Visual Analysis of Viseme Dynamics

A Turkmani

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey

Centre for Vision, Speech and Signal Processing
Faculty of Engineering and Physical Sciences
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

September 2007

© A Turkmani 2007
Abstract

Face to face dialogue is the most natural mode of communication between humans. The combination of human visual perception of expression and perception in changes in intonation provides semantic information that communicates idea, feelings and concepts. The realistic modelling of speech movements, through automatic facial animation, and maintaining audio-visual coherence is still a challenge in both the computer graphics and film industry. A common approach to producing visual speech is to interpolate parameters that describe mouth variation in sequence, known as visemes. A viseme corresponds to a phoneme in an utterance. Most talking head systems use sets of static visemes, represented by a single mouth shape image or 3D model. However, discretising visemes in this way does not account for context-dependent dynamic information, coarticulation.

This thesis presents several visual analysis and dynamic modelling techniques for visual phones. This spans several areas of work, from capture and representation through to analysis and synthesis of speech movements and coarticulation. A novel method is reported for the automatic extraction of inner-lip contour edges from sequences of mouth images in speech. The proposed detection technique is a key-frame exemplar-based method that is not dependent on any prior frame information for intitialisation allowing for reliable and accurate inner-lip localisation for large frame to frame changes in lip-shape inherent in 25Hz video of visual speech.

Visual analysis of phonemes in continuous speech is performed, that involves the investigation of mouth representations as well as a comparative analysis between static and dynamic representations of visemes. The analysis shows the need to analyse and model the underlying dynamics of visemes due to coarticulation. Finally, visual analysis of lip coarticulation in Vowel-Consonant-Vowel (VCV) utterances is presented. Based on ensemble statistics a novel approach to analysis and modelling of temporal dynamics is presented. Results show that the temporal influence of coarticulation is significant both in lip shape variation and timings of lip movement during coarticulation. This work shows that the effect of temporal variation due to coarticulation is statistically significant and should be taken into account in modelling visual speech synthesis.

The work in this thesis provides the foundation for further research towards achieving perceptually realistic animation of a talking head and the understanding of visual dynamics of shape and texture during speech.

Key words: Viseme, Dynamics, Coarticulation
Acknowledgements

I’d like to thank all my colleagues and supervisors for their support over the last few years. I’d like to specifically thank my supervisor Adrian Hilton for his great support and encouragement. I’d also like to thank my friend Gregor Miller for sharing the experience of a PhD with me. We shared many moans but it really did get me through. I’d like to say thank you to God. The past 5 years have been quite a journey both academically and personally but I know that my God was with me through it all. I’d like to say thank you to my father Adel Turkmani for pushing me academically since I was young. You were always such a great inspiration to me. Finally, I’d like to say a very big thankyou to my wonderful husband Andy who’s love and support is invaluable to me.
## Nomenclature

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D</td>
<td>Two Dimensions</td>
</tr>
<tr>
<td>3D</td>
<td>Three Dimensions</td>
</tr>
<tr>
<td>AAM</td>
<td>Active Appearance Model</td>
</tr>
<tr>
<td>ASM</td>
<td>Active Shape Model</td>
</tr>
<tr>
<td>CCD</td>
<td>Charge-Coupled Device</td>
</tr>
<tr>
<td>EMA</td>
<td>Electro-Magneto Articulograph</td>
</tr>
<tr>
<td>FACS</td>
<td>Facial Action System</td>
</tr>
<tr>
<td>FAP</td>
<td>Facial Animation Parameter</td>
</tr>
<tr>
<td>FAU</td>
<td>Facial Action Unit</td>
</tr>
<tr>
<td>FDP</td>
<td>Facial Definition Parameter</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>HTK</td>
<td>Hidden Markov Model toolkit</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent Component Analysis</td>
</tr>
<tr>
<td>IPA</td>
<td>International Phonetic Alphabet</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear Discriminant Analysis</td>
</tr>
<tr>
<td>MOCHA</td>
<td>Multi-CChannel Articulatory</td>
</tr>
<tr>
<td>MPEG-4</td>
<td>Moving Picture Experts Group, 4 standard</td>
</tr>
<tr>
<td>PC</td>
<td>Principal Component</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>RGB</td>
<td>Red, Green and Blue Colour Model</td>
</tr>
<tr>
<td>RMS</td>
<td>Root Mean Squared</td>
</tr>
<tr>
<td>SSD</td>
<td>Sum of Squared Difference</td>
</tr>
<tr>
<td>SVD</td>
<td>Singular Value Decomposition</td>
</tr>
<tr>
<td>TIMIT</td>
<td>Texas Instruments and Massachusetts Institute of Technology Speech Corpus</td>
</tr>
<tr>
<td>TTP</td>
<td>Text-To-Phoneme</td>
</tr>
<tr>
<td>TTS</td>
<td>Text-To-Speech</td>
</tr>
<tr>
<td>VCV</td>
<td>Vowel-Consonant-Vowel</td>
</tr>
</tbody>
</table>
## Contents

1 Introduction .............................................................................................................. 1

2 Background ............................................................................................................... 7
   2.1 Visual Speech ........................................................................................................... 8
       2.1.1 Speech Synthesis ......................................................................................... 9
       2.1.2 Visual Speech Synthesis ........................................................................... 10
   2.2 Learning Face Dynamics ....................................................................................... 24
   2.3 Coarticulation ......................................................................................................... 25
       2.3.1 Modelling Coarticulation ........................................................................... 26
   2.4 Lip Extraction ......................................................................................................... 30
       2.4.1 Feature-based Tracking ............................................................................... 30
       2.4.2 Contour-based tracking ............................................................................... 31
       2.4.3 Model-based tracking ............................................................................... 32
   2.5 Summary ............................................................................................................... 34

3 Appearance-based Inner-lip Extraction .................................................................. 35
   3.1 Method .................................................................................................................... 39
       3.1.1 Viseme Extraction ...................................................................................... 40
       3.1.2 Training ....................................................................................................... 42
       3.1.3 Outer-lip tracking ...................................................................................... 46
       3.1.4 Inner-lip detection ...................................................................................... 49
   3.2 Results and Discussion ......................................................................................... 56
       3.2.1 Comparison ................................................................................................. 57
   3.3 Conclusion ............................................................................................................. 66
## Contents

### 4 Analysis of Visual Phones in Continuous Speech

4.1 Initial Work ............................................. 68
4.2 Lip Parameterisation ..................................... 70
4.3 Data Acquisition ........................................... 72
   4.3.1 Aligning the Training Set ............................... 73
4.4 Statistical Modelling ...................................... 73
   4.4.1 A Statistical Model of Shape ......................... 75
   4.4.2 A Statistical Model of Texture ....................... 76
   4.4.3 A Statistical Model of Appearance ................... 77
   4.4.4 Finding the Modes of Variation ....................... 78
4.5 Analysis of Visual Phones ................................ 79
   4.5.1 Linear Discriminant Analysis ......................... 80
4.6 Results and Discussion ................................... 81
   4.6.1 Analysis of Outer-lip Shape Information .............. 83
   4.6.2 Analysis of Inner-lip Shape Information ............. 85
   4.6.3 Analysis of Outer and Inner-lip Shape Information  90
   4.6.4 Analysis of Texture Information ..................... 92
   4.6.5 Analysis of Appearance Information .................. 94
4.7 Conclusion ................................................... 98

### 5 Visual Analysis of Lip Coarticulation in VCV Utterances

5.1 Data Acquisition ......................................... 103
5.2 Alignment and Resampling of the Speech Fragments ........ 105
5.3 Dynamic Analysis .......................................... 106
   5.3.1 Analysis of Phoneme /p/ ............................ 108
   5.3.2 Comparing /p/ and /m/ ............................... 116
   5.3.3 Conclusions .......................................... 121
5.4 Synthesis ................................................... 122
   5.4.1 Gaussian Mean Offset ............................... 123
5.4.2 Non-parametric Statistics ........................................ 124
5.4.3 PCA-based synthesis ............................................. 127
5.4.4 Conclusions ...................................................... 130
5.5 Conclusions and Future Work ................................. 130

6 Conclusions and Future Work ........................................ 133

Bibliography .............................................................. 137
List of Tables

2.1 Viseme and related phonemes MPEG-4 [62] . . . . . . . . . . . . 11

3.1 This table shows the number of failed frames for each method. . . 61
3.2 This table shows the average RMS error (in pixels) for each method. 63

4.1 A table showing phoneme to viseme mappings . . . . . . . . . . . 81

5.1 This table shows measures of separation, between /apa/, /ipi/ and /upu/ when projected into a VCV lip space, Figure 5.7. . . . . . . 114
5.2 This table shows measures of separation, between /apa/, /ipi/,
/upu/, /ama/, /imi/ and /omo/. . . . . . . . . . . . . . . . . . . . . 120
List of Figures

2.1 Modelling coarticulation: (a) dominance functions (after [52]) representing the temporal influence of a segment over an utterance for different articulators, (b) dominance functions (after [16]) - above is the final trajectory generated by a combination of the below dominance functions. ........................................... 29

3.1 This figure shows: a) An example of two frames with similar outer-lip shape and different inner-lip shapes and textures. b) An example of two frames with similar inner-lip shape and different outer-lip shape and texture. c) An example of two frames with similar inner-lip and outer-lip shape and different texture. .................. 39

3.2 The outer-lip is the lip-skin boundary and the inner-lip is the boundary between the lip and the inside of the mouth. .......... 40

3.3 An overview of how visemes are extracted from video. ........... 42

3.4 An example of iterative image segmentation based on Delaunay triangulation. Data points from a reference lip shape is triangulated into two regions. An iteration step consists in adding a point on the barycenter of each triangle, resulting in three new triangles. Iteration continues until convergence. ......................... 44

3.5 Two areas of texture are extracted from a lip image using the reference triangulation indices, where each index is equivalent to a pixel value in the image ..................... 45

3.6 An example of a simple classification of mouth appearance. .... 45

3.7 The 2D template matching algorithm [6]. This algorithm is designed to draw landmark points closer to the lip edge from an initial estimate. ..................... 48

3.8 At each iteration of the pattern matching algorithm outliers are found and brought closer to their desired position based on their neighbouring landmark position estimates. ............. 49
3.9 A sequence of the speaker saying the word 'top'. The change from the 2nd and 3rd frame demonstrates the rapid change in appearance of the inner mouth between frames that can occur during speech. ......................................................... 49

3.10 An example of the closest shape in the training set, $s_{i,\text{test}}$, represented as white contour points. The black contour points represent the input outer-lip contour, $s_{o,\text{test}}$. ......................................................... 53

3.11 Some edges between features of the mouth may appear more visible in colour images where they may not in gray-scale images. ................................................. 54

3.12 The top row shows, from left to right: the original RGB image, the hue of the image and result of a Sobel edge detection on the hue image. The bottom row shows, from left to right: the normalised highlight invariant image ($I_{\text{invariant}}$), the hue of $I_{\text{invariant}}$ and result of a Sobel edge detection on the hue image. ................................................. 55

3.13 This figure shows the detected edges (white) in the normalised hue domain. Each red dot represents a point in $s_{\text{new},i}$. The black line represent the normals of those points. The green dots represent the intersection(s) of the normal with any detected edges along the normals. The intersections closest to the outer-lip, $s_{\text{new},o}$ are chosen as the lip-edges. ......................................................... 56

3.14 This figure shows a case were colour transformation fails and the inner-lip boundary is not highlighted. ................................................................. 57

3.15 Graph (a), (b), (c) shows error (in pixels) relative to ground-truth from the AAM, constrained AAM and the novel detection approach with refinement for three sentences. The black lines are placed at the starting frame of a phone, with a phone label per interval. Labels in red are not included in the training set. ................................................. 61

3.16 Graph (a), (b), (c) shows error (in pixels) relative to ground-truth from novel detection approach with and without refinement for three sentences. The black lines are placed at the starting frame of a phone, with a phone label per interval. ................................................. 63

3.17 This figure shows a sequence of the frames 63 to 76. The left column of images are the result of the AAM tracker; the middle shows the proposed detection approach without refinement; the images in right column are a result of the proposed detection approach method with refinement. The red boxes show were the AAM hand to be re-aligned by hand after drift. The yellow box shows an example where refinement brings the inner-lip estimate close to the actual inner-lip edges. ......................................................... 65
4.1 Seven articulatory positions from EMA data from an utterance from a female speaker in the MOCHA-Timit corpus, \textit{fsew0}. Positions are relative to the reference articulator, the upper incisor. Data is sampled at a rate of 500Hz and in this example there are 96 samples. Only articulators highlighted in blue are considered for the experiments in this chapter. ........................................... 69

4.2 Articulator positions over time; (a) shows the upper lip’s vertical movement relative to the upper incisor and (b) shows the lower lip’s vertical movement relative to the upper incisor. .................. 70

4.3 This figure shows the EMA data of the three articulators, the upper-lip, lower-lip and lower-incisor for instances of three phones groups: /p,b,m/ (red), /l/ (blue) and /ch,jh,sh/ (green), projected onto the first 3 Principal Components of the EMA phone space. .. 71

4.4 (a) \textit{O} landmarks of a parameterised outer-lip; (b) \textit{I} landmarks of a parameterised inner-lip; (c) An example of a ‘shape-free’ texture found using Delaunay triangulation. .................. 71

4.5 Each column shows the several images associated with an instance of a phoneme (dynamic). Images highlighted with red borders are images selected as \textit{static} representations. .......................... 80

4.6 Cross section of a head. Labelled items influence the production of sound in one way or another. .................. 82

4.7 The first nine principal modes of outer-lip shape variation captured in the training set. .......................... 84

4.8 Dynamic instances of outer-lip data projected on the first 3 principal components of the outer-lip point distribution model. ........ 84

4.9 Static instances of outer-lip data projected onto the first 2 PCs of the outer-lip point distribution model. .................. 85

4.10 This table shows quantitative measures of separability found through LDA between all seven viseme groups in the outer-lip shape space. 86

4.11 This figure shows the contribution of the first three principal components in the separability of viseme /w/ vs all seven viseme groups. 86

4.12 This figure shows the contribution of the first three principal components in the separability of viseme /pbm/ vs all seven viseme groups. .................. 87

4.13 The first five principal modes of inner-lip shape variation captured in the inner-lip distribution model. .................. 87
4.14 Dynamic instances of inner-lip data projected on the first 3 principal components of the inner-lip point distribution model. . . . . 88
4.15 Static instances of inner-lip data projected onto the first 3 PCs of the inner-lip point distribution model. .................... 89
4.16 This table shows quantitative measures of separability found through LDA between all seven viseme groups in the inner-lip shape space. 89
4.17 The first nine out of eleven principal modes of mouth shape variation captured in the mouth shape distribution model. ........ 90
4.18 Dynamic instances of mouth shape (outer and inner-lip) data projected on the first 3 principal components of the mouth shape space. 91
4.19 Static instances of the mouth shape data projected onto the first 3 PCs of the mouth shape model. .................. 91
4.20 This table shows quantitative measures of separability found through LDA between all seven viseme groups in the mouth shape space. 92
4.21 The first 3 principal modes of texture variation captured in the texture space. Texture of the lips is included in this figure for visualisation purposes only. Columns show variation between ±3 standard deviations from left to right. ...................... 93
4.22 Dynamic instances of texture data projected on the first 3 principal components of the texture space. ...................... 93
4.23 Static instances of the texture data projected onto the first 3 PCs of the texture model. .......................... 94
4.24 This table shows quantitative measures of separability found through LDA between all seven viseme groups in the texture space. 95
4.25 The first 5 principal modes of appearance variation captured in the appearance space. Texture of the lips is included in this figure for visualisation purposes only. Columns show variation between ±3 standard deviations from left to right. ...................... 96
4.26 Dynamic instances of appearance data projected on the first 3 principal components of the appearance space. .................. 96
4.27 Static instances of the appearance data projected onto the first 3 PCs of the appearance model. .......................... 97
4.28 This table shows quantitative measures of separability found through LDA between all seven viseme groups in the texture space. 97
4.29 This figure shows static examples of visemes /t/ (green), /o/ (orange) and /p/ (red) projected onto the first 3PCs of the appear (all other visemes are coloured in grey). The mean static representation of these three visemes is found (black cross) and intermediate points (black crosses) between these means is found through linear interpolation. ........................................ 99

4.30 The top row shows the appearance information from a sequence of frames of the speaker saying “top” from the original database. The bottom row shows the appearance information of a synthesised sequence of frames found through a linear trajectory that passes through static examples of the visemes /t/, /o/ and /p/. ........................................ 99

5.1 $N$ landmarks of a parameterised lip. ........................................ 103

5.2 International Phonetic Alphabet (IPA) vowel chart with English examples. Vowels at the right and left of the circles are rounded and unrounded. The chosen cardinal vowels are marked with red circles. ........................................ 104

5.3 An example waveform and wide-band spectrogram for the phrase “Is apa it?” ........................................ 105

5.4 Alignment of fragment (b) to segment (a). The fragment and corresponding sampling is warped (squashed) during the alignment, (c). ........................................ 106

5.5 The first two principal modes of mouth shape variation for the VCV lip-space. ........................................ 107

5.6 A sequence of the reconstructed lip shapes, frame rate, along the VCV time interval, for (a) 1st PC, (b) 2nd PC. Red represent variation ($\pm$ 3 s.d.) for /apa/; green for /ipi/ and blue for /upu/. ........................................ 109

5.7 All repetitions of /apa/, /ipi/ and /upu/ projected onto the first 3 PC's. Each line in this space represents a VCV repetition, each one being on average 10 frames long. ........................................ 110

5.8 This figure shows all the data samples over time: (a) is for the 1st PC and (b) for the 2nd PC. ........................................ 111

5.9 The mean (continuous lines) and variance for $\pm$3 standard deviations(vertical bars) of the data, over time: (a) is for the 1st PC and (b) for the 2nd PC. ........................................ 113

5.10 Variation, in terms of Fisher's class discrimination measure, over time, when the data is projected onto (a) the 1st PC and (b) the 2nd PC. ........................................ 115
5.11 All repetitions of /apa/,/ipi/,/upu/,/ama/,/imi/,/omo/ projected onto the first 3 PC’s of the VCV lip space. Each line in this space represents a VCV repetition, each one being on average 10 frames long.

5.12 This figure shows all the data samples over time: (a) is for the 1st PC and (b) for the 2nd PC.

5.13 This figure shows all the data samples over time: (a) is for the 1st PC and (b) for the 2nd PC.

5.14 The mean (continuous lines) and variance for \( \pm 3 \) standard deviations (vertical bars) of the data, over time: (a) is for the 1st PC and (b) for the 2nd PC.

5.15 The mean (continuous lines) and variance for \( \pm 3 \) standard deviations (vertical bars) of the data, over time: (a) is for the 1st PC and (b) for the 2nd PC.

5.16 This figure shows an example of a trajectory tube where at each time a Gaussian distribution can be estimated.

5.17 This figure shows 30 synthetic trajectories of VCV /apa/, each one weighted according to a different Gaussian random variable.

5.18 An example of repetitions of /apa/ in a trajectory tube at time \( t \) following a multi-modal distribution.

5.19 Figure (a) shows an example of 15 utterances at time step \( t \) along a VCV trajectory tube sampled into 6 bins. The histogram generated is then translated into intervals in the range \([0,1]\), Figure (b).

5.20 This figure shows where the new points lie in the trajectory tube.

5.21 This figure shows 30 synthetic trajectories of VCV /apa/ generated with non-parametric statistics.

5.22 This figure shows 15 repetitions of VCV /apa/ projected onto the first 3 principal components of it’s temporal space. The red dots represent a single VCV trajectory and the blue cross represents a synthetic point in this temporal space.

5.23 This figure shows 30 synthetic trajectories of VCV /apa/ generated with PCA based statistics.
Chapter 1

Introduction

One of the most common approaches for speech synthesis is based on the concatenation of speech units, most commonly diphones. In the same way visual synthesis of speech can be achieved by the concatenation of rendered visemes, the visual shape and appearance of the face associated with the pronunciation of a phoneme. These visual speech units can be acquired through analysis of a person’s articulation and can be parameterised and stored for future use. Numerous application areas could benefit from visual speech synthesis including facial animation, human-computer interfaces, research into audiovisual speech perception, speech therapy and telecommunication.

The first computer generated images of faces were created in the early 1970’s. Since then the level of understanding of facial dynamics and the sophistication of technology has improved dramatically. This is evident in recent films such as the Lord of the Rings trilogy. Such animation still requires considerable manual animation that is both time consuming and expensive, requiring highly-skilled animators. The automatic production of realistic facial animation still remains a challenge.

The animation of faces can be performed at different levels of complexity. Simply
in-betweening targets can be effective for global changes in expression, given appropriate blending functions which taper the movement between extrema. However, speech animation is a particular example where such gross simplification is inadequate. The lips, tongue and jaw do not move in a linear fashion between extrema. Animations where this approach is taken typically appear sped up (over-articulated) and unrealistic. The reason for this is that there is a causal relationship between the speech audio, which we hear, and the articulatory movements, which we see. The audio is produced, in part, by the movements of the lips and tongue, and there is a direct perceptual link between the two. In fact, experiments show both that seeing someone speak improves the recognition rate of the audio [72] and that incorrect visual movements can change its perception [58]. This necessitates a more thorough handling of speech movements in facial animation.

The major contributory factor to the difficulties in animating speech movements is the physical phenomenon of coarticulation. Speech is often segmented into atomic units known as phonemes, representing constituent elementary sounds and their related vocal tract state. Given that each of these sounds is related to a shape or transitional movement of the vocal tract, coarticulation describes the motion of the articulators as they transition between states. In fact coarticulation is difficult to simulate because some phonemes are less important than others and disappear in the final transitional movement. Numerous models have been reported to describe specific effects of coarticulation [52], [45], [56], [78], [61].

Systems for generating speech trajectories can typically be split into three categories: target-based synthesis, concatenative synthesis, and model-based synthesis. Target-based synthesis uses combinations of static poses to structure a trajectory, usually with some form of approximating curve [22], [16], [77]. Concatenative synthesis, similarly to concatenative audio synthesis (e.g. Festival [9]), uses combinations of captured units (speech movements) to generate trajecto-
ries [14]. Model-based synthesis attempts to find a relationship between the audio speech signal and the movements of the vocal articulators, usually using a finite-state machine [33] [12].

All of these methodologies rely on the visual information of articulators given that the perception of speech involves visual as well as audio cues. Thus often the visual phonetic classification of speech has been attempted. This classification describes speech in terms of visual-phonemes (often shortened to viseme). Phonemes can form a many to one mapping with visemes since some sounds look visually the same. Examples of this are the two bilabial stops /b/ and /p/. A phoneme to viseme mapping can be used, similar to that shown in Table 2.1. For synthesis of a talking head to a novel soundtrack, the identified phonemes in that utterance can be mapped to their corresponding visemes. There are many mouth shapes associated with one phoneme, an effect due to coarticulation (flow from one phoneme to another). The sound /d/ in “did” differs in lip shape to that of /d/ in “do”. Using static visemes does not account for this effect, which is important for realistic articulation.

This thesis presents several visual analysis and modelling techniques for visual phones. This spans several areas of work, from capture and representation through to analysis and synthesis of speech movements and coarticulation. A novel method is reported for the automatic extraction of outer and inner-lip contour edges from sequences of mouth images in speech. An extensive visual analysis of phonemes in continuous speech is performed, that involves the investigation of mouth representations as well as a comparative analysis between static and dynamic representations of visemes. Finally, novel techniques are presented for the analysis and modelling of coarticulation in VCV contexts.

The novel contributions of this thesis to the area of facial animation and visual speech synthesis are:
• A novel inner-lip detection system which allows reliable and accurate localisation. The proposed detection technique is a key-frame exemplar-based method that is not dependent on any prior frame information for initialisation allowing improved inner-lip localisation for large frame to frame changes in lip-shape inherent in 25Hz video of visual speech. This method has been published in Turkmani and Hilton [74] (CVMP’06).

• An extensive analysis of visemes in continuous speech using a combination of Principal Component Analysis and Linear Discriminant Analysis. A comparative study between four different mouth representations is performed. An extensive analysis of static representations of visemes is also performed. The aim of this work is to motivate the need to analyse and model the underlying dynamics of visemes due to coarticulation.

• A visual analysis of lip coarticulation in VCV utterances. Based on ensemble statistics a novel approach to analysis and modelling of temporal dynamics is presented. This work is published in Turkmani et al. [75] (Interspeech’07).

• Research has also contributed to other joint publications in visual speech analysis and synthesis (Ypsilos et al. [82]).

The thesis is structured as follows:

• Chapter 2, Background, contains an overview of background information pertaining to speech and visual aspects of speech in particular. Mostly this concerns phonetic categorisation of speech sounds and the relationship between phonemes and articulatory movement. Also, this chapter gives an overview of coarticulation and previous work in numerical modelling to represent its effects. An overview of some methods of lip extraction techniques is included.
• Chapter 3, *Appearance-based Inner-lip Extraction*, presents a novel inner-lip detection system based on a hierarchical examplar search. A novel colour space normalisation is presented to enhance colour contrast for lip-edge detection. A comparative study is performed against tracking that uses the standard Active Appearance Model [19] often used in the tracking of points on the face.

• Chapter 4, *Analysis of Visual Phones in Continuous Speech*, presents a comparison of four mouth representations using a combination of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). The grouping of visual phonemes is presented and an extensive analysis is performed to find the best separability between these viseme classes. A comparison between static representation of viseme and the underlying dynamics is also performed.

• Chapter 5, *Visual Analysis of Lip Coarticulation in Vowel-Consonant-Vowel (VCV) Utterances*, presents a novel approach to the analysis and modelling of temporal dynamics based on ensemble statistics. The statistical significance of the effect of temporal variation due to coarticulation is investigated in detail for VCV utterances.
Chapter 2

Literature Review

The first computer-generated images of faces were created in the early 1970’s, with F. I. Parke’s [63] computer generated faces being one of the earliest developed. This work used a crude polygon representation of the head, which resulted in a flip-pack animation of the mouth and eyes opening and closing. Combining this with a smooth polygon-shading algorithm developed by Henri Gouraud, Parke produced several segments of rigid facial animation. This was done by collecting facial expression data from real faces using photo-grammetric techniques and interpolating between them. By 1974, Parke completed the first parameterised facial model by grouping vertices together to perform specified non-rigid deformations.

In the 1980’s, the first physically based facial muscle model was developed. In 1987, Waters [76] reported a new muscle model approach to facial expression animation. This approach allowed a variety of facial expressions to be created by controlling the underlying muscle structure of the face. Waters defined 22 muscle groups that are parameterised using a Facial Action Coding System, where single or small groups of muscles are under the control of a single Action Unit parameter. Ekman and Friesen [30] developed the original Facial Action System
(FACS) [34] that has become a universal standard for measuring and describing visually distinguishable facial movements for facial expressions. This the analysis of facial movement into anatomically based minimal Action Units. Animation using Action Unit parameters deforms the polygonal skin model to mimic the effects of contraction on the skin. This method of facial animation is significant to the evolution of facial animation and has since been the basis of several commercial animation implementations.

Toward the late 1980’s and early 1990’s, the use of computer facial animation was first being used as a key story telling component by incorporating facial expressions as well as speech into film animation [2]. The first “life-like” character was a computer-generated baby, Billy, in “Tin Toy”, a short film developed in 1988 by Pixar Animation Studios [71]. A muscle model, similar to that of Waters [76] was used to animate a set of 3-D points, which represented Billy’s skin. By “moving” groups of muscles, the animator had control over the aesthetics and choreography of the face. This kind of animation requires a highly skilled animator to achieve a believable animation. In addition it is time consuming and manual animation is expensive. One of the difficulties is realistic animation of talking heads which is important in both communication and entertainment. The problem of achieving animation that is not only realistic but automatic has received much interest. Research to find solutions to these problems is reviewed in this section.

2.1 Visual Speech

Speech has evolved over thousands of years such that today humans are capable of producing a vast array of complex sounds [13]. These sounds are produced using the jaw, lips, tongue, velum and larynx, and both the nasal cavity and oral cavity. These are all speech articulators and when we see and hear someone speaking, it is produced by the intricate interplay of these articulators. Speech
analysis research has developed an abstract set of symbols to represent speech. These symbols include:

- **Phonemes**, a family of similar sounds that a language treats as being the “same”. A minimal unit that serves to distinguish between meanings of words. Diphones are two consecutive phonemes while triphones are three consecutive phonemes

- **Allophones**, members of the phoneme family. A phonetic variant of a phoneme in a particular language

- **Monophthongs**, a single un-compounded vowel sound or combination of two written vowels pronounced as one

- **Diphthongs**, a complex speech sound or glide that begins with one vowel and gradually changes to another vowel in the same syllable

### 2.1.1 Speech Synthesis

There are 42 phonemes in the English Language and the concatenation of these sounds can make up any word in the language. Speech synthesisers, otherwise known as text-to-speech (TTS) systems, take advantage of this and can generate artificial human speech without directly using a human voice.

The system typically takes an input of raw text and assigns phonetic transcriptions to each word, a process called text-to-phoneme (TTP) conversion. The text is also divided into prosodic units, such as phrases, clauses, and sentences. The combination of the phonetic transcripts and information about prosodic units make up the symbolic linguistic representation, considered the front-end of the TTS. The back-end, or synthesiser, takes the symbolic linguistic representation and converts it into sound.
One of the most successful technologies for generating synthetic speech waveforms is concatenative speech synthesis [9]. This is based on the concatenation of segments of recorded speech. A subtype of concatenative synthesis is unit selection synthesis. A large speech database is created made up of utterances segmented into individual phones, syllables, words, phrases and sentences. For synthesis the appropriate segments from the database are picked out and concatenated.

### 2.1.2 Visual Speech Synthesis

The ideas behind speech synthesis can be used as a basis for the animation of a talking head. Basic facial animation can be considered just as the concatenation of the visual equivalent of utterance segments like phonemes, otherwise known as visemes. The term viseme was first coined in 1968 by Fisher [37] as an amalgamation of the words “visual” and “phoneme”.

Visual speech is produced as a continuous flow of articulatory movements, written as a discrete set of symbols. The sound /d/ in “did” differs in lip shape to the /d/ in “do”. These differences in facial movements, due to context of the word, are due to coarticulation (flow from one phoneme to another). For realistic facial animation the complex effects of coarticulation must be modelled accurately.

There is no simple one-to-one relation between the 42 phonemes in the English language and visemes, as different sounds may look the same. Therefore, there are many phonemes to one viseme. An example of this is with the two bilabial stops /b/ and /p/. The only difference between them is /b/ is voiced while /p/ is voiceless. Visually there is no difference in fluent speech and therefore the two phonemes can be placed in the same visemic category [35].

There are also many visemes to one phoneme, as illustrated above with the phoneme /d/, which are due to effects of coarticulation. The relationship between phonemes and visemes can become rather complex when trying to model
2.1. Visual Speech

...it in animation of speech. Table 2.1, shows a very basic phoneme to viseme mapping but there is no standard set of visemes. Approaches aimed at realistic animation of speech have used from as few as 16 visemes [33] and to as many as 50 in [69] to try to model effects of coarticulation.

<table>
<thead>
<tr>
<th>#</th>
<th>Phonemes (MPEG-4)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>p, b, m</td>
<td>put, bed, mill</td>
</tr>
<tr>
<td>2</td>
<td>f, v</td>
<td>voice, far</td>
</tr>
<tr>
<td>3</td>
<td>T, d</td>
<td>the, think</td>
</tr>
<tr>
<td>4</td>
<td>t, d</td>
<td>tip, doll</td>
</tr>
<tr>
<td>5</td>
<td>k, g</td>
<td>call, glass</td>
</tr>
<tr>
<td>6</td>
<td>tS, dZ, S</td>
<td>chair, join, she</td>
</tr>
<tr>
<td>7</td>
<td>s, z</td>
<td>sir, zeal</td>
</tr>
<tr>
<td>8</td>
<td>n, l</td>
<td>lot, not</td>
</tr>
<tr>
<td>9</td>
<td>r</td>
<td>red</td>
</tr>
<tr>
<td>10</td>
<td>A:</td>
<td>car</td>
</tr>
<tr>
<td>11</td>
<td>e</td>
<td>bed</td>
</tr>
<tr>
<td>12</td>
<td>I</td>
<td>tip</td>
</tr>
<tr>
<td>13</td>
<td>Q</td>
<td>top</td>
</tr>
<tr>
<td>14</td>
<td>U</td>
<td>book</td>
</tr>
</tbody>
</table>

Table 2.1: Viseme and related phonemes MPEG-4 [62]

2.1.2.1 2D Performance Driven Animation

In the early 1990’s an approach emerged which takes photographic images of real individuals enunciating different words and concatenating between them by using 2D image morphing techniques [7]. This technique to date seems to be the most effective approach in achieving photo-realistic facial animation. Facial animation can be considered as the concatenation of visemes, the visual equivalent...
of phonemes. In order to make the transitions between visemes smooth, 2D image morphing techniques can be used.

This section describes one of the most popularly used methods of 2D morphing, developed by Beier and Neeley [7]. Then some of the most successful research done on 2D face animation of talking heads is reviewed.

2.1.2.2 Morphing Between Photographic Images

The term morphing stands for metamorphosing and refers to an animation technique in which one graphical object is gradually turned into another [7]. Morphing can affect both the shape and attributes of graphical objects. It defines a transformation, which maps every point in the source coordinates to the corresponding point in the destination coordinates. Warping is another common technique used on animation and this refers to the geometric transformation of graphical objects (images, surfaces or volumes) from one coordinate system to another [7].

One example of the use of these techniques, is in the animation of the dancers in the Michael Jackson video “Black and White” created in 1990, where they used a 2D morphing technique developed by Beier and Neeley [7]. Beier and Neeley introduced a morphing technique that is a combination of generalised image warping with a cross-dissolve between image elements. They used multiple pairs of manually drawn lines to map from one image to another. For animation, a set of line segments are manually marked at key frames for each sequence of images. Then metamorphosis is performed on the two frames, which involves producing intermediate images that are generated from forward and inverse warping and the warped images are cross-dissolved. Marking the key frame is very time-consuming for the animator who needs to mark many key features, such as drawing a line down the nose, for both video sequences. However, this produces photo-realistic results, as seen in the Michael Jackson video.
2.1.2.3 2D Video Concatenation

Facial animation of a talking head requires the facial dynamics to be synchronised to an audio soundtrack. In 2D video concatenation, video is segmented into visemes and reused to generate animation often of a talking head speaking to a new soundtrack.

One of the first examples of video reuse was the special effects used in the 1994 film “Forrest Gump” that was manually created by an animator. A system, called “Video Rewrite”, developed by Bregler et al. [14] was the first facial animation system to automate the tasks of resynchronising existing footage to a new soundtrack. It learns from example footage how a person’s face changes during speech by capturing dynamics and idiosyncrasies of the articulations.

For analysis an annotated database of video clips is created. This database models how the subject’s mouth and jaw move during speech, a “video model”. They automatically segment the speech into tri-phones (three subsequent phonemes). For example the word “teapot” is split into a sequence of triphones as /SIL-T-IY/, /T-IY-P/, /IY-P-AA/, /P-AA-T/, /AA-T-SIL/, where /SIL/ represents a silence. Thus, the video is segmented into sequences depending on the triphone timings. The corresponding facial movements are tracked automatically by using eigen-points to locate the fiduciary points around the mouth and jaw. The combination of the fiduciary-point locations and the triphones, describe the visemes (or in this case trisemes) and the triseme sequences are stored in the “video model”.

For synthesis, the system then selects from the video database, as dictated by the new utterance, the appropriate triseme sequence. A phoneme to viseme mapping is used with 26 viseme classes. The triseme sequences of the mouth and jaw are spatio-temporally aligned and blended into neighbouring visemes to account for coarticulation effects. These triseme sequences are then stitched into the background sequence. All morphs are done using the Beier and Neeley morphing
Chapter 2. Background

The output is a video-realistic (comparable to the original video quality) video sequence of the person uttering newly synthesised speech.

Since videos of triphones are used as the basic unit of for synthesis, this results in a large library of tens of thousands of video segments. To create the eigen-point models a number of images need to be hand-labelled to model a person’s mouth and jaw movement during speech. This system is limited to specific people with a large video corpus needed for each person. This system does not animate for any head movement.

Another system that generates photo-realistic 2D video animations of talking heads based on image samples is that developed by Cosatto and Graf [21]. This system automatically extracts samples from videos of talking people and stores them in a library. In comparison to the system developed by Bregler et al. [14], the number of samples is reduced by several orders of magnitude because the head is decomposed into a hierarchy of facial parts such as the mouth, teeth, jaw, eyes and brows. These parts can be animated independently.

They use 12 static visemes and all other phonemes are mapped to this set. The images are first aligned and then mouth shapes are modelled with a set of parameters namely: mouth width, position of upper lip and position of the lower lip. This is a crude parameterisation that does not cover the finer nuances in the mouth shapes so multiple examples of the mouth shapes are stored in their parameter space, “viseme library”, to account for this. Based on these parameters the images samples can be ordered in the corresponding libraries. They parameterise the other facial parts in a similar way creating libraries for the eyes, mouth etc.

For synthesis, they use a TTS system to get phoneme information. A mapping is defined between each phoneme and the parameters for the corresponding mouth shape. The parameters of the mouth in the video sequence are tracked and thus characteristic lip motions are extracted. The lip parameters from a person speak-
2.1. Visual Speech

The most common tri-phones are stored in a “coarticulation library”. When a new string of phonemes is extracted from the TTS, a sliding window of three phonemes is scanned across this string. At each time interval, the mouth parameters are computed by adding the weighted influence of the current, previous and following phonemes. Thus a trajectory is defined through the space of mouth parameters. Then the “viseme library” is used to find the corresponding mouth shapes depending on the mouth parameters found. This trajectory is sampled at video rate with a mouth shape generated at each time interval by merging the two samples with parameters closest to the sample point. All face parts are warped and blended to a base head using a feathering mask to avoid boundary artifacts.

The number of samples needed for synthesis is reduced by several orders magnitude, in comparison to the system developed by Bregler et al. [14], because the head is decomposed into a hierarchy of facial parts. However, the “viseme space” needs to be densely populated to ensure the produced animation is not jerky, thus several thousand of images are needed in the database.

More recently, a video-realistic text-to-audiovisual speech synthesizer was developed by Ezzat et al. [35] [33]. They created a visual corpus that captures a large range of the visual instances of American English phonemes. One image or viseme for each phoneme is manually extracted from this corpus. Ignoring coarticulation this viseme set is reduced to a 16-viseme set. Given two viseme images, a correspondence map, that ensures correspondence between geometric attributes on the faces, is computed. Optical flow is used to measure the apparent motion of the face and allow for automatic determination of correspondence maps between consecutive frames by finding the warp parameters. The first image is warped along the computed flow vectors, effectively moving the pixels along from one image to the next. Forward warping is also performed on the second image toward the first. Both processes generate intermediate images. Morphing is then
used to combine the texture found in both forward warps. Thus a viseme morph is generated.

Concatenating appropriate visemes morphs creates a visual stream. So, the word “one” or /w-uh-n/ can be composed of two viseme morphs /w-uh/ and /uh-n/. They use the Festival Text-To-Speech system (TTS) [9] to convert input text into a diphones (2 consecutive phonemes), along with duration information for the diphones. Using this information, the appropriate sequence of viseme morphs is created, as well as the rate of the transformations.

This system to date has produced the most video-realistic animation. Although morphing produces smooth transitions between the new mouth shapes, the system does not model the whole head and is limited to animating faces from the original video corpus only. However, despite view morphing [70], the main problem with all the aforementioned 2D face animation techniques is the fact that they do not allow for much freedom in face orientation, relighting and interaction with 3D objects.

2.1.2.4 3D Face Modelling and Animation

To date 2D facial animation produces the most photo-realistic results but is limited in face orientation and integration into a 3D scene. The most obvious solution to achieve such freedom is to use 3D techniques to animate the face.

Before any kind of animation can be done an animatable 3D face model is needed. Types of 3D face model can be separated into two main categories: those fully synthesised, these may integrated with 3D objects such as eyes and teeth, and those generated from “real-data”, that model the surface of the face.

Synthesised face models include parameterised facial models that are a polygon representation of a face. Often these form a generic model who’s vertices can
be deformed for animation. In the early 1970’s, Parke [63] developed the first parameterised polygonal 3D face that is still used in some animations to date.

An extension to parameterised facial models are physics-based muscle models. The face is represented as a multi-layered polygon, where layers may represent the skin, muscles or the skull. In 1987 Waters [76] proposed the first muscle model, that has been used by, among others, Pixar in many of their animations. More recently a more accurate biomedical muscle model has been proposed to include the “skull” in the muscle model developed by Lee et al. [50]. Static 3D facial models can also be generated from geometrically accurate data or measurements of real people. The main method of retrieving this data is through the use of commercially available scanners with CyberwareTM [1] colour laser scanner the most commonly used for facial modelling. It produces a densely sampled geometry of several thousand vertices per scanned face with a corresponding texture map. The data provided can be manipulated to perform many animations as will be described later.

Another technique for retrieving “real-data” of a dynamic face is by active structured light stereo reconstruction. Projected light patterns are used to provide a dense surface texture by computing pixel correspondences to derive dense depth maps instant independently. Commercially available products based on these techniques include the Eyetronics Shape SnatcherTM scanner and the 3DMD face scanner. These systems capture either shape only or shape interlaced with colour.

Once 3D face models have been created they can be animated using several techniques dependent on the 3D model. There are two main approaches to animation that are reviewed here:

- Synthetic animation - using Facial Animation Parameters (FAPs) or Facial Action Units (FAUs) that control muscle models
• Performance driven animation - using morph targets or motion capture

2.1.2.5 3D Face Modelling from Images

Blanz and Vetter [10] developed a parametric technique for modelling textured 3D faces from single images. They use laser scans of 200 heads taken using a Cyberware™ scanner. They represent the geometry of a face with a shape-vector and the texture is represented by a texture-vector that correspond to the shape vertices. Principal Component Animation (PCA) is performed on the set of shape and texture vectors to give a low dimensional approximation representation of the “Face Shape”, where each point in the space represents a mode of variation in shape and appearance. New shape and texture vectors can be created by linearly combining the shape and texture vectors from the example faces. Facial attributes, such as femininity and weight, defined by a hand labelled set of examples, can also be mapped onto the morphable model.

The synthesis of a new face involves matching a “morphable” model to one or more images. An average 3D morphable head model is manually aligned to the image(s) and then the closest matching 3D head in the vector space spanned by the database is found. By varying the shape, texture and rendering parameters this head can be morphed and rendered back to the image(s), effectively creating a new image.

The result is photo-realistic reconstructions (static reconstructions are comparable to the original image of the person) from un-calibrated images that are easy to animate. However, the morphable model is restricted by the size of the database. The linear combinations of shape and appearance modes do not guarantee a valid face.

Fua et al. [38] [26] use the face models developed by Blanz and Vetter [10] for tracking in image sequences using the shape vectors only for robustness to illumi-
nation. The position and orientation of the model is assumed to be fixed and the position and orientation of each camera in the video sequence with respect to the model is computed. Approximate positions of five feature points in one reference image are manually supplied and the camera pose parameters are estimated to initialise the process. The model is projected into this image and the face area is sampled. Using normalised cross correlation, corresponding points can be found in the second image. Triangulation is used to estimate the transfer function that maps the coordinates from the first image into the second is found. This transfer function is dependent on camera pose and shape, so with a large enough set of samples the shape parameters and position parameters of the face for each camera can be found. The shape parameters are used to deform the closest matching 3D head from the database and the camera position indicates the orientation of the face. This head is then texture mapped to produce photo-realistic 3D models that are robust in challenging lighting conditions.

2.1.2.6 Synthetic 3D Animation

The Moving Pictures Expert Group, MPEG, created an MPEG-4 standard in 1999. For facial animation, MPEG-4 provides a low bit rate representation of faces in the form of Facial Animation Parameters (FAPs) and Facial Definition Parameters (FDPs). A face is rendered and animated after these parameters have been decoded. FAPs are closely related to muscle actions and represent a complete set of basic facial actions. The FDP set is used to personalise a generic face model and contain a set of 3D feature points that are used to deform the generic face model.

Escher et al. [31] use a generic head as a 3D polygonal mesh composed of 1500 vertices with fixed set of vertices corresponding to the feature points defined by the MPEG-4 standard. The generic head is deformed using a Dirichlet Free Form Deformation method so that volume deformations can be done while maintaining
the continuity of the surface. Once all the positions of the feature points are known the generic head is fitted to the FDP points, creating a calibration model. All the points on a generic 3D face mesh and the calibration 3D face mesh are both represented by a cylindrical projection. The projection of the generic map is mapped to the projection of the calibration map, where the generic mesh now has feature point with a corresponding new 3D location. Texture is mapped to this new cylindrical map. Escher et al. [32] used the above method in a video conferencing system since MPEG-4 provides a low bit-rate compression ideal for such an application. However, the low-bit rate animation does not produce photo-realistic results.

Lee et al. [50] developed a physics-based facial model. They presented a novel triangular layered synthetic tissue model consisting of two layers, making up the skin model, that are connected by a muscle layer to just a few skull surface nodes. They model 28 of the primary facial muscles that are embedded into the muscle layer on the tissue model. Forces can be exerted on the springs in the sin/muscle model to deform it, within a constrained volume for realism, by automatically scaling and positioning muscle vectors. The skull surface is estimated to be the surface normals in the range data image and any deformations in the tissue model are constrained not to effectively “penetrate” the skull. The final face model is a combination of the skin/muscle/skull model with other pre-synthesised geometric models, such as eyes, teeth and neck.

Using the Cyberware\textsuperscript{TM} scanner to scan 360 degrees around an individual. This scan produces two images, a range image and a reflectance image. Contours of facial features needed for adaptation, such as boundaries of lips and eyes, are located using a Laplacian field function. A generic face mesh is adapted to range data to allow for automatic scaling and positioning of facial muscles and reduce the amount of data acquired from the scanner. All facial features, such as nose tip and mouth contour, are labelled so the generic mesh can be fitted to the range
map and texture map. The range image at the location of the nodes on the face mesh is sampled in order to find the geometry of the face. This model can then be easily animated using FAPs.

Physically-based models aim to model intricate dynamics of the human face that other modelling techniques cannot achieve. However, this is difficult to achieve due to the complexity of skin folding and the difficulty of simulating inter-reflections and self-shadowing [64].

2.1.2.7 Performance Driven 3D Animation

Morphing Techniques  Pighin et al. [64] produced realistic 3D facial animation of emotional expressions from multiple view images. They capture multiple views of people and manually mark an initial set of corresponding points on the images, typically around the mouth and corners of eyes etc. These morph targets are then used to recover the camera parameters corresponding to each photograph, automatically, as well as the 3D positions of the marked points in space. These 3D points are used to deform a generic 3D mesh. Texture maps are extracted for the 3D model from the photos. The whole process is repeated for several different facial expressions of the same person.

In order to create an animation of concatenated facial expressions, they developed a morphing algorithm that interpolates between these facial expressions. Since all the 3D expressions models have the same topology, due to a consistent marking of major facial features across different facial expression models, so there is a natural correspondence between their vertices.

This system produces realistic animation of facial expressions that is able to capture the subtle changes in illumination and appearance due to facial creases. This method of animation is widely used in the film industry. Despite its advantages, it requires manual marking of the set of input images which is time consuming.
and markers to be places on the subjects faces.

**3D Video Concatenation** Kalberer *et al.* [41] combined markers with structured light to acquire dense facial shape deformations for animation. They captured 3D shape models of subjects using the *Eyetronics Shape Snatcher*\textsuperscript{TM} System at a temporal sampling rate of 25 3D snapshots-per-second. A generic head template of 2268 vertices and the standard MPEG-4 feature points (FAPs) as a subset is fitted to the 3D models automatically by using prominent features of the face, found by using shape and texture cues, as anchor points.

Faces of differing age, race and gender were captured. Principal Component Analysis, PCA, is performed on the masks to give a compact low dimensional representation of the masks, which in turn is used to determine the main deviations from the average face, by finding weight vectors between them. These principal components span a space called “Face Space”, a database of neutral reference faces each representing a plausible change, similar to “Face Space” used by Blanz and Vetter [10]. Faces are represented by points in this space. The “Face Space” can be exploited to create new faces by adapting the values of individual principal components of an average face to make them look female/male or old/young.

A “Viseme Space” is generated for the purposes of animation. 3D models were captured for 10 reference persons reading sentences with several instances of all visemes. 116 black dots are drawn on the face and these are tracked throughout the 3D speech sequences to measure the displacements between these faces and a neutral face. It was found that using PCA these modes of displacement could be represented by 16 components but several represented a very similar expression, such as the movements of the cheeks. Independent Component Analysis (ICA) is used in this case to represent a set of modes that are each realistic in their own right. ICA looks for modes in the PC space that correspond to linear combinations
of the PC’s that are maximally independent. The space spans 16 Independent Components (that represent 16 static visemes).

The neutral faces and speech dynamics of 10 reference subjects were scanned. These 10 faces are represented by 10 points in Face Space, forming a hyperplane. When a new face enters the space, it can be approximated by a linear combination of these faces by projecting it orthogonally onto the hyperplane. This approximated face will have its own personalised speech dynamics and any visemes of speech that has not been captured are created as a linear combination of the closest reference persons visemes to create new, personalised visemes. So visemes are adapted to the physiognomy of a face, yielding more realistic results.

Text was fed into a TTS system that generates a file that contains an ordered list of allophones and their timings. These allophones are translated to visemes using a phoneme to viseme mapping, discussed in [42]. They use 20 visemes all with two distinct forms, that of a round mouth and of a wide mouth, to try to allow for effects of coarticulation. Visual speech is represented by trajectories through “Viseme Space”, which is path from one viseme to another, over the time intervals generated by the TTS. To ensure smooth navigation through “Viseme Space” splines are fitted to “Viseme Space” coordinates of the visemes.

The result of this system is an almost automatic high-quality facial animation of previously scanned faces and newly created faces with new personalised dynamics for speech to achieve realism. However, due to the use of visible markers for the creation of the databases there can not be simultaneous capture of faces and expressions.

All the aforementioned 3D animation techniques lack video-realism when modelling the dynamics of the face during speech. This is mainly due to the sparse data around the mouth area [64]. Kalberer et al. [41] overcome this problem by combining markers with structured light to acquire dense facial shape deformations for animation. A space of different neutral 3D face scans is created that
can be manipulated to create new faces. Each 3D face scan in the Face Space has a corresponding Viseme Space that represents how the face changes in speech found by tracking markers in 3D video. Animation produced from a smoothed trajectory through this space. Coarticulation effects are not modelled accurately due to the visemes in the Viseme Space being static. Also this technique is limited to modelling shape only.

2.2 Learning Face Dynamics

Approaches to viseme creation have mainly dealt with static visemes, representing a single mouth shape by a 2D image \[33\] [21] or a 3D image or model [10] [26]. Discretising visemes in this way results in loss of information [12] such as the dynamics of the face due to coarticulation. Work on dynamic visemes has been limited to date. Bregler et al. [14] partially model the dynamics of vocal coarticulation with triphones but do not model the dynamic of facial appearance.

Brand [12] introduced a method of generating facial animation from expressive information in an audio soundtrack. Using entropy estimation, a model of a face’s observed dynamics during speech is learned from video, effectively creating a “puppet” that can be used to animate a face. The video is analysed to yield a probabilistic finite state machine, a mapping from states into regions of facial configuration space (facial states) and an occupancy matrix that gives the state probabilities at each frame for the training video sequence. The occupancy matrix is combined with synchronised audio to give each state a dual mapping into audio feature space. The state machine and vocal distributions are combined to form a vocal Hidden Markov Model (HMM). Using this HMM, given a new vocal track, the Viterbi algorithm is applied to the model to find the most likely sequence of predicted facial states. The facial output probabilities of the Viterbi states are then mapped to actual facial configurations. The “puppet” can be used to
control a 3D animated head model or to warp 2D input face images to synthesise new video. Animators can choose from a range of “puppet” models trained on sequences of different speech styles and facial mannerisms, to create animation from novel audio in the style of the training performance. The HMM models context across an entire utterance to account for coarticulation. However, this model is heavily dependent on the amount and quality of training data.

2.3 Coarticulation

Coarticulation is the physical phenomenon which describes the blurring of boundaries between atomic units of speech, both visibly and audibly. Transitions between articulatory gestures are brought about by a physical system of muscles. The constraints of that physical system prevent instantaneous transitions between gestures.

The result of coarticulation is that the articulatory gesture formed for a certain speech unit (and the resulting sound itself) will vary during the production of natural speech. Some aspects of the gesture will vary less (e.g. lip contact in bilabial stops), and some more (e.g. jaw rotation in vowels.) In this regard phonemes have varying influence over a speech utterance. This varying dominance has been described in Recasens et al. [65] that uses scaled phonemes ranked according to degree of articulatory constraint; such a scale can be used in conjunction with coarticulation rules to determine final trajectories through a parameterisation (space) representing vocal articulatory states.

Coarticulation does not only regard the extent to which a gesture is realised, but also the influence of that gesture over a period of a speech act. Coarticulation can be anticipatory, i.e. the vocal tract is preparing for an upcoming important gesture (forward coarticulation, e.g. lip rounding in $t\overline{w}o$), and also can reflect the effect carried over from a previous gesture (backwards coarticulation, e.g. lip
protrusion in \textit{boots}.) Contextual effects of coarticulation have been observed up to seven segments preceding a gesture in the French vowel /y/ from \textit{istr	extolinebreak[4]y} in the phrase \textit{une sinistre structure} \cite{8}.

In order to account for the nature of coarticulation, several theories have been proposed. Kent and Minifie \cite{45} categorise these into the following: learnt allophonic models; target based models; and hierarchical models. Allophonic models, such as \cite{78}, propose that the lowest level units for speech production are allophones. These units are context dependant and exhibit less variation than phones. Target-based models \cite{56} argue that speech production is a goal-oriented task, where neuromotor commands are generated in a lookahead manner to attain contextually-invariant targets. Finally, hierarchical models place coarticulation as a part of an overall speech production strategy, for example Kent and Minifie propose a hierarchy which covers the broad range of speech tasks from neuromotor control up to syllabic grouping. Whilst there are many proposals, with matching supporting arguments and evidence, few are concrete enough to be put to practical use (e.g. in a synthesis system.) In the following section the most common are discussed in more detail.

2.3.1 Modelling Coarticulation

In order to both understand and reproduce the effects of coarticulation on natural speech, numerical models have been applied. Such models must reliably reproduce the variation seen in speech, which means accounting for the physical constraints of articulatory movement. Öhman \cite{61} describes a numerical model which accounts for the effects of coarticulation on non-symmetric vowel-consonant-vowel syllables (V1CV2.) In this model the movement of the tongue body in x-ray sequences is predicted by as:

\[
s(x; t) = v(x; t) + k(t)[c(x) - v(x; t)]w_c(x)
\]  

(2.1)
In this equation, $s(x; t)$ represents the shape of the vocal tract at a point $x$ on the tongue body at a time $t_{V1} \leq t \leq t_{V2}$ (i.e. between the centre of the initial and final vowels $V1$ and $V2$). $v(x; t)$ and $c(x)$ represent the surrounding vowel and consonant vocal tract shapes respectively. The vowel shape is related to the current time because it is a transitional function between the initial and final vowel shapes. Between the initial and final vowels the influence of the central consonant is represented by a combination of $k(t)\epsilon[0, 1]$, which represents the location of the central consonant, and $w_c(x)$ which scales the influence of the consonant according to its dominance. $k(t)$ varies from 0 at time $t_{V1}$ to 1 at $t_C$ and back to 0 at $t_{V2}$, and is a smoothly varying function of time. As a result the consonant has maximum influence at time $t_C$ which occurs at some point between $t_{V1}$ and $t_{V2}$.

This model is a simple extension of interpolation for the modelling of complex coarticulation. However, the model is simplified for the purposes of general modelling and the application to synthesis. For example there is no way to model consonant-consonant coarticulation, and scaling the solution to non-VCV syllables provides significant challenges. Regardless of these shortcomings this model has been applied to general coarticulation in the “Mother” visual-speech synthesis system [66].

Löfqvist extends the ideas from Öhmans simplified coarticulation model to general speech [52]. In this model each articulator (lips, tongue, jaw etc.) has a number of related dominance functions which determines the influence a segment (phoneme) exerts over its trajectory, Figure 2.1. The dominance a segment exerts will vary with each articulator; for example bilabial plosives, such as pat, will exert a greater influence over the motion of the lips than that of the tongue.

In Löfqvists model the shape of the dominance functions will directly determine the trajectory of a speech utterance. Although only loosely defined these functions are maximal at the centre of a segment and decrease with temporal
distance. The width of a dominance function will determine the section of an utterance over which the segment will have an influence, and thus must be no more than seven segments wide to maintain consistency with previously reported results [8]. Dominance functions of this form easily compare with Öhmans equivalent $w_c(x)k(t)$ term in 2.1; both describe time-varying influence of one segment over its neighbours.

Cohen and Massaro [16] describe a model which implements Löfqvists model of coarticulation. In this model negative exponential functions are used to represent the time varying dominance functions, Figure 2.1. The final speech trajectory formed by this method can be fitted to real speech motions, demonstrating a relationship between this technique and speech coarticulation. Several limitations of this method for generating trajectories with specific types of speech targets (e.g. bilabial stops) have been reported [49]. Despite this the Cohen and Massaro/Löfqvist model is the most commonly used by the visual synthesis community [22] [49] [16].

Another method of modeling coarticulation is the visual speech unit selection from clustered data, as used by Krnoul et al. [46]. In this method several instances of each visual speech unit are stored and are clustered using decision trees. Static target vectors, representing points on the face, are used to represent each phone. For synthesis these target vectors are interpolated forming a continuous track.

Visual speech synthesis provides examples of coarticulation modelling in a practical setting. Bregler et al. [14] use audio-visual triphone segments extracted from a larger audio-visual corpus. Cosatto and Graf [21] parameterise the space of lip positions populated by images from a recorded corpus. Synthesis is performed by traversing trajectories in this lip space, created using Cohen and Massaro’s [16] coarticulation rules. Ezzat et al. [33] use viseme-based alignment from incoming audio, where visemes are static distinguishable lip shapes. Synthesis is produced via image morphing driven by pixel-flow analysis. Brand [12] proposed a model
2.3. Coarticulation

Figure 2.1: Modelling coarticulation: (a) dominance functions (after [52]) representing the temporal influence of a segment over an utterance for different articulators, (b) dominance functions (after [16]) - above is the final trajectory generated by a combination of the below dominance functions.

to learn observed dynamic parameters during speech from labeled audio-visual data.

In [16] [21] static phonetic units are blended to generate synthetic articulatory trajectories. The parameters of the blending functions used in these models determine how the lips move between viseme targets. Such models are similar to the theoretical models of [61] [52] in how they attempt to synthesize speech movements from discrete phonetic targets. Another popular method, used in both audio and visual speech synthesis, is the concatenative approach. In these models, short segments of real speech (e.g. syllables or triphones) are blended to generate synthetic trajectories. In terms of visual synthesis this approach has been demonstrated using both video [14] and motion-captured point trajectories [47] as the underlying speech data. Finally, in [33] an optimisation approach is used to fit a trajectory through targets represented as the mean and covariance of each viseme at its center.
2.4 Lip Extraction

Automatic and accurate analysis of facial features has motivated intensive research in the field of computer vision and is important for several applications such as data-driven animation, morphing and re-animation of novel sequences for photo-realistic video or graphics and increasingly is used for facial recognition and identification. Extraction of lip motion is increasingly used to aid the speech community: to aid automatic audio-visual speech recognition and for speech science studies of speech production, co-articulation and dysfunction.

A number of methods have been proposed for extracting lip contours from images. Most of the methods reviewed in this section fall into one of three categories, feature-based, contour-based, and model-based methods. Feature-based tracking techniques extract local regions of interest (features) from images and identify the corresponding features in each subsequent image in the sequence. With contour-based tracking techniques the contour of a moving object is represented by a ”snake” which is dynamically updated. Both feature-based and contour-based techniques, generally, are highly dependent on initialisation. Model-based tracking techniques exploit a priori knowledge of typical objects in a given scene. Tracking is localised by matching a projected model to the image data. Appearance model-based techniques, specifically, track objects as a whole rather than features.

2.4.1 Feature-based Tracking

One of the most common methods of visual feature extraction is the deformable template method [83], where the outline of the lips is modelled by a set of hand coded polynomials. The feature of interest, such as the lips, is described by a parameterised template. A parametric deformable template is a parameterised mathematical model used to track the movement of the given object. An energy
function is defined which links edges, peaks, and valleys in the image intensity to corresponding properties of the template. The model interacts dynamically with the image by altering parameter values to minimise the energy function, thus drawing the shape template onto salient features. Image search is performed by fitting the deformable template to image gradients, assuming consistent strong edges at the lip-contour. Several parametric models have been proposed for lips. Tian et al. [73] uses a simple three state geometric model made of parabola. Coianiz et al. [17] uses a lip model with two parabola instead of one for the upper boundary. Hennecke et al. [40] proposes a lip model that comprises of two types of curve, parabola and quartic. These methods are not able to resolve fine contour details because the parametric models do not account for asymmetry consistent of lips, especially during speech.

### 2.4.2 Contour-based tracking

Another common approach for shape modelling is based on active contour models or snakes which are parameterised energy minimising splines that converge to an object contour within an image. The snake technique was first introduced by Kass et al. [43]. Modifications of this early technique have been applied to the lip in [25]. Snakes are able to resolve fine contour details, however, is sensitive to salient regions, due to shadow and reflections, close to the desired lip boundary and thus align to undesirable local minima. Bregler et al. [15] developed a contour tracking method based on a combination of snakes and deformable models. A configuration of lips are represented as points in a feature space and a set of all possible lip configurations is a surface or manifold in this space. The snakes are controlled using the learned manifold while a gradient based image search is performed.

Barnard et al. [6] developed a lip-tracking system that uses a combination of a
modified snake algorithm [79] and a 2D template matching technique that does not require any prior training. In this case the snakes are controlled by using two dimensional pattern templates of the lip edge contour instead of the image gradient. The snake points and thus 2D pattern templates are manually initialised around the outer-lip edge for the first frame. During tracking, the templates of the snake points are updated by using a weighted average of the initial pattern template and the template extracted from the previous image of the sequence. Image energy is defined as the 2D correlation between a 2D patch from the image and the expected template for the specified snake point. A parametric model similar to Yuille et al. [83] is used.

2.4.3 Model-based tracking

Where as deformable templates and snakes align to strong gradients for locating the desired feature in an image regardless of the actual appearance in the image, appearance models learn the grey level appearance, within the feature, from a training set and use them for image search. For the active shape model (ASM) [20] only a small area of texture around each landmark point, usually the appearance perpendicular to the contour, is used to iteratively minimise the distance between model points and corresponding image points. It has been applied to lips by [55] [57]. Lip shape is constrained using a reduced statistical shape model that is constructed from hand-labelled training data. Each example shape model is represented by the 2D cartesian coordinates of it’s landmark points. A global profile texture model is built to create a statistical model of concatenated grey-level profiles from the normals of each landmark. Given a global profile model, and a corresponding statistical shape model the shape and pose parameters that best fit the image based on landmark profiles is minimised using a downhill simplex function [54]. The simplex algorithm [59] does not require calculation of the gradient error surface image but may require many iterations.
to converge to a local minimum. Again this method is dependent on strong contour edges, consistent in training data, and requires a lot of landmark around the boundary of a lip in order to maximise search area.

The active appearance model (AAM) [19] is the dominant appearance-based face-tracking approach at present. AAM’s manipulate a full model of appearance of the whole region of the mouth [57] [29] [4] [3] [27]. The active appearance model consists of the combination two principal component (PCA) models. The first is a statistical shape model as built for ASM’s. The second principal component model is trained on the variations in shape-free mouth textures, sampled within the landmark point, that are morphed to a normalised lip shape. The appearance model is built by applying a further PCA to identify the correlation between the shape model and grey-level texture model. An appearance model is fitted to an image by iteratively finding the appearance parameters (shape and texture) that seeks to minimise the difference between the synthesised image and target image. AAM’s require fewer landmark points than ASM’s, however, the active appearance model requires construction before searching can begin and frequent updating as new images are labelled. This is time consuming process.

Lip-tracking techniques that use appearance spaces are continuous generative appearance models that are highly dependent on initialisation for correct local minimisation. The highly deformable appearance of the inner mouth during speech can make appearance models highly complex so would result in a reduced likelihood that aforementioned trackers would minimise to a correct position in the appearance space. Unlike other lip-tracking approaches that incorporate a-priori knowledge and previous frame positions for initialisation of the current frame, the proposed approach is a system where the inner-lip for each input frame is detected independently of all previous frame history. The proposed technique is a discrete exemplar-based method that minimises to key frames in the training data and is not dependent on initialisation.
2.5 Summary

A common approach to produce visual speech is to interpolate parameters that describe mouth variation in sequence, known as visemes. A viseme corresponds to a phoneme in an utterance. The interpolation process should consider the issue of context dependent representations (coarticulation) in order to produce realistic looking talking heads. There has been limited work in the physical (appearance of the face) modelling of coarticulation during speech.

