

# **The Application of Manifold based Visual Speech Units for Visual Speech Recognition**

By

**Dahai Yu**

A dissertation submitted in partial fulfillment of the  
requirements for the award of Doctor of Philosophy  
(PhD)

Supervisors: Dr. Alistair Sutherland  
Prof. Paul F. Whelan

Dublin City University

Faculty of Engineering and Computing, School  
of Computing

September 22<sup>nd</sup>, 2008

# Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: \_\_\_\_\_

Student ID: \_\_\_\_\_

Date: \_\_\_\_\_

-- **Dahai Yu**

## **Acknowledgment**

Many helps and encouragements of many people across my study time since I started a new life in Ireland. It is a pleasure to convey my gratitude to them all in my humble acknowledgement.

In the first place I would like to express my sincere appreciation and gratitude to Dr. Alistair Sutherland and Prof. Paul F. Whelan for their supervision, advice and guidance from the very early stage of my research as well as giving me extraordinary experiences through out the work. Their truly scientist intuitions, experiences and passions in science, which exceptionally enrich my growth of research study. They always provide me unflinching encouragement and support in various ways. This work would not be possible without their help.

Also, I gratefully acknowledge Dr. Ovidiu Ghita for his advice, support and crucial contribution. His originality has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come. I am grateful in every possible way and hope to keep up our collaboration in the future.

Many thanks go to my former colleagues Dr. George Awad, Dr. Junwei Han and Tommy Coogan for sharing with me their research experience and all the postgraduate students from Vision Systems Group for their friendship and unreserved support.

I am also grateful to the Faculty of Engineering and Computing in DCU for funding my PhD and supporting my research.

Finally, I would like to thank my wife Hu Bo for her dedication, love and persistent confidence in me. My words cannot express my appreciation to my father Yu Jiang and my mother Zhang Yuzhi, for their love, support, encouragement and endless care about my studies and life.

***Dahai Yu***

## **Abstract**

This dissertation presents a new learning-based representation that is referred to as Visual Speech Unit for visual speech recognition (VSR).

The automated recognition of human speech using only features from the visual domain has become a significant research topic that plays an essential role in the development of many multimedia systems such as audio visual speech recognition (AVSR), mobile phone applications, human-computer interaction (HCI) and sign language recognition. The inclusion of the lip visual information is opportune since it can improve the overall accuracy of audio or hand recognition algorithms especially when such systems are operated in environments characterized by a high level of acoustic noise.

The main contribution of the work presented in this thesis is located in the development of a new learning-based representation that is referred to as Visual Speech Unit for Visual Speech Recognition (VSR). The main components of the developed Visual Speech Recognition system are applied to: (a) segment the mouth region of interest, (b) extract the visual features from the real time input video image and (c) to identify the visual speech units. The major difficulty associated with the VSR systems resides in the identification of the smallest elements contained in the image sequences that represent the lip movements in the visual domain.

The Visual Speech Unit concept as proposed represents an extension of the standard viseme model that is currently applied for VSR. The VSU model augments the standard viseme approach by including in this new representation not only the data associated with

## ***ABSTRACT***

---

the articulation of the visemes but also the transitory information between consecutive visemes. A large section of this thesis has been dedicated to analysis the performance of the new visual speech unit model when compared with that attained for standard (MPEG-4) viseme models. Two experimental results indicate that:

1. The developed VSR system achieved 80-90% correct recognition when the system has been applied to the identification of 60 classes of VSUs, while the recognition rate for the standard set of MPEG-4 visemes was only 62-72%.
2. 15 words are identified when VSU and viseme are employed as the visual speech element. The accuracy rate for word recognition based on VSUs is 7%-12% higher than the accuracy rate based on visemes.

# Content

<b>DECLARATION</b> .....	<b>I</b>
<b>ACKNOWLEDGMENT</b> .....	<b>II</b>
<b>ABSTRACT</b> .....	<b>IV</b>
<b>CONTENT</b> .....	<b>VI</b>
<b>LIST OF FIGURES</b> .....	<b>XI</b>
<b>LIST OF TABLES</b> .....	<b>XV</b>
<b>Chapter 1</b> .....	<b>1</b>
<b>INTRODUCTION</b> .....	<b>1</b>
1.1 INTRODUCTION.....	1
1.2 PROBLEM OUTLINES .....	2
1.3 OVERVIEW OF THE PROPOSED VSR SYSTEM .....	6
1.4 THESIS OVERVIEW .....	9
<b>Chapter 2</b> .....	<b>11</b>
<b>LITERATURE REVIEW</b> .....	<b>11</b>
2.1 INTRODUCTION.....	11
2.2 LIP REGION LOCALIZATION .....	12
2.3 FEATURE EXTRACTION .....	14
2.3.1 <i>Shape-based Feature Extraction</i> .....	14
2.3.2 <i>Intensity-based Feature Extraction</i> .....	17
2.4 CLASSIFICATION .....	18
2.4.1 <i>Visual Speech Classes</i> .....	18
2.4.2 <i>Classifiers</i> .....	20
2.5 SUMMARY .....	23
<b>Chapter 3</b> .....	<b>24</b>
<b>FEATURE EXTRACTION: LIP SEGMENTATION AND MANIFOLD REPRESENTATION</b> .....	<b>24</b>
3.1 INTRODUCTION.....	24
3.2 INTENSITY-BASED LIP SEGMENTATION .....	26
3.2.1 <i>Colour Model</i> .....	27
3.2.2 <i>Proposed Lip-segmentation Algorithm</i> .....	28
3.2.2.1 <i>Colour Models for Face Skin and Lips</i> .....	29
3.2.2.2 <i>Lip Detection Based on Histogram Thresholding</i> .....	31
3.2.2.3 <i>Image Normalization</i> .....	34
3.2.3 <i>Lip Segmentation Results</i> .....	34
3.3 EM-PCA ALGORITHM .....	35
3.4 PROPOSED APPROACH: EM-PCA MANIFOLD REPRESENTATION.....	41
3.4.1 <i>Manifold Calculation from Input Data</i> .....	41
3.4.2 <i>Manifold Interpolation</i> .....	45
3.5 SUMMARY .....	47

<b>Chapter 4 .....</b>	<b>48</b>
<b>VISUAL SPEECH MODELING .....</b>	<b>48</b>
4.1    INTRODUCTION.....	48
4.2    VISEME REVIEW .....	49
4.2.1 <i>Viseme Introduction</i> .....	49
4.2.2 <i>Viseme Representation</i> .....	50
4.2.3 <i>Visemes Limitations</i> .....	56
4.3    VISUAL SPEECH UNIT REPRESENTATION.....	61
4.3.1 <i>Generation of Visual Speech Unit Models</i> .....	63
4.3.2 <i>Registration between VSU Model and Word Manifold</i> .....	72
4.3.2.1 <i>Dynamic Time Warping Review</i> .....	72
4.3.2.2 <i>Registration between VSU and Word Manifold</i> .....	73
4.3    SUMMARY .....	78
<b>Chapter 5 .....</b>	<b>83</b>
<b>EXPERIMENTAL RESULTS .....</b>	<b>83</b>
5.1.    INTRODUCTION.....	83
5.2.    DESCRIPTION OF DATABASE .....	84
5.3.    HIDDEN MARKOV MODELS .....	86
5.4.    HIDDEN MARKOV MODEL CLASSIFICATION.....	87
5.5.    ANALYSIS OF THE EXPERIMENTAL RESULTS .....	89
5.5.1. <i>Experiment 1: Performance Evaluation for Visual Speech Units and Visemes</i> .....	89
5.5.2. <i>Experiment 2: Performance Evaluation for Visual Speech Units with the Variation in the Number of Training Examples</i> .....	95
5.5.3. <i>Experiment 3: Performance Evaluation for Visual Speech Units and Visemes in the Context of Word Recognition</i> .....	97
5.6.    SUMMARY .....	104
<b>Chapter 6 .....</b>	<b>106</b>
<b>CONCLUSIONS AND FUTURE WORK.....</b>	<b>106</b>
6.1.    CONCLUSIONS .....	106
6.1.1 <i>Thesis Summary</i> .....	106
6.1.2 <i>Contributions</i> .....	109
6.2.    FUTURE WORK.....	110
<b>APPENDIX A .....</b>	<b>1</b>
<b>VISEME MODELS .....</b>	<b>112</b>
<b>APPENDIX B.....</b>	<b>114</b>
<b>ORIGINAL IMAGES DATASET .....</b>	<b>114</b>
<b>APPENDIX C .....</b>	<b>116</b>
<b>CONTINUOUS MANIFOLD REPRESENTATION .....</b>	<b>116</b>
<b>APPENDIX D .....</b>	<b>120</b>
<b>VISEME REPRESENTATION .....</b>	<b>120</b>
<b>APPENDIX E.....</b>	<b>126</b>
<b>VISUAL SPEECH UNIT REPRESENTATION 1 .....</b>	<b>126</b>
<b>APPENDIX F .....</b>	<b>129</b>
<b>VISUAL SPEECH UNIT REPRESENTATION 2 .....</b>	<b>129</b>

***CONTENT***

---

**APPENDIX G ..... 135**  
**VISUAL SPEECH UNIT REPRESENTATION 3 ..... 135**  
**PUBLICATIONS RESULTING FROM THIS RESEARCH..... 136**  
**REFERENCES ..... 138**  
**BIBLIOGRAPHY ..... 144**

## **Glossaries**

1. 3D: 3 Dimensional
2. AAM: Active Appearance Model
3. ASMs: Active Shape Models
4. AVSR: Audio Visual Speech Recognition
5. DHT: Discrete Hartley Transform
6. DTW: Dynamic Time Warping
7. EM: Expectation Maximization
8. HMM: Hidden Markov Model
9. ML: Maximum Likelihood
10. PC: Principal Component
11. PCA: Principal Component Analysis
12. RGB: Red/Green/Blue
13. ROI: Region of Interest
14. VSU: Visual Speech Units.
15. VSR: Visual Speech Recognition
16. PSR: Probability Synthesis Rule
17. VA: Viterbi Algorithm

## **Definitions**

1. **Isolated Word:** A sequence of visual speech with a limited number of visemes that start and end with silence.
2. **Sentence:** a sequence of visual speech with infinite number of visemes.
3. **Continuous Visual Speech:** Isolated word and sentence
4. **Word Manifold:** A 3 dimensional EM-PCA vectors that describes the visually spoken words. Each vector presents the particular mouth shape or lip movements.
5. **Viseme:** The representation in the visual domain of the mouth shapes that correspond to one or more phonemes.
6. **Visual Speech Unit:** A new representation is manually constructed from the EM-PCA Vectors that describe visually spoken words and it has three distinct states: (a) articulation of the first viseme, (b) transition to the next viseme, (c) articulation of the second viseme.
7. **MEPG-4:** An international audiovisual object-based video representation standard.

## List of Figures

<b>Fig. 1.1: Overview of the General VSR System Architecture.....</b>	<b>2</b>
<b>Fig. 1.2 An overview of the Visual Speech Recognition system.....</b>	<b>7</b>
<b>Fig. 3.1 Lip segmentation process.....</b>	<b>24</b>
<b>Fig. 3.2 Manifold generation process.....</b>	<b>25</b>
<b>Fig. 3.3 RGB histogram profile for selected skin and lip regions.....</b>	<b>30</b>
<b>Fig. 3.4 RGB, Pseudo-Hue and Hue images. ....</b>	<b>31</b>
<b>Fig. 3.5 Histogram-based selection of the threshold value .....</b>	<b>32</b>
<b>Fig. 3.6 Lip detection process.....</b>	<b>33</b>
<b>Fig. 3.7 Lip-segmentation results.....</b>	<b>34</b>
<b>Fig. 3.8 Sequences of lip segmentation results.....</b>	<b>35</b>
<b>Fig. 3.9 The EM-PCA and Standard PCA when applied to a large dataset (6900 images).....</b>	<b>40</b>
<b>Fig. 3.10 Matrix conversion to one-dimensional vector.....</b>	<b>42</b>
<b>Fig. 3.11 EM-PCA Manifold representation. ....</b>	<b>43</b>
<b>Fig. 3.12 Manifold Examples.....</b>	<b>44</b>
<b>Fig. 3.13 Manifold interpolation.....</b>	<b>46</b>
<b>Fig. 4.1 Mapping table for 6 visemes associated with Standard English consonants [43].....</b>	<b>51</b>
<b>Fig. 4.2 Phoneme to viseme mapping [40].....</b>	<b>51</b>
<b>Fig. 4.3 The representation of the visemes <math>[b]</math>, <math>[a:]</math> and <math>[t]</math> in the EM-PCA manifolds of the word <math>[ba:t]</math> .....</b>	<b>54</b>
<b>Fig. 4.5 The viseme feature space constructed for two different words.....</b>	<b>58</b>
<b>Fig. 4.6 Limitations of the viseme-based approach.....</b>	<b>60</b>

## *LIST OF FIGURES*

---

<b>Fig. 4.7 Examples of Visual Speech Units</b> .....	61
<b>Fig. 4.8 Manifold examples for VSUs containing the viseme [silence]</b> .....	62
<b>Fig. 4.9 Examples of Visual Speech Units</b> .....	64
<b>Fig. 4.10 Examples of Visual Speech Units. The EM-PCA manifolds of VSUs: [silence-b], [b-o], [b-u], [b-i], [b-e]</b> .....	65
<b>Fig. 4.11 VSU Manifold re-sampling process</b> .....	67
<b>Fig. 4.12 The calculation of VSU Mean Models</b> .....	69
<b>Fig. 4.13 The VSU Mean Models and the VSU extracted from different word manifolds</b> .....	71
<b>Fig. 4.14 Registration using Dynamic Time Warping between the mean model manifold of VSU [silence-b] (purple line) and the word manifold [ba:t] (red line)</b> ..	74
<b>Fig. 4.15 Complete registration using Dynamic Time Warping between the VSU mean models and the word manifold, [silence-b] (purple line), [b-a:] (blue line), [a:-t] (green line) and the word manifold [ba:t] (red line)</b> .....	74
<b>Fig. 4.16 Step-by-Step VSU registration and classification</b> .....	76
<b>Fig. 4.17 The complete registration and matching between the VSU mean models contained in the database and the manifold of the word [ba:t]</b> .....	77
<b>Fig. 5.1 HMM topology for VSU and viseme</b> .....	88
<b>Fig. 5.2 Viseme vs. VSU classification for speaker one</b> .....	91
<b>Fig. 5.3. Viseme vs. VSU classification for speaker two</b> .....	92
<b>Fig. 5.4. Correct and incorrect VSU registration</b> .....	94
<b>Fig. 5.5 Visual Speech Unit classification with respect to the number of training examples</b> .....	97
<b>Fig. 5.6. Word-based recognition when the VSUs and visemes are used to model the visual speech</b> .....	99
<b>Fig. 5.7. Word Recognition Process</b> .....	100
<b>Fig. 5.8. The application of the Viterbi algorithm for word recognition</b> .....	102
<b>Figure B.1: Samples of original frames from video sequence 1(Speaker One)</b> .....	114
<b>Figure B.2: Samples of original frames from video sequence 2 (Speaker One)</b> .....	114

## *LIST OF FIGURES*

---

<b>Figure B.4: Samples of original frames from video sequence 4(Speaker Two) .....</b>	<b>115</b>
<b>Figure C.1: Two continuous manifolds of the word [bu:t].....</b>	<b>116</b>
<b>Figure C.2: Two continuous manifolds of the word [ba:bi].....</b>	<b>117</b>
<b>Figure C.3: Two continuous manifolds of word [chu:s].....</b>	<b>117</b>
<b>Figure C.4: Two continuous manifolds of the word [hot] .....</b>	<b>118</b>
<b>Figure C.5: Two continuous manifolds of the word [bi:t].....</b>	<b>118</b>
<b>Figure C.6: Two continuous manifolds of the word [fäst]. .....</b>	<b>119</b>
<b>Figure D.1: Viseme [b], [o] and [t] - word manifold- [bot]. .....</b>	<b>120</b>
<b>Figure D.2: Viseme [b], [u] and [t] - word manifold-[bu:t].....</b>	<b>121</b>
<b>Figure D.3: Viseme [ch], [e] and [k] - manifold-[chek]. .....</b>	<b>121</b>
<b>Figure D.4: Representation of same class of viseme [ch] and [dg] extracted from the word manifolds [cha:dg] (four examples).....</b>	<b>122</b>
<b>Figure D.5: Representation of viseme [b] and viseme [o] extracted from different words manifolds - [bot] and [bobi] (2 examples each word).....</b>	<b>123</b>
<b>Figure D.6: Representation of viseme [b] and viseme [a:] extracted from different word manifolds- [ba:t] and [ha:t] (2 examples each word).....</b>	<b>124</b>
<b>Figure D.7: Representation of viseme [u:] extracted from different word manifolds-[chu:s] and [hu:k] (2 examples each word).....</b>	<b>125</b>
<b>Figure E.1: Re-sampled VSU Manifolds. Five VSUs which all start with viseme [silence] (two samples for each VSU).....</b>	<b>126</b>
<b>Figure E.2: Re-sampled VSU Manifolds. Five VSUs which all start with viseme [b] (two samples for each VSU). .....</b>	<b>127</b>
<b>Figure E.3: Re-sampled VSU Manifolds. Five VSUs which all end with viseme [silence] (two samples for each VSU). .....</b>	<b>128</b>
<b>Figure F.1: The mean model of VSU [b-a:] and the VSUs extracted from different words (two examples each word).....</b>	<b>130</b>
<b>Figure F.2: The mean model of VSU [b-i] and the VSUs extracted from different words (two examples each word).....</b>	<b>131</b>
<b>Figure F.3: The mean model of VSU [b-o] and the VSUs extracted from different words (two examples each word).....</b>	<b>132</b>

## *LIST OF FIGURES*

---

<b>Figure F.4: The mean model of VSU [n-a:] and the VSUs extracted from the word- 'banana' (two examples).....</b>	<b>133</b>
<b>Figure F.5: The mean model of VSU [ch-silence] and the VSUs extracted from different words (two examples each word).....</b>	<b>134</b>
<b>Figure G.1: VSU modeling using three state HMMs.....</b>	<b>135</b>

## List of Tables

<b>Table 5.1: Words Database</b> .....	84
<b>Table 5.2: The set of MPEG-4 visemes</b> .....	85
<b>Table 5.3: 60 classes of Visual Speech Units</b> .....	85
<b>Table 5.4. Word Correct Recognition Rate</b> .....	103
<b>Table A.1 Viseme Model of MPEG-4 standard for English [27, 39-40, 51]</b> .....	112
<b>Table A.2 44 Phoneme to 13 Viseme Mapping using the HTK phone set [1, 76]</b> ...	113
<b>Table A.3 Representation of six major viseme classes [57]</b> .....	113

# Chapter 1

## Introduction

### 1.1 Introduction

Automatic Visual Speech Recognition (VSR) has become a significant research topic that plays an essential role in the development of many multimedia systems such as audio-visual speech recognition (AVSR) [18, 19], mobile phone applications, human-computer interaction [58] and sign language recognition [22]. Visual speech recognition can also be applied in the development of systems for person identification, machine control or game animation.

In general, a VSR system consists of five steps: face localization, lip segmentation, visual feature extraction, visual speech modeling and recognition. The standard system architecture of a VSR system is shown in Fig 1.1. The first task of a VSR system is to locate the face. This is usually carried out based on the analysis of various skin models. Following the localization of the face, the region of interest surrounding the lips is extracted in each image of the video sequence. The third step deals with the calculation of the visual features that are extracted in order to produce a compact representation that describes either the visual appearance or the shape of the lips in each image. The result of the feature extraction is used to generate feasible visual speech models that represent the lip motions during the speech process. The last step of the VSR system performs the visual speech recognition task in order to register and match the visual speech elements present in the input video sequence and those contained in a database.

