).