Work in visual speech synthesis provides examples of coarticulation modelling in a practical setting. Most commonly this involves the representation of visemes as static poses and coarticulation is the interpolation between these poses. There has been limited work in the modelling of dynamic visemes. The work presented in this thesis addresses the issue of modelling coarticulation in dynamic contexts during speech through statistical ensemble analysis.

The extraction of lip motion in speech has motivated intensive research. However, the accurate extraction of lip motion for the purposes of analysis, in image sequences of speech, still forms a challenge. This particularly true for speech sequences captured at 25Hz. The work presented in this thesis proposes a method of accurately extracting lip information, that in turn is used to represent visemes for the modelling of coarticulation.
Chapter 3

Appearance-based Inner-lip Extraction

Automatic and accurate analysis of facial features has motivated intensive research in the field of computer vision and is important for several applications such as data-driven animation, morphing and re-animation of novel sequences for photo-realistic video or graphics and increasingly is used for facial recognition and identification. Extraction of lip motion is increasingly used to aid the speech community: to aid automatic audio-visual speech recognition and for speech science studies of speech production, coarticulation and dysfunction. Accurate extraction of lip motion, for the purposes of analysis, in image sequences of speech is a particular challenge addressed in this chapter.

There has been a lot of successful work in the accurate extraction of the outer-lip contour (the boundary between the outer-lip and skin). However, accurate extraction of inner contour of the lip in long image sequences of speech is difficult, particularly for speech sequences captured at 25Hz. This is because the inner-mouth is highly deformable in terms of both shape and texture, as shown in Figure 3.9. During speech the mouth goes from fully closed to open in less
than 20 milliseconds, i.e. within a single frame. The mouth varies in colour across individuals (spatial variability), due to locutions (temporal variability) and specularity due to lighting (spatiotemporal variability); in addition, they are subject to rigid motion (head rotation). Obstruction and occlusion of the lip, teeth, tongue and velum can be extreme during speech leading to severe distortions in the image. The boundary between the lip and inner-mouth can be weak and protrusions often cause shadows to form. An ideal inner lip-tracking method could be described by the following:

- It is able to deal with variability across individuals due to the locutions and lighting changes explained above
- It is able to accurately track novel speech sequences from a minimal training set
- It is not dependent on manual initialisation

The aim of the approach presented in this chapter is to automatically extract the outer and inner-lip contour edge from sequences of mouth images in speech. An automatic outer-lip landmark tracker was implemented that is an iterative contour fitting method that consists of a pattern matching algorithm, Section 3.1.3. It uses the previous frame positions of the landmark points as initialisation for the current frame. It was found that this tracker localised to an incorrect minima on average every 40 frames. However, it was found that tracking could not be applied to extract inner-lip contour edge, since changes in appearance are less gradual for the inner mouth and causes the tracker to drift after only a few frames.

The inner-lip is automatically extracted using a novel appearance-based detection system. The proposed technique uses a set of discrete key frame exemplar to detect inner-lip shape, where initialisation is independent of information from
prior frames. In order to make detection accurate a refinement based on edge detection is performed. The aim is that the proposed inner-lip detection system will be able to accurately landmark the inner-lip contour for previously unseen sequences with high deformities that naturally occur in speech.

**Background** The most successful approaches proposed for extracting lip contours from images fall into three categories, *feature-based, contour-based,* and *model-based* methods. Feature-based tracking techniques extract local regions of interest (features) from images and identify the corresponding features in each subsequent image in the sequence. With contour-based tracking techniques the contour of a moving object is represented by a “snake” which is dynamically updated. Both feature-based and contour-based techniques are highly dependent on initialisation. Model-based tracking techniques exploit *a priori* knowledge of objects in a given scene. Tracking is localised by matching a projected model to the image data. Appearance model-based techniques, specifically, track objects as a whole rather than features.

The Active Appearance Model or AAM [19] is the most dominant appearance-based tracking approach at present. AAM’s manipulate a full model of appearance of the whole region of the mouth [57] [29] [4] [3] [27]. The active appearance model consists of the combination of two models found through Principal Component Analysis (PCA). The first is a statistical shape model as built for active shape model’s (ASM) [20]. Lip shape is constrained using a reduced statistical shape model that is constructed from a hand-labelled training data. Each example shape model is represented by the 2D cartesian coordinates of it’s landmark points. The second principal component model, in an AAM, is trained on the variations in shape-free mouth textures, sampled within the landmark points. Textures are morphed to a normalised lip shape to become shape-free. The appearance model is built by applying a further PCA to identify the correlation
between the shape model and grey-level texture model. An appearance model is fitted to an image by iteratively finding the appearance parameters (shape and texture) that minimise the difference between the synthesised and target image. The active appearance model requires construction before searching can begin and frequent updating as new images are labelled. This is a time consuming manual process.

Lip-tracking techniques that use appearance spaces are continuous generative appearance models that are highly dependent on initialisation for correct local minimisation. Unlike other lip-tracking approaches that incorporate a-priori knowledge and previous frame positions for initialisation of the current frame, the proposed approach is a system where the inner-lip for each input frame is detected independently of previous frames. The proposed technique is a discrete exemplar-based method that minimises to key frames in the training data and is not dependent on initialisation.

With AAMs a constrained shape model is built from labelled training data. This model provides a compact description of a space of valid lip-shapes. It can be seen in Figure 3.1 that the outer-lip lip shape cannot always uniquely infer the shape or texture of the inner mouth. For these reasons the lip-tracking system, proposed in this chapter, does not constrain the lip-shape training data in a combined shape model, but rather has been developed to automatically landmark points along the outer-lip separately to the inner-lip.

The chapter is organised in the following way: Section 3.1 explains our proposed method to track the outer-lip and detect the inner-lip. In Section 3.2, results are presented that include a comparison between the state-of-the-art Active Appearance Model (AAM). Conclusions are given in Section 3.3.
3.1 Method

Colour video of a front-facing speaker, under uniform lighting conditions, is captured at a rate of 25Hz. It is assumed that the face is always front-facing and there is minimal head movement. A rectangular area, initialised manually, around the mouth is extracted from the video sequence. A system has been developed to automatically extract the outer-lip and inner-lip contour, represented by a set of contour points, from the mouth in this video sequence, as shown in Figure 3.2. A system was developed that considers the outer-lip shape, inner-lip shape and texture separately.

1. A training set is built, consisting of three sets of data, Section 3.1.2. A set of tracked outer-lip data, $S_o$, a set hand-labelled inner-lip data, $S_i$ and a set of texture data, $T_g$. The textures are categorised into seven different classes.

Figure 3.1: This figure shows: a) An example of two frames with similar outer-lip shape and different inner-lip shapes and textures. b) An example of two frames with similar inner-lip shape and different outer-lip shape and texture. c) An example of two frames with similar inner-lip and outer-lip shape and different texture.
2. The outer-lip contour is extracted, Section 3.1.3, for the entire test sequence using a simple 2D pattern matching tracking algorithm, similar to that in [6]. This algorithm only requires the initialisation of control points in the first frame of the sequence. For each frame the tracked outer-lip, \( s_{o,test} \) along with the training data, forms the input to the third part of the system.

3. The third part of the system land-marks the inner-lip, Section 3.1.4. The aim is to find the closest estimate of an inner-lip shape from the training data of inner-lip shapes. A search for the best inner-lip shape is constrained using the tracked outer-lip shape, \( s_{o,test} \), and texture classes.

4. A local optimisation is performed to bring this estimate closer to the inner-lip boundary.

### 3.1.1 Viseme Extraction

There are two classes of tools commercially available that automatically generate phonemes and their timings: Text-To-Speech (TTS) synthesisers and Speech Recognisers. TTS synthesisers can synthesise novel speech from input text while providing the phonemes it has used and their timings. Speech recognisers can
take any input utterance and recognise the phonemes uttered and provide their timings. The latter method was chosen for phoneme segmentation since video of people speaking known utterances will be used for analysis, thus there is no need to create synthesised speech.

A Hidden Markov Model (HMM) toolkit (HTK version 3.2.1) [81] speech recogniser was used to segment input utterances into its phonemes and timings. HTK is a portable toolkit for building and manipulating HMMs, primarily used for speech recognition. It consists of a training tool that estimates the parameters of a set of HMMs using training utterances and their associated transcriptions. Unknown utterances can be transcribed using a recognition tool, HVite.

HTK uses *isolated word recognition* where each spoken word is represented by a sequence of speech vectors. Given a set of training examples corresponding to a particular HMM, the parameters of the model can be determined automatically. To recognise some unknown utterance, the likelihood of each model is calculated and the most likely model is chosen to represent the word. Recognition is based on the maximum likelihood of a state sequence of the HMMs.

Continuous speech involves connecting HMMs together in sequence. For *continuous speech recognition* each model in the sequence corresponds to phonemes. HTK processes feature vectors extracted from the speech data along with symbolic transcriptions to identify occurrences of phonemes and find the best fit HMMs and the corresponding sequence that best models the input utterances. The result is a system that, given a novel utterance, can recognise the chosen underlying symbols and estimate their timings. In this research we require accurate phoneme timings rather than a general speech recognition system. Therefore, a forced alignment system was developed trained on training data consisting of speech files, their phonetic transcriptions and the British English Pronunciations (BEEP) dictionary [67]. Rather than having an utterance as input to the the search engine and thus a set of possible phonemes to search for in the speech
recogniser, the search engine is given an exact transcription of what is being 
spoken. The system then aligns the transcribed data with the speech data, iden-
tifying which time segments in the data correspond to particular phonemes in the 
transcribed data.

3.1.2 Training

A corpus of speech utterances is captured for each subject for training. The 
audio is labelled into phonemes and their timings using a forced alignment that 
is trained using the MOCHA (Multi-CHannel Articulatory) [80] database which 
consists of 460 phonetically balanced sentences from both a northern speaking 
male and a southern speaking female. This data-set accommodates for a test sets 
of any gender and some variation in dialect. The phone is extracted in the visual 
domain (sequence of frames), visemes. An overview of how visemes are extracted 
from video is shown in Figure 3.3. A number of key training frames, $F$, were 
automatically selected as the mid-frame of a number of visemes uttered by the 
speaker.

Figure 3.3: An overview of how visemes are extracted from video.

The outer-lip contour is tracked in the sequences of captured speech using a 

simple 2D pattern matching tracking algorithm, Section 3.1.3, represented by
3.1. Method

$N_o$ landmarks. Each landmark on the outer-lip is represented as a 1D vector of cartesian coordinates $(x, y)$, and the landmark coordinates are concatenated to form a 1-dimensional vector, $s_o$. The outer-lip of the previously chosen key-frames for a set of training outer-lip shapes, $S_o$, where

$$S_o = \{ \hat{s}_n \}_{f=1}^{N_o}$$

(3.1)

For each of these key-frames $N_i$ landmark points are manually selected on the inner-lip contour, $s_i$, forming a a labelled set of key-frame shape templates, $S_i$.

$$S_i = \{ \hat{s}_i \}_{f=1}^{N_i}$$

(3.2)

Using the outer-lip and inner-lip contour data for each frame the grey-level pixel values, texture, within these contours can be extracted. It was observed that the textures should be shape free and aligned. The texture of the mouth was separated, for each frame, into two separate texture areas; the lips and inside of the mouth. This is done for alignment between frames so that the texture of, for example, the lips is sampled in the same way for frames that have the mouth open or closed lips.

Shape-free texture is found from each mouth image using an iterative image segmentation based on Delaunay triangulation of inner and outer-lip sample points. A lip shape is triangulated for the data points, as shown in Figure 3.4, resulting in a set of triangles such that no data points are contained in any triangle’s circumscribed circle. An iteration step consists in adding a point on the barycenter of each triangle. Iteration continues until convergence. Thus it stops when either the triangles are homogeneous or the surface of the triangles are less than a threshold. Each triangle is represented as a set of index values, forming reference triangulation indices. This order of indices is applied to any lip-shape of the same coordinate structure, where each index is equivalent to a pixel value in the image, Figure 3.5. This mapping process removes spurious texture variation due to shape differences.
For a single frame, a concatenated 1-dimensional vector, $t$, is formed of intensity pixel values for the lips and then the inner mouth area. For the $F$ frames the texture, represented in a vector $t$ of length $N_t$, was extracted and accumulated to form a key set of texture templates, $T$, where

$$T = \{t_f\}_{f=1}^{N_t}$$  \hspace{1cm} (3.3)

Figure 3.4: An example of iterative image segmentation based on Delaunay triangulation. Data points from a reference lip shape is triangulated into two regions. An iteration step consists in adding a point on the barycenter of each triangle, resulting in three new triangles. Iteration continues until convergence.

### 3.1.2.1 Classification of Texture Data

Each frame from the training data was classified (given a class label) according to $N_c$ number of different texture classes, resulting in $N_c$ class matrices, each column vector of the matrix being a texture from $T$. Classification can happen in two ways:

3.1. Method

Figure 3.5: Two areas of texture are extracted from a lip image using the reference triangulation indices, where each index is equivalent to a pixel value in the image.

2. Manual selection of a small number of frames, for each class. Example frames for each class are shown in Figure 3.6. An average of these texture is found for each class that is used to classify all other frames in the training data using a similarity measure presented in Section 3.1.4.1.

The mean of each class matrix is found, $\bar{t}_{\text{class}}$, creating a set of $N_c$ class textures,

$$T_{\text{class}} = \{\bar{t}_{\text{class},f}\}^{N_c}_{f=1} \quad (3.4)$$

Figure 3.6: An example of a simple classification of mouth appearance.
3.1.3 Outer-lip tracking

The system developed automatically landmarks points along the outer-lip edge separately to the inner-lip. An automatic outer-lip tracker based on 2D template matching techniques [6] was implemented. The system is an iterative fitting method that consists of two main steps. Firstly, there is an initialisation of the 2D templates. Secondly, an iterative contour fit is performed that consists of a pattern matching algorithm, similar to that in [6], that finds the outer-lip edge. The outer-lip edge is then filtered to remove outliers and higher frequency noise. This system requires no previous training and is only dependent on an initial set of outer-lip contour patches.

3.1.3.1 Initialisation

The lip model, \( S_o \), is represented by \( N_o \) landmark points on the boundary between the skin and the outer-lip contour. The tracking system is initialised with the manual selection of points in the first frame. Initial pattern template surface image patches for each landmark are automatically extracted. These templates are not updated over the tracking, as in [6], as it was found this allowed the tracker to drift resulting in convergence to an incorrect local minima.

For all other frames, the position of landmarks from the previous frame, \( S_o(-1) \), were used as an initial estimate, \( S_{o,\text{init}}(t) \) of the position of the current frame:

\[
S_{o,\text{init}}(t) = S_o(t - 1)
\]  

3.1.3.2 Iterative contour fit

An iterative contour fit is performed to estimate the location of the outer-lip boundary in the current frame. An error function, \( \epsilon \), is minimised at each iteration
3.1. Method

with respect to \( s_o \):

\[
\arg \min \epsilon(s_o) \quad (3.6)
\]

\[
\epsilon(s_o) = \epsilon_D + \lambda \epsilon_S \quad (3.7)
\]

where \( \epsilon_D \) is the error of the observations at the current frame \( s_o(t) \) and \( \epsilon_S \) is the distance to landmarks in the previous frame. \( \lambda \) is a regularisation term.

**Observations at the current frame** To find an estimate for the position of the landmarks at each iteration pattern matching is performed. For each landmark point a scan line is generated along the normal of the lip edge. The scan-line is sampled at \( L \) discrete points and at each of these samples a patch is extracted. The sampled image patch is compared to the landmark template using a 2D cross-correlation metric, as illustrated in Figure 3.7. The sample along the scan-line that gives the highest correlation, is considered to be the outer-lip edge, \( S_{\text{max}} \) for all \( N_o \) landmark points.

\[
S_{\text{max}} = [s_{\text{max},f}]_{f=1}^{N_o} \quad (3.8)
\]

For each iteration the distance between the observed outer-lip edge and the landmark point at the previous iteration is minimised as:

\[
\epsilon_D = \sum_{f=1}^{N_o} (S_{\text{max},f} - s_{o,f})^2 \quad (3.9)
\]

All landmark points, across the sequence of frames, are tracked using the initial pattern templates as a basis. For the corners of the mouth it was found that the search area needed to be extended as it was observed these landmarks, generally, moved in directions other than along their normals.

**Smoothing (regularisation term)** The 2D pattern matching is used to estimate landmark positions at each iteration, \( S_{\text{max}} \). However, due to changes in appearance, mainly due to shading and specularity some landmark points
Figure 3.7: The 2D template matching algorithm [6]. This algorithm is designed to draw landmark points closer to the lip edge from an initial estimate.

may not find the correct lip edge. The Euclidian distance between points in the current and previous frame is found as:

\[ S_{o,diff} = S_o(t-1) - S_{max}(t) \]  

(3.10)

The median of \( S_{o,diff} \) is found and those landmark points in \( S_{max} \) that are above this median are considered outliers. It was found that approximately 12% of landmark points are considered as outliers per frame. Once these outliers have been found they are re-estimated as the mean of their two neighbouring landmarks in Euclidean space, Figure 3.8. At each iteration the this smoothing is minimised as:

\[ \epsilon_S = \frac{\sum_{f=1}^{N_o} (s_{o,f}(t-1) - s_{max,f}(t))^2}{N_o} \]  

(3.11)

As a final post-processing step, a non-uniform rational B-spline is made to pass through each estimated landmark point and is sampled so the the final landmark points remain equidistant. This results in a outer-lip contour \( S_{o,test} \).
3.1. Method

Figure 3.8: At each iteration of the pattern matching algorithm outliers are found and brought closer to their desired position based on their neighbouring landmark position estimates.

3.1.4 Inner-lip detection

Tracking outer-lip shapes in a sequence is based on a local search, initialised using the shape and position from the previous frame. However, the appearance of the inner mouth changes rapidly in speech, especially with video captured at a rate of 25Hz, as shown in Figure 3.9.

Figure 3.9: A sequence of the speaker saying the word 'top'. The change from the 2nd and 3rd frame demonstrates the rapid change in appearance of the inner mouth between frames that can occur during speech.

With this in mind, it was observed that a detection rather than tracking methodology would be more appropriate to deal with this variability. In the proposed system, the position of landmarks on the inner-lip shape for each frame, in the test sequence, are found independently of the position of landmarks in the previous frame. A hierarchical search, based on appearance data, is performed to find the best estimate inner-lip shape for each frame. Inner-lip detection is performed in two parts:
1. Exemplar Detection

The first is an automatic conditional search for the best inner-lip shape from the set of key-frame outer-lip shape data \( S_o \), inner-lip shape data \( S_i \) and texture data \( T \). The implementation of detection is explained in details in Section 3.1.4.1.

2. Refinement

The second is a refinement process to bring the landmark points of the inner-lip shape estimate closer to the actual edge in the mouth image. The implementation of this refinement is explained in details in Section 3.1.4.2.

### 3.1.4.1 Exemplar Detection

For each input frame in the test sequence, the best inner-lip shape is found from the key-frame exemplars. There are four stages to finding this best estimate of the inner-lip shape. The first is finding the initial estimate of the inner-lip shape.

#### Step 1: Find initial estimate of the inner-lip shape

At this stage the information available is the outer-lip shape, \( s_{o,\text{test}} \), of the input frame and the key-frame training data. To perform an efficient search of the estimate of the inner-lip, \( s_{i,\text{initial}} \), from \( S_i \) (inner-lip key-frame set) two tests are performed:

1. **Test similar outer-lip shapes**

   We need to identify a set of outer-lip shapes from \( S_o \) that are closest to the input outer-lip shape to \( s_{o,\text{test}} \). For each outer-lip shape in \( S_o \) the similarity with \( s_{o,\text{test}} \) is found. Those outer-lip shape that have a similarity measure less than a threshold are deemed closest to \( s_{o,\text{test}} \). The similarity measure is found first by translating a training shape so that it’s centroid lies on top of the centroid of input inner-lip shape and
second by finding the Euclidean difference between their corresponding landmark points. For two corresponding landmark points along two separate lip contours, \( P = (p_x, p_y) \), and \( Q = (q_x, q_y) \), the Euclidean distance, \( e \), is computed as:

\[
e = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}
\]

(3.12)

The Euclidean distance is found for the \( N_o \) landmark points along lip-contour, resulting in a set of \( N_o \) Euclidean distances:

\[
E = \{e_f\}_{f=1}^{N_o}
\]

(3.13)

The median, \( m \), of \( E \) is found. Those lip shapes in \( S_o \) whose median, \( m \), value is less than a threshold are considered close to \( s_o \). A heuristic threshold was set at a value of 2 pixels. Through experimentation this threshold was found to be optimum. This step creates a set of the \( L \) closest outer-lip shapes from the training data.

\[
S'_o = \{s_{o,f}\}_{f=1}^{L}
\]

(3.14)

If \( S'_o \) is not large enough (two or fewer examples), the search for valid outer-lip shapes is extended by relaxing the threshold.

2. **Test for valid inner-lip shape**

For each outer-lip shape in, \( S'_o \) there is an associated inner-lip shape, resulting in a subset of \( S_i \) of the same length as \( S'_o \). So each example in the subset of inner-lip shapes, \( S'_i \), and the subset of outer-lip shapes, \( S'_o \) come from the same exemplar. As discussed earlier, for the similar outer-lip shapes there are likely to be multiple inner-lip shapes of varying sizes. This is captured in \( S'_i \).

The initial estimate of the inner-lip shape, \( s_{i,initial} \), is an example in \( S'_i \) that has an associated outer-lip shape in the set \( S'_o \) that is the closest to the test outer-lip shape, \( s_{o,test} \).
Step 2: Find the exemplar with the closest texture

The initial estimate of the inner-lip shape, \( s_{i,initial} \), may not fit to the actual inner-lip boundary in the input frame. Texture information is used to further refine the search of the inner-lip shape from the key-frame training data.