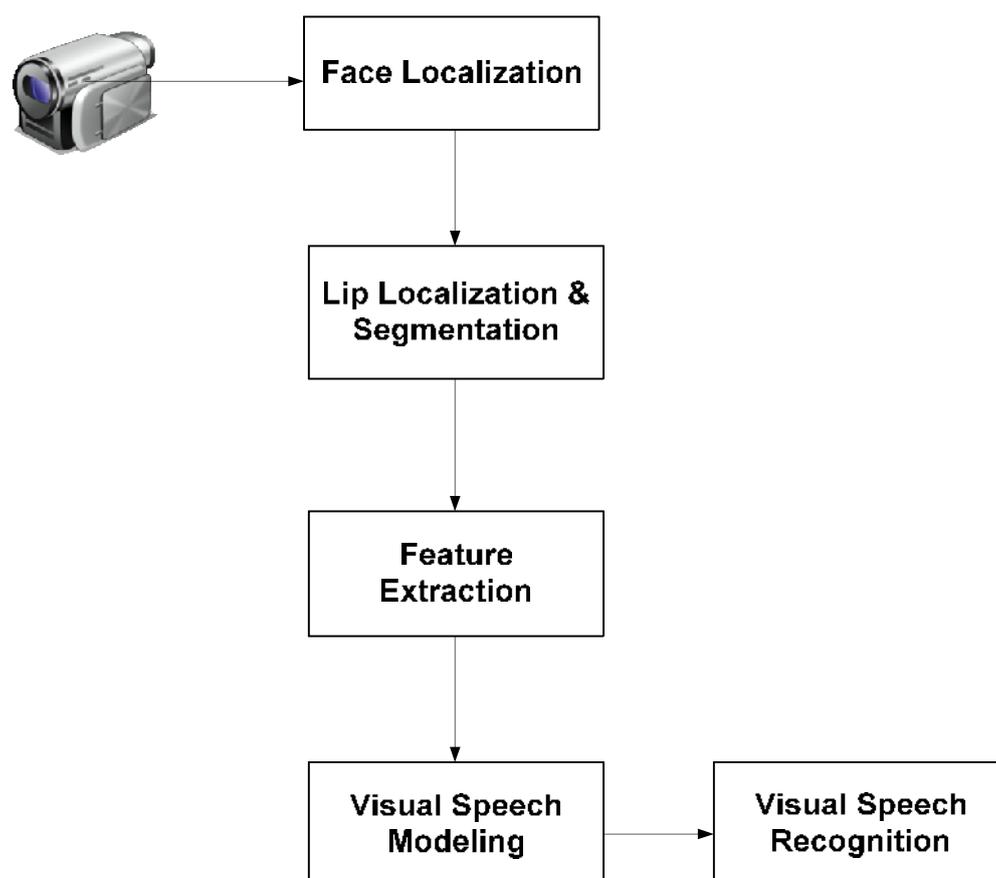


Fig. 1.1: Overview of the General VSR System Architecture.

## 1.2 Problem Outlines

Visual speech perception is inherently a multi-process, whose aim is to provide and interpret the information necessary to establish communication at perceptual level between humans and computers. It is well known that the lip visual information is opportune since it can improve the overall accuracy of audio or hand recognition algorithms especially when such systems are operated in environments characterized by a high level of acoustic noise [1, 23].

In recent years, visual speech information has been exploited to increase the robustness of the conventional Audio-Speech Recognition system [18, 19]. In this regard, several Audio Visual Speech Recognition (AVSR) systems that are able to recognize

complex video-speech patterns from multiple speakers are being reported [20, 21]. While such AVSR systems are useful when operated in noisy environments, it is worth mentioning that they are not suitable to be used in the development of a sign language recognition system [22] since the users of the sign language recognition systems are people with hearing or speech impairment. Due to this reason, there is a need to research and develop visual-only, audio-less recognition systems, generically called visual speech recognition (VSR) systems.

The task of solving visual speech recognition using computers proved to be more complex than initially envisioned. The visual speech recognition has been carried out on discrete or continuous visual domains. In the discrete visual domain the main emphasis was placed on the evaluation of the independent mouth shapes or lip movements (i.e.: viseme). Continuous visual domain deals with the analysis of sequences of visual speech that correspond with multiple context-dependent mouth shapes or lip movements (i.e.: words/sentences). In this thesis, the main focus is placed on the analysis of the discrete speech elements based on “isolated words”. The isolated word in the database contains a limited number of visemes (more than two visemes in general). It is useful to notice that we assume a sentence can be formed as a large number of visemes. Both isolated words and sentence can be referred as “continuous visual speech” which includes two more visemes (mouth shapes) and this study will be addressed in the future work.

Since the first automatic visual speech recognition system was reported by Petajan [23] in 1984, abundant VSR approaches have been reported in the literature over the last two decades. While the systems reported in the literature have been in general concerned with advancing theoretical solutions to various subtasks associated with the development

of VSR systems, this makes their categorization difficult. However the major trends in the development of VSR can be divided into three distinct categories:

1. Feature extraction techniques. The feature extraction techniques applied in the development of VSR systems can be divided into two categories: shape based [3, 7-10, 24-27] and intensity (appearance) based [2, 6, 28-30] approaches. Based on a detailed literature review we can conclude that the intensity-based approaches limited geometrical errors and in general produce better results than shape based feature extraction techniques.
2. Classification algorithms. A number of visual classifiers are proposed to solve the visual recognition task including weighted distance in visual feature space [5], neural network [8, 33], support vector machines [23, 32] and HMM [38-41, 43]. By far though, HMMs have proved to be the most widely used classifier in the development of VSR systems.
3. Recognition tasks. In this process, common recognition tasks include the recognition of visemes [1, 42, 55, 57], isolated words [7], connected digits [9, 35] and sentences [36], mostly in English, but also in French, Chinese and other languages. The literature on VSR indicates that most systems were focused on the robust identification of small independent speech elements (visemes) while the word recognition task has been viewed as a simple combination between standard visemes.

Based on the aforementioned categorization, we can notice that numerous methods have been proposed to address the problem of feature extraction and visual speech classification, but very limited research has been devoted to the identification of the most

discriminative visual speech elements that are able to model the speech process in the continuous visual domain. As mentioned earlier, most works on VSR focused on the identification of visemes, but in practice the viseme identification proved problematic since visemes have a limited visual support when analysed for continuous lip motions and as a result different visemes may overlap in the feature space, a fact that makes their identification difficult.

To address the problems associated with the standard viseme recognition approach, this thesis will provide a theoretical evaluation and quantitative answers to the following issues:

- How to extract the information associated with the lips motions from the frames that define the input video sequence?
- What is the appropriate set of visual speech element that can be applied for VSR by including not only the data associated with the visemes but also the transitional information between consecutive visemes?
- What criteria can be applied to register the new visual speech element into the continuous visual speech sequence and how to apply them to word recognition?

In order to answer these questions, a new set of visual speech elements for VSR, referred to as Visual Speech Units (VSU), is proposed in this thesis. Other contributions of this work include the development and evaluation of several techniques such as Pseudo-Hue based lip segmentation, lip-feature extraction based on EM-PCA manifold representation and HMM based classification. The main contribution of this dissertation is located in the theoretical studies that lead to the development of a new set of speech elements (VSUs) for VSR. Another important task is to evaluate the performance of the

VSU representation when applied to the recognition of isolated words. Based on experimentation, it is demonstrated that the inclusion of the new set of speech elements improves the overall performance of the VSR system when compared with the performance offered by the analysis of the standard set of visemes.

### **1.3 Overview of the proposed VSR System**

In order to achieve robust visual speech recognition, the process of visual speech recognition is formulated as shown in Fig. 1.2. The new system presented in this thesis consists of four major components: lip segmentation, feature extraction, Visual Speech Units modeling and Visual Speech Units registration and Classification.

- **Intensity-based Lip Segmentation**

For any given image from the input video sequence, a generic skin colour model is applied to extract the initial facial skin areas. In order to extract the lips from skin regions, the pseudo-hue is calculated based on RGB component values and the lips are segmented by applying a histogram-based thresholding scheme. The image area describing the lips is extracted in each frame from the input video sequence.

- **Manifold Representation**

A representation using 3-dimensional (3D) PCA vectors that describe the visually spoken words is proposed. These PCA vectors are referred to “word manifold”. In this regard, the image data contained in the region of interest (ROI) surrounding the lips is extracted from the previous step and it is converted into a matrix form. The converted data is compressed using Expectation Maximization PCA (EM-PCA) into a 3-dimensional feature space, where each image area describing the lips in the input sequence is projected onto the low-dimensional EM-PCA space.

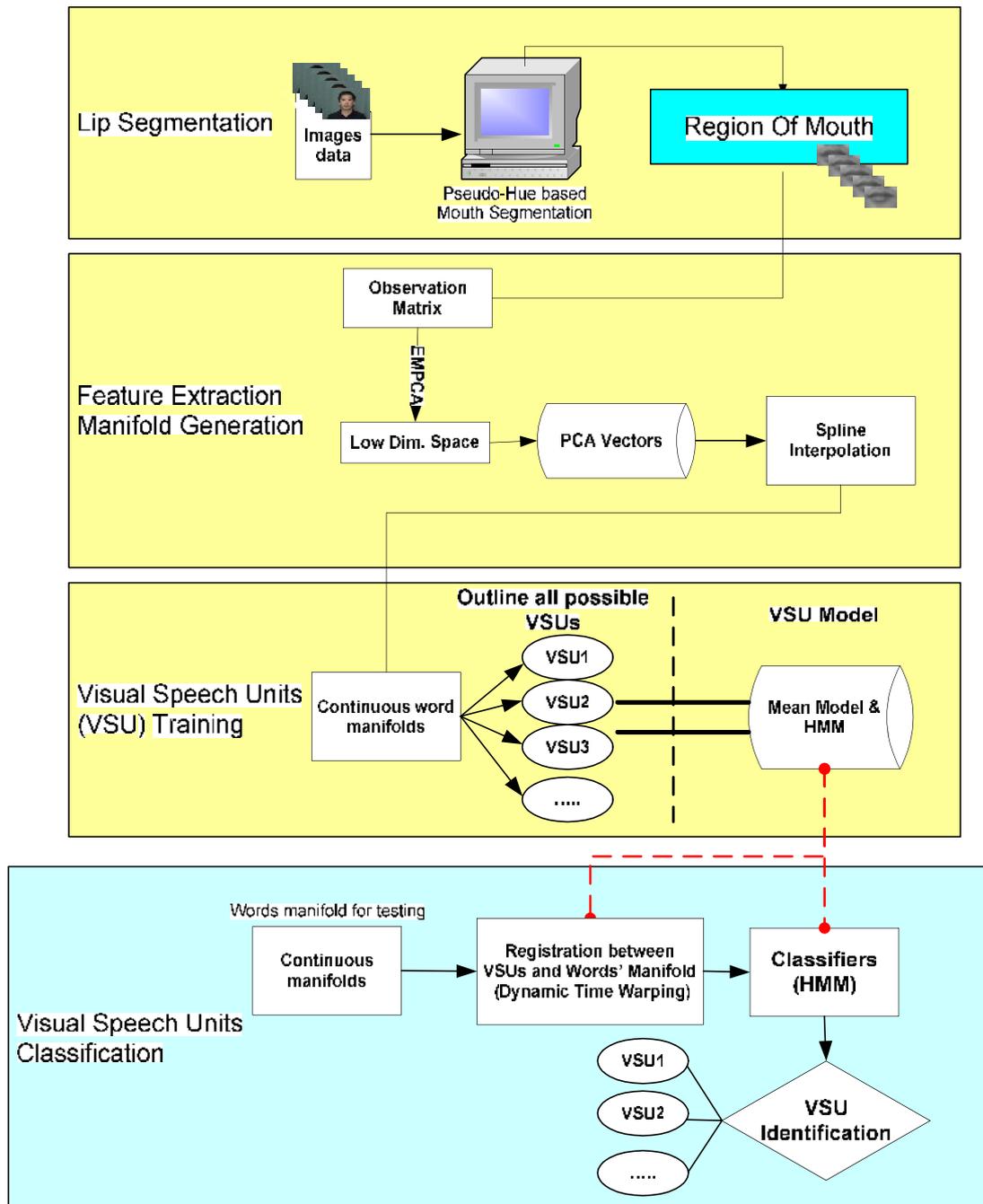


Fig. 1.2 An overview of the Visual Speech Recognition system.

The aim of this procedure is to obtain an EM-PCA “trajectory” where for each mouth shape a low dimensional vector is assigned. The projections of these images form a trajectory that truly extends to different mouth shapes and lip movements. To obtain a

continuous representation, the manifold is interpolated using cubic splines and re-sampled based on equal distances on the interpolated manifold surface. The re-sampled manifolds are used for Visual Speech Units modeling.

- **Visual Speech Units Modeling**

This is an off-line component of the system that is applied to generate a database of VSUs. The proposed VSU extends the standard viseme model by including in the new representation the transition between consecutive visemes. In the training process, the VSU are constructed from the training data and for each class of VSU a mean model is generated based on the EM-PCA representation.

- **Visual Speech Units Registration and Classification**

In the final phase of the VSR system, the registration process between the VSU mean models and the continuous manifold calculated from the input video sequence is carried out using Dynamic Time Warping (DTW). In this way, the VSU recognition process is viewed as a two-step approach. In the first step, the VSU mean models are registered to the continuous manifold calculated from the input video sequence manifold using DTW. Then in the second step, the matching cost between the VSU mean models and the registered sections of the continuous manifold are measured using HMM classification. This process is applied in an iterative manner until the entire surface of the continuous manifold is covered by an ordered sequence of VSUs.

In conclusion, the main goal of this thesis is to advance theoretical and practical solutions in the field of feature extraction, visual speech modeling and visual speech recognition based on a flexible framework that analyses the lip movements in the visual domain.

## **1.4 Thesis Overview**

Chapter 2 is a literature review of the related techniques proposed by different research groups to solve the problem of VSR.

Chapter 3 explains the proposed lip segmentation algorithm with various evaluation results. The EM-PCA manifold representation for visual speech feature extractions is also discussed.

Chapter 4 gives a particular analysis of the viseme model and introduces our proposed VSU model.

Chapter 5 details a large number of experimental results where the performance of the new VSU model is compared against that offered by the standard set of MPEG-4 visemes.

Chapter 6 concludes with a summary and advances some future work directions

# Chapter 2

## Literature Review

### 2.1 Introduction

Recognition of human speech by computers using only visual information is a significant research area that spans across multiple disciplines such as linguistics and speech modeling [27]. In the past decades, a great deal of research effort has been devoted to the development of robust visual speech recognition (VSR) systems that are able to localize the region of interest (ROI) around the lips, extract visual information from lip movements and emulate human cognitive ability in recognizing speech based on the dynamic deformation of the lips outlines. The aim of a VSR system is to provide valuable aid to the acoustical [3] or gesture recognition [22] under degraded conditions.

There has been much progress in automatic VSR over the past decades and various visual speech recognition techniques are reported. However, it is useful to note that most of the research in automatic VSR has been concentrated around two major topics: feature extraction and visual speech classification. In this regard, the feature extraction process requires robust lip segmentation and extraction of suitable features that are able to encode the lip movements in a low dimensional representation. Thus, the visual speech features provide a rich source of information that can be used in the development of computer vision systems able to understand human actions and behaviors. For example, the speech visual features have been currently applied to solve a large range of practical problems

such as face or facial expression recognition [59], control the car environment [58] and image animation and coding [60, 62].

Once features become available, the next step involves visual speech classification that is applied to identify the visual elements in the input video sequence. The literature on VSR indicates that visual speech classification has focused on two major issues: visual speech classes and the design of visual speech classifiers. The visual speech class is the choice of the speech model that is assumed to generate the observed features. This class is organized as the basic unit of visual speech that can be concatenated to form words and sentences, thus providing the flexibility for the proposed visual speech recognition systems to be extended to cover large vocabulary representations [27]. The visual speech classifier is the statistical classification approach of the automatic VSR process. The classifier is applied to model and classify the speech classes.

In the next section, the most relevant approaches in the area of lip segmentation, feature extraction and classification will be analysed.

## **2.2 Lip Region Localization**

Lip segmentation has become an important issue in both automatic VSR processing and automatic face recognition. In such systems, the region of interest (ROI) around lips must be detected in each frame of the image sequence. This procedure is normally carried out by fitting a range of colour models to the image and this is followed by face detection and extraction of the ROI surrounding the lips.

Early VSR systems performed the lip segmentation in conjunction with the application of artificial markers (lipstick) on the lips [11]. The application of lipstick enables the system to detect precisely the lips in the image data, but this procedure is

inappropriate since it is uncomfortable for users and such VSR systems can be operated only in constrained environments. Thus, the main research efforts have been concentrated in the development of vision-based lip segmentation algorithms. Many studies have shown that colour information can be successfully applied to identify the skin or face in digital images [2]. The main idea behind this approach is to transform the RGB signal into a new representation where the mouth is clearly visible, so that it can be easily segmented. To this end, a large number of colour representations have been proposed. In 1996, Coiaiz et al [5] used the hue component of the HSV representation to highlight the red colors which are assumed to be associated with the lips in the image. Later, the HSV colour space is further used by Zhang and Measereau [4] for lip detection. They used prominent peaks in the hue signal as an indicator to locate the position of the lips. Then based on the identified lip area, the interior and exterior lip boundaries are extracted using both colour and spatial edge information using a Markov Random Field (MRF) framework. Other approaches carried out the lips detection task in the YCrCb colour space since that facial skin covers a small area of the CrCb subspace [12, 13].

In 2001, Eveno et al [2] propose a new colour mixture and chromatic transformation for lip segmentation. In their approach, a new transformation of the RGB colour space and a chromatic map was applied to increase the discrimination between the lips and facial skin. They demonstrated that the proposed approach is able to achieve robust lip detection under non-uniform lighting conditions. Later, Eveno et al [3] introduced a different method where the pseudo-hue [6] was applied for accurate lip segmentation that has been embedded in an active contour framework. They applied the proposed algorithm

for visual speech recognition and the results show significant improvement in terms of accuracy in lip modeling.

Another method for mouth segmentation has been proposed by Liew [14] in 2003. In their approach, the colour image is transformed into the CIE-Lab and CIE-Luv colour spaces, and then a lip membership map is computed using the spatial fuzzy clustering algorithm. After morphological filtering, the ROI around the mouth can be identified from the face area.

In 2006, Guan [17] improved the contrast between lip and the other face regions using the Discrete Hartley Transform (DHT). In this paper, lips are extracted by applying wavelet multi-scale edge detection across the  $C_3$  component of the DHT which takes both the colour information and the geometric characteristic into account.

## **2.3 Feature Extraction**

As indicated in the first chapter of this dissertation the feature extraction techniques developed for VSR can be categorized into two major groups, namely shape-based and intensity-based feature extraction approaches.

### **2.3.1 Shape-based Feature Extraction**

The shape-based approaches rely on the extraction of geometrical features from the outline of the lips. This information is used to encode a standard set of mouth shapes that are applied to model the lip motions during the speech process.

This approach was applied by Petajan [23, 24] in the development of a lip-reading system where simple shape features such as height, width and mouth area are used to encode the shape of the region described by the lips contour. In 1994, Hennecke et al [15] used a deformable template to model lips dynamics. This template is generated based on

a model of the lips defined by a set of parameters which are chosen by minimizing a criterion based on the distance between the edges of the model and the edges of the lips. The proposed approach shows good results in tracking the height and widths of lips, but it has some problem on the lower edges of lips under various lighting conditions.

Using a different approach, Silveira et al [7] employed the horizontal and vertical features extracted from the mouth shape. In this study, the difference between the two consecutive frames of the sequence under analysis is calculated and an entropy-based threshold is computed to detect the mouth region. One horizontal distance and three vertical distances calculated from the lip data are extracted and used for visual speech recognition on a subset of words.

It is important to note that the approaches detailed above use a limited number of geometric features and their performance proved to be inappropriate when applied to image data affected by non-constant illumination conditions. To circumvent this problem, other approaches apply Active Shape Models (ASM), Active Appearance Model (AAM), or snakes to extract the lip outlines [3, 25]. But the application of these techniques to VSR proved to be problematic since they require a complex initialization procedure. For instance in 1995, Luetin et al [8] developed an ASM method that was able to learn the grey-level profile around the lip contours. They applied additional constraints to ensure that the detected boundary belongs to the possible lip shapes only, but to achieve this they used a large training set that is able to cover a high variability range of lip shapes. Moreover, the images contained in the training set have to be cautiously calibrated. The initial mouth shapes associated with different articulation conditions have to be constant. Otherwise the ASM method leads to unreliable results [3].

This approach was further advanced by Li and Ai [9] when they applied the ASM approach for mouth contour extraction in conjunction with the Ada-Boost classification scheme to characterize the local texture. ASM approaches identify the lips in the image by fitting a statistical shape model of the lips to the video frames. Such model-based approaches are less sensitive to image noise as they only use the lip contour information, but they are not very useful in describing the continuous speech process [1, 27, 77].

Other implementations use the Active Appearance Model (AAM) approach [26] as to extract the lip shapes where the shape model is combined with a statistical model in the intensity domain [16]. The AAM is a generalization of the widely used ASM approach since it uses all the information in the image region covered by the target object, rather than just that near modeled edges [78]. Although the performance of AAM is demonstrated to outperform ASM in lip tracking [26-27], it still has two disadvantages when applied to motion tracking. First, the estimated out-of-plan motions are not very well accommodated since AAMs encode the lip shapes using a 2D representation. Second the convergence of the optimization process to desired minima is not guaranteed [79]. However, both AAM and ASM techniques are sensitive to tracking and modeling error [27].

More recently, Tian et al [10] combined shape, colour and motion for lip tracking. They developed a method for tracking lip contours in colour images by applying a multi-state model that is able to represent different mouth shapes such as open, relatively open and tightly closed across individuals. The lip state transitions were determined by the lip shapes and colour. Given the initial location of the lip template in the first frame, the algorithm tracks the lip key points using the Lucas-Kanade method where the lip

contours are detected by enforcing the corresponding lip template parameters. This method proved to be able to track lips even in the presence of vertical and horizontal head rotations. The main limitation of this method is based on limited lip templates (open, relatively closed and tightly closed). For non-symmetrical facial expressions and complex lip shapes which are not included in the training set, errors between the tracking lip contour and actual lip shapes are encountered.

### **2.3.2 Intensity-based Feature Extraction**

One limitation associated with the shape-based approaches (e.g.: ASM, AAM) resides in the fact that only geometrical information is used to encode the mouth shapes. Such shape based approaches only analyze the lip contour information and they do not encode the speech articulation [27]. For example, lip contours cannot describe the information related to the oral cavity and the protrusion of the lips. In addition these approaches are sensitive to tracking errors and they are not able to encompass the information contained in consecutive frames efficiently. Their performance is depended on the initial conditions and they are not able to directly handle cases well outside the training templates.

To address these issues, intensity-based approaches [2, 6, 27, 28] have been proposed. Their major advantage is that they use the entire grayscale (or colour) information available to sample the spectrum of mouth shapes. Intensity-based features are capable of encompassing the visual information within the mouth cavity and the surrounding face regions that are not included in the high-level shape-based features [27]. The intensity-based features are demonstrated to produce better results than features extracted using ASM and AAM algorithms in [21]. In this regard the image area around the lips is

extracted for each frame in the video sequence and this information can be compressed to obtain a low-dimensional representation using PCA [29, 31], DCT [28, 29], and LDA [30]. The representation of the mouth shapes in a low-dimensional feature space proved to be opportune and the performance of these methods in general is better than that attained by the shape-based VSR techniques [21]. Moreover, intensity-based approaches do not require *a priori* statistical lips models and this fact allows the development of computationally efficient VSR systems [27].

## **2.4 Classification**

### **2.4.1 Visual Speech Classes**

The literature review on VSR systems indicates that researchers have attempted speech recognition for individual words (digits, letters, etc) or sentence level [7, 29, 35-37]. The main disadvantage of these approaches resides in the fact that an extensive database is necessary to model all words contained within the English dictionary. In recent years, the main investigations have focused on the robust identification of visemes. The basic unit that describes how speech conveys linguistic information is the phoneme [1]. In visual speech, the smallest distinguishable unit in the visual domain is called viseme [1, 42, 55, 57]. A viseme can be viewed as a cluster of phonemes and a model for English phoneme-to-viseme mapping has been proposed by Pandzic and Forchheimer [54] (see Appendix A). In this regard, static and dynamic visemes were both used for visual speech synthesis. A static viseme can be conceptualized as a still human face picture with the visual configuration represented by the mouth shape, tongue and jaw that is visually correlated with the speech sound corresponding to a phoneme [83]. Dynamic visemes represent the process of the visual lip movements during the speech articulation. These

dynamic visual speech elements can be produced by independent phonemes, or constructed from continuous visual speech such as words or sentences [39-40]. More recently researchers have approached visual speech recognition using the dynamic visemes concept.

Goldschen et al [84] proposed a continuous optical automatic speech recognizer (OASR) that uses 13 dynamic features for optical information from the oral-cavity shadow of a speaker. In this system, 150 sentences are tested using Hidden Markov Models based on visemes, trisemes and generalized trisemes. In 1999, V. Matousek [61] developed one of the first viseme-based classification systems where a time-delayed neural network is applied to classify 14 classes of visemes. This work has been further advanced by Foo et al [38-40, 55], where adaptive boosting and HMM classifiers were applied to recognize visual speech visemes. Yau et al [56] initially examined the recognition of 3 classes of viseme using motion history image (MHI) segmentation and later they increased the number of visemes up to 9 classes. In this system, 2D spatio-temporal templates (STT) combined with the discrete stationary wavelet transform and Zernike moments were used to describe the lip movements in the temporal domain and HMM were used for classification [41].

In the literature on VSR, a viseme is regarded as the smallest unit that can be identified using the visual information from the input video data. Word recognition (or other continuous speech recognition) is viewed as a simple combination of standard visemes. Although words can be theoretically formed by a combination of standard visemes, in practice viseme identification within words is problematic since different visemes may overlap in the feature space a fact that makes their identification difficult.

### **2.4.2 Classifiers**

A large number of classifiers have been proposed for automatic VSR. One of the most simplistic classifier evaluates the Euclidean distance between the pre-stored visual features and those extracted from the input video sequence [23, 24]. The main advantage of this approach resides in its simplicity but it proves to be inaccurate when applied to discriminate a large number of mouth shapes.

In 2002, Gordan et al [32] introduced Support Vector Machine (SVM) to recognize temporal sequences of visemes. They trained one SVM for each viseme in the database and in their approach they used SVMs with 3<sup>rd</sup> degree polynomial kernels. They reported a recognition rate of 90% when they applied their system to recognize a small set of visemes.

Artificial neural networks (ANN) have also been used for visual speech classification. In this regard, Yau [27] proposed an ANN based learning algorithm to classify the moment-based features that were used to describe a small number of visemes. In this approach, an ANN is trained for each viseme class contained in the database. The experimental results reported in the paper show that they achieved 84% recognition rate when applied to the recognition of 9 classes of visemes.

In 2006, Ravyse et al [83] introduced a multi-stream Dynamic Bayesian Network (DBN) model to analyze either audio and video streams for AV automatic speech recognition. They applied the proposed DBN based system and the classical Hidden Markov Model in order to recognize 50 independent sentences. The experiments indicated that the DBN model is more robust than HMM when applied to noisy data.

However, Hidden Markov Models [33, 73] are the most widely used classification scheme for VSR. In 1998, Potamianos et al [35] applied an HMM classifier whose

parameters were optimized by maximum likelihood Viterbi training for automatic lip-reading. In their work, both lip contour and image transform-based visual feature are considered for HMM training. The performance showed significant improvement compared to standard techniques that analysis the lip motions only in the intensity domain.

Yu and Bunke [36] combined HMMs with grammar to recognize visual speech sentences of email commands and words describing integers. In this paper, a set of basic words is used to generate pre-defined sentences based on some grammar knowledge. For each basic word, a HMM is constructed. After training the HMMs for each individual word in the database, a complex HMM is obtained by concatenating the individual HMMs according to the grammar. The complex HMM is applied to recognize any sentence generated in agreement with the pre-defined grammar. This VSR system achieved 80% correct words recognition but it attained only 54% successful recognition when the system was applied for sentence recognition.

In the same year, Chan [37] developed an HMM based audio-visual speech recognition system that combines geometric and appearance based visual features. Initially, geometric features such as the height and width of the lips are extracted using a contour-based lip tracking algorithm. Then, the pixel-based features that are robust to variation in scale and translation are extracted. To achieve this goal, a subset of pixels located in the center of the inner mouth was selected. This cluster of pixels was found effective in capturing sufficient details of the appearance of the teeth and tongue to be used in the discrimination of the spoken words. In the final stage, an HMM is applied to recognize word-models in the input video sequence. The experimental data indicates that

this approach is able to produce sufficiently accurate results up to 90% in 9 isolated digits recognition of a single speaker.

In recent years, coupled HMMs, factorial HMMs and Ada-boosted HMMs have been explored. Foo and Dong [38] applied a boosted multi-HMM classifier to recognize visual speech elements. The main novelty of this approach is the Baum-Welch training algorithm that is used to classify the visemes in English. Later, they improved their initial approach by combining adaptive boosting and HMMs to build AdaBoosting-HMM classifiers [39]. This classifier is trained to cover different groups of visemes. In 2005, they further improved the HMM using a novel two-channel training strategy [40]. In this classification strategy, a separable-distance function that measures the difference between a pair of training samples is adopted. The symbol emission matrix of an HMM is split into two channels: a static channel to maintain the validity of the HMM and a dynamic channel that is modified to maximize the separable distance. This approach achieved an 80% recognition rate.

In 2007, Yau et al [41] propose the use of image moments and multi-resolution wavelet images for visual speech recognition. In their approach, the input video data is represented by a general spatio-temporal template that is decomposed by applying the discrete stationary wavelet transform and HMMs are used for viseme modeling. The preliminary results show that this system achieved about 88% correct recognition when applied to recognize 14 classes of visemes.

By far though, the most widely used classifiers are traditional HMMs that statistically model transitions between the visual speech classes and assume a class-dependent generative model for the observed features.

## **2.5 Summary**

In this chapter, a large number of VSR systems have been reviewed with the main focus being on the lip segmentation, feature extraction and classification. Based on this review, I conclude that lip segmentation methods based on skin models are the most promising approaches. Among all feature extraction and classification algorithms, intensity-based feature extraction techniques used in conjunction with HMM are the best approaches to model and analyze temporal processes for VSR.

I also noticed that visemes are widely used as the basic speech element by many research groups, but they have the main shortcoming that visemes cover only a small subspace of the mouth motions represented in the visual domain. In addition to this, the viseme model cannot represent transitions between visemes in continuous speech (words) recognition system. To address this problem, in this thesis a new VSR model called Visual Speech Unit (VSU) is proposed. An application based on this VSU model is developed to recognize group of words and the experimental results demonstrate the validity of the adopted approach.

In the next chapter, all stages of the adopted lip segmentation methods are described. In the next chapter also a new Expectation-Maximization Principal-Component-Analysis (EMPCA) manifold representation that is applied to encode the mouth shapes is detailed.

# Chapter 3

## Feature Extraction: Lip Segmentation and Manifold Representation

### 3.1 Introduction

The task of feature extraction entails two steps namely lip segmentation and manifold representation. Lip segmentation requires several computational procedures that are applied to enhance the presence of the facial skin in the image, to find the color difference between the face skin and lips in the image and finally to identify the region of interest (ROI) around the lips. Fig. 3.1 outlines the developed lip-extraction algorithm. In this process, the lip-segmentation procedure is applied individually to captured images contained in visual speech videos. To enhance the presence of skin in the image, the pseudo-hue component is calculated based on the RGB values of tracking images and the region around the lips is extracted by applying a histogram-thresholding scheme. This algorithm is used by human annotation for mouth alignment. The images resulting from lip segmentation are used as input data for manifold representation.

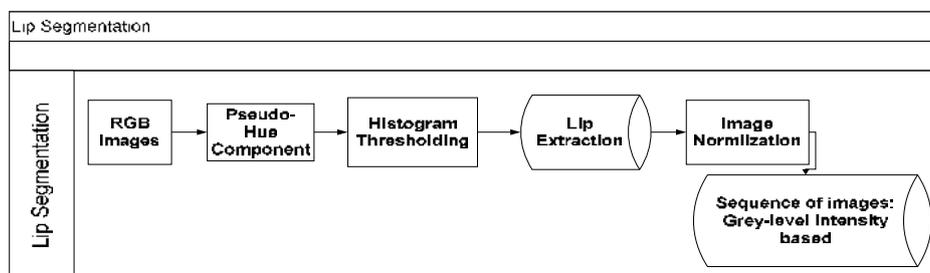


Fig. 3.1 Lip segmentation process



dimensional EM-PCA vector referred as “word manifold” that is subjected to an interpolation procedure that is applied to obtain a continuous representation.

## **3.2 Intensity-based Lip Extraction**

Lips are highly deformable objects where their shape varies significantly during the speech process. Lips also vary in shape, colour, reflection and their relation to surrounding features such as tongue and teeth [3]. As indicated before, the first step of the lip extraction algorithm involves pre-processing the input data to enhance the presence of the facial skin in the image and find the colour difference between the face skin and lips. Such techniques are often referred in the literature to as “skin detection”.

Skin detection has received a lot of research interest in recent years where the main aim of the developed systems is the identification of human skin regions in a colour image. The skin detection algorithms have been applied to face detection [44], [45], visual speech analysis [7] and lip tracking [10]. In particular, skin detection algorithms play an important role in the development of face detection techniques since the search space for feature of interest such as eyes and mouth can be greatly reduced through the detection of skin regions.

The main challenge that has to be addressed by skin detection algorithms is to accommodate the large variations that may occur in the skin appearance [7]. In general, the skin detection is achieved using either pixel-based classification methods or region-based methods. In pixel-based classification, the algorithms divide the image content into two disjoint classes, namely the skin and non-skin pixels, while region-based methods evaluate the spatial differences between the preceding frames and current frame by

evaluating the motion in consecutive frames. Since region-based techniques are sensitive to background motions, this thesis will focus on the analysis of the pixel-based methods.

In the last two decades, a large number of techniques have been proposed for skin detection that analyse the pixel distribution in colour images [44, 46]. In this regard, five colour spaces and non-parametric skin-modelling methods (lookup table and Bayes skin probability map) have been evaluated in [47]. In [48] two popular parametric skin models have been compared in chrominance-separated colour spaces and a new skin-detection algorithm has been proposed. Building on this, in [3] and [2] a pseudo-hue colour model has been successfully applied for lip-detection and this approach will be followed in this thesis since it offers an elegant and accurate skin-detection framework.

### **3.2.1 Colour Model**

Colour provides strong visual cues and plays important roles in various aspects of biological vision [B2]. Historically, computer vision techniques have been applied to monochromatic data where changes in the intensity map are used to identify the objects present in the image [B1]. Many investigations indicate that the difference between human skins is better captured by the chrominance components than the luminance [49, 80-81]. For example, human lips are defined by a darker colour than the colour of the surrounding skin. Thus, the choice of colour models can be considered as the primary step in lip segmentation.

A colour model is an abstract mathematical formulation that describes the way colours can be represented as tuples of numbers, typically as three or four colour components (e.g. the RGB and CMYK colour models). RGB is the default colour model for most available image formats. It has three primary colours red(R), green (G) and blue

(B). A typical camera always provides images of tri-chromatic pixels with RGB components.

Any other colour models can be obtained from a linear or non-linear transformation from RGB. In this regard, Hue based colour models for skin detection represent accurate highlight between lips and skin. It is described as follows.

- ❖ **HSI & HSV:** Colours are described by the chrominance (Hue) - the property of a colour that varies in passing from red to green, followed by the strength of colour (Saturation) - the property of a colour that varies in passing from red to pink and the brightness (Intensity) - also called lightness or value, the property that varies in passing from black to white. Hue corresponds to intuitive notion of “colour” while saturation is the vividness or purity of colour. HSI attempts to produce a more intuitive representation of colour than the RGB colour space but it cannot be described directly by RGB. While the transformation from RGB to HSV is invariant at white lights, ambient light and surface orientations relative to the light source and hence, the HSV color space may form an optimal representation for skin detection methods [80]. Other similar colour models are HSL and TSL.

### **3.2.2 Proposed Lip-segmentation Algorithm**

The aim of the developed algorithm for lip segmentation is to increase the discrimination between lips and facial skin. Then using this primary information, we attempt to identify the mouth position by employing a histogram-thresholding scheme

that separates the lips from the facial skin. The algorithm that has been developed consists of three main steps:

- ❖ Colour models for face skin and lips.
- ❖ Histogram-based thresholding for lip-detection.
- ❖ Image normalization.

### **3.2.2.1 Colour Models for Face Skin and Lips**

Many studies have indicated that colour plays a key role in the development of skin detection algorithms. This observation is motivated by the fact that the skin is better characterized by the chromatic components than by the brightness component [2]. Our experiments have also indicated that the skin and lip pixels can be separated in the RGB space. This can be observed in Fig. 3.3 where the histograms calculated for selected skin and lips regions are illustrated.

In Fig. 3.3 it can be observed that the skin and lip pixels have quite different components in the RGB space. For both regions the red colour is dominant. Based on the colour distributions shown in Fig. 3.3 it can be concluded that the skin colour is more yellow than the colour of the lips because the difference between red and green is greater for lips than for skin and as a result the pseudo-hue [6, 49] component is best suited to sample this difference.

The pseudo-hue component is demonstrated better results than lip segmentation using classic Hue component in skin detection and lip segmentation [3]. The pseudo-hue component presents more accurate distinguishable between lips and skin, that is able to speed up the segmentation when thresholding is applied.

The pseudo-hue is computed as follows:

$$H(x, y) = \frac{R(x, y)}{G(x, y) + R(x, y)} \quad (3.1)$$

Where  $R(x, y)$  and  $G(x, y)$  are the red and green components of the pixel with coordinates  $(x, y)$ , and  $H(x, y)$  is the pseudo-hue value. As can be observed in Fig 3.4, the lip areas can be better observed in the pseudo-hue image than in the hue image.