With \( s_{i,initial} \) and the input outer-lip shape \( s_{o,test} \), the texture from the input frame can be re-sampled to the canonical form using Delaunay triangulation of the contour points, as described in Section 3.1.2. This results in a shape-free texture, \( t_{initial} \).

For every outer-lip and inner-lip shape in the training data there is an associated texture, \( T \). Thus, a subset of \( T \) can be found, \( T' \) with each texture in this subset coming from the same exemplar as in \( S'_o \) and \( S'_i \). Each exemplar in \( T' \) has an associated texture class label. The texture \( t_{initial} \) is classified into one of the \( N_c \) classes and the subset of \( T' \) with the same class is found, \( T_{valid} \). The closest exemplar texture, \( t_{closest} \) in the subset \( T_{valid} \) to \( t_{initial} \) is found.

Step 3: Classifying \( t_{initial} \)

\( t_{initial} \) is compared to each representative texture in \( T_{class} \) and the closest class is assumed to be the class that \( t_{initial} \) belongs to. Sum of Squared Difference (SSD) was used to find the closest texture class. The SSD between \( t_{initial} \) and all class texture representatives in \( T_{class} \) is found,

\[
SSD = \sum_{f=1}^{N_t} \frac{(T_{class}(f) - t_{initial})^2}{N_t}
\]  

(3.15)

The class texture representative that gives the lowest SSD value is taken as the best class and \( t_{initial} \) is given this class label.

Step 3: Find the best estimate of inner-lip
Using $s_{o,test}$ and each inner-lip shape exemplar in $S'_i$ a shape-free texture can be found from the input frame and compared to the closest exemplar texture, $t_{closest}$, using SSD. The inner-lip shape that contributes the lowest SSD value is taken as best inner-lip estimate $s_{i,test}$ to represent the input frame, as shown in Figure 3.10.

![Figure 3.10: An example of the closest shape in the training set, $s_{i,test}$, represented as white contour points. The black contour points represent the input outer-lip contour, $s_{o,test}$.](image)

### 3.1.4.2 Refinement

The landmark placement process has so far found the closest inner-lip shape from the exemplar data to best fit the input lip frame. At this stage there is no guarantee that these landmark points fit to the actual inner-lip boundary edge. Thus, a local refinement is performed on each landmark point, in $s_{i,test}$, to bring them closer to the desired boundary. This is a post-processing step.

Shadowing and shading causes changes in appearance of the inner-lip when trying to track the inner-lip contour reliably, as shown in Figure 3.11(a). The definition between the inner-lip contour and the inside of the mouth can become ambiguous causing difficulty in localising the inner-lip. It has been observed that definition of boundaries affected by shadowing are particularly evident in gray-scale images.
of the mouth. It was observed that colour provides additional information needed to define the inner-lip boundary more accurately in the presence of self shadows, as depicted in Figure 3.11(b).

![Figure 3.11: Some edges between features of the mouth may appear more visible in colour images where they may not in gray-scale images.](image)

In order to reduce the effects of specularities and shadows, a highlight invariant transformation [36] is performed on the RGB image, $I_{rgb}$. Each RGB channel, for each pixel in $I_{rgb}$, is transformed into the following coordinates:

$$
\begin{align*}
    r' &= r - \overline{rgb} \\
    g' &= g - \overline{rgb} \\
    b' &= b - \overline{rgb} \\
    \overline{rgb} &= \frac{r + g + b}{3}
\end{align*}
$$

Where $r, g,$ and $b$ represent the three channels of a pixel and $\overline{rgb}$ is the intensity of that pixel. This produces negative values so a global normalisation, Equation 3.20 is performed on each pixel:

$$
px' = \left( \frac{px_c - \text{min}_c}{\text{max}_c - \text{min}_c} \right), c = R, G, or B
$$

where $px_c$ represents the value of channel $c$, for a single pixel in $I_{rgb}$; $\text{min}_c$ is the global minimum value of channel $c$ in $I_{rgb}$ and $\text{max}_c$ is the global maximum.
3.1. Method

These two steps result highlights differences in colour by re-normalising colour to a unit cube, as shown in the first column of Figure 3.12. This creates a normalised image, $I_{\text{norm}}$. The hue (pure colour) is found from $I_{\text{norm}}$ and an edge detection is performed on this image. It can be seen in the second column of Figure 3.12 the hue image found from $I_{\text{norm}}$ segments the lips, teeth, skin and tongue into distinct regions. The hue of the original RGB image does not find such distinction. A simple Sobel edge detection is performed on this normalised hue image. It can be seen from the third column that both the outer-lip and inner-lip edge can be identified.

![Figure 3.12: The top row shows, from left to right: the original RGB image, the hue of the image and result of a Sobel edge detection on the hue image. The bottom row shows, from left to right: the normalised highlight invariant image ($I_{\text{invariant}}$), the hue of $I_{\text{invariant}}$ and result of a Sobel edge detection on the hue image.](image)

Using this edge image each point from the estimated inner-lip shape, $s_{i,\text{test}}$, can be brought closer to the lip-edge. This is done by finding the normals of each
landmark point in \( s_{i, test} \) and looking along them for an intersection with the edge detected lip-boundary. The intersection closest to the outer-lip, \( s_{o, test} \), is the point chosen, as illustrated in Figure 3.13. Some small regions of lip-edge may still not be detected. To deal with this smoothing, as discussed in Section 3.1.3.2, is performed on \( s_{i, test} \). This creates a new set of inner-lip landmark points, \( s_{i, opt} \).

For frames of a closed mouth, where the boundary between the lips is not well defined, as shown in Figure 3.14, this colour transformation does not enhance inner-lip boundaries, thus no edge is detected. In these cases landmark placement is fully dependent on the inner-lip detection system and refinement is redundant.

Figure 3.13: This figure shows the detected edges (white) in the normalised hue domain. Each red dot represents a point in \( s_{new,i} \). The black line represent the normals of those points. The green dots represent the intersection(s) of the normal with any detected edges along the normals. The intersections closest to the outer-lip, \( s_{new,o} \) are chosen as the lip-edges.

### 3.2 Results and Discussion

A speech corpus [33] of 4000 frames of a talking front-facing female was captured using a Sony 9100P 3CCD colour camera. Due to constraints on data capture this we were only able to capture a full database for one speaker. Images are captured at a rate of 25Hz, progressive scan SD resolution (720x576 pixels) and audio
3.2. Results and Discussion

Figure 3.14: This figure shows a case where colour transformation fails and the inner-lip boundary is not highlighted.

captured at a rate of 16KHz. A rectangular area, (100x170 pixels), initialised manually, around the mouth is extracted for the captured frames.

As in Section 3.1.1 static viseme extraction is performed on the 4000 test frames to find key-frame exemplar frames. This dissects the audio into the phonemes uttered and their timings. These timings are used to find their corresponding frames in the video, visemes (visual phonemes). The first 1000 frames formed the test sequence. From the other 3000 frames, 400 key training frames, $F$, were automatically selected as the mid-frame of 23 out of 47 phonemes uttered by the speaker and the inner-lip was manually labelled. The outer-lip was automatically tracked, using the method presented Section 3.1.3. Then inner-lip landmark placement was performed on the test sequence using the novel appearance-based detection presented in Section 3.1.4. The next subsection shows a comparison between the detection system presented in this chapter and the standard AAM method [19].

3.2.1 Comparison

The inner-lip contour for the 1000 frame test sequence was hand-labelled for ground truth, $s_{i,\text{ground}}$. Four tests were performed on the test sequence, using the same training data:
Chapter 3. Appearance-based Inner-lip Extraction

1. Detection without refinement

The inner-lip contour was detected using the landmark placement step presented in Section 3.1.4.1. For each frame, the input to the algorithm is a test outer-lip shape, $s_{o,test}$ and training data, $S_o$, $S_i$ and $T$; the output is a test inner-lip shape, $s_{i,detect}$.

2. Detection with refinement

The inner-lip contour was detected on the test sequences using both the landmark placement and refinement step, presented in Section 3.1.4.2. For each frame, the input to the algorithm is the test RGB image, the test outer-lip shape, $s_{o,test}$ and training data, $S_o$, $S_i$ and $T$; the output is a test inner-lip shape, $s_{i,detectrefine}$.

3. Active Appearance Model [19]

The inner-lip contour for the test sequence was tracked using AAM, $s_{i,aam}$. The AAM is trained on the 400 exemplar training data: $S_o$, $S_i$ and $T$. For each frame in the test sequence a search is performed using the AAM until convergence. When a failure occurs due to drift the search is re-initialised using the outer-lip shape data as a starting point. For each frame, a failure is classified as when any landmark points for the inner-lip are outside the outer-lip contour, $s_{o,test}$. The output is a test inner-lip shape, $s_{i,aam}$.

4. Constrained AAM

The inner-lip contour for the test sequence was tracked using a constrained AAM. Each test frame was cropped to the width and height of it’s corresponding outer-lip shape estimate, $s_{o,test}$. At each frame the search is re-initialised to this outer-lip shape and iterated until convergence. The AAM is trained on the training RGB colour images, the test outer-lip shape, $s_{o,test}$ and the same training shape data as for the proposed detection system, $S_o$. 
3.2. Results and Discussion

and $S_i$. The input to the algorithm is test outer-lip shapes and training data; the output is a test inner-lip shape, $s_{i,aamconst}$.

For each frame the estimated inner-lip shapes $s_{i,aam}$, $s_{i,aamconst}$, $s_{i,detect}$ and $s_{i,detectrefine}$ are compared to the ground truth $s_{i,ground}$ using a Root Mean Squared (RMS) Euclidean error. This measure results in a value (in pixels) where the higher the value the less similar the detected/tracked inner-lip shape is to the ground truth. Figures 3.15(a)- 3.15(c) shows $s_{i,aam}$, $s_{i,aamconst}$, $s_{i,detectrefine}$ vs $s_{i,groundtruth}$ for three sentences: ‘checking our top stories’, ‘the fire is being allowed to burn itself out’, ‘the probe will orbit the moon’. These sentences contain five manners of articulation: stops, nasal stops, fricatives, central and lateral, that account for 25 phones, 10 of which are not included in the training data. It can be seen that the $s_{i,detectrefine}$ error for the three sentences is consistently lower than the $s_{i,aam}$ error and the $s_{i,aamconst}$ for 80% of frames in the three sentences, by an average error of 2.63 pixels.

Results for the AAM methodology show that even with manual re-initialisation of the initial estimate of the inner-lip shape the localisation may still fail resulting in convergence to incorrect local minima. For the constrained AAM method the search is re-initialised at each frame and iterated until convergence. Despite this re-initialisation at each frame large errors are still produced. The AAM had to be re-initialised for 79 out of the 1000 test frames and 31 out of 1000 for the constrained-AAM. The inner-lip detection approach introduced in this work did not require re-initialisation.

The number of failed frames (in the 1000 frame test sequence) for all four methods was counted and tabulated, Table 3.1. A failure is classified as when any landmark points for the inner-lip drifts anywhere outside the outer-lip contour. Figure 3.17 shows 3 examples of failures. Table 3.2 shows average RMS error (in pixels) produced for the 1000 test frames for the four approaches.
Chapter 3. Appearance-based Inner-lip Extraction

(a) ‘Checking our top stories’

(b) ‘The fire is being allowed to burn itself out’
3.2. Results and Discussion

Figure 3.15: Graph (a), (b), (c) shows error (in pixels) relative to ground-truth from the AAM, constrained AAM and the novel detection approach with refinement for three sentences. The black lines are placed at the starting frame of a phone, with a phone label per interval. Labels in red are not included in the training set.

<table>
<thead>
<tr>
<th>Method</th>
<th>Num. failed frames (per 1000 frames)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive AAM</td>
<td>79</td>
</tr>
<tr>
<td>Constrained AAM</td>
<td>31</td>
</tr>
<tr>
<td>Detection with or without refinement</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.1: This table shows the number of failed frames for each method.
Chapter 3. Appearance-based Inner-lip Extraction

(a) ‘Checking our top stories’

(b) ‘The fire is being allowed to burn itself out’
3.2. Results and Discussion

(c) ‘The probe will orbit the moon’

Figure 3.16: Graph (a), (b), (c) shows error (in pixels) relative to ground-truth from novel detection approach with and without refinement for three sentences. The black lines are placed at the starting frame of a phone, with a phone label per interval.

<table>
<thead>
<tr>
<th></th>
<th>AAM</th>
<th>Constrained AAM</th>
<th>Detection with refinement</th>
<th>Detection without refinement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average RMS error</td>
<td>21.43</td>
<td>18.56</td>
<td>4.78</td>
<td>3.56</td>
</tr>
<tr>
<td>(per 1000 frames)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: This table shows the average RMS error (in pixels) for each method.
Figures 3.16 shows a comparison between the novel refined detection and non-refined detection approaches presented in this Chapter. For the sentence Figure 3.16(a) 60% of frames with refinement produced a lower error (relative to ground-truth) than without refinement. For the sentence Figure 3.16(b) 69% of frames with refinement produced a lower error than without refinement. For the sentence Figure 3.16(c) 80% of frames with refinement produced a lower error than without refinement. It was found that, for the 1000 test frames, 68% of frames with refinement produced a lower error than without. Thus, it can be concluded that refinement improves the accuracy of the detection technique. For 1000 frames the average error of detection without refinement was found to be 4.78 pixels; for detection with refinement average error was found to be 3.56 pixels.

In some cases detection with refinement produced errors either the same or higher than detection without refinement because lip (for 32% of the test frames). This is thought to be due to two conditions:

1. Points along the inner-lip contour of localising to an incorrect edge. This is mainly due to noise in the image causing the edge detection to incorrectly highlight edges. This often results in the detection with refinement to produce a higher error than detection with refinement. An example of this can be seen in frame 74 of Figure 3.16(a).

2. Where the lips are closed and there no visually distinct inner-lip edge resulting in the colour space normalisation failing. In this case detection with and without perform the same. This usually occurs for frame associated with plosive, such as frame 92 of Figure 3.16(a).

Figure 3.17 shows examples of where tracking fails with the standard AAM tracker due to drift and where the inner-lip was re-aligned manually. This figure also shows an example where refinement really makes a significant difference.
Figure 3.17: This figure shows a sequence of the frames 63 to 76. The left column of images are the result of the AAM tracker; the middle shows the proposed detection approach without refinement; the images in right column are a result of the proposed detection approach method with refinement. The red boxes show were the AAM hand to be re-aligned by hand after drift. The yellow box shows an example where refinement brings the inner-lip estimate close to the actual inner-lip edges.
3.3 Conclusion

The aim of the work presented in this chapter was to introduce a novel inner-lip detection system which allows for reliable and accurate localisation. The proposed detection technique was a key frame exemplar-based method that is not dependent on any prior frame information for initialisation, allowing improved inner-lip localisation for large frame to frame changes in lip shape, which occur in 25Hz video of visual speech. A hierarchical search method was proposed which identifies the exemplars with the nearest shape and appearance from a set of exemplars. A novel colour space normalisation approach to enhance colour contrast for lip edge detection was presented. In order to make detection accurate this refinement was performed as a post-process and it was shown that this improved accuracy by 68%.

An evaluation on a 1000 frame sequence was performed. A comparison was performed against the standard AAM method in two states: a naive state and a constrained state. It was proved that tracking fails by localising to an incorrect minima and so has to be manually re-initialised. This did not occur with the detection system.

Due to constraints on data capture only a full database for one speaker was captured. In the future we would like to extend this work to multiple speakers.
Chapter 4

Analysis of Visual Phones in Continuous Speech

One of the most common approaches to producing visual speech is to morph between visemes (visual phonemes). Such approaches assume the complete set of mouth shapes associated with human speech may be spanned by a finite set of visemes. To date there is no specific definition for visemes but in general it is agreed that visemes are representations of speech that are visually contrasting. In this chapter, the representation of mouth shapes is investigated with the aim to find a set of visemes that best represents lip shape and motion.

In order to find visual separations between phonemes a combination of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) is performed. The motivation behind using PCA is to study the dynamics of the facial feature points during fluent speech, in addition to reducing the dimensionality of the data. Points on the lip are highly correlated to each other, as they cannot be physically moved independently of each other. However, as shown in Figure 3.1, the shape of the inner-lip contour is not as highly correlated, in some cases, to the shape of the outer-lip during lip-movement in speech. This inter-relation, that
occurs due to natural constraints, can be extracted using PCA. LDA provides a quantitative analysis of this inter-relation.

4.1 Initial Work

In this section, viseme data from the freely available MOCHA database [80] is investigated, which includes recordings of the 460-sentence British TIMIT corpus. The corpus includes two speakers: a male with a northern English accent and a female speaker with a southern English accent. For each of its speakers, the MOCHA database supplies audio files, Electro-Magneto Articulograph (EMA) files, laryngograph files, and electroglottograph files for the 460 sentences in the British TIMIT corpus. The EMA recordings consisted of eight sensors placed in the midsagittal plane of the vocal tract, attached to the following locations: the vermillion border of the upper lip, the vermillion border of the lower lip, the upper incisor, the lower incisor, the tongue tip, tongue body, tongue dorsum and the velum, of each speaker.

The analysis performed on this data are based on the sparse features derived from the EMA files. The EMA files contain samples at a rate of 500Hz of the x and y coordinates of the positions of the 7 different articulators, for a total of 14 values per sample. The EMA files also contain additional coordinates for the upper incisor that is only used for aligning the positions of the other articulators relative to it. In these experiments only those articulators that are visible in a front facing speaker are considered; the upper-lip, lower-lip, lower-incisor. Figure 4.1 shows EMA articulatory position data of the 7 articulators, relative to the reference articulator, for one utterance in the corpus. Figure 4.2 shows the vertical movement of the upper lip and lower lip, relative to the reference articulator (upper incisor), over time. It can be seen that the vertical movement of the upper-lip and lower-lip articulator over-time clearly shows extreme lip closure.
4.1. Initial Work

Greater variation over time can be seen in the lower-lip movement.

Figure 4.1: Seven articulatory positions from EMA data from an utterance from a female speaker in the MOCHA-Timit corpus, *f sew0*. Positions are relative to the reference articulator, the upper incisor. Data is sampled at a rate of 500Hz and in this example there are 96 samples. Only articulators highlighted in blue are considered for the experiments in this chapter.

In total there are 44 phones captured in this database (excluding silence phones). The data spanning each instance of each phone for the three articulators was extracted from the database. PCA is performed on this data creating an EMA phone space, Section 4.4. Phones were clustered according to the phoneme to viseme mapping used in Ypsilos *et al.* [82]. It was found only 3 main visemes groups could be separated: voiceless alveopalatal, bilabials and alveolar laterals, as shown in Figure 4.3. In the PCA representation there is significant overlap between phonemes with distinct visual dynamics. This indicates that the EMA data does not contain sufficient information on the mouth shape to separate the underlying visemes. The articular points used do not include any information on the lip-widening or rounding. This analysis indicates the a more detail tracking of the lip-shape is required to characterise lip dynamics.
Figure 4.2: Articulator positions over time; (a) shows the upper lip’s vertical movement relative to the upper incisor and (b) shows the lower lip’s vertical movement relative to the upper incisor.

4.2 Lip Parameterisation

Following the observations on sparse articulators other information on lip shape and texture is investigated to characterise visemes.

1. Outer-lip contour

The outer-lip contour is represented by a set of $O$ contour points around the outer-lip edge of the mouth, as shown in Figure 4.4(a). The outer-lip contour landmark points are automatically extracted using the method presented in Section 3.1.3.

2. Inner-lip contour

The inner-lip contour is represented by a set of $I$ contour points around the inner-lip edge of the mouth. In this chapter the inner-lip is considered to be the visible tangent to the lip edge, as shown in Figure 4.4(b). Both the
4.2. Lip Parameterisation

Figure 4.3: This figure shows the EMA data of the three articulators, the upper-lip, lower-lip and lower-incisor for instances of three phones groups: /p,b,m/ (red), /l/ (blue) and /ch,jh,sh/ (green), projected onto the first 3 Principal Components of the EMA phone space.

Figure 4.4: (a) $O$ landmarks of a parameterised outer-lip; (b) $I$ landmarks of a parameterised inner-lip; (c) An example of a ‘shape-free’ texture found using Delaunay triangulation.
outer-lip and inner-lip points change physically over time. The inner-lip is automatically extracted using a novel appearance-based detection system as presented in Section 3.1.4.

3. Texture

A ‘shape-free’ texture is found from each mouth image using an iterative image segmentation based on Delaunay triangulation, Section 3.1.2. This removes spurious texture variation due to shape differences. Triangulation is dependent on the outer-lip and inner-lip contour points. For a single frame, intensity pixel values for the lips and then the inner mouth area are concatenated to form a texture vector, \( \hat{g} \), Figure 4.4(c).

4.3 Data Acquisition

A speech corpus [33] of 4000 frames of a talking southern English speaking female was captured using a Sony CCD colour camera, under uniform lighting conditions. The speaker is always facing towards the camera and there is minimal head movement. A rectangular area, (100x170 pixels), initialised manually, around the mouth is extracted from the video sequences.

The landmark points on both inner and outer lip boundaries are represented as a two-dimensional Cartesian coordinate, \((x_i, y_i)\). For a single lip example all landmark coordinates for the two boundaries, \(I\) contour points for the inner-lip and \(O\) for the outer-lip, are concatenated to form a \(O + L\) element vector, \( \hat{x} \), where

\[
\hat{x} = (x_1, y_2, ..., x_N, y_N)^T, \tag{4.1}
\]

Given \(F\) training examples, an \(N \times F\) matrix, \(X\), can be generated:

\[
X = (\hat{x}_1, \hat{x}_2, ..., \hat{x}_F), \tag{4.2}
\]
where $N$ is the number of dimensions and $F$ are the number of variables.