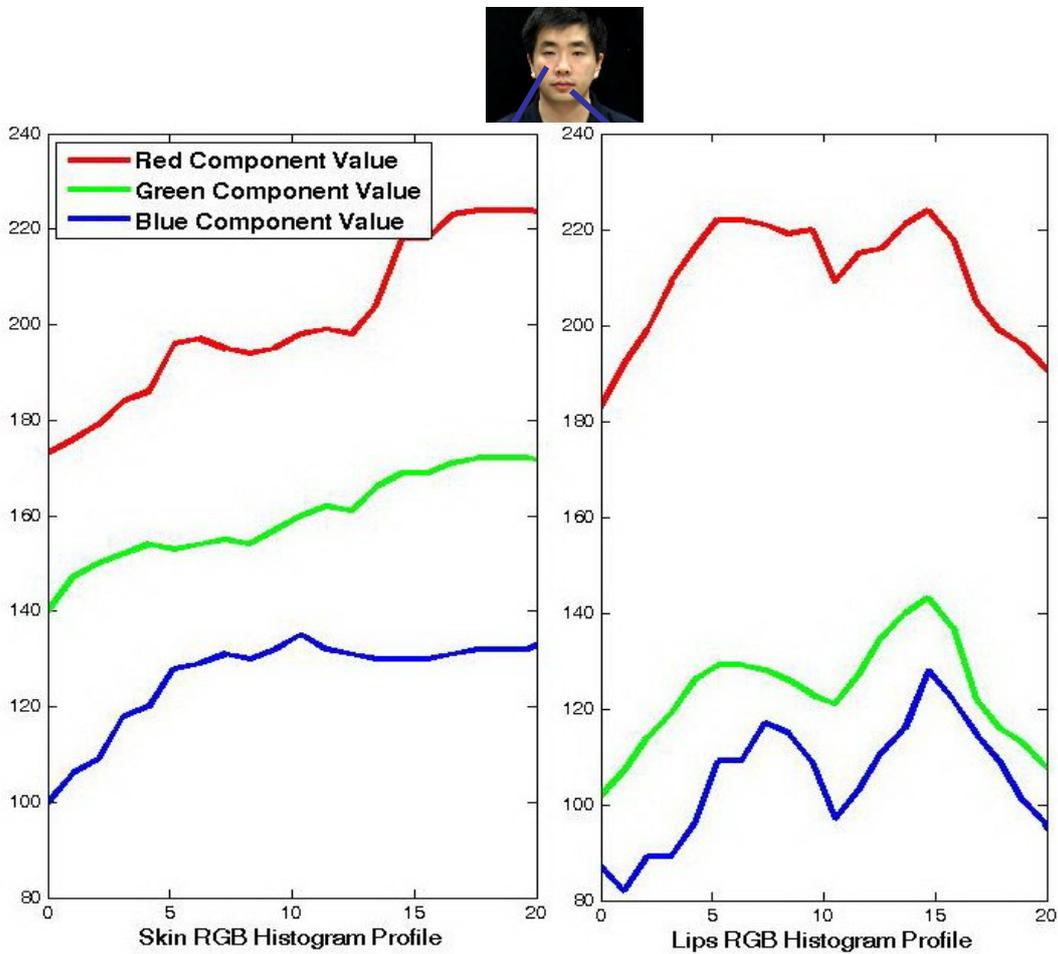


Fig. 3.3 RGB histogram profile for selected skin and lip regions.

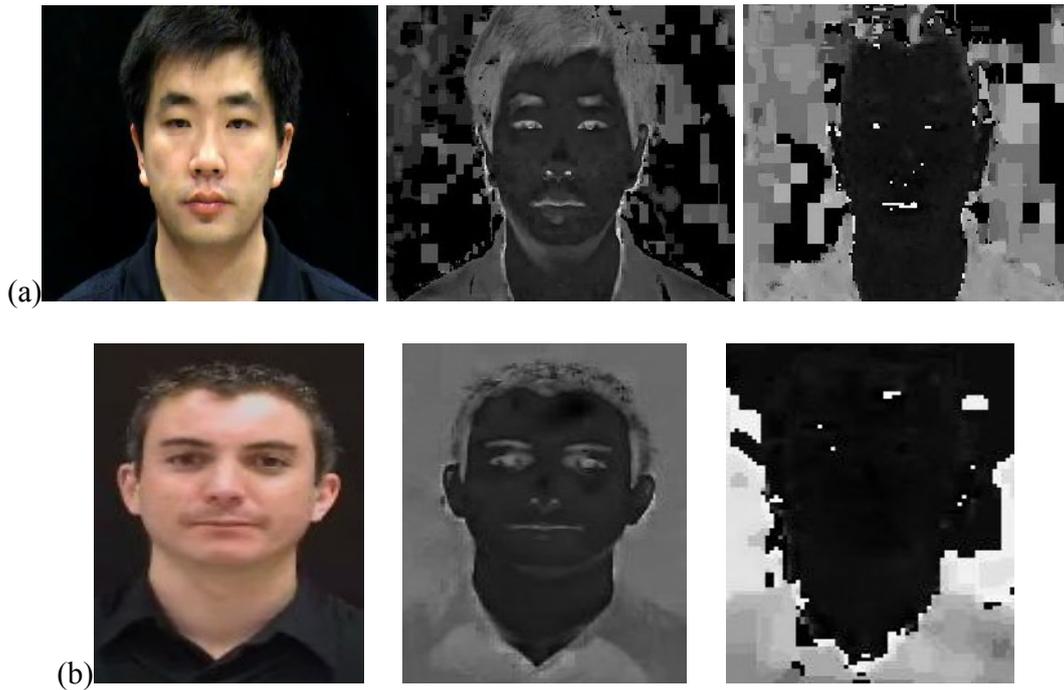


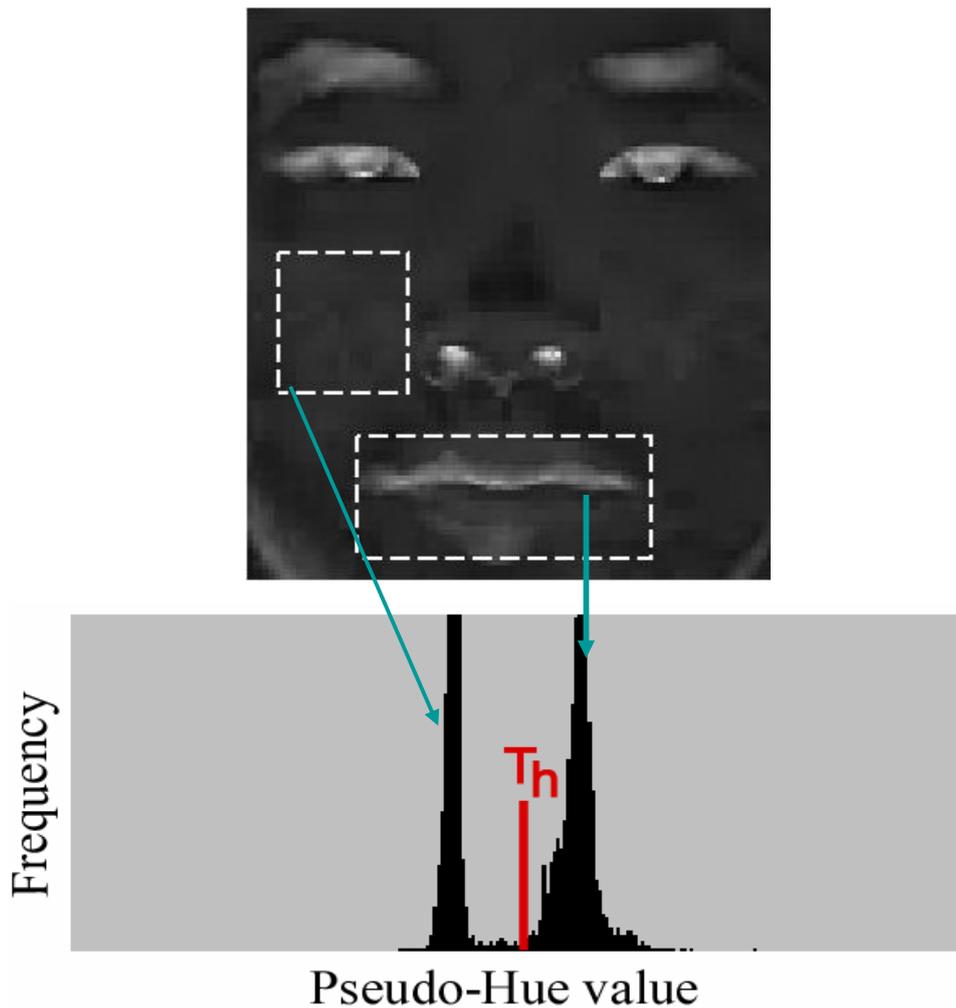
Fig. 3.4 RGB, Pseudo-Hue and Hue images.

### 3.2.2.2 Lip Detection Based on Histogram Thresholding

Thresholding is a basic segmentation technique that has been applied to remove the background information that is associated with the face skin and retain the mouth area as a uniform region in the image [49-50]. The aim of this operation is to binarise the pseudo-hue image into two values as follows:

$$g(i, j) = \begin{cases} 0K & \text{if } f(i, j) \leq Th \\ 1K & \text{otherwise} \end{cases} \quad (3.2)$$

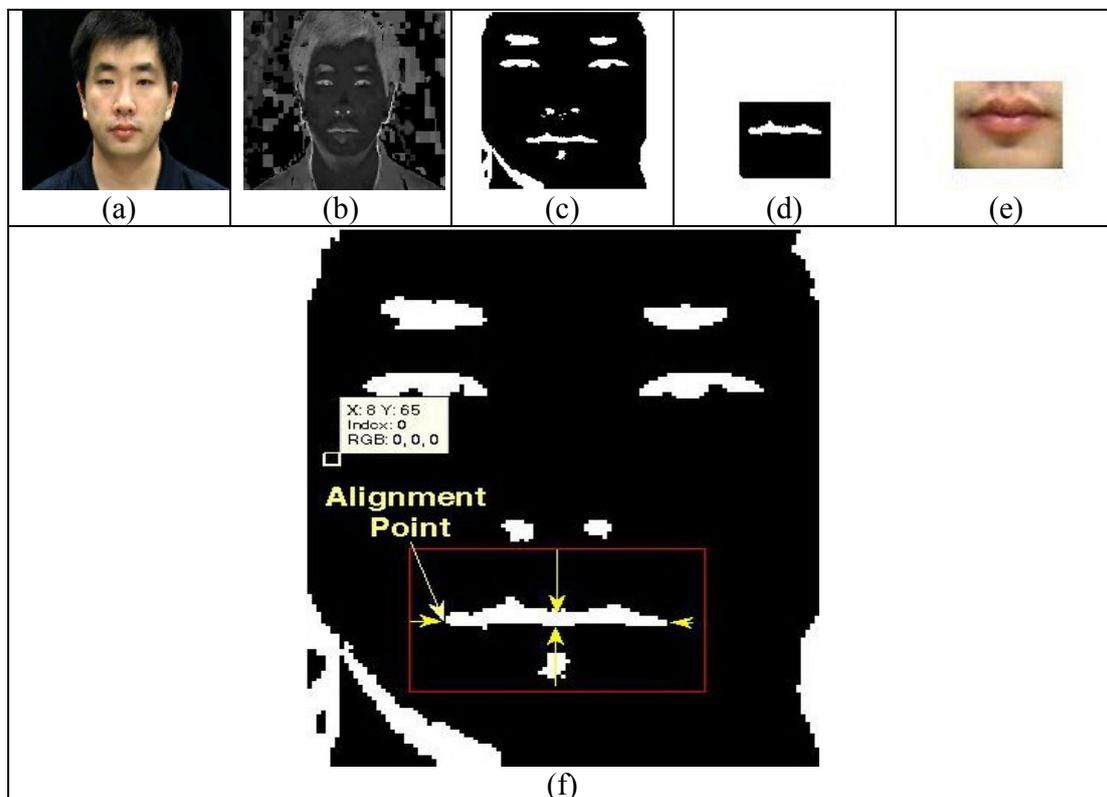
Where  $f(i, j)$  and  $g(i, j)$  are the input and output images respectively and  $Th$  is a threshold value. In this implementation, the threshold  $Th$  is selected based on the knowledge that the histogram calculated from pseudo-hue image has two apparent peaks as illustrated in Fig. 3.5.



**Fig. 3.5 Histogram-based selection of the threshold value**

While the second peak of the histogram is generated by bright image areas (lips), the threshold ('Th') is automatically detected as the local minimum with respect to the second peak in the histogram as illustrated in Fig. 3.5. This operation will identify the mouth area in the image and a ROI around the lips is constructed as the bounding box that encompasses the extreme corners of the upper lips as depicted in Fig. 3.6. Morphological techniques were applied to close the gaps between the segmented pixels and eliminate the isolated pixels generated by noise (see Fig. 3.6(c)). The fixed geometric structure of the face has been used to identify the final lips position in the image. In this

regard, the most left corner of the lips is used for the mouth alignment. The region of interest around mouth is extracted based on the area between nose and jaw. This process is illustrated in Fig. 3.6(f). It is important to notice that the approach used on lip segmentation is semi-manual (i.e.: the lip location is automatically identified based on the structure of the face, then the extraction is corrected by manually alignment for some images). It is motivated to learn and adopt other automate approaches (e.g.: AAM, MHI) to improve the robust of lip segmentation in the future work.



**Fig. 3.6 Lip detection process.**

(a) Original RGB Image (b) Pseudo-Hue Component (c) Image resulting after thresholding and the application of morphological operators. (d) Image describing the mouth region. (e) ROI extracted from the original image. (f) Alignment example of mouth detection

### 3.2.2.3 Image Normalization

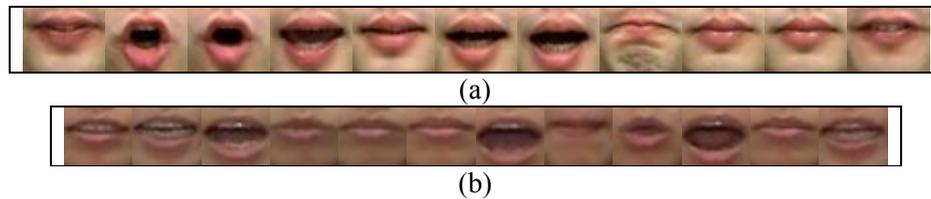
Image normalization is often applied to compensate for uneven illumination that is generated by the image acquisition procedure. In our implementation, the mean flow technique is applied to normalize the image intensities and to remove the undesired illumination effect of the skin. This image normalization technique is defined as follows:

$$N(R_n, C_1 \dots C_{30}) = \frac{P(R_n, C_1 \dots C_{30})}{\text{mean}(P(R_n, C_1 \dots C_{30}))} \quad (3.3)$$

Where  $N$  is the normalized image,  $P$  is the original raw image,  $R$  is row and  $C$  is column,  $n$  is the row index ( $n = 1 \dots 40$ ). Images resulting from the normalization procedure are used as input data for the VSR system.

### 3.2.3 Lip Segmentation Results

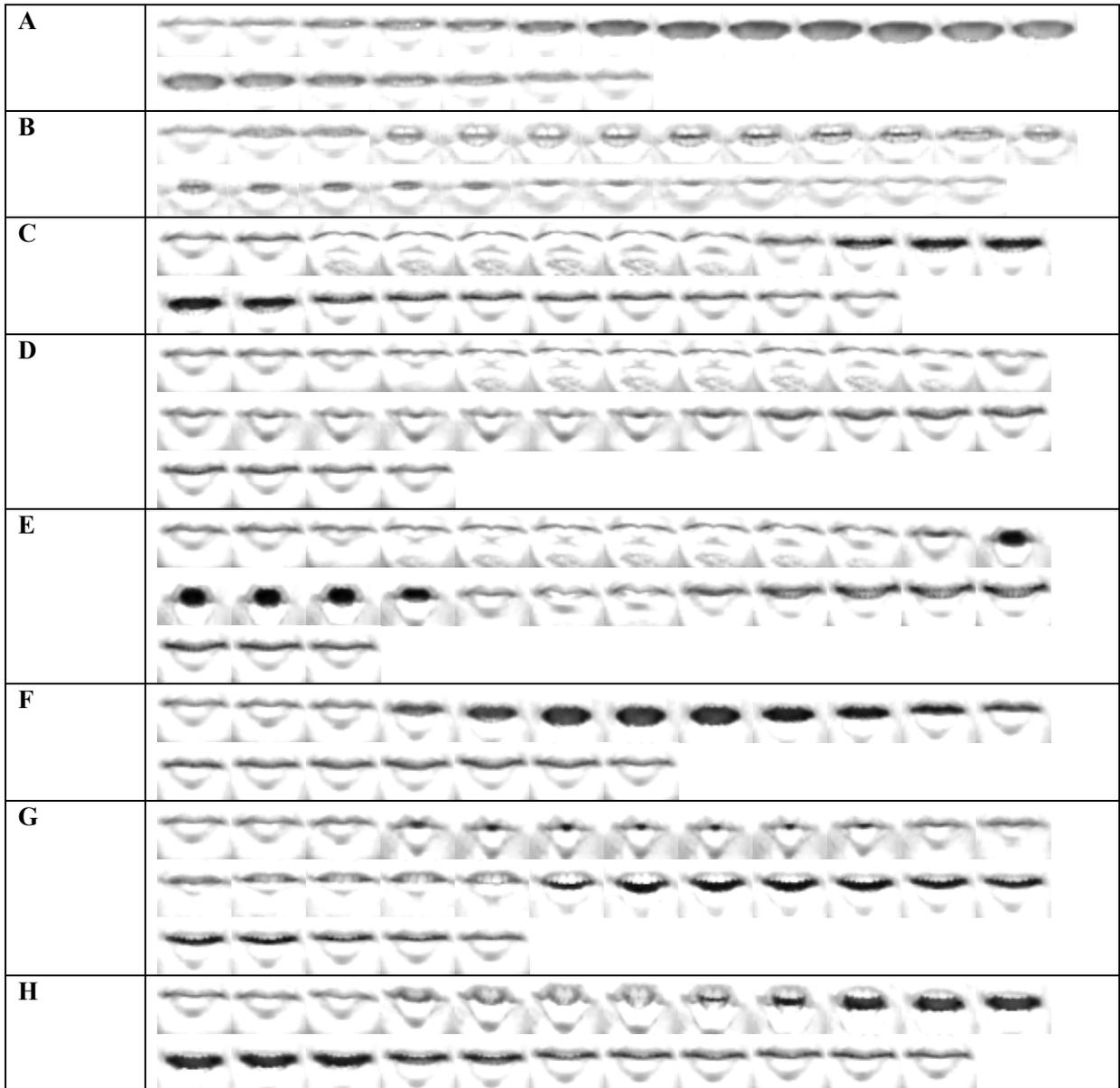
The proposed lip-segmentation method has been tested on data generated by two speakers and a number of experimental results are depicted in Fig. 3.7.



**Fig. 3.7 Lip-segmentation results.**

(a) Speaker One (b) Speaker Two.

In our experiments we have used a database of 700 visual speech sequences associated with 50 words. Fig. 3.8 shows eight sequence examples where each sequence describes the lip movements for one word.



**Fig. 3.8 Sequences of lip segmentation results.**

(a) Word 'I' (b) Word 'You' (c) Word 'But' (d) Word 'Boot' (e) Word 'Barbie' (f) Word 'Heart'  
 (g) Word 'Hoover' (h) Word 'Chard'.

### 3.3 EM-PCA Algorithm

Principal Component Analysis (PCA) is a transform that is widely applied to reduce the dimensionality of the input data. The main idea behind PCA is to identify a

compressed representation for input data in order to highlight the similarities and differences between input patterns. Since the input patterns have high dimensions, the application of exhaustive search procedures to identify the similar patterns is a time consuming procedure. Thus, in order to represent the input data efficiently the PCA is applied to generate orthogonal (eigenvector) decomposition.

Although PCA is a powerful technique for image compression it has several shortcomings. The first is the fact that it is a naïve method for finding the principal component directions and it is cumbersome to be applied to data defined by a large numbers of data points. Another shortcoming of standard PCA is that it is not efficient when applied to sparse data [51].

The Expectation-Maximization (EM) is a probabilistic framework that is usually applied to learn the principal components of a dataset using a probabilistic space partitioning approach. Its main advantage resides in the fact that it does not require computing the sample covariance as PCA and has a complexity limited to  $O(knp)$  where  $k$  is the number of leading eigenvectors to be learned. This redundant parameterization of the models gives us a more robust procedure when applied to sparse data. It can be formulated in terms of estimating the maximum likelihood values for missing information at the each iteration [52, 53]. The EM algorithm has the following steps:

■ Initialization

- Assume some initial models. The better the initial models sample the modes of the data, the better the estimated result. The initial parameters are used to evaluate the expectation, as indicated in the next step.

- Expectation Step (see Equation 3.4)
  - Use the current estimate of the parameters and the observed data to estimate the unknown factors that will minimise the distance between the patterns to the closest models (i.e. compute the expected value of the data for the next step based on the estimate of the parameters and observed data).
  
- Maximization Step
  - Based on this information we need to compute the Maximum Likelihood (ML) estimate of the parameters using the data from the expectation step.
  
- Convergence
  - Iterate the expectation and maximization steps until a convergence criterion is met. It is useful to note at each iteration, an increase in the log-likelihood is obtained and the algorithm is guaranteed to converge to a local maximum.

Expectation-Maximization PCA (EM-PCA) is an extension of the standard PCA technique by incorporating the advantages of the EM algorithm in terms of estimating the maximum likelihood values for missing information. This technique has been originally developed by Roweis [51] and its main advantage over the standard PCA is the fact that it is more appropriate to handle large high dimensional datasets especially when dealing

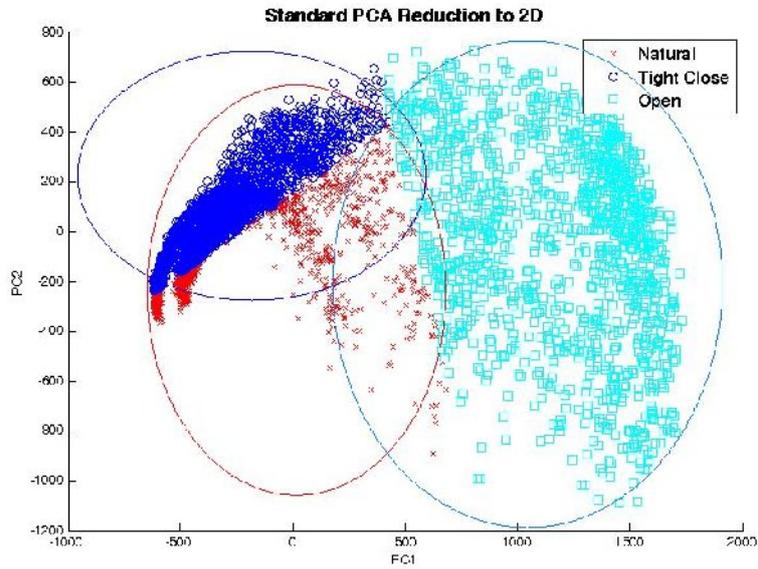
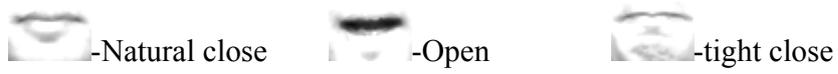
with missing data and sparse training sets. The EM-PCA procedure has two distinct stages, the E-step and M-step:

$$\begin{aligned}
 \text{E-step: } W &= (V^T V)^{-1} V^{-1} A \\
 \text{M-step: } V_{new} &= A W^T (W W^T)^{-1}
 \end{aligned} \tag{3.4}$$

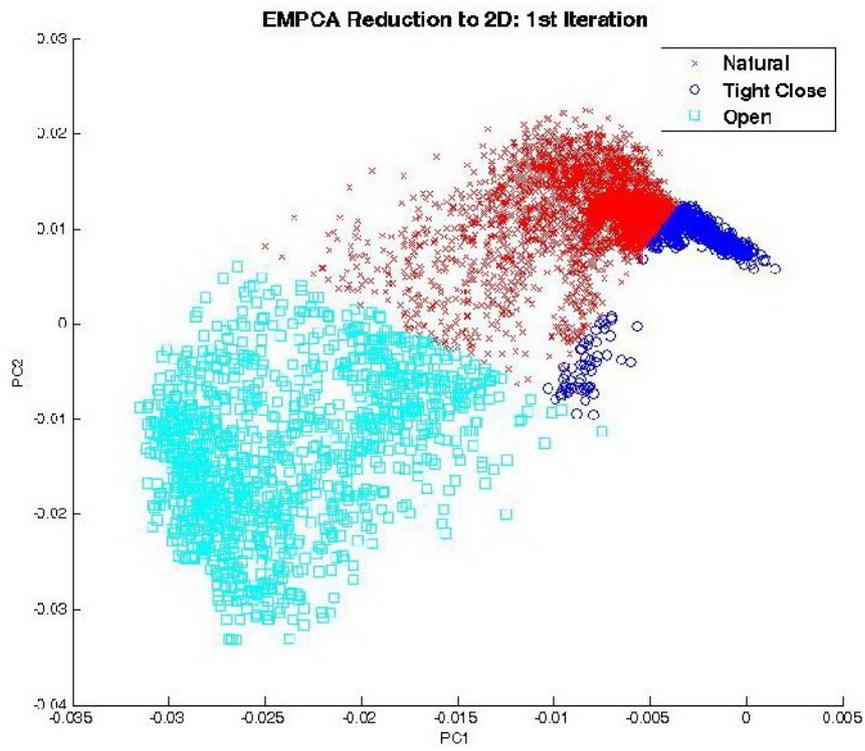
Where ‘ $W$ ’ is the matrix of unknown states, ‘ $V$ ’ is the test data vector, ‘ $A$ ’ is the observation data and  $^T$  is the transpose operator. The columns of ‘ $V$ ’ span the space of the first  $k$  principal components.

To illustrate the superior performance of the EM-PCA when compared to that attained by the standard PCA, both algorithms were applied to 6200 images associated with 3 classes of mouth shape (natural close, open and tight close) with the aim to reduce the high dimensions of the input data to a 2-dimensional space (see Fig 3.9).

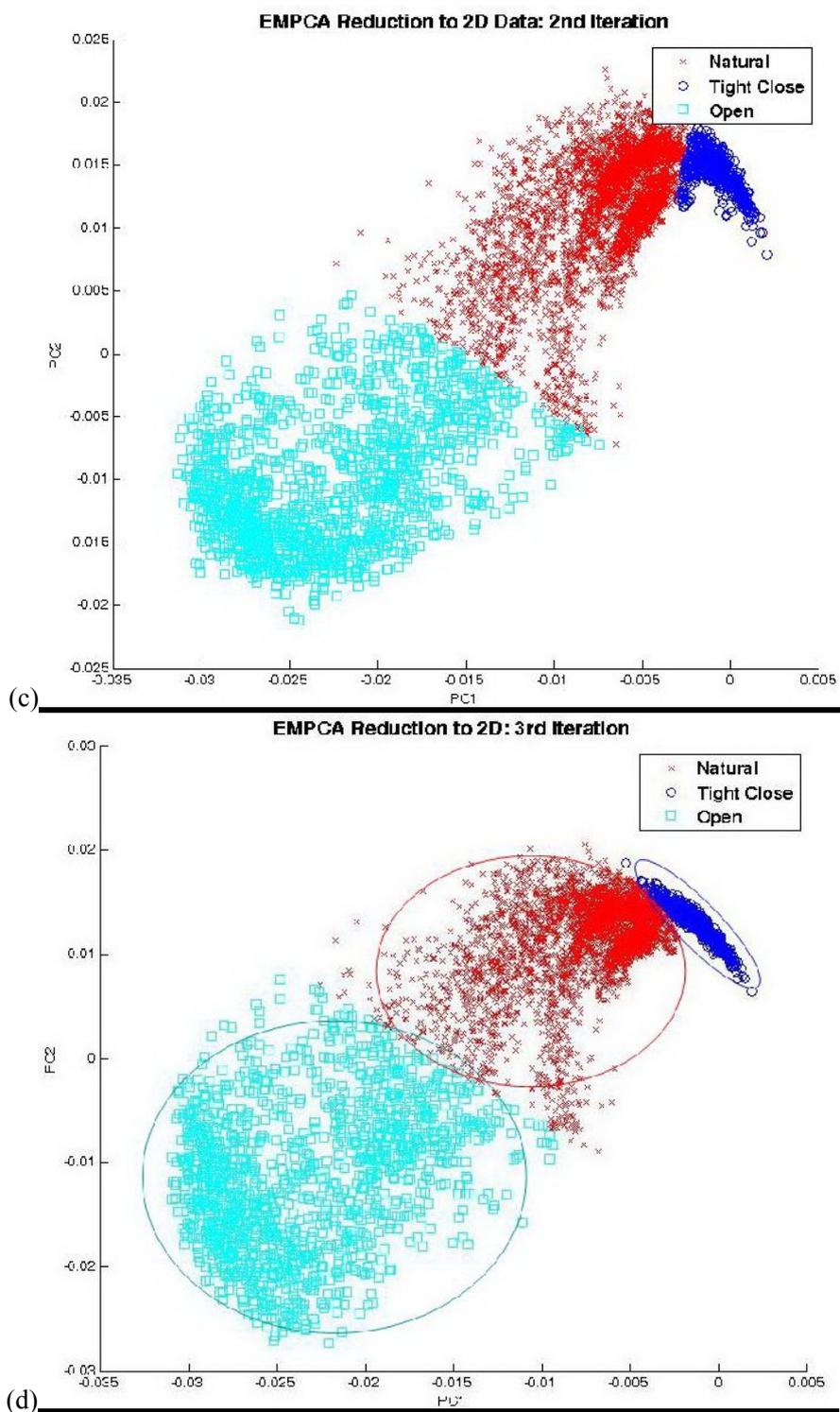
All images are randomly selected from video speech sequences and they are manually labeled into three classes. The experimental results in Fig. 3.9(b-d) indicate that EM-PCA algorithm converges to the expected solution in only three steps and the compression result presents a better data distributions among three classes of mouth shape than the standard PCA (Fig. 3.9a). In another words, the EM-PCA data compression reduces the class overlapping with a larger extent than the standard PCA when the training mouth shape images are generated in a very high-dimensional (1200 dimensions) space.



(a)



(b)



**Fig. 3.9** The EM-PCA and Standard PCA when applied to a large dataset (6900 images). (a) Standard PCA. (b) EM-PCA, 1<sup>st</sup> iteration (c) EM-PCA, 2<sup>nd</sup> iteration. (d) EM-PCA, 3<sup>rd</sup> iteration.

### 3.4 Proposed Approach: EM-PCA Manifold Representation

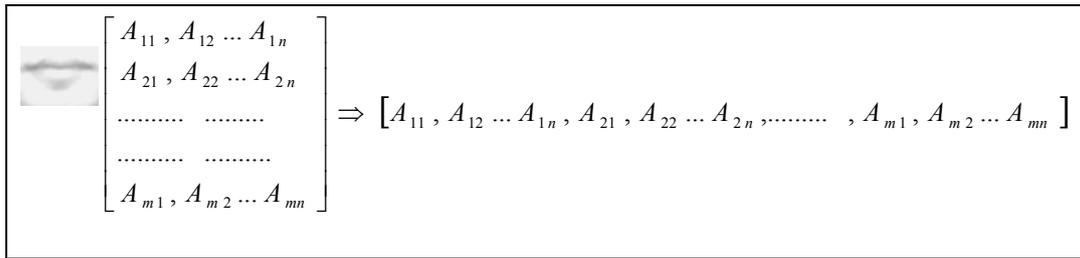
Visual speech feature extraction is a key component required by the VSR system. As indicated before, intensity-based and shape-based feature extractions are two of the most commonly used algorithms in literature (see Section 2.3). One of the major drawbacks of the shape-based feature extraction is that needs large training sets to cover the large range of mouth shapes and this fact make the inclusion of these feature extraction schemes into VSR applications difficult.

In our implementation, intensity-based feature extraction and EM-PCA data compression algorithm have been deemed to be the most appropriate. These methods are used to encode the appearance of the lips in each frame as a point in a low-dimensional feature space that is obtained by projecting the input data onto the eigenvector space generated by the EM-PCA procedure.

#### 3.4.1 Manifold Calculation from Input Data

For visual speech recognition purposes, the gray-level images describing the lip motions were extracted from the input data in order to provide a more efficient data structure for feature extraction (Section 3.3). In our approach, the gray-level pixels from all segmented frames (see Fig. 3.7) are arranged in one large vector. From this vector a low-dimensional space is calculated using the EM-PCA algorithm. The matrix conversion procedure applied to generate the one-dimensional vector  $A$  is depicted in Fig. 3.10.

In our database, each image resulting from the lip segmentation algorithm is  $[40 \times 30]$  size normalized and it is converted into a matrix of intensity values by reading the image in a raster scan mode.

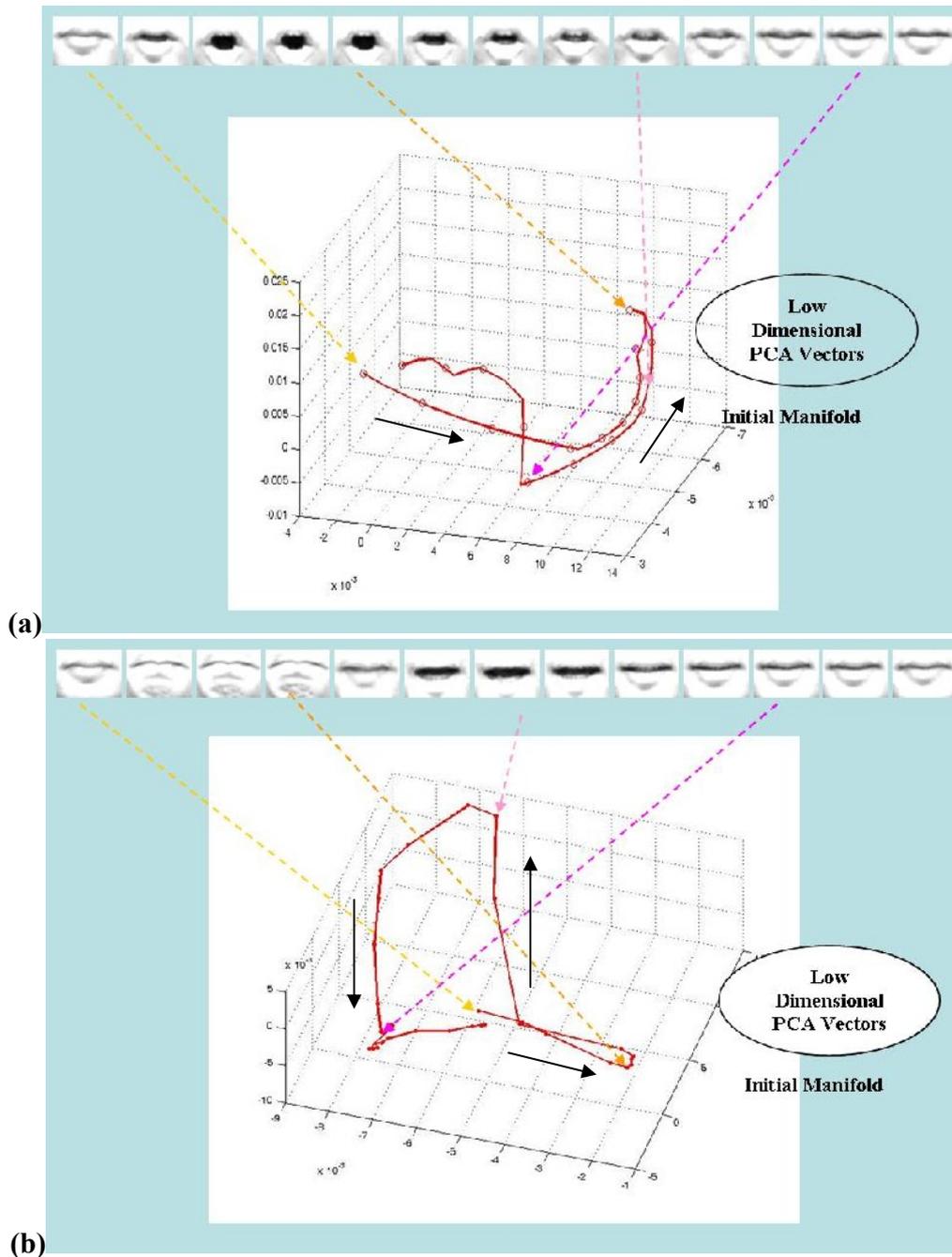


**Fig. 3.10 Matrix conversion to one-dimensional vector.**

The next step involves data compression using the EM-PCA procedure. There are different ways to select the number of components. This selection is very dependent on how much information you are going to present (i.e.: some researchers [89] use a 32-D subspace or a 10-D subspace for image interpolation comparisons). In this implementation we used only the first three EM-PCA components since they are able to capture approximately 87% of the 40,000 images contained in the database. Then, the lip images extracted for each frame are projected onto the EM-PCA low-dimensional space and for each image a low dimensional feature point (vector) is obtained. The feature points obtained after data projection on the low-dimensional EM-PCA space are joined by a poly-line by ordering the frames in ascending order with respect to time (see Fig. 3.11). As mentioned in Section 1.3, a surface is generated based on the trajectories of the feature points in the 3D EM-PCA space where different mouth shapes/lip movements generate a compressed representation of the visual speech that is referred to as “manifold”

Each feature point on the manifold surface presents a particular mouth shape and the whole manifold encodes the entire lip movements of the visual speech sequence. It is useful to notice that three EM principal components (PC) are strongly related to the features that describe the mouth shape. In this case, the 1<sup>st</sup> PC captures the skin information around lips while the 2<sup>nd</sup> PC captures more localized information such as the geometry of the mouth shapes (closed, opened, etc.). The 3<sup>rd</sup> dimensional PC captures

finer details (the presence of teeth and tongue). We can tell this by looking at original images in EM-PCA representation.



**Fig. 3.11 EM-PCA “Word Manifold” representation.**

(a) “hot” (b) “bart”. Each feature point of the manifold is obtained by projecting the image data onto the low-dimensional EM-PCA space.