### 4.3.1 Aligning the Training Set

Before statistical analysis is performed on the vectors in $S$, the shapes represented are aligned to a common coordinate reference or pose. Translation, scale and rotational effects are filtered out, so that the shape information are aligned according to an affine transform. Generalised Procrustes alignment is used, however, it is performed without scaling, since it is presumed there is minimal head movement in the training set. The scaling constraint means that the aligned shapes lie on a hyper-sphere, which can introduce significant non-linearities, [18]. Procrustes alignment is an iterative approach:

1. Choose an initial estimate of the mean shape of the data set
2. Align all the remaining shapes to the mean shape (translate and rotate).
3. Re-calculate the estimate of the mean from the aligned shapes.
4. If the estimated mean has changed return to step 2

Convergence is declared when the mean shape does not change significantly within an iteration. It was found that two iterations was enough. The Procrustes mean is:

$$\bar{x} = \frac{1}{F} \sum_{i=1}^{F} \hat{x}_i,$$  \hspace{1cm} (4.3)

### 4.4 Statistical Modelling

Principal Component Analysis or PCA is a common technique for data processing and dimensionality reduction. The purpose of a PCA is to reduce the number
Chapter 4. Analysis of Visual Phones in Continuous Speech

of factors of a data set while retaining as much information as possible. It seeks
the linear combination of the original variables such that the derived orthogonal
dimensions capture maximal variance. It performs a rotation of the data to an
orthogonal coordinate system formed by eigenvectors.

PCA and Independent Component Analysis (ICA) are related statistical tech-
niques. They both provide a linear decomposition of the sampled data. The
fundamental difference is that PCA assumes the latent variables are uncorrelated
whereas ICA assumes they are independent. ICA tries to explain data as a linear
combination of maximally independent basis signals, the Independent Compo-
nents. The goal of PCA is to find a sequence of uncorrelated random variables
(components) where each variable covers as much of the variance of the data as
possible. The resulting sequence is ordered by decreasing variance coverage. For
this reason, PCA is often an effective compression technique: by keeping the first
few components most of the variance in the data can be covered. The independent
components produced by ICA provide a separation mechanism between sources
that are assumed independent rather than a compression mechanism.

Blanz and Vetter [10] [11] performed PCA on a set of shape and texture vectors
that represent the 3D scan of example faces. This provides an estimate of the
probability density within a face space. Kalberer and Van Gool [42] extract the
natural modes of the face during speech and reduce the dimensions of their mask
shape space using PCA. PCA is used in this chapter for the modeling of lip-shape
variation in continuous speech. The phone data used in this chapter is all from the
same speaker and so ICA cannot be used to find separation between the phones
as they do not come from independent sources. If data from multiple speakers
was analysed, ICA would be considered. PCA produces a parameterised viseme
space that allows for statistical analysis of the actual speech data, specifically
phones.
4.4.1 A Statistical Model of Shape

The lip shape data obtained, Section 4.3 is subjected to Principal Component Analysis (PCA). By calculating the principal components and removing those corresponding to low variance, the dimensionality of the feature vectors can be reduced. Each lip shape, \( \hat{x} \), consists of \( N \) lip points and the number of total frames is \( F \).

The empirical mean lip shape, \( \bar{x} \), is found from \( S \) by finding the mean along the number of variables for each dimension.

\[
x_n = \frac{1}{F} \sum_{f=1}^{F} \hat{x}_{n,f} \quad \text{(4.4)}
\]

\[
\bar{x} = (x_1, ..., x_N) \quad \text{(4.5)}
\]

This mean is then subtracted from the each lip shape, \( \hat{x} \), in \( X \).

\[
X' = X - \bar{x} \quad \text{(4.6)}
\]

The covariance matrix, \( C \), for \( X' \) is obtained as

\[
C = \frac{1}{F-1} \sum_{f=1}^{F} X'X'^T, \quad \text{(4.7)}
\]

where \( T \) is the transpose.

PCA is performed by calculating the eigenvectors of the \( F \times F \) covariance matrix. Eigenvalues and unit eigenvectors for this matrix can be calculated using singular value decomposition, SVD. This decomposes the covariance matrix as the product of a rotation, a scaling and the inverse rotation:

\[
C = P \lambda Q^T \quad \text{(4.8)}
\]

where \( \lambda \) is the diagonal scaling matrix of the eigenvalues (singular values) and \( P \) is the rotation matrix whose columns correspond to \( p \) eigenvectors. The highest
Chapter 4. Analysis of Visual Phones in Continuous Speech

eigenvalues are the principal components that characterise the most important dimensions of the data set.

PCA can be summarised as the following linear equation:

$$\hat{x} = \bar{x} + P_s b_s,$$  \hspace{1cm} (4.9)

where $P_s = (p_1|p_2|...|p_t)$ is a set of $t$ orthogonal modes of variation, that can then approximate any of the training set; $b_s$ is a set of shape parameters, given by:

$$b_s = P_s^T(\hat{x} - \bar{x}),$$  \hspace{1cm} (4.10)

that when varied can change the mean shape, $\hat{x}$ using Equation 4.9.

The eigenvectors in $P$ are re-arranged in order of decreasing eigenvalue. The eigenvalues represent the distribution of the source data’s variance (or energy) among each of the eigenvectors, where the eigenvectors form a basis for the data. The cumulative energy content $\hat{e}$ for the $t^{th}$ eigenvector is the sum of the energy content across all of the eigenvectors from 1 through $t$:

$$\hat{e}[t] = \sum_{q=1}^{t} \lambda_{p,q} \text{ for } p = q \text{ and } t = 1...T$$  \hspace{1cm} (4.11)

Using $\hat{e}$ as a guide a subset of $L$ eigenvectors, with the largest eigenvalues, are selected from $P$ as basis vectors:

$$85\% \leq \hat{e}[t = L] \leq 90\%$$  \hspace{1cm} (4.12)

The output of this is a set of $L$ principal components and the variance of each component.

4.4.2 A Statistical Model of Texture

‘Shape-free’ texture is found across the training set based on the outer-lip and inner-lip shape for each frame. For a single frame, intensity pixel values for the
lips and then the inner mouth area are concatenated to form a texture vector, \( \hat{g} \). Given \( F \) training examples, \( F \) number of texture vectors of length \( N \) can be generated, resulting in an \( N \times F \) matrix \( G \):

\[
G = (\hat{g}_1, \hat{g}_2, ..., \hat{g}_F),
\]

where \( N \) is the number of dimensions and \( F \) are the number of variables.

PCA is applied to \( \hat{g} \) to obtain the linear model:

\[
\hat{g} = \bar{g} + P_g b_g,
\]

where \( \bar{g} \) is the mean normalised grey-level vector, \( P_g \) is a set of orthogonal modes of variation and \( b_g \) is a set of grey-level parameters.

### 4.4.3 A Statistical Model of Appearance

In order to analyse a lips structure, the variation in shape and variation in texture of the lip needs to be modeled. Lip shape variation is modelled, as shown in Section 4.4.1. ‘Shape-free’ texture is obtained from the these lip shapes and the texture variation is modelled, as shown in Section 4.4.2. A combined statistical model of both shape and texture can be generated using Appearance models [18]. Appearance models make the assumption that there is correlation between parameters in the shape model and parameters in the texture model across the training set.

From the shape models and texture models discussed earlier, the shape and texture of any example can be summarised by the parameter vectors \( b_s \) and \( b_g \) respectively. A further PCA is applied to these vectors, generating a concatenated vector

\[
b = \begin{pmatrix} W_s b_s \\ b_g \end{pmatrix} = \begin{pmatrix} \bar{x} \\ P_s^T (\bar{x} - \bar{x}) \\ P_g^T (\hat{g} - \bar{g}) \end{pmatrix}
\]

(4.15)
where $\mathbf{W}_s$ is a diagonal matrix of weights for each shape parameter in $\mathbf{b}_s$. This weighting is required because of the difference in units between texture (intensity values) and shape (cartesian coordinate values or distance) and so they cannot be compared directly. The weight matrix $\mathbf{W}_s$ can be set as $r\mathbf{I}$ where $r^2$ is the ratio of the total intensity variation to the total shape variation.

A further PCA is applied on $\mathbf{b}$ to give the following model:

$$\mathbf{b} = \mathbf{P}_a \mathbf{b}_c$$

(4.16)

where $\mathbf{P}_a$ are the eigenvectors and $\mathbf{b}_c$ is a vector of appearance parameters controlling both the shape and grey-levels of the model.

The shape model (Equation 4.9) and texture model (Equation 4.14) can now be expressed linearly as:

$$\hat{x} = \bar{x} + (Q_s c), \quad \hat{g} = \bar{g} + (Q_s c)$$

(4.17)

where $Q_s = \mathbf{P}_s \mathbf{W}_s^{-1} \mathbf{P}_{as}$ and $Q_g = \mathbf{P}_g \mathbf{P}_{ag}$ and where

$$\mathbf{P}_a = \begin{pmatrix} \mathbf{P}_{as} \\ \mathbf{P}_{ag} \end{pmatrix}$$

(4.18)

### 4.4.4 Finding the Modes of Variation

For the shape, texture and appearance models the modes of variation for each principal component can be found. For simplicity each model can be summarised as a linear approximation:

$$\hat{\mathbf{v}}(j) = \bar{\mathbf{v}} + \sum_{k=1}^{t} \mathbf{p}(k) \mathbf{b}(k,j),$$

(4.19)

where $\mathbf{p}(k)$ are the $t$ most significant eigenvectors of the sample covariance matrix, i.e. those having the largest variance $\sigma^2(k)$ as obtained from the eigenvalues of
the covariance matrix. The weight coefficient of each mode, $b(k, j)$, may be found by using the orthonormality of the eigenvectors $p(k)$.

For visualisation deviations from the sample mean, $\bar{v}$, along each of the principal modes of variation $p(k)$, were evaluated at $\pm 3$ times the standard deviation $\sigma^2(k)$.

$$\hat{v}(j) = \bar{v} + \sum_{k=1}^{t} p(k)\sigma(k),$$

where

$$\sigma(k) = b(k, j) \ast w, \quad -3sd < w < +3sd$$  \hspace{1cm} (4.21)

### 4.5 Analysis of Visual Phones

Visemes are segmented from a database using Forced Alignment, as shown in Section 3.1.1, that takes as input the audio signal and phoneme transcriptions. This aligns the transcribed data with the speech data, resulting in time segments per phoneme. These timings can then be used to extract the corresponding lip parameterisation data, as obtained earlier. The audio is captured at a sample rate of $16kHz$ and the video frames are captured at a rate of $25Hz$.

There are two ways in which the extracted viseme data can be visualised, static and dynamic. For the static case the mid-point of the phoneme timings is found and the lip data frame closest to this mid-point is extrapolated. So each phoneme is represented by a single frame and the phonemes can be compared directly. For the dynamic case all frames within each phoneme timings represent a phoneme. Example images of these can be shown in Figure 4.5.

In both states the lip data (being shape and/or texture) is modelled using PCA, Section 4.4, resulting in a reduced dimensionality “viseme space”. K-Means clustering is used to help identify key viseme classes in this projected viseme space for visualisation only. Linear Discriminant Analysis is then performed to quantitively measure the difference between the viseme classes.
Figure 4.5: Each column shows the several images associated with an instance of a phoneme (dynamic). Images highlighted with red borders are images selected as static representations.

### 4.5.1 Linear Discriminant Analysis

To quantitively measure which principal component gives the greatest separation between the visemes, Fisher’s linear discriminant analysis (LDA) is applied [5]. LDA, a widely used technique for pattern classification, finds the linear boundary that yields optimal discrimination between two classes. Data is projected onto a line and classification is performed in the 1D space. The projection maximises the distance between the means of the two classes, while minimising the variance within class. LDA uses the projection that maximises the following ratio:

\[
J(\Theta) = \arg \max_{\Theta} \frac{\Theta^T S_B \Theta}{\Theta^T S_W \Theta}
\]  

(4.22)

where \( S_B \) is the between classes scatter matrix, \( S_W \) is the within classes scatter matrix and \( \Theta \) is the covariance matrix of all the data. The between classes scatter matrix is defined as:

\[
S_B = \sum_{c=1}^{N_c} p_c (\mu_c - \overline{x})(\mu_c - \overline{x})^T
\]  

(4.23)

where, \( \mu_c \) is the mean vector for class \( c \), \( p_c \) is the fraction of data belonging to class \( c \) and \( \overline{x} \) is the mean of all data. The within classes scatter matrix can
defined as:

\[ S_W = \sum_{c=1}^{N_c} p_c \Theta_c \]  \hspace{1cm} (4.24)

where \( \Theta_c \) is the covariance matrix of class \( c \).

\[ \Theta_c = \frac{1}{N_c} \sum_{c=1}^{N_c} (\mathbf{x}_c - \mu_c)(\mathbf{x}_c - \mu_c)' \]  \hspace{1cm} (4.25)

where \( \mathbf{x}_c \) is the mean of the data associated with class \( c \).

## 4.6 Results and Discussion

The lip frame data for thirteen phones (/p/, /b/, /m/, /ch/, /sh/, /jh/, /w/, /f/, /v/, /dh/, /th/, /t/, /l/) are extracted from the database of a single speaker. In total the thirteen phones account for around 1,500 frames or 60 seconds worth of lip frame data. Table 4.1 shows the 13 phones grouped into 7 consonant viseme groups based on those used by Ypsilos et al. [82].

<table>
<thead>
<tr>
<th>Viseme Group</th>
<th>Phonemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viseme Group 1</td>
<td>/p,b,m/</td>
</tr>
<tr>
<td>Viseme Group 2</td>
<td>/f,v/</td>
</tr>
<tr>
<td>Viseme Group 3</td>
<td>/t,d/</td>
</tr>
<tr>
<td>Viseme Group 4</td>
<td>/th,dh/</td>
</tr>
<tr>
<td>Viseme Group 5</td>
<td>/w/</td>
</tr>
<tr>
<td>Viseme Group 6</td>
<td>/ch,jh,sh/</td>
</tr>
<tr>
<td>Viseme Group 7</td>
<td>/l,n/</td>
</tr>
</tbody>
</table>

Consonants may also be classified according to the manner of articulation and the point of articulation: that is, how and where the flow of air is stopped or impeded when the consonant is articulated. The seven viseme groups chosen have
been categorised by the point of articulation (physical). Examples of the classes that lie within this category are Bilabials, Labiodentals, Interdentals, Alveolar, Alveopalatal and Velar. Bilabials (/p,b,m/) are consonants for which the flow of air is stopped or restricted by the two lips, thus producing the narrowest lip shape out of the seven visemes, Figure 4.6. Labiodentals (/f,v/) are consonants for which the flow of air is restricted by the lips and teeth (usually upper teeth). Alveolars (/t,d/) are consonants for which the flow of air is stopped or impeded by creating a block or a small aperture between the tongue and the alveolar ridge.

Interdentals (/th,dh/) are consonants for which the flow of air is restricted by catching the tongue between the teeth, resulting in both the teeth and tongue being equally visible during articulation. Alveopalatals (/ch,jh,sh/) are consonants for which the flow of air is stopped or impeded by creating a block or a small aperture between the tongue and the region of the hard palate just behind the alveolar ridge. Due to the positioning of the tongue, the alveopalatals produce the most open mouth shape during articulation (resulting in the greatest teeth exposure) of the seven viseme groups investigated here. Viseme /w/ is a

Figure 4.6: Cross section of a head. Labelled items influence the production of sound in one way or another.
velar semi-vowel. Semivowels are vowel-like consonants: that is, the air-flow is not stopped or impeded so as to cause a friction-sound, but the aperture through which the air passes is smaller than the aperture of any vowel.

4.6.1 Analysis of Outer-lip Shape Information

The outer-lip shape data for the 1,500 frames is extracted and PCA is performed, creating a statistical outer-lip point distribution model. Dimensionality is reduced and can be represented in 9 modes of variation as shown in Figure 4.7. Figure 4.8 shows the outer-lip shape data projected onto the first 3 principal components of the PCA model. Each phone instance spans several frames, the shortest phone being 2 frames long and the longest being 7 frames long. There are 382 instances of the thirteen phones in total. Each phone instance, consisting all frames within the phone, is represented as a line that is coloured according to which viseme group it belongs to.

It can be seen in Figure 4.8 that instances of the same visemes tend to cluster within the same area, particularly for the viseme group /p,b,m/. However, separation between visemic groups cannot be seen in this space.

For each instance of the thirteen phones the mid-point is found, based on its phoneme timings found through Forced Alignment, and the data associated with the lip frame closest to this mid-point is extrapolated. Figure 4.9 shows the static representation of each phone, projected onto the first two principal components. Clustering of static instances of the viseme groups looks tighter than in the dynamic space. However, again the separation between the viseme groups cannot be easily determined.

LDA is performed on the outer-lip shape data in order to quantitatively measure the separability between the viseme groups in this PCA space. The results can be seen in Figure 4.10. The greatest separation can be found between the
Chapter 4. Analysis of Visual Phones in Continuous Speech

Figure 4.7: The first nine principal modes of outer-lip shape variation captured in the training set.

Figure 4.8: Dynamic instances of outer-lip data projected on the first 3 principal components of the outer-lip point distribution model.
Figure 4.9: Static instances of outer-lip data projected onto the first 2 PCs of the outer-lip point distribution model.

visemes /w/ and /l/. It was found that PC2 (rounding of outer-lip) contributed the highest proportion (64%) to this separation, Figure 4.11. The outer-lip shape associated with viseme /w/ produces the most extreme rounding out of all the other viseme groups investigated in this chapter. The second greatest separation is between visemes /p,b,m/ and /ch,sh,jh/. It was found that PC1 (opening and closing of the outer-lip) contributed the most (85%) to this separation, Figure 4.12. Again, out of the seven viseme groups investigated it can be observed that for viseme /ch,sh,jh/ the outer-lip shape data associated produces the biggest open mouth shapes and viseme /p,b,m/ (bilabials) produces the narrowest. These assumptions are reflected in the LDA results.

4.6.2 Analysis of Inner-lip Shape Information

The inner-lip shape data for the 13 phones is extracted and PCA is performed. This creates an inner-lip distribution model. Dimensionality is reduced and can
<table>
<thead>
<tr>
<th>/p-b-m/</th>
<th>/ch-jh-sh/</th>
<th>/w/</th>
<th>/f-v/</th>
<th>/dh-th/</th>
<th>/t/</th>
<th>/l/</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.8665</td>
<td>1.0763</td>
<td>1.4527</td>
<td>0.99313</td>
<td>2.2383</td>
<td>2.3738</td>
</tr>
<tr>
<td>3.8665</td>
<td>0</td>
<td>2.831</td>
<td>1.8263</td>
<td>1.4545</td>
<td>0.6643</td>
<td>3.1437</td>
</tr>
<tr>
<td>1.0763</td>
<td>2.831</td>
<td>0</td>
<td>1.8515</td>
<td>1.4885</td>
<td>1.9793</td>
<td>4.0706</td>
</tr>
<tr>
<td>1.4527</td>
<td>1.8263</td>
<td>1.8515</td>
<td>0</td>
<td>0.88123</td>
<td>0.43977</td>
<td>2.9994</td>
</tr>
<tr>
<td>0.99313</td>
<td>1.4545</td>
<td>1.4885</td>
<td>0.88123</td>
<td>0</td>
<td>0.32213</td>
<td>1.4585</td>
</tr>
<tr>
<td>2.2383</td>
<td>0.66443</td>
<td>1.9793</td>
<td>0.4397</td>
<td>0.32213</td>
<td>0</td>
<td>1.0583</td>
</tr>
<tr>
<td>2.3738</td>
<td>3.1437</td>
<td>4.0706</td>
<td>2.9994</td>
<td>1.4585</td>
<td>1.0583</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 4.10: This table shows quantitative measures of separability found through LDA between all seven viseme groups in the outer-lip shape space.

Figure 4.11: This figure shows the contribution of the first three principal components in the separability of viseme /w/ vs all seven viseme groups.
4.6. Results and Discussion

Figure 4.12: This figure shows the contribution of the first three principal components in the separability of viseme /pmb/ vs all seven viseme groups. These visemes can be represented in 5 modes of variation as shown in Figure 4.13. Figure 4.14 shows the inner-lip shape data projected onto the first 3 principal components of the PCA model. This figure shows that overall clearer separation can be seen, when compared to Figure 4.8, between dynamic instances of the viseme groups projected in the first 3 PCs for the inner-lip shape space in comparison to the outer-lip shape space. This can be particularly seen between groups /p,b,m/ and /ch,sh,jh/.

Figure 4.13: The first five principal modes of inner-lip shape variation captured in the inner-lip distribution model.
Chapter 4. Analysis of Visual Phones in Continuous Speech

Figure 4.14: Dynamic instances of inner-lip data projected on the first 3 principal components of the inner-lip point distribution model.

Figure 4.15 shows the static representation of each phone, projected onto the first three principal components in the inner-lip shape space. It can be seen that each viseme group clusters more closely in this space in comparison with those in Figure 4.9 and there is also increased separation between the viseme groups.

LDA is performed on the inner-lip shape data in order to quantitatively measure the separability between the viseme groups in this PCA space. The results can be seen in Figure 4.16. Again the greatest separation between is found between visemes /w/ and /l/ with PC2 contributing 49% to this variation. Variation in PC1 and PC3 also contribute 41% and 10% respectively to the separation between /w/ and /l/. So it is the contribution of the inner-lip rounding, opening and diagonal variation that create this separation.
4.6. Results and Discussion

Figure 4.15: Static instances of inner-lip data projected onto the first 3 PCs of the inner-lip point distribution model.