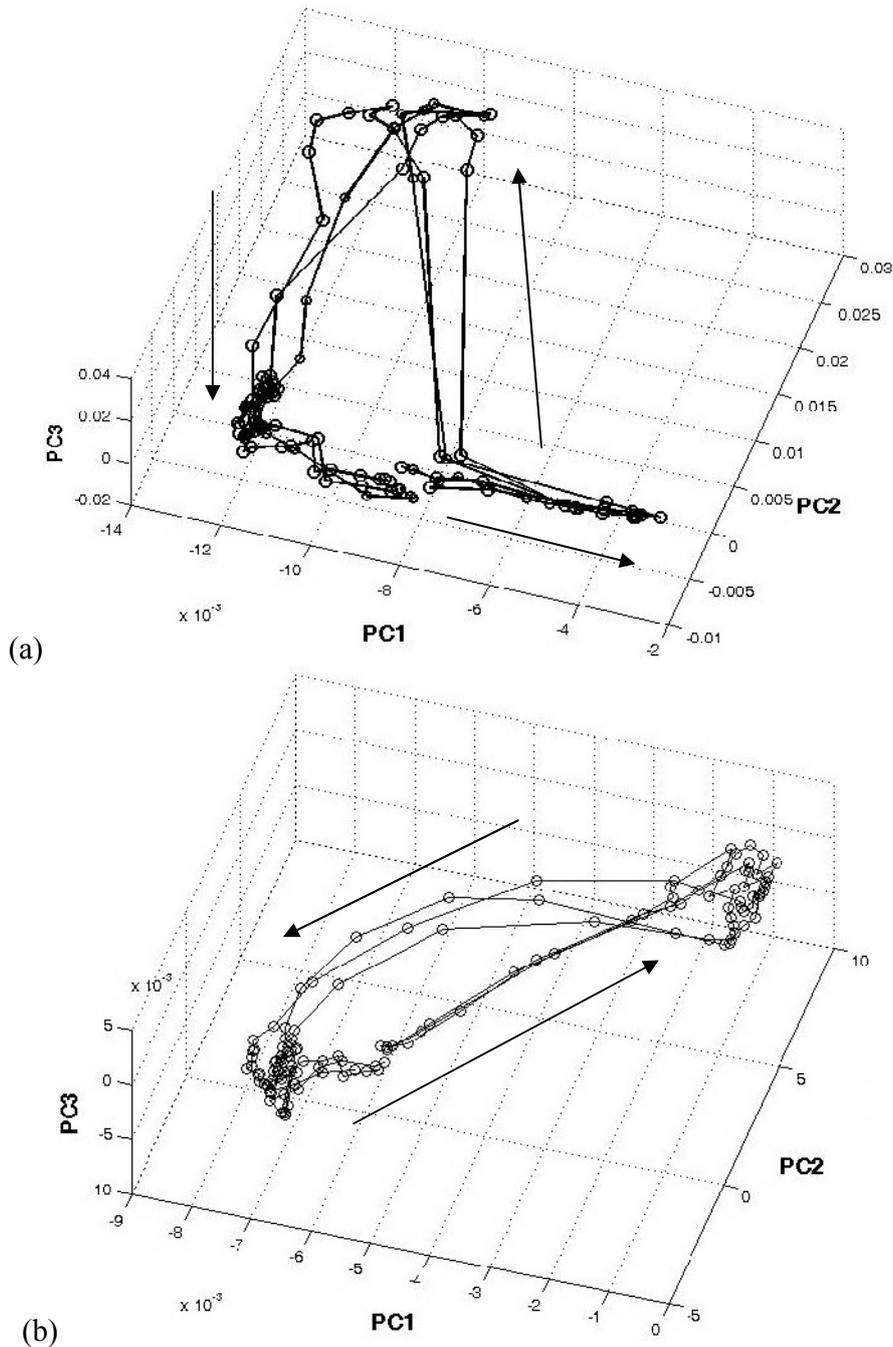


Fig. 3.12 “Word Manifold” Examples.

(a) “Word Manifold” generated from three image sequences representing the word “Bart”. (b) “Word Manifold” generated from three image sequences representing the word “Hook”. Note: The EM-PCA space is represented by the first three principal components: PC1, PC2 and PC3. It is important to notice that the appearances of manifolds for each word indicate that their shapes are similar and contains information in regard to the word spoken.

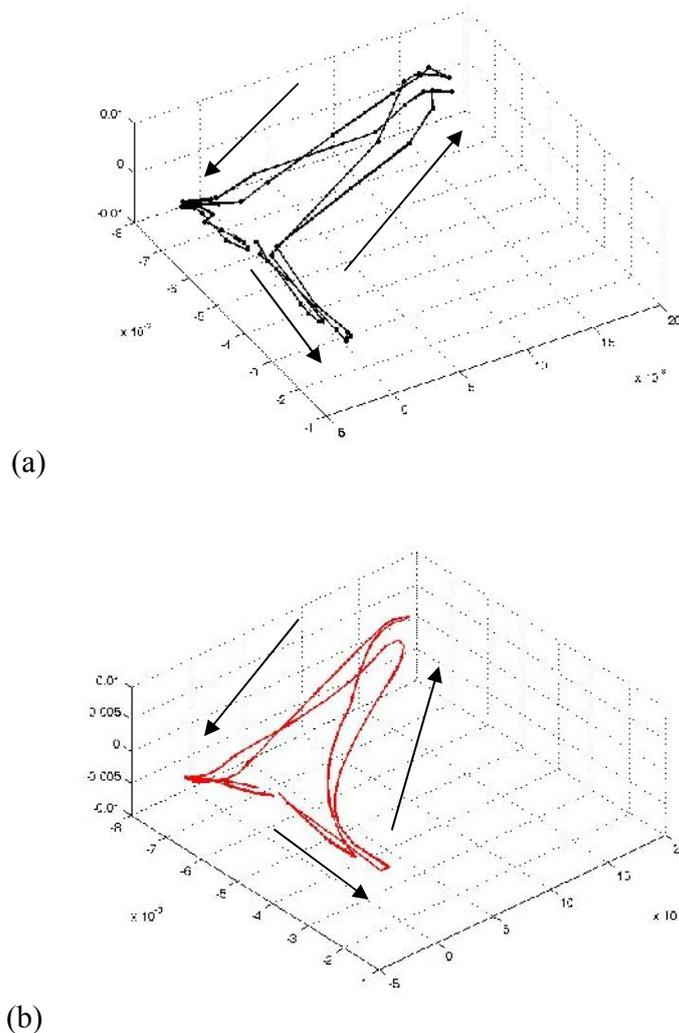
Since the EM-PCA “Word Manifold” encode the lip motion through image compression, the shape of the manifold will be strongly related to the words spoken by the speaker and recorded in the input video sequence. Fig. 3.12 illustrates the manifolds calculated for three independent image sequences (describing two words) in the EM-PCA feature space. It can be noted that the shapes of the manifolds are very similar and can be interpreted as word “signatures”.

### **3.4.2 Manifold Interpolation**

As illustrated in Fig. 3.12 the shape of the “word manifold” can be potentially used to discriminate between different words. While “word manifold” can be interpreted as a word “signature”, they cannot be used directly to train a classifier and to recognize an unknown input image sequence since the number of feature points that generate the “word manifold” is not constant (the number of frames contained in the input image is variable and depends on the complexity of the word spoken by the speaker). In this way, short words such as “bart”, “hot”, etc. have associated a small number of frames and as results the manifolds will be defined by a small number of feature points. Conversely, longer words such as “beautiful” and “banana” have associated larger image sequences and the number of feature points that defines the manifolds is larger. This is a real problem when these manifolds are used to train a classifier as the number of feature points is different.

This “word manifold” representation is not convenient due to the fact that the spoken words are sampled by a different number of frames that may vary when the video data is generated by different speakers. To address this issue, the feature points that define the “word manifold” are interpolated using a cubic spline to obtain a continuous manifold

representation. The application of the cubic spline interpolation has two main advantages. Firstly, it allows the generation of smooth EM-PCA “word manifold” and secondly it reduces the effect of noise (and the influence of objects surrounding the lips such as teeth and tongue). This is clearly shown in Fig. 3.13 where the appearance of the manifolds obtained after the application of cubic interpolation is illustrated. Fig. 3.13 illustrates the interpolated manifolds generated for the two examples of the word “bart”.



**Fig. 3.13 “Word Manifold” interpolation.**

(a) Initial manifolds - word “bart”; (b) Interpolated manifolds - word “bart”.

### **3.5 Summary**

This chapter describes the process of lip segmentation and EM-PCA “word manifold” representation. In this regard, the mouth region is segmented after the pseudo-hue component is subjected to histogram-based thresholding that is applied to separate the face skin and mouth regions in the image. Afterwards, the “word manifold” is generated from the lip gray-level intensity images and this data is compressed into a low-dimensional feature space using an EM-PCA procedure. Since these “word manifolds” are defined by a different number of frames, they cannot be used directly as inputs for classification. To address this problem, the “word manifolds” are interpolated to generate a continuous representation that will be further analysed to identify the visual speech units (VSU) that will be detailed in the next chapter of this thesis.

# Chapter 4

## Visual Speech Modeling

### 4.1 Introduction

Visual speech recognition (VSR) is a difficult task that involves the identification of visual speech models. Visual speech models are required to generate speech classes that are typically constructed from observed mouth shapes. In general, each speech class is defined as a basic unit and these units can be concatenated to form words and sentences, thus allowing the VSR systems to be applied to continuous speech sequences [27].

The selection of the appropriate visual speech model is the key issue for any VSR system. The literature review detailed in Section 2.4 indicates that visemes play an important role in the development of VSR systems and many researchers have approached continuous speech (e.g. at word level) recognition as a process of sequential viseme recognition [1, 27, 43, 55, 81]. Although words can be theoretically formed by a time-ordered combination of standard visemes, in practice due to various pronunciation styles, similar visemes can be associated with different visual signatures. In addition to this, the articulation (pronunciation) phase plays an important role in the process of defining each viseme [38-40] and as a result the viseme representation is not able to model the transitions between consecutive visemes. In order to alleviate the shortcomings associated with standard visemes, a new Visual Speech Unit (VSU) model is proposed in this thesis. This new speech representation includes not only the information associated with standard visemes but also the transitory information between consecutive visemes.

In the approach discussed in this thesis, the registration process between the VSU mean models and the continuous word manifolds (see Chapter 3) is carried out using Dynamic Time Warping (DTW).

## **4.2 Viseme Review**

### **4.2.1 Viseme Introduction**

The basic unit that describes the audio speech process is represented by the phoneme [1]. In the case of the visual speech, the basic units that correspond to the visually distinguishable phonemes are referred to as visemes [63]. A viseme can be regarded as the smallest element that describes a phoneme or a group of phonemes in the visual domain. In this thesis, viseme is seen as the representation in the visual domain of the mouth shapes that correspond to one or more phonemes. In order to represent visemes in the EM-PCA feature space, the images that correspond to a particular viseme are manually selected based on the appearance of the mouth shapes and the presence of teeth and/or tongue. Then these manually selected images will be projected on the EM-PCA eigenspace and the low-dimensional points are used to represent visemes based on the manifold representation. (A number of examples are shown in Section 4.2.2)

In recent years, the theory of viseme modeling has been actively researched and found applications in the areas of Automatic Speech Recognition (ASR) [1, 23], Visual Speech Recognition (VSR) [27, 38-43, 55-57] and computer animation [60, 63]. Most researchers converged to the conclusion that visemes should be constructed using basic visual lip motions that are observed during the speech process. The relationship between phonemes and viseme is a many-to-one mapping because phonemes do not generate an exact correspondence between lip position and acoustic sounds. In another words,

phonemes are easy to “hear” but hard to “see”. For example, although phonemes [b], [m] and [p] are acoustically distinguishable, they are always grouped [27, 40, 43] into one viseme category as they are described by similar sequences of mouth shapes.

In this thesis, English is used as the language for visual speech recognition. Although there is a reasonably strong consensus about the set of English phonemes, there is less unanimity about the selection of most representative visemes [60]. Since phonemes and visemes cannot be mapped directly, the total number of visemes is much lower than the number of standard phonemes. In practice, various viseme sets have been proposed with their sizes ranging from 6 [57] to 50 visemes [64]. Actually this number is by no means the only parameter in assessing the level of sophistication of different schemes applied for viseme categorization [60]. For example, some approaches propose small viseme sets based on English consonants [27], while others propose the use of 6 visemes that are obtained by evaluating the discrimination between various mouth shapes (closed, semi-opened and opened mouth shapes [57]). A list of proposed viseme categories is provided in Appendix A. This thesis adopts the viseme model established for facial animation by an international object-based video representation standard known as MPEG-4 [54]. Based on the MPEG-4 viseme model, there are nine visemes associated with English consonants and five visemes associated with English vowels. The representation of the MPEG-4 viseme categories using EM-PCA manifolds is discussed in the next section.

### **4.2.2 Viseme Representation**

In the visual speech processing domain, a viseme consists of a time-ordered sequence of lip shapes. In practice, VSR systems are trained with either static visemes where each viseme is generated separately by the speakers (the speaker is asked to speak each viseme

individually), or with visemes that are manually constructed by isolating the frames of interest from continuous video speech sequences.

The static visemes are mapped based on mouth shapes and placement of tongue during phoneme articulation. For example, Lee and Yook [43] developed a viseme mapping table that is shown in Fig. 4.1.

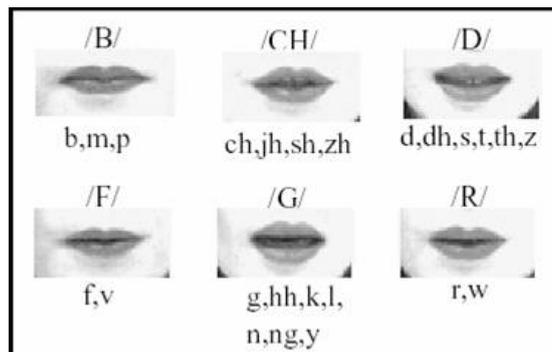


Fig. 4.1 Mapping table for 6 visemes associated with Standard English consonants [43].

The static visemes are favored by researchers since they are easy to generate and identify. In this way, the speaker is asked to articulate each isolated viseme and the images extracted from the video sequence are used to generate a viseme representation. An example that shows the mapping between phonemes and visemes is introduced by S. Foo and Lian [40], which is illustrated in Fig. 4.2.

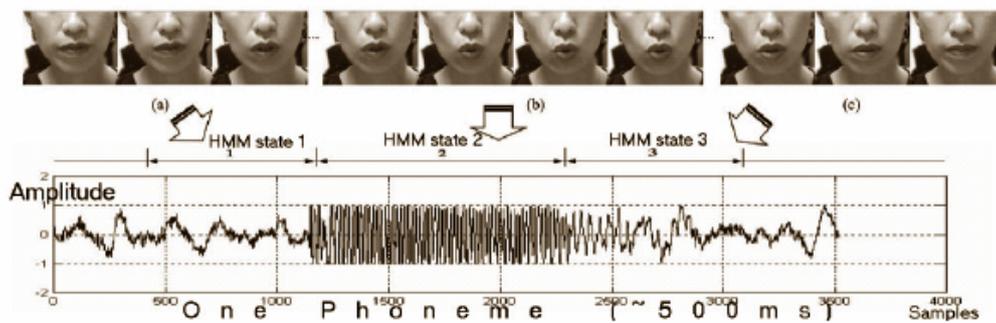
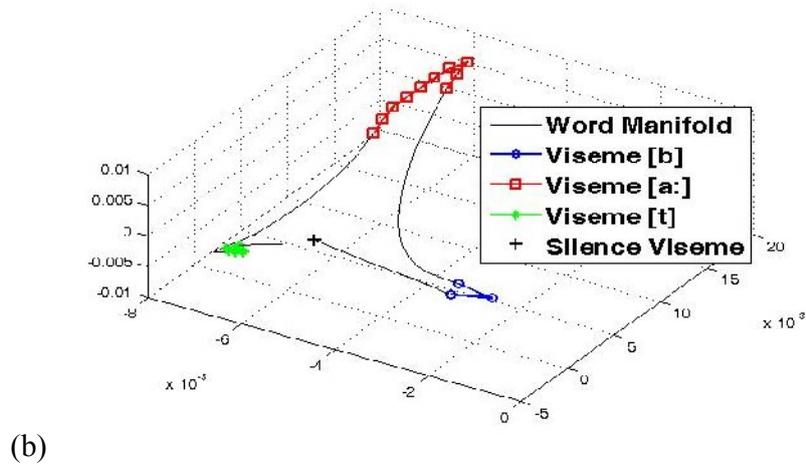
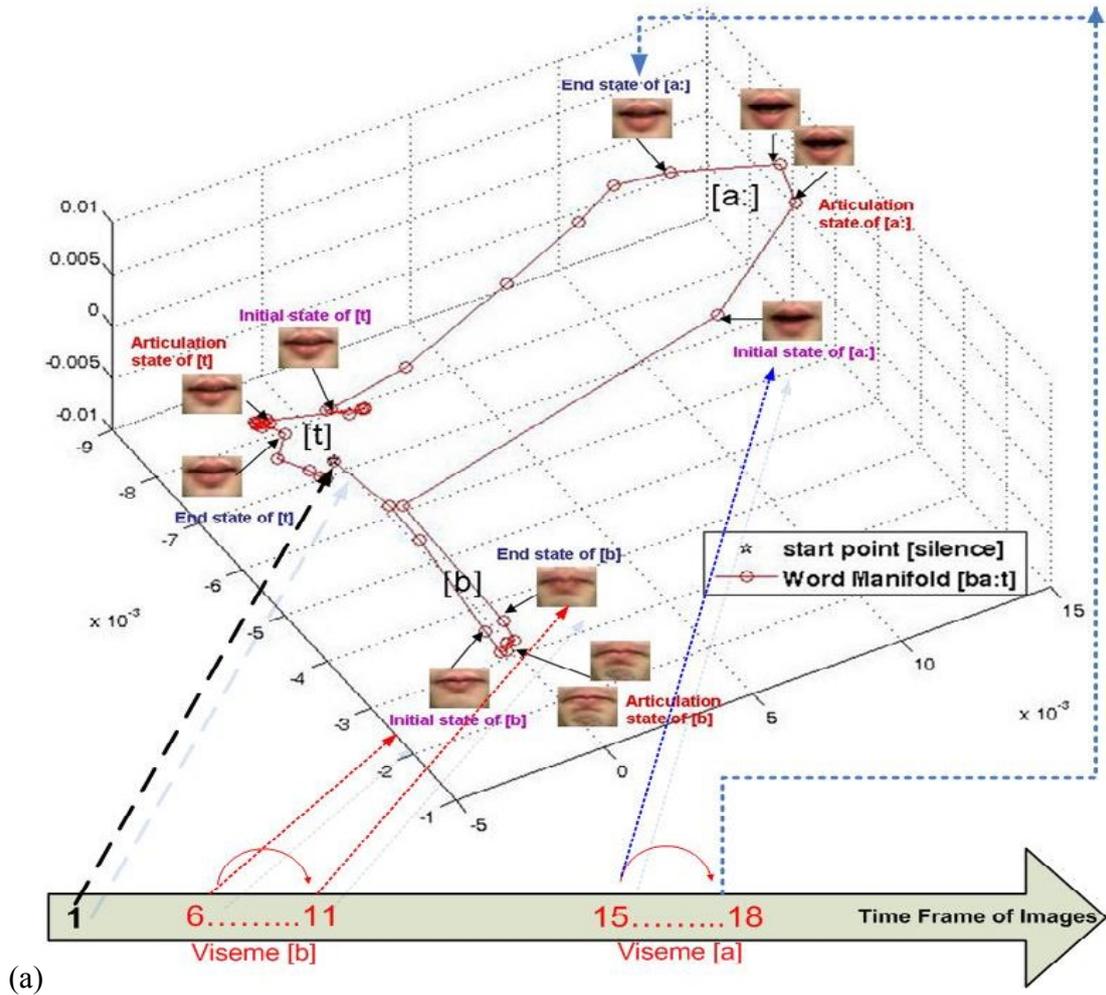


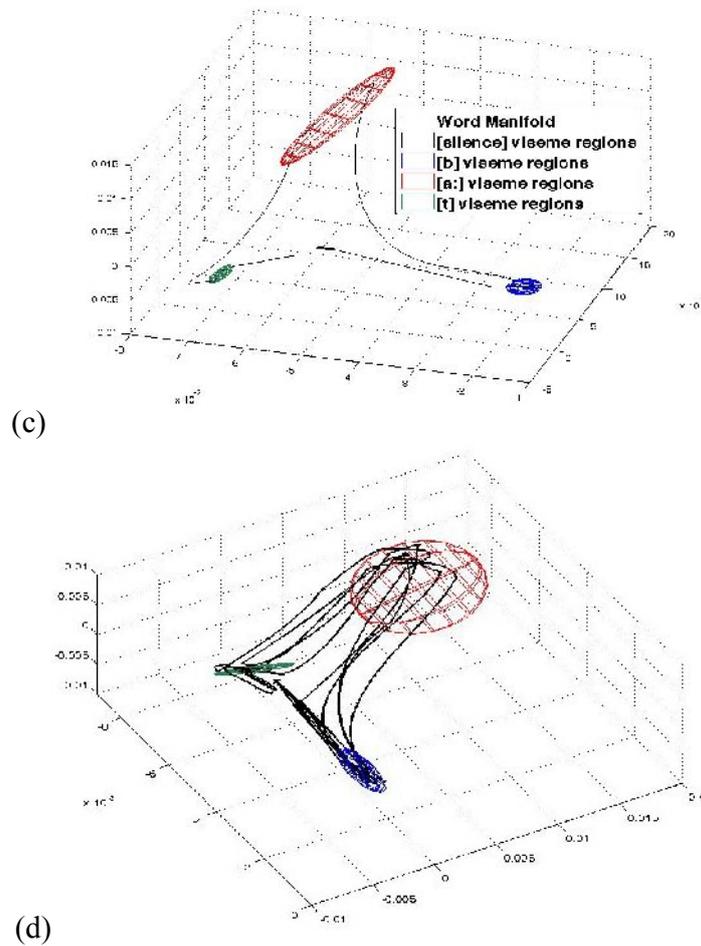
Fig. 4.2 Phoneme to viseme mapping [40].

Many researchers [27, 38-43, 55-56, 61] have applied this static viseme generation approach in the development of VSR systems, but the identification of static visemes is better suited for recognition of isolated visemes than their recognition in continuous speeches that is seen as a process of sequential viseme recognition. Humans do not speak in discrete units and as a result speech recognition has to be formulated in terms of viseme identification from a continuous flow of lip movements. This fact indicates that static visemes may not be directly applicable to word recognition, the viseme changes gradually in varied speech environment. For example, within a small segment of continuous speech such as a word, the previous viseme affects the initial mouth shapes associated with the next viseme while the middle portion of the viseme is relatively stable. In order to handle the dynamic characteristics of lip motions for automatic visual speech recognition, visemes are more realistically generated by isolating the frames from continuous video speech sequences.

In this approach, the set of visemes is extracted from input video sequences associated with different words. For instance, frames describing viseme  $[b]$  are extracted from words such as ‘Bart’, ‘blue’ etc., while frames describing viseme  $[s]$  are extracted from words such as ‘slow’, ‘snow’, etc. As indicated in Chapter 3, an EM-PCA manifold encodes the lips motions through image compression where the shapes of the manifolds are strongly related with the words spoken.

The feature points on the manifold surface describe particular mouth shapes or lip movements and indicated earlier they are manually selected to construct visemes from spoken words. An example is provided in Fig. 4.3(a).





**Fig. 4.3** The representation of the visemes  $[b]$ ,  $[a:]$  and  $[t]$  in the EM-PCA manifolds of the word  $[ba:t]$  (the initial manifold is represented using a red line, continuous (interpolated) manifold is represented using a black line).

(a) Projection points associated with images from the video sequence on initial manifold

(b) Feature points are displayed in blue for viseme  $[b]$ , in red for viseme  $[a:]$  and in green for viseme  $[t]$  in one instance of word ( $[ba:t]$ ) manifold. The initial state of the video sequence (silence state) is shown in the diagram with a black cross. The interpolated manifold is plotted with a black line.

(c) The regions in the EM-PCA feature space for visemes  $[b]$ ,  $[a:]$  and  $[t]$  are constructed from five instances of the word ('bart') manifold. The region describing the  $[silence]$  state is represented in the diagram with a black star. The word manifold is plotted with a black line.

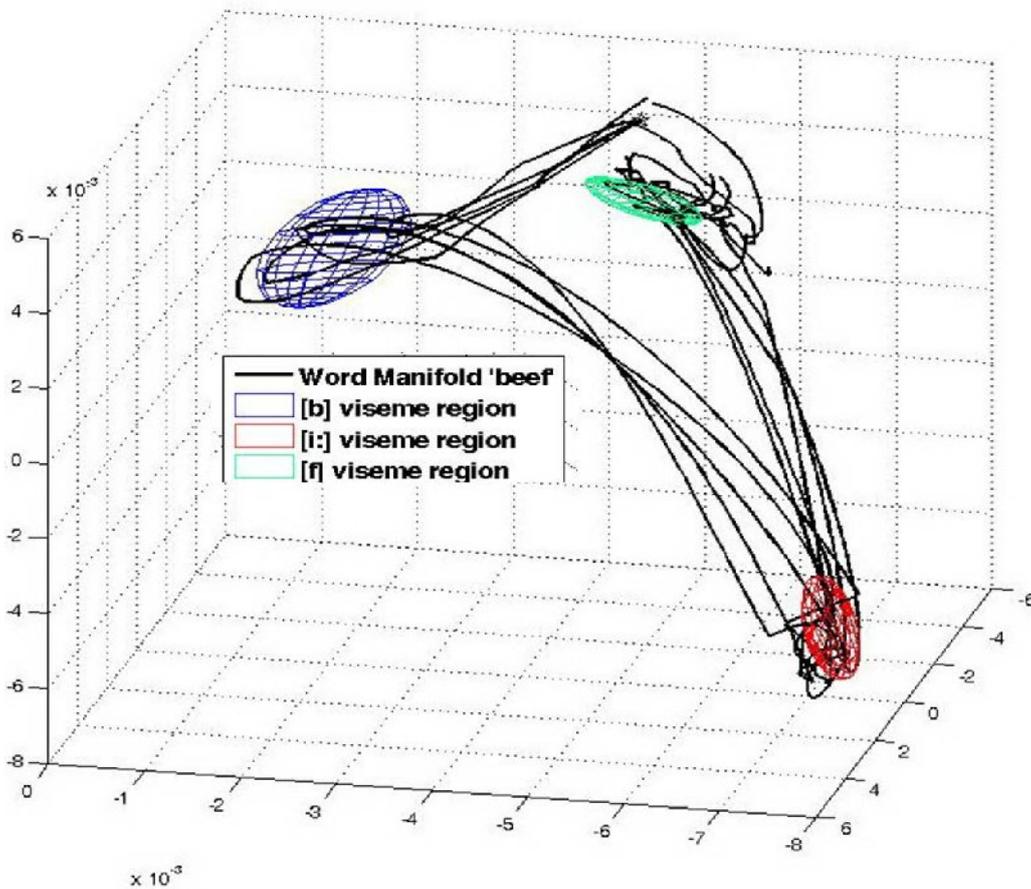
(d) The regions in the EM-PCA feature space for visemes  $[b]$ ,  $[a:]$  and  $[t]$  are constructed from five instances of the word 'bart'. The region describing the  $[silence]$  state is represented in the diagram with a black star. All word manifolds are plotted with a black line.

Note: All samples are represented in same EM-PCA space with different angles of view.

Fig. 4.3 (a) shows the association between feature points that form the manifolds and the corresponding images that define visemes. Three sets of images are shown for the word manifold ‘*Bart-[ba:t]*’. From this diagram, it can be observed that frames describing standard visemes include three independent states. The first state is the initial state of the viseme; the second state describes the articulation process and the last state models the mouth actions associated with the relaxed state. These frames are projected onto the EM-PCA space and as a result each viseme is defined by a number of feature points as illustrated in Fig. 4.3 (b). The feature points for visemes  $[b]$ ,  $[a:]$  and  $[t]$  on the EM-PCA manifold are constructed from video sequences describing the word ‘*Bart-[ba:t]*’. By analyzing different instances of the same word  $[ba:t]$ , a group of features points for visemes  $[b]$ ,  $[a:]$  and  $[t]$  is constructed based on the manifold representation. These feature points are manually drawn in the EM-PCA space as ellipsoids to indicate the space covered by particular visemes. The example of this ellipse is shown in Fig. 4.3(c) and (d).

Fig. 4.4 depicts another example for the word ‘beef’ where is illustrated the representation of visemes  $[b]$ ,  $[i:]$  and  $[f]$  in the EM-PCA feature space.

Based on these examples, it can be concluded that visemes can be theoretically applied to identify the words spoken, but they only cover a small part of the word manifold (see Appendix D for more examples). Visemes are too small entities to fully characterize the entire word information since the transitions between visemes are not used in the viseme-based speech representation.



**Fig. 4.4** The representation of the visemes  $[b]$ ,  $[i:]$  and  $[f]$  in the continuous EM-PCA manifold.

The region describing the [silence] state is represented in the diagram with a black star. Five instances interpolated manifold of the word 'beef' are plotted with a black line.

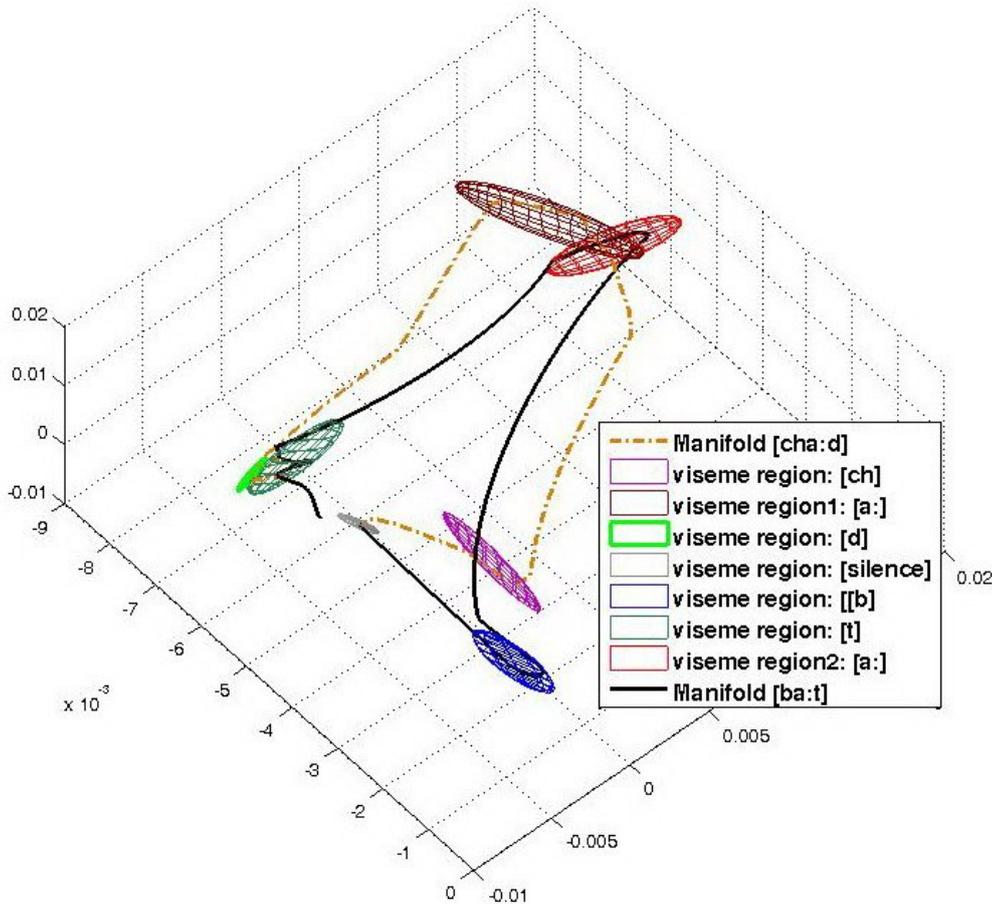
Note that visemes  $[b]$ ,  $[i:]$  and  $[f]$  cover only a small part of the word manifold.

### 4.2.3 Visemes Limitations

The previous section demonstrates that visemes are able to describe partially the word manifolds. While the viseme representation detailed in Figs. 4.3 and 4.4 is intuitive and easy to be applied in the development of VSR systems, it still has associated several drawbacks. The main shortcoming associated with the viseme representation is given by the fact that a large part of the word manifold (i.e. transitions between visemes) is not

used in the recognition process. This approach is inadequate since the inclusion of more instances of the same viseme extracted from different words would necessitate larger regions required to describe the feature space for each viseme (see Fig. 4.5) and this will lead to significant overlaps in the feature space when describing different visemes.

To circumvent this problem most of the developed VSR systems applied the viseme recognition process to a reduced set of visemes and to a relatively small number of words [27, 36, 40-41, 56, 61, 63, 65]. This problem is clearly shown in Fig. 4.5 where the process of constructing viseme spaces for two different words is illustrated. It is important to note that in the manifold representation of the word ‘chard’ the viseme  $[a:]$  is distorted when compared with the viseme  $[a:]$  of the word ‘Bart’ and the consonant  $[r]$  cannot be distinguished. The viseme  $[t]$  and viseme  $[d]$  are in the same category of viseme model and they require larger regions in the feature space. (See Appendix D for more examples.)



**Fig. 4.5 The viseme feature space constructed for two different words.** Word ‘Bart’ – visemes  $[b]$ ,  $[a:]$  and  $[t]$ . Word ‘chard’ – visemes  $[ch]$ ,  $[a:]$  and  $[d]$ .

Note 1: the viseme  $[a:]$  (dark red ellipsoid) is distorted in the word  $[cha:d]$  when compared with viseme  $[a:]$  (red ellipsoid) in the word  $[ba:t]$ . A large region is required to describe the viseme  $[a:]$  in these two different words.

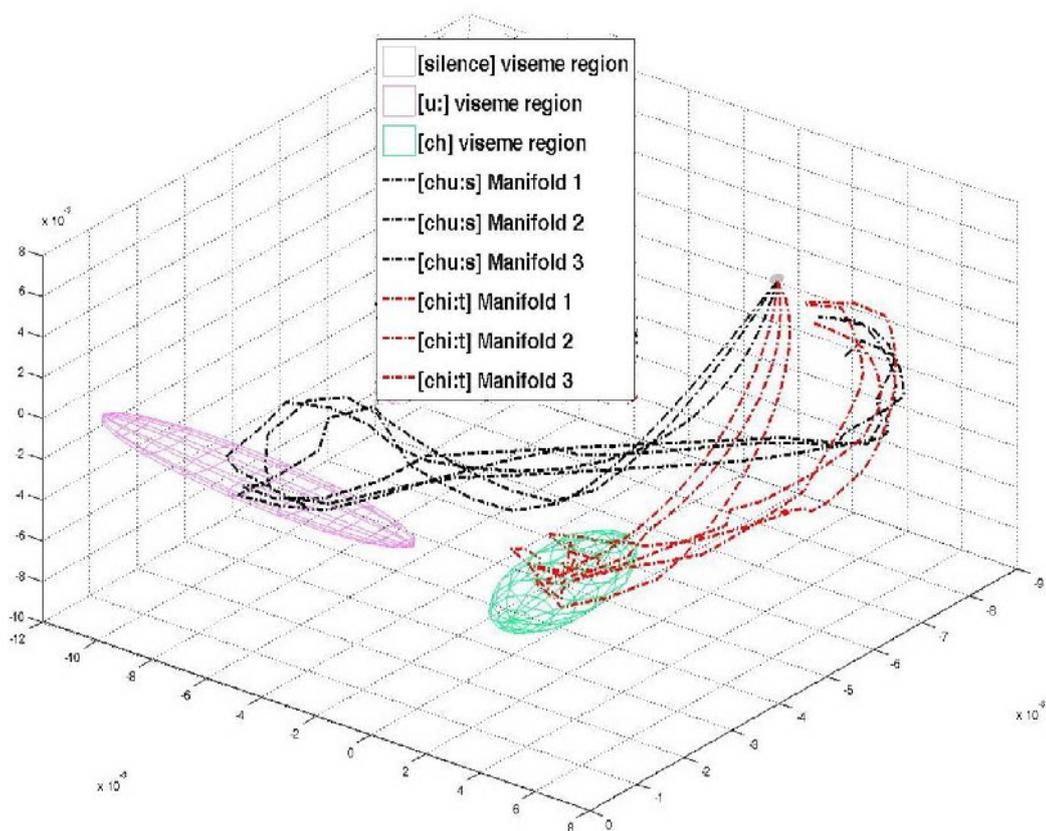
Note 2: viseme  $[d]$  (green) in word  $[cha:d]$  and viseme  $[t]$  (dark green) in word  $[ba:t]$  are in the same category of visemes and they require a larger region in the feature space.

Note 3: in the manifold representation of the word ‘chard’, the viseme  $[a:]$  is distorted and the consonant  $[r]$  cannot be distinguished.

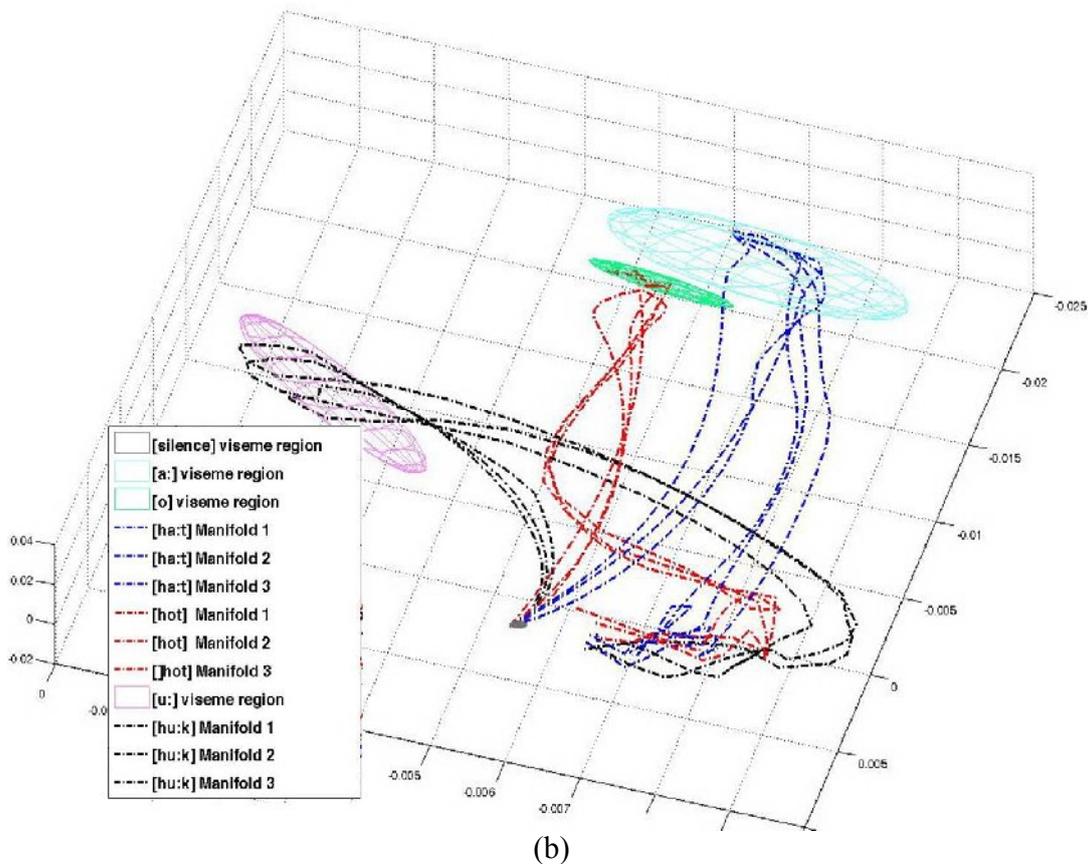
Another limitation of the viseme-based representation is that some visemes may be severely distorted and even may disappear in the video sequences that describe visually the spoken words [41, 65-66]. As mentioned above, the viseme may suffer distortions during continuous speech (see section 4.2.2) and in addition the mouth shapes that define some visemes may be difficult to detect in the EM-PCA space. In other words, some

visemes may be severely distorted when the next or previous visemes are intentionally accentuated in continuous spoken.