<table>
<thead>
<tr>
<th></th>
<th>/p-b-m/</th>
<th>/ch-jh-sh/</th>
<th>/w/</th>
<th>/f-v/</th>
<th>/dh-th/</th>
<th>/t/</th>
<th>/l/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p-b-m/</td>
<td>0</td>
<td>2.4471</td>
<td>0.93836</td>
<td>1.1038</td>
<td>0.79888</td>
<td>2.0657</td>
<td>1.9129</td>
</tr>
<tr>
<td>/ch-jh-sh/</td>
<td>2.4471</td>
<td>0</td>
<td>2.9853</td>
<td>1.7845</td>
<td>0.72776</td>
<td>0.66883</td>
<td>3.7054</td>
</tr>
<tr>
<td>/w/</td>
<td>0.93836</td>
<td>2.9853</td>
<td>0</td>
<td>1.8236</td>
<td>1.1725</td>
<td>1.8614</td>
<td>4.107</td>
</tr>
<tr>
<td>/f-v/</td>
<td>1.1038</td>
<td>1.7845</td>
<td>1.8236</td>
<td>0</td>
<td>0.85081</td>
<td>0.45879</td>
<td>2.7224</td>
</tr>
<tr>
<td>/dh-th/</td>
<td>0.79879</td>
<td>0.72776</td>
<td>1.1725</td>
<td>0.85081</td>
<td>0</td>
<td>0.03795</td>
<td>1.8654</td>
</tr>
<tr>
<td>/t/</td>
<td>2.0657</td>
<td>0.66883</td>
<td>1.8614</td>
<td>0.45879</td>
<td>0.03795</td>
<td>0</td>
<td>0.88616</td>
</tr>
<tr>
<td>/l/</td>
<td>1.9129</td>
<td>3.7054</td>
<td>4.107</td>
<td>2.7224</td>
<td>1.8654</td>
<td>0.88616</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 4.16: This table shows quantitative measures of separability found through LDA between all seven viseme groups in the inner-lip shape space.
4.6.3 Analysis of Outer and Inner-lip Shape Information

The outer and inner-lip shape data for the 13 phones is concatenated and PCA is performed. This creates a mouth shape distribution model. Dimensionality is reduced and can be represented in 11 modes of variation accounting for 98% of the variance, as shown in Figure 4.17. Figure 4.18 shows the mouth shape data projected onto the first 3 principal components of the PCA model. The viseme groups span a greater area in this space.

Figure 4.19 shows the static representation of each phone, projected onto the first three principal components in this mouth space. There is no clearer separation when looking at the data projected onto the first 3PCs of the mouth shape space in comparison to the inner-lip shape space.

Figure 4.17: The first nine out of eleven principal modes of mouth shape variation captured in the mouth shape distribution model.

LDA is performed on the mouth shape data and the results can be seen in Figure 4.20. It can be seen that there is higher separation between all viseme groups
4.6. Results and Discussion

Figure 4.18: Dynamic instances of mouth shape (outer and inner-lip) data projected on the first 3 principal components of the mouth shape space.

Figure 4.19: Static instances of the mouth shape data projected onto the first 3 PCs of the mouth shape model.
Chapter 4. Analysis of Visual Phones in Continuous Speech

This indicates that representing mouth shape as a concatenation of outer-lip and inner-lip data helps to distinguish between viseme groups more significantly than if just the outer-lip or inner-lip data were considered.

4.6.4 Analysis of Texture Information

The texture information for the 13 phones is extracted and a statistical model of texture is created, Section 4.4.2. In initial experiments it was found that inner mouth texture alone provided all the information needed to distinguish between viseme groups, variation in the texture of the lips alone added nothing significant and so was excluded from any further analysis. Dimensionality is reduced to 220 modes of variation. Figure 4.21 shows the modes of variation first 3 PCs. The first mode of variation shows the change from no visible teeth to just teeth.

Figure 4.22 shows the texture data projected onto the first 3 principal components of the PCA model. Significantly more intra and inter-viseme separation can be seen in the static representation of viseme in this texture space, Figure 4.23.

LDA is performed on the texture data in order to quantitatively measure the separability between the viseme groups in this PCA space. The results can be seen in Figure 4.20.
4.6. Results and Discussion

Figure 4.21: The first 3 principal modes of texture variation captured in the texture space. Texture of the lips is included in this figure for visualisation purposes only. Columns show variation between $\pm 3$ standard deviations from left to right.

Figure 4.22: Dynamic instances of texture data projected on the first 3 principal components of the texture space.
Figure 4.23: Static instances of the texture data projected onto the first 3 PCs of the texture model.

in Figure 4.24. It can be seen that compared to the LDA results for mouth shape, Figure 4.20, there is greater separation between viseme groups /ch-jh-sh/ and /w/. Tooth information is not accounted for with the shape representation of the mouth but is visible with the inner-lip texture. For the alveopalatals /ch-jh-sh/ the teeth is visually prominent and for the semi-vowel there are generally no teeth visible during speech, Figure 4.5. It can also be seen that shape representations show greater separation between visemes /p-b-m/ and /ch-jh-sh/ than for the texture representation.

4.6.5 Analysis of Appearance Information

The shape model and texture model for all instances of the 13 visemes are concatenated and a statistical model of appearance is created, Section 4.4.3. PCA reduces dimensionality that can be represented in 150 modes of variation, for 98% variance. Shape and texture data are represented in different units and so cannot be compared directly. In order to compensate for this a weighting matrix is ap-
4.6. Results and Discussion

Figure 4.24: This table shows quantitative measures of separability found through LDA between all seven viseme groups in the texture space.

<table>
<thead>
<tr>
<th></th>
<th>/p-b-m/</th>
<th>/ch-jh-sh/</th>
<th>/w/</th>
<th>/f-v/</th>
<th>/dh-th/</th>
<th>/t/</th>
<th>/l/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p-b-m/</td>
<td>0</td>
<td>3.7787</td>
<td>0.14298</td>
<td>1.6536</td>
<td>1.3337</td>
<td>3.7196</td>
<td>2.3536</td>
</tr>
<tr>
<td>/ch-jh-sh/</td>
<td>3.7787</td>
<td>0</td>
<td>5.3842</td>
<td>1.35</td>
<td>1.0519</td>
<td>0.14651</td>
<td>3.2579</td>
</tr>
<tr>
<td>/w/</td>
<td>0.14298</td>
<td>5.3842</td>
<td>0</td>
<td>2.8087</td>
<td>2.6703</td>
<td>3.9004</td>
<td>4.4933</td>
</tr>
<tr>
<td>/f-v/</td>
<td>1.6536</td>
<td>1.35</td>
<td>2.8087</td>
<td>0</td>
<td>0.4692</td>
<td>0.61966</td>
<td>3.091</td>
</tr>
<tr>
<td>/dh-th/</td>
<td>1.3337</td>
<td>1.0519</td>
<td>2.6703</td>
<td>0.4692</td>
<td>0</td>
<td>0.32367</td>
<td>2.3677</td>
</tr>
<tr>
<td>/t/</td>
<td>3.7196</td>
<td>0.14651</td>
<td>3.9004</td>
<td>0.61966</td>
<td>0.32367</td>
<td>0</td>
<td>1.2009</td>
</tr>
<tr>
<td>/l/</td>
<td>2.3536</td>
<td>3.2579</td>
<td>4.4933</td>
<td>3.091</td>
<td>2.3677</td>
<td>1.2009</td>
<td>0</td>
</tr>
</tbody>
</table>

LDA is performed on the texture data in order to quantitatively measure the separability between the viseme groups in this PCA space. The results can be seen in Figure 4.28. The results in this table show that there is generally greater separation between all viseme groups, particularly between the viseme groups that produced the greatest separations in Figures 4.20 and 4.24. It can be concluded that the combination of shape and texture information in an appearance space allows for the best distinguishable separation between viseme groups.
Figure 4.25: The first 5 principal modes of appearance variation captured in the appearance space. Texture of the lips is included in this figure for visualisation purposes only. Columns show variation between $\pm 3$ standard deviations from left to right.

Figure 4.26: Dynamic instances of appearance data projected on the first 3 principal components of the appearance space.
4.6. Results and Discussion

Figure 4.27: Static instances of the *appearance* data projected onto the first 3 PCs of the appearance model.

<table>
<thead>
<tr>
<th>/p-b-m/</th>
<th>/ch-jh-sh/</th>
<th>/w/</th>
<th>/f-v/</th>
<th>/dh-th/</th>
<th>/t/</th>
<th>/l/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p-b-m/</td>
<td>0</td>
<td>3.7787</td>
<td>0.14298</td>
<td>1.6536</td>
<td>1.3337</td>
<td>3.7196</td>
</tr>
<tr>
<td>/ch-jh-sh/</td>
<td>3.7787</td>
<td>0</td>
<td>5.3842</td>
<td>1.35</td>
<td>1.0519</td>
<td>0.14651</td>
</tr>
<tr>
<td>/w/</td>
<td>0.14298</td>
<td>5.3842</td>
<td>0</td>
<td>2.8087</td>
<td>2.6703</td>
<td>3.9004</td>
</tr>
<tr>
<td>/f-v/</td>
<td>1.6536</td>
<td>1.35</td>
<td>2.8087</td>
<td>0</td>
<td>0.4692</td>
<td>0.61966</td>
</tr>
<tr>
<td>/dh-th/</td>
<td>1.3337</td>
<td>1.0519</td>
<td>2.6703</td>
<td>0.4692</td>
<td>0</td>
<td>0.32367</td>
</tr>
<tr>
<td>/t/</td>
<td>3.7196</td>
<td>0.14651</td>
<td>3.9004</td>
<td>0.61966</td>
<td>0.32367</td>
<td>0</td>
</tr>
<tr>
<td>/l/</td>
<td>2.3536</td>
<td>3.2579</td>
<td>4.4933</td>
<td>3.091</td>
<td>2.3677</td>
<td>1.2009</td>
</tr>
</tbody>
</table>

Figure 4.28: This table shows quantitative measures of separability found through LDA between all seven viseme groups in the texture space.
4.7 Conclusion

In this chapter, thirteen phonemes are represented by four mouth representations, outer-lip and inner-lip shape, texture and the combination of both shape and texture (appearance). These thirteen phonemes were then categorised into seven consonant viseme groups. Using PCA these representations were modelled. This reduces dimensionality and also helped to identify the most significant variations for these representations. LDA was used to quantitatively measure the difference between the seven viseme groups in this reduced space. It was found that appearance data provided the best representation for the database captured of the mouth during speech.

It was shown that, on the whole, significant separation was found between the seven viseme groups, particularly when considering static visemes. However, there is also large in class variation due to both differences in shape and sampling resulting in significant overlap of the distributions. As a result dynamic information for each viseme cannot be easily distinguished in the context of continuous speech.

To illustrate this further, synthesis via interpolation of the word “/t/-/o/-/p/” is performed. For this example the utterances of the viseme /o/ were projected into the appearance space. As shown in Figure 4.29 the mean static representation of the three visemes is found and intermediate representations are found between these means via linear interpolation. Synthesised frames can be found by traveling through the linear trajectory (as shown by the arrows) and back-projecting each point along this trajectory. This results in a sequence of frames. This sequence is compared to a sequence of frames of the speaker saying the word “top” from the original database, Figure 4.30.

It can be seen from Figure 4.30 that the synthesis does well to show the variation in texture along the sequence in comparison to the original sequence. This is
4.7. Conclusion

Figure 4.29: This figure shows static examples of visemes /t/ (green), /o/ (orange) and /p/ (red) projected onto the first 3 PCs of the appear (all other visemes are coloured in grey). The mean static representation of these three visemes is found (black cross) and intermediate points (black crosses) between these means is found through linear interpolation.

Figure 4.30: The top row shows the appearance information from a sequence of frames of the speaker saying “top” from the original database. The bottom row shows the appearance information of a synthesised sequence of frames found through a linear trajectory that passes through static examples of the visemes /t/, /o/ and /p/. 
particularly the case with the first 3 frames in the sequence. The texture of the inner mouth with the teeth is almost the same as that for the real-life example. However, significant differences in texture can be seen with the remaining two frames in the sequence. There is also a significant difference in terms of the mouth shape between the synthesised sequence and real-life sequence.

It can be seen that the underlying dynamics of mouth movement in speech cannot be easily found in the context of continuous speech.
Chapter 5

Visual Analysis of Lip Coarticulation in VCV Utterances

Coarticulation is the variability of an articulators pose, dependent on context, caused by the assimilation of a speech unit to a preceding unit. Variability due to articulatory planning affects the subsequent speech unit. Thus, coarticulation is bidirectional. Many theories of coarticulation have sought to encode the relationship between articulatory planned sequence of speech gestures and their physical realisation [51] [39] [60] [24] [44] [68]. The constraints of physiology, effort minimisation, linguistic contrast and inter-articulator coordination all affect the average articulatory behavior and its variability. Most of these studies have been based on audio data and analysis of formant transitions. In recent years, as greater amounts of articulatory data have become widely available, researchers have concentrated on statistical approaches that allow the properties of articulatory configurations to be learnt from annotated measurements [61] [53] [23].

Dynamics is especially crucial to the multi-modal perception of spoken utterances
because our visual perception is highly sensitive to movement on human faces. We can detect minor occurrences of unnatural motion, e.g., from a discontinuity, from poor interpolation or from physiologically implausible gestures. The lips, jaw, teeth and occasionally the tongue are the only parts of the human vocal apparatus that are obviously visible in the face, so not all articulators are relevant in the study of visual speech. Variation in visual dynamics are important cues for non-verbal communication.

In this work the underlying dynamics of labelled visual speech units, represented as lip shape, is investigated from VCV utterances. The aim of this work is to provide detailed quantitative statistical analysis, from ensemble data, of variation in lip dynamics due to coarticulation. Here, tracking of both inner and outer lip contours is used to describe the lip configuration and its visual appearance in video.

In the articulatory and acoustic domains, there have been attempts to provide statistical models of the dynamics of speech movements that account for coarticulatory effects. Coarticulations can exhibit dependency on neighbouring phonetic context, this can be modelled in three ways:

- in the articulator’s 3D coordinates,
- in the relative configuration of articulators,
- in terms of the articulators’ motion over time.

The study presented in this chapter investigates all three aspects for the bilabial plosive consonant /p/ which, together with /b/ and /m/, is arguably the most crucial speech gesture in English from the perspective of visual impact. The pronunciation of /p/ involves context dependent combinations of lip movement, an interaction of lip and jaw positions and a rapid transition at release of the
5.1 Data Acquisition

Data is acquired from color video, captured at a rate of 25Hz, of an English speaker, under uniform lighting conditions. The speaker is always facing towards the camera and there is minimal head movement. A rectangular area, (100x170 pixels), initialised manually, around the mouth is extracted from the video sequences. In the experiments the lips are parameterised (outer and inner lip contour) as a set of $N$ landmarks, Figure 5.1. Each landmark on the lip boundaries is represented as a Cartesian coordinate. All landmark coordinates around the lip contours are concatenated to form a 1D vector, $x$.

![Figure 5.1: N landmarks of a parameterised lip.](image)

The landmarks are labeled manually for half the number of frames, training data. The lip contours for the remaining frames are extracted using a standard Active Appearance Model (AAM) tool [19]. Each lip shape is translated, rotated and scaled so as to minimise the sum of squared distances with respect to the first frame.

For the work presented in this chapter the analysis is based on the case of a voiceless bilabial /p/ and the voiced bilabial /m/ in three VCV contexts. Three
cardinal monophthong vowels /a/ (low back), /i/ (high front) and /u/ (high back and rounded) have been chosen. Vowels are open mouthed sounds in speech and can be best defined by tongue position (height and location) and roundedness of lips. Cardinal vowels occur at the extrema of tongue positioning. The chosen cardinals are the extreme front and back vowels, Figure 5.2.

Figure 5.2: International Phonetic Alphabet (IPA) vowel chart with English examples. Vowels at the right and left of the circles are rounded and unrounded. The chosen cardinal vowels are marked with red circles.

The three VCV nonsense words are placed in the carrier phrase “Is ... it?” The consonants in the carrier phrase form visually consistent stops (mouth open, wide and teeth showing and together). All three phrases are repeated 15 times for both /p/ and /m/, giving a total of 90 speech tokens. Lip data for frames associated with the VCV utterances are considered; frames associated with the carrier phrase are ignored. For this experiment there are a total of 932 frames, \( T \), of lip shape data, \( x \), accumulated into a matrix, \( X \).

\[
X = [x_t]', t = 1 \rightarrow T
\]  

Time synchronised audio is captured at 16 kHz. The audio stream is manually labelled into the known phonemes of the carrier phrases, using the audio waveform, wide-band and narrow-band spectrograms as a reference, Figure 5.3. It was found that the manual audio labelling has an error of no more than a frame.
5.2 Alignment and Resampling of the Speech Fragments

Given the appropriate selection of phonemes these are aligned to a common reference (the first carrier phrase utterance), Figure 5.4. The time alignment allows for comparison of lip data across all the utterances at any time instant. Given a speech fragment of length $T_i$, this can be warped to the length of the reference speech fragment $T_o$, as shown in Equation 5.2.

$$T'_i = T_i \ast \rho = T_0$$  \hspace{1cm} (5.2)

where $\rho$ is the warp ratio:

$$\rho = T_0/T_i$$  \hspace{1cm} (5.3)

The largest adjustment of any phone required a 18% scaling onto the reference utterance. This alignment can be achieved by evenly distributing frame data between the repositioned phonetic labels. However, that will lead to an uneven distribution in the sampling of the phonemes, which gives an inconsistent frame
5.3 Dynamic Analysis

Ensemble analysis is performed on lip shape data acquired from a speaker for two phonemes /p/ and /m/ in three contexts; /aa-p-aa/, /uu-p-uu/, /ii-p-ii/ and /aa-m-aa/, /uu-m-uu/, /ii-m-ii/. All lip shape data obtained, $X$, is subjected to Principal Component Analysis (PCA), resulting in a VCV lip-space. By calculating the principal components and removing those corresponding to low variance,
the dimensionality of the feature vectors can be reduced. The lip shape consists of \( N \) lip points and the total number of frames is \( F \) resulting in an \( N \times F \) matrix \( S \). PCA is performed by calculating the eigenvectors of the \( N \times N \) covariance matrix of \( S \). The output of this is a small set of principal components (PC) and the variance of each component. In this experiment 9 PC’s with the largest variance were used, accounting for 98% of the total variation in the observed data.

The work presented in this chapter is focused on the analysis of the first two principal components, that account for 89% of the variance, Figure 5.5. The first principal component represents the general variation of the opening and closing and the widening and rounding of the mouth, 80% of the variance. The second PC represents a local variation of the inner-lip that displays protrusions associated with the rounding of the mouth, representing 9% of the total variation. All other components represent subtle local variations, which are less significant for analysis, but may be considered for modelling in future work.

Figure 5.5: The first two principal modes of mouth shape variation for the VCV lip-space.
5.3.1 Analysis of Phoneme /p/

In this section of the work the dynamics of VCV utterances for phoneme /p/ is analysed in three coarticulation contexts. Figure 5.7 shows all repetitions of /apa/, /ipi/ and /upu/ projected onto the first 3 principal components of the VCV lip space. In this representation clear separation between the three VCV classes can be seen.

Using PCA, the lip shapes can be reconstructed as a weighted sum of the principal components. Figure 5.6 shows how the lip geometry varies over time, sampled at nine time steps, with three lip representations per vowel and consonant of the VCV. This provides a qualitative visual analysis of the difference between the lip shapes associated with /apa/, /ipi/ and /upu/. Figure 5.6(a) shows how the lips generally vary over time for the first principal component. The difference in lip shape between the coarticulations is visible, particularly at time stamps within the regions of the vowels. Figure 5.6(b) shows variation in lip shape for the second component. There is a distinct difference in inner-lip shape over time between /upu/ and /apa/ and also between /upu/ and /ipi/. There is no distinct difference between /apa/ and /ipi/, over all time steps, for this principal component.

The time-aligned shape data is projected onto the first two principal components, Figure 5.8. A polynomial is fitted onto the data to allow for continuous interpolation. For dynamic analysis continuous interpolation of the data is considered more appropriate than sampling the data in discrete time bins as in standard ensemble analysis. It can be seen that there is variation between different coarticulation contexts for the same phoneme. Intra-coarticulation variation (variance of utterances of the same VCV context) is observed to be lower than the inter-coarticulation variation (variance of utterances across VCV contexts).
Figure 5.6: A sequence of the reconstructed lip shapes, frame rate, along the VCV time interval, for (a) 1st PC, (b) 2nd PC. Red represent variation (±3 s.d.) for /apa/; green for /ipi/ and blue for /upa/.
Chapter 5. Visual Analysis of Lip Coarticulation in VCV Utterances

The associated lip shape for each VCV context was generated at three time steps (the 2nd, 5th and 8th lip shapes in Figure 5.6. It can be seen in Figure 5.8 that utterances between time intervals $t_1$ and $t_2$ (first vowel) for VCV /apa/ exhibit open inner and outer-lip shapes. For /ipi/ lip shapes are narrower in height and wider in width and for /opo/ lip shapes are rounder.

At the mid-point of the time interval $t_2$ and $t_3$ (consonant) it can be seen that the inner-lip shape for /opo/ does not reach full closure like it does with /apa/ and /ipi/. Between time intervals $t_3$ and $t_4$ (second vowel) once again /apa/ produces open lip shapes. However, variation between the three coarticulation contexts is lower than for the $t_1$ and $t_2$ time interval, indicating that all realisations of all three VCV context are not symmetrical in nature.