These problems can be observed in Fig. 4.6(a), where the viseme  $[ch]$  can be clearly located in the manifold of the word ‘cheat’, but it cannot be located in the manifold of the word ‘choose’. The articulation of  $[ch]$  is produced only by the vocal cords using air-stream and as a result the viseme  $[ch]$  is not visible. In Fig. 4.6 (b) the viseme  $[h]$  is silent (cannot be observed) in words ‘heart’  $[ha:t]$ , ‘hat’  $[hæt]$  and ‘hook’  $[hu:k]$ . The articulation of  $[a:]$ ,  $[æ]$  and  $[u:]$  are typical emphasized in these words and the viseme  $[h]$  is not visible in the words manifolds.



(a)



**Fig. 4.6 Limitations of the viseme-based approach**

(a) The EM-PCA manifolds for words ‘cheat’ [chi:t] (red) and ‘choose’ [chu:s] (black). The viseme [ch] displayed in green is visible in the manifold of the word ‘cheat’, but it cannot be distinguished in the manifold of the word ‘choose’.

(b) The EM-PCA manifold for words ‘heart’ [ha:t] (blue), ‘hat’ [hæt] (red) and ‘hook’ [hu:k] (black). The feature space for viseme [a:] is depicted in cyan, for viseme [æ] in green and for viseme [u:] in purple. Viseme [h] cannot be distinguished.

These limitations indicate that visemes do not map accurately the lip motions and they are subjected to a large degree of distortion when evaluated in continuous speech sequences. In conclusion, the viseme model is not optimal when applied to continuous visual speech recognition. Thus, in this thesis a new representation is proposed that extends the viseme model by including the transitions between visemes. This new representation is called Visual Speech Unit and will be detailed in the next section.

### 4.3 Visual Speech Unit Representation

As indicated in Section 4.2 there is no consensus among vision researchers about how the sets of visemes in English are constituted [1, 27, 60] and in the previous section it has been shown that visemes are not efficient models to be used for continuous visual speech recognition. This is the fact that they cover only a small portion of the words manifolds and they may be severely distorted by the preceding visemes during the continuous speech process (see Section 4.2.3).

In this thesis, a new representation called Visual Speech Unit (VSU) is proposed. Each VSU is manually constructed from the word manifolds and it has three distinct states: (a) articulation of the first viseme, (b) transition to the next viseme, (c) articulation of the second viseme. This can be observed in Fig. 4.7.

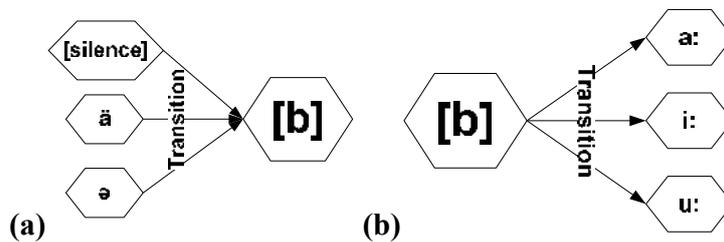
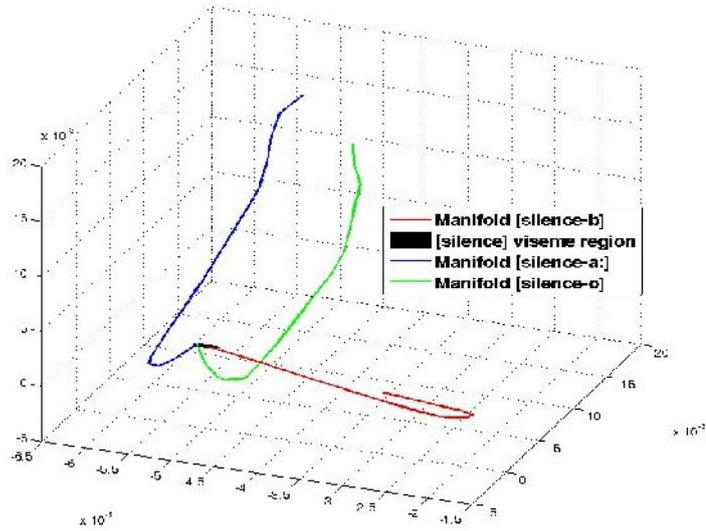


Fig. 4.7 Examples of Visual Speech Units

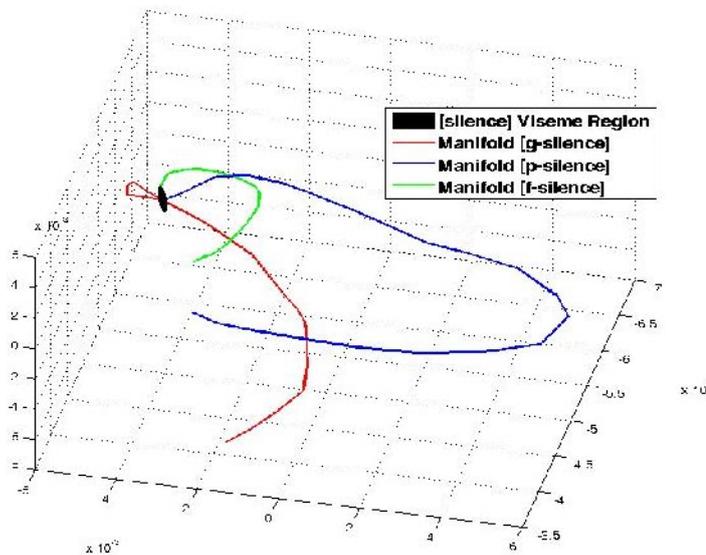
(a) VSUs: [silence -b], [ä-b] and [ə-b] (b) VSUs: [b-a], [b-i] and [b-u].

It is important to note that in the approach detailed in this thesis the MPEG-4 viseme set is used to construct VSU models. In this approach, the state [silence] is considered as an independent class of viseme. This is motivated by the fact that the speech process starts from [silence] and then the word is articulated (consisting of one viseme or more) and ends in [silence]. Fig. 4.8 (a) shows the manifolds constructed for VSUs [silence-b], [silence-a:] and [silence-o:] which are extracted from words ‘bart’, ‘heart’ and ‘hot’. Fig.

4.8(b) shows the manifold examples for VSUs  $[g\text{-silence}]$ ,  $[p\text{-silence}]$  and  $[f\text{-silence}]$  extracted from words ‘charge’, ‘cheap’ and ‘beef’.



(a)



(b)

**Fig. 4.8** Manifold examples for VSUs containing the viseme [silence].

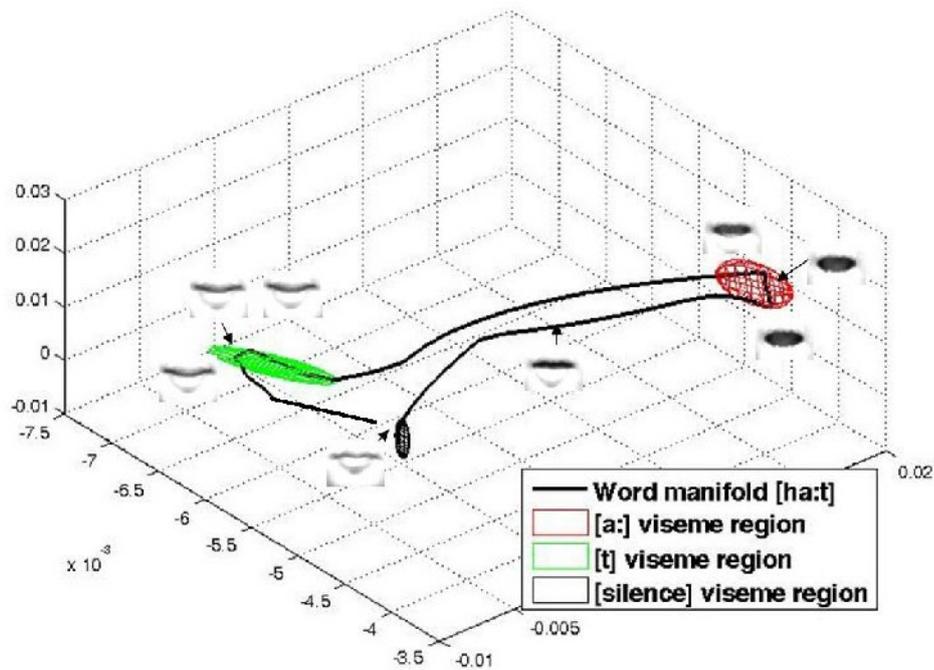
(a) Manifold examples for [silence] to visemes [b] ([ba:t]), [a:] ([ha:t]) and [o] ([hot]).

(b) Manifold examples for visemes [g] ([cha:g]), [p] ([chi:p]) and [f] ([bi:f]) to [silence].

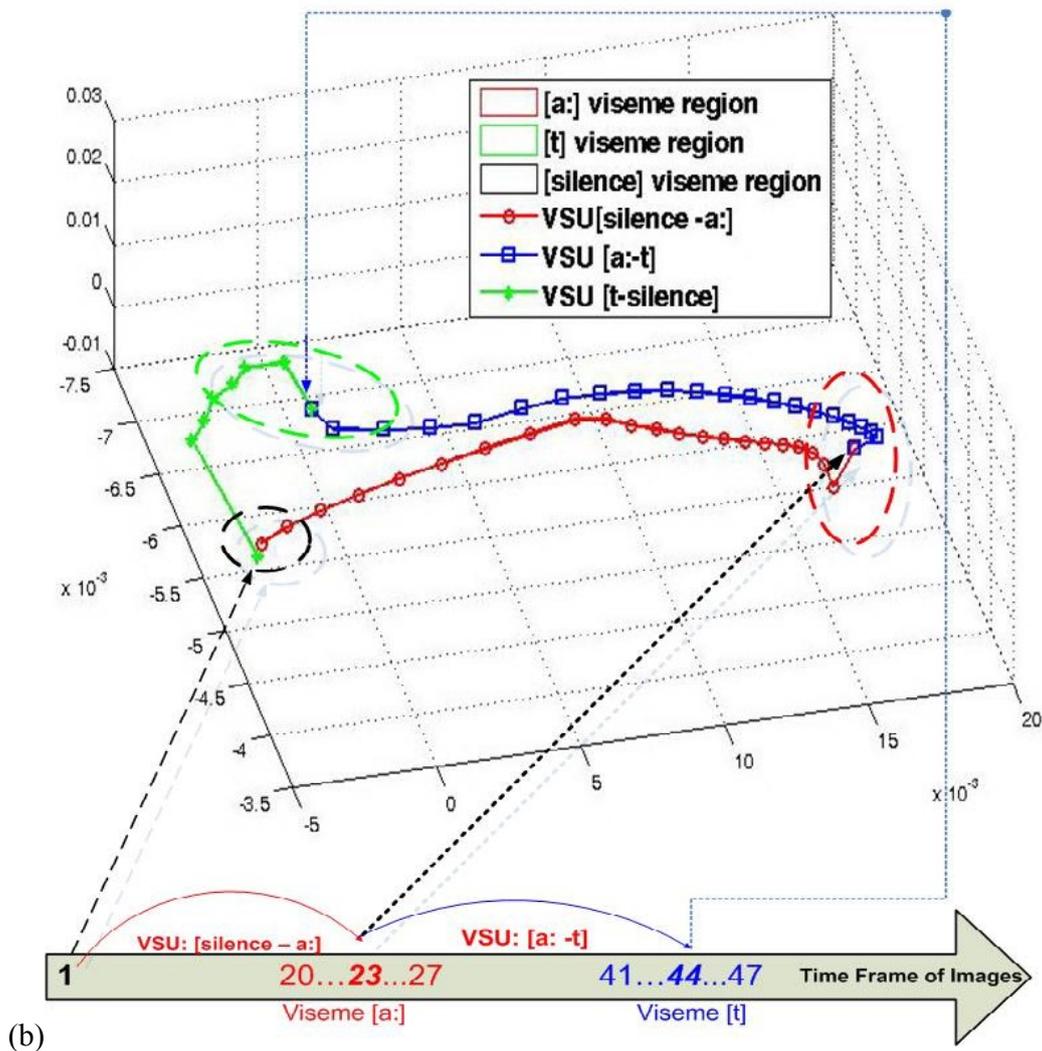
The diagrams depicted in Fig. 4.8 indicate that transitions from *[silence]* or transitions to *[silence]* can be used to detect the start or the end section of the words that are described visually in the video sequence. As a result, this information is used to perform the registration between VSUs and the word manifold. This will be described in Section 4.3.2.

### 4.3.1 Generation of Visual Speech Unit Models

Each VSU is manually constructed from word manifolds using the viseme information and the transition information between consecutive visemes. In this regard, the corresponding feature points for consecutive visemes are first segmented based on word manifold representation (see Section 4.2.2). The start point of VSU is estimated as the center feature point that is related to the articulation of the first viseme and the end point of VSU is estimated as the center feature point related to the articulation of the second viseme. An example that illustrates the construction of the VSU is in Fig. 4.9.



(a)



**Fig. 4.9 Examples of Visual Speech Units**

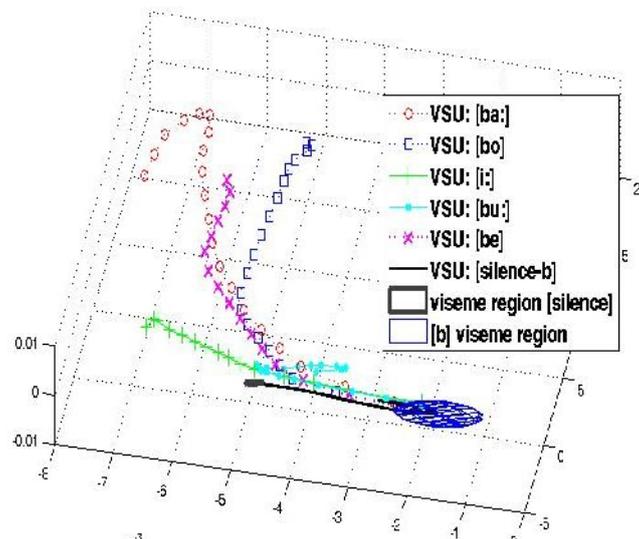
(a) The word ‘heart’ ([ha:t]), word manifold (black line) and all visemes [silence] (black ellipsoid), [a:] (red ellipsoid) and [t] (green ellipsoid). Note that viseme [h] is not visible (see Section 4.2.3 and Fig. 4.6a).

(b) VSUs Segmentation: [silence-a:] (red manifold), [a:-t] (blue manifold) and [t-silence] (green manifold).

As mentioned before, visemes may be distorted during the continuous speech process and this generates a real problem when visemes are applied to construct VSUs. For instance, the word ‘heart’ [ha:t] can be constructed using the following viseme sequence: [silence], [h], [a:] and [t]. Using the VSU representation the word ‘heart’ is constructed

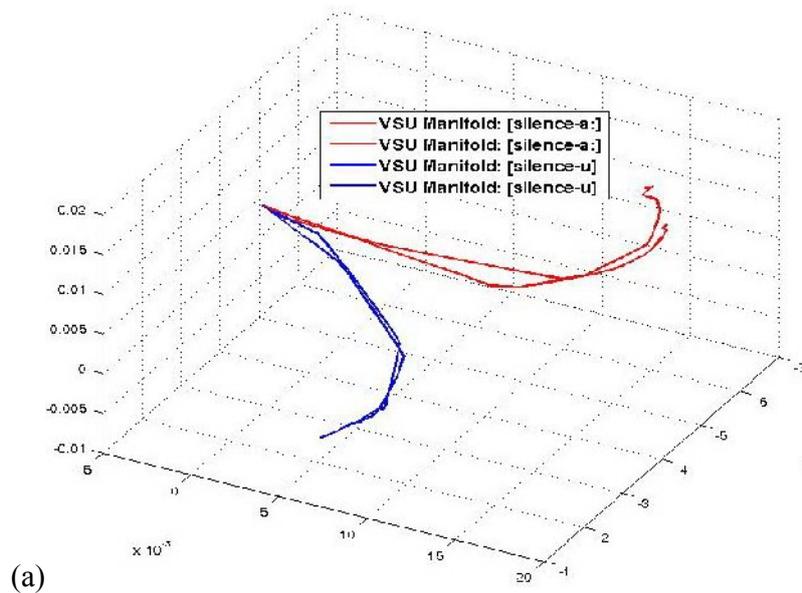
using the following sequence of VSUs:  $[silence-h]$ ,  $[h-a:]$ ,  $[a:-t]$  and  $[t-silence]$ . In practice, viseme  $[h]$  cannot be identified in the visual domain and all we observe is a continuous articulation from viseme  $[silence]$  to  $[a:]$ . To address this problem, in the approach detailed in this thesis, the construction of VSUs is based on adjacent the visemes that can be identified in the word manifolds (or visemes describe the articulation process (lip movements) that can be observed in the visual domain). In the manifold representation, the visemes that can be observed in the visual domain are represented as a unique region in the EM-PCA feature space. Using this approach, the VSUs associated with word  $[ha:t]$  are:  $[silence-a:]$ ,  $[a:-t]$  and  $[t-silence]$  and they are displayed in Fig. 4.9.

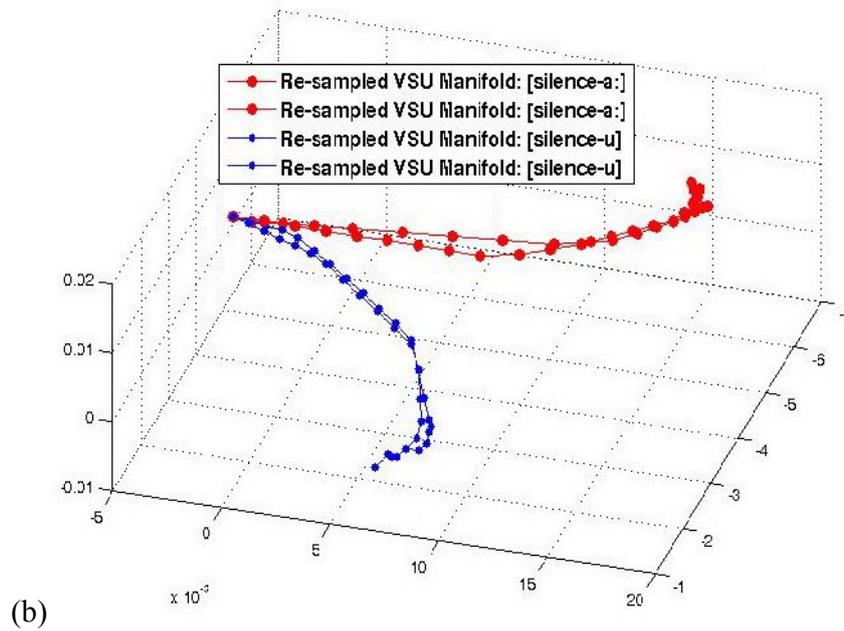
To further illustrate the construction of VSUs, a number of additional examples are depicted in Fig. 4.10. From these examples it can be clearly observed that VSUs do not include only the lip motions associated with particular visemes but also the transitions between adjacent visemes.



**Fig. 4.10 Examples of Visual Speech Units. The EM-PCA manifolds of VSUs:  $[silence-b]$ ,  $[b-o]$ ,  $[b-u]$ ,  $[b-i]$ ,  $[b-e]$ .**

To apply the VSU representation to visual speech recognition, we construct a mean model for each class of VSU. Given a testing sequence (a “word manifold”) that describes one word, and a set of VSUs, we can not compare them directly to each other since they are different objects (VSU is an element of a word). Due to this reason, the manifold has to be divided into a number of sub-sections, and corresponding regions of each sub-section are registered between mean model of all possible VSUs and word manifold (registration VSUs will be detailed in section 4.3.2). To facilitate this process, the interpolated word manifolds (see Chapter 3) are re-sampled uniformly into a fixed number of feature-points. In order to generate standard manifolds for training and recognition tasks, the re-sampling procedure will generate a pre