Figure 5.7: All repetitions of /apa/, /ipi/ and /upu/ projected onto the first 3 PC’s. Each line in this space represents a VCV repetition, each one being on average 10 frames long.
Figure 5.8: This figure shows all the data samples over time: (a) is for the 1st PC and (b) for the 2nd PC.
Each VCV context in Figure 5.8 can be represented by a mean trajectory, $\bar{x}$ (Equation 5.5), and intra-viseme ensemble variance around the mean over time.

$$
\bar{x}(t) = \frac{1}{N} \sum_{i=1}^{N} x_i(t)
$$

(5.5)

At any time instant these variances can be sampled, as shown in Figure 5.9. Figure 5.9 (a) shows how the data for the three VCV phrases, /apa/, /ipi/ and /upu/, vary over time for the first component. It can be seen that there is a separation between /apa/ and /upu/, particularly at the peaks and troughs that lie within the intervals of all three phonemes in the VCV. There is also separation between /ipi/ and /upu/ in the $t_2$ and $t_3$ time interval (consonant), when lip motion is leading to closure. It can also be seen that the extreme lip position of phoneme /p/, in the /apa/ context, has a slower onset than for the other two utterances. This indicates that it takes longer for the lips to reach a closed mouth position from the most extreme mouth position in time interval for the first vowel ($t_1$ and $t_2$). However, this is expected since the variation between lip shapes in the time interval $t_1$ and $t_2$ and those in $t_2$ and $t_3$ is greater than for /ipi/ and /upu/. Thus when going from /a/ to /p/ the lips have to do more work to reach closure. A separation also occurs between /apa/ and /ipi/ in the region of time for the first vowel.

Figure 5.9 (b) shows how the data for the three VCV phrases, varies over time for the second component. It can be seen that there is greater separation /upu/ and the other two contexts in the time intervals for the vowels. As discussed earlier the second component of the VCV lip space describes the variation in lip protrusions associated with the rounding of the mouth. Protrusions are most prominent with the rounded vowel /u/. For this component the spread of data for /apa/ and /ipi/ is almost identical.
Figure 5.9: The mean (continuous lines) and variance for ±3 standard deviations (vertical bars) of the data, over time: (a) is for the 1st PC and (b) for the 2nd PC.
Table 5.1: This table shows measures of separation, between /apa/, /ipi/ and /upu/ when projected into a VCV lip space, Figure 5.7.

<table>
<thead>
<tr>
<th></th>
<th>/a-p-a/</th>
<th>/i-p-i/</th>
<th>/u-p-u/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a-p-a/</td>
<td>0</td>
<td>1.3906</td>
<td>12.4118</td>
</tr>
<tr>
<td>/i-p-i/</td>
<td>1.3906</td>
<td>0</td>
<td>12.6505</td>
</tr>
<tr>
<td>/u-p-u/</td>
<td>12.4118</td>
<td>12.6505</td>
<td>0</td>
</tr>
</tbody>
</table>

5.3.1.1 Linear Discriminant Analysis

To quantitively measure which principal component gives the greatest separation between the three contexts of the phoneme /p/, Fisher’s linear discriminant analysis (LDA) is applied, Chapter 4. Separation has been shown between the three VCV contexts, Figure 5.7 and the separation between these classes can be quantified using LDA.

LDA is applied to the data projected on all principal components. Table 5.1 shows the total measure of discrimination between the three classes of /p/. It can be seen that the greatest discrimination occurs when /upu/ is compared to /ipi/. This is a measure of overall class separability, irrespective of the underlying dynamics.

In order to measure dynamic separation, separability over time between the three classes is found. Figure 5.10 shows this dynamic discrimination for the first two principal components of the VCV lip-space. Figure 5.10(a) shows how the measure of discrimination varies over time between the three contexts when data is projected onto the first principal component. It can be seen in the time interval for the first vowel in the VCV contexts (between $t_1$ and $t_2$) the greatest dynamic separation occurs between /apa/ and /ipi/, as was shown in Figure 5.9. For the time interval associated with the consonant (between $t_2$ and $t_3$) dynamic separation occurs between /ipi/ and /upu/ with the most extreme separation occurring at the time instant (0.21s) like in Figure 5.9. Once again, it can be...
5.3. Dynamic Analysis

seen for VCV /upu/ compared to both /ipi/ and /apa/ the greatest separations occur in the time interval that corresponds to lip movement leading to closure (between $t_2$ and $t_3$).

Figure 5.10(b) shows a similar discrimination when data is projected into the second principal component. However, in this case there is greater separation between the classes in the regions corresponding to the vowels (between $t_1$ and $t_2$ and $t_3$ and $t_4$) in the VCV utterances, particularly for the time interval associated with the first vowel (between $t_1$ and $t_2$).

![Figure 5.10: Variation, in terms of Fisher’s class discrimination measure, over time, when the data is projected onto (a) the 1st PC and (b) the 2nd PC.](image_url)
5.3.2 Comparing \(/p/\) and \(/m/\)

Figure 5.11: All repetitions of \(/apa/, /ipi/, /upu/, /ama/, /imi/, /omo/\) projected onto the first 3 PC’s of the VCV lip space. Each line in this space represents a VCV repetition, each one being on average 10 frames long.

In this section a comparative dynamic analysis between phonemes \(/p/\) and \(/m/\) in three vowel contexts is performed. Figure 5.11 shows all 15 utterance repetitions for all six VCVs projected onto the first three principal components of the VCV lip-space. It can be seen that utterances for \(/apa/\) and \(/ama/\) cluster together; \(/ipi/\) and \(/imi/\) form a cluster and \(/opo/\) and \(/omo/\) form another.

Figure 5.12 shows time-aligned shape data for the six VCVs projected onto the first principal component of the VCV lip space, represented over time. Once again the associated lip-shape for each VCV context was found at three time steps. It can be seen that for the same viseme clusters as in Figure 5.11, in the time interval \(t_1-t_2\), have similar lip shapes. In the time interval \(t_2\) and \(t_3\) both \(/opo/\) and \(/omo/\) exhibit lip shapes that do not quite reach full inner-lip closure. VCVs \(/apa/\) and \(/ipi/\) produce similar lip shapes as do \(/ama/\) and \(/imi/\), which
is unlike the viseme clustering previously found, with phoneme /p/ reaching the most extreme inner-lip closures.

Figure 5.13 shows time-aligned shape data for the two phonemes /p/ and /m/, in their three contexts, projected onto the second principal component of the lip space over time. Once again, it can be seen from the lip shapes that for the consonant time interval \( t_2 \) and \( t_3 \) /apa/ and /ipi/ tend to exhibit similar behaviour and /ama/ and /imi/ a different inner-lip shape. This indicates that in this time interval, inner-lip shape in particular is not context dependent for /p/ and /m/, when in the contexts of vowels /a/ and /i/. The is not the case when the vowel is /m/.

Figure 5.14 shows the ensemble data in Figures 5.12 and 5.13 represented as a mean and variance (per VCV) sampled at equal time steps. This figure highlights the differences in the onset of lip closure (during time interval \( t_2, t_3 \)) between the visemes in the six contexts. Although /upu/ and /umu/ reach the same minima, it takes longer for /umu/ to reach this minima. /ipi/, /imi/ and /apa/ all reach a similar minima but reach this lip-closure at different speeds. Phoneme /m/ in the context of /a/ is the last to reach inner-lip closure.

5.3.2.1 Linear Discriminant Analysis

To quantitively measure the class separation shown in Figure 5.11, Linear Discriminant Analysis (LDA) is performed. Table 5.2 shows the total measure discrimination between the three classes of /p/ and /m/. For phoneme /m/ the greatest separation between it’s three context occurs between /imi/ and /umu/. For both /p/ and /m/ the intra-viseme separability is greatest between contexts /i/ and /u/. When comparing /p/ and /m/, the greatest discrimination also occurs between these two contexts.
Figure 5.12: This figure shows all the data samples over time: (a) is for the 1st PC and (b) for the 2nd PC.
Figure 5.13: This figure shows all the data samples over time: (a) is for the 1st PC and (b) for the 2nd PC.
Chapter 5. Visual Analysis of Lip Coarticulation in VCV Utterances

Figure 5.14: The mean (continuous lines) and variance for ±3 standard deviations (vertical bars) of the data, over time: (a) is for the 1st PC and (b) for the 2nd PC.

Table 5.2: This table shows measures of separation, between /apa/, /ipi/, /upu/, /ama/, /imi/ and /omo/.

<table>
<thead>
<tr>
<th></th>
<th>/a-p-a/</th>
<th>/i-p-i/</th>
<th>/u-p-u/</th>
<th>/a-m-a/</th>
<th>/i-m-i/</th>
<th>/u-m-u/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a-p-a/</td>
<td>0</td>
<td>1.3906</td>
<td>12.4118</td>
<td>0.23768</td>
<td>2.728</td>
<td>12.99</td>
</tr>
<tr>
<td>/i-p-i/</td>
<td>1.3906</td>
<td>0</td>
<td>12.6505</td>
<td>1.8878</td>
<td>0.5026</td>
<td>12.378</td>
</tr>
<tr>
<td>/u-p-u/</td>
<td>12.4118</td>
<td>12.6505</td>
<td>0</td>
<td>10.4201</td>
<td>12.2541</td>
<td>0.3966</td>
</tr>
<tr>
<td>/a-m-a/</td>
<td>0.23768</td>
<td>1.8878</td>
<td>10.4201</td>
<td>0</td>
<td>3.0114</td>
<td>9.1105</td>
</tr>
<tr>
<td>/i-m-i/</td>
<td>2.728</td>
<td>0.5026</td>
<td>12.2541</td>
<td>3.0114</td>
<td>0</td>
<td>11.362</td>
</tr>
<tr>
<td>/u-m-u/</td>
<td>12.99</td>
<td>12.378</td>
<td>0.3966</td>
<td>9.1105</td>
<td>11.362</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 5.15 shows separability between coarticulations of /p/ and /m/ in terms of dynamics for the first two principal components of the VCV lip space. Similar dynamic discrimination curves are found for both principal components. When comparing /apa/ and /ama/ separability, the greatest is found in the consonant time interval (between \( t_2 \) and \( t_3 \)). This is due to the later onset of /ama/ as shown earlier in Figure 5.14. Due to coarticulation this intra-viseme separation occurs.

5.3.3 Conclusions

Based on ensemble statistics a novel approach to analysis and modelling of the temporal dynamics is presented. Two phonemes /p/ and /m/ in three coarticu-
lation contexts were investigated. Analysis has shown that there is inter-viseme variation due to coarticulation but there is also intra-viseme variation due to dynamics although variation is lower. The temporal influence of coarticulation is significant both in lip-shape variation and the timings of lip-shapes during coarticulation. It was shown through LDA that dynamic separation exists most crucially at time intervals associated with lip closure for both /p/ and /m/ in the same contexts. This is significant since these phonemes are often grouped together into the same viseme group. During visual synthesis of the mouth the viseme group is often represented by one static representation. But it has been shown that this in fact does not model the underlying influence of coarticulation on these two phonemes.

5.4 Synthesis

Analysis has shown that temporal influence of coarticulation is significant both in lip shape variation and timing of lip movement. For each VCV, all it’s repetitions were represented as mean polynomials and variances sampled over time, forming a trajectory tube per VCV. These tubes form natural constraints and it is assumed that any VCV trajectory for the captured speaker will never exceed these constraints.

As discussed earlier there are several existing methods of modeling coarticulation in visual speech synthesis. These methods do not address the physical properties of motion. Edge et al. [28] and Lazalde et al. [48] present a constraint-based approach to visual speech synthesis. To account for coarticulation viseme are modeled as distribution around an ideal target. In [28] visemes are regarded as normally distributed static poses. In [33] visual phonemes are represented as a localised Gaussian target regions. Each distribution is regarded as a constraint that must be satisfied by a final speech trajectory. Through constrained optimi-
sation the trajectory goes through each viseme according to pose constraints and context.

In this next section of work several approaches are presented for automatically generating VCV trajectories that fit within the natural constraints (trajectory tubes) found through the analysis of VCV utterances.

### 5.4.1 Gaussian Mean Offset

In the first approach it is assumed that at any time along a trajectory tube a normal distribution can be estimated to model the intra-viseme variance due to coarticulation, as shown in Figure 5.16.

\[ x(t) = \mu(t) + \omega \sigma(t) \]  

\(^{(5.6)}\)

Figure 5.16: This figure shows an example of a trajectory tube where at each time a Gaussian distribution can be estimated.

At any time step along the trajectory tube a new value can be found according to this gaussian weight, as shown in Equation 5.6.
where $\omega$ is $N[0, 1]$ a Gaussian random variable with zero mean and unit variance. $\mu(t)$ is the mean trajectory value at time $t$ and $\sigma^2(t)$ is the variance of the spread of data at time $t$. $\omega$ is a constant for $x(t)$. A new value can be calculated along the trajectory tube (for each $t$) and a polynomial can be fitted, resulting in a new trajectory, as shown in Figure 5.17.

![Figure 5.17](image)

Figure 5.17: This figure shows 30 synthetic trajectories of VCV /apa/, each one weighted according to a different Gaussian random variable.

It can be seen in Figure 5.17 that with a constant weight for the $x(t)$ trajectories are generated that are always parallel to the mean of the tube, which is unlike the trajectories in Figure 5.18.

### 5.4.2 Non-parametric Statistics

The Gaussian mean offset approach presented in Section 5.4.1 makes the assumption that the variation of samples at a given time, can be represented as a
Gaussian distribution. It was observed that the data tended to follow a multi-modal distribution, Figure 5.18. However this cannot be confirmed statistically with the data captured in this work due to the relatively small number of samples (15 for each utterance); a greater number of VCV utterances would need to be captured to statistically verify this.

In order to model this non-Gaussian behaviour a non-parametric offset is applied using a histogram representation. At each time along the trajectory tube the data can be sampled into $N$ bins, Figure 5.19(a). The number of data samples that fall within each bin is found and so the proportion of data points that fit within each bin can be calculated, Figure 5.19(b).

![Figure 5.18](image.png)

Figure 5.18: An example of repetitions of /apa/ in a trajectory tube at time $t$ following a multi-modal distribution.

A random uniformly distributed variable $r$, in the range $[0, 1]$ is generated and the bin that the variable lies within is found with the range $[Y_{\text{min}}(t), Y_{\text{max}}(t)]$. 
Figure 5.19: Figure (a) shows an example of 15 utterances at time step $t$ along a VCV trajectory tube sampled into 6 bins. The histogram generated is then translated into intervals in the range $[0,1]$, Figure (b).

This variable is then scaled relative to the size of the bin it lies within as:

$$ r'(t) = r - Y_{min}(t) $$  \hspace{1cm} (5.7)

From this a weight, $\omega$, relative to the original histogram at time $t$ can be found as:

$$ \omega(t) = Y_{min}(t) + r'(Y_{max}(t) - Y_{min}(t)) $$  \hspace{1cm} (5.8)

Samples of the trajectory $x(t)$ are then generated as:

$$ x(t) = X_{min}(t) + \omega(t)(X_{max}(t) - X_{min}(t)) $$  \hspace{1cm} (5.9)

where $X_{min}(t)$ and $X_{max}(t)$ are the minimum and maximum variance of the trajectory tube at time $t$. So at each time along the trajectory tube a different $\omega(t)$ is found thus creating synthesised samples that can lie at either side of the mean, Figure 5.20. Thus synthesised trajectories are not parallel to the mean of
the trajectory tube, unlike the Gaussian mean offset approach in Section 5.4.1. Examples of synthesised /apa/ trajectories can be seen in Figure 5.21. It can be seen that synthesised trajectories are still restricted by the variances set and tend to follow similar shapes. The temporal variation between trajectories shown in Figure 5.18 are not produced with this method.

Figure 5.20: This figure shows where the new points lie in the trajectory tube.

5.4.3 PCA-based synthesis

For each principal component of a VCV utterance, all $N$ repetitions within it’s trajectory tube can be sampled at $T$ equal time stamps. Resulting in an $N \times T$ temporal observations per principal component. A further PCA can then be applied to this matrix to model the intra-VCV temporal variation, thus each principal component in the original lip space a trajectory can be represented as a linear combination of it’s mean and scaled eigenvectors, Equation 5.10.

$$x(t) = \bar{x} + \sum_j \omega_j \cdot p_j \sigma_j,$$  \hspace{1cm} (5.10)
where \( x(t) \) is a new trajectory \( (t = [1, T]) \) and \( \omega_j \) is a random variable sampled from a zero mean and unit variance Gaussian distribution. So for each principal component vector \( j \) there is a separate \( \omega_j \). Figure 5.18 shows an example fifteen repetitions of the VCV /aa-p-aa/ projected onto the first principal component of the VCV lip space. Figure 5.22 shows an example of these repetitions being projected into their own space, where each point in the space represents a trajectory. A new trajectory is a back-projection of a point in the new space. A set of synthetic sequences for VCV /apa/ can be seen in Figure 5.23.

When comparing Figure 5.23 to Figure 5.18 these synthesised lip-shape trajectories exhibit the same temporal correlation as the sample data. Thus, it is expected that this method will produce more realistic (closer to captured sequences) synthesised lip shapes. Figure X shows examples of lip-shapes over time, comparing an original sequence with a synthetic one.
Figure 5.22: This figure shows 15 repetitions of VCV /apa/ projected onto the first 3 principal components of its temporal space. The red dots represent a single VCV trajectory and the blue cross represents a synthetic point in this temporal space.
Chapter 5. Visual Analysis of Lip Coarticulation in VCV Utterances

5.4.4 Conclusions

Based on the ensemble statistics three methods of synthesis were presented. The first two methods use Gaussian or non-parametric statistics but do not model temporal coarticulation over time. In the third method a temporal PCA is introduced to synthesise variation due to coarticulation. Results are presented for synthesis of lip shape trajectories which exhibit the same temporal coarticulation as the sample data.

5.5 Conclusions and Future Work

The work presented in this chapter investigates the visual variation on the bilabial plosive consonants /p/ and /m/ in three coarticulation contexts. The effect of coarticulation was analysed based on ensemble analysis of repeated utterances of symmetric VCV coarticulation to derive the temporal mean and variance characteristics. Results show that temporal influence of coarticulation is significant
both in lip shape variation and timings of lip movement during coarticulation. Linear discriminant analysis of different VCV utterances shows that dynamic separations exist, in terms of lip shape. It can be concluded that the effect of temporal variation due to coarticulation is statistically significant and should be taken into account in modelling visual speech synthesis.

The ensemble analysis of a small number of viseme coarticulations illustrates the non-linear effects of coarticulation on lip trajectories. This work introduces a novel approach to analysis and modelling of temporal dynamics. To validate this work two steps are required:

- extension to a large set of viseme coarticulations
- analysis of inter-person variation

Due to constraints on data capture this analysis was not possible but is intended for future work to model viseme dynamics.
Chapter 6

Conclusions and Future Work

This thesis presented several visual analysis and modelling techniques for captured visual phones. Several areas of work were presented, from capture and representation through to analysis and synthesis of speech movements and coarticulation. A novel method was reported for the automatic extraction of outer and inner-lip contour edges from sequences of mouth images in speech. An extensive visual analysis of phonemes in continuous speech was performed, that involves the investigation of mouth representations as well as a comparative analysis between static and dynamic representations of visemes. Finally, novel techniques were presented for the analysis and modelling of coarticulation in VCV contexts.

Inner-lip Detection The aim of the work presented in this chapter was to introduce a novel inner-lip detection system which allows for reliable and accurate localisation. The proposed detection technique is a key frame exemplar-based method that is not dependent on any prior frame information for initialisation, allowing improved inner-lip localisation for large frame to frame changes in lip shape, which occur in 25Hz video of visual speech. A hierarchical search method is proposed which identifies the exemplar with the nearest shape and appearance from a set of inner-lip shape
examples. A novel colour space normalisation approach to enhance colour contrast for lip edge detection has been presented. In order to make detection accurate this refinement is performed as a post-process and it is shown that this improved accuracy by 68%.

An evaluation on a 1000 frame sequence was performed. A comparison was performed against the standard AAM method in two states: a naive state and a constrained state. It was shown that AAM tracking fails due to localising to an incorrect minima and so has to be manually re-initialised. This did not occur with the detection system.

Due to constraints on data capture only the full database for one speaker was captured. In the future it aimed that this work will be extended to multiple speakers.

**Visual analysis of phonemes in continuous speech** In this work, thirteen phonemes are represented by four mouth representations, outer-lip and inner-lip shape, texture and the combination of both shape and texture (*appearance*). These thirteen phonemes were then categorised into seven consonant viseme groups. Using PCA these representations were modelled. This reduces the dimensionality and also helped to identify the most significant variations for these representations. LDA was used to quantitatively measure the difference between the seven viseme groups in this reduced space. It was found that appearance data provided the best representation for the database captured of the mouth during speech.

It was shown that, on the whole, significant separation was found between the seven viseme groups, particularly when considering static visemes. However, there is also a large in class variation due to both differences in shape and sampling resulting in significant overlap of the distributions. As a result dynamic information for each viseme cannot be easily distinguished in the context of continuous speech.
A visual analysis of lip coarticulation in VCV utterances. The work presented in this chapter investigates the visual variation on the bilabial plosive consonants /p/ and /m/ in three coarticulation contexts. The effect of coarticulation is analysed based on ensemble analysis of repeated utterances of symmetric VCV coarticulation to derive the temporal mean and variance characteristics. Results show that the temporal influence of coarticulation is significant both in lip shape variation and timings of lip movement during coarticulation. Linear discriminant analysis of different VCV utterances shows that dynamic separations exist, in terms of lip shape. It can be concluded that the effect of temporal variation due to coarticulation is statistically significant and should be taken into account in modelling visual speech synthesis.

The ensemble analysis of a small number of viseme coarticulations illustrates the non-linear effects of coarticulation on lip trajectories. This work introduces a novel approach to analysis and modelling of temporal dynamics. To validate this work two steps are required:

1. extension to a large set of viseme coarticulations
2. analysis of inter-person variation

Due to constraints on data capture this analysis was not possible but is intended for future work to model viseme dynamics.

The research presented in this thesis identifies, through statistical ensemble analysis of visual speech, the importance of coarticulation on lip dynamics. The analysis presented and techniques introduced for modelling and synthesis of lip dynamics during speech provide the foundation for further research towards achieving perceptually realistic animation of a talking head and the understanding of visual dynamics of shape and texture during speech.
Bibliography


[71] Pixar Animation Studios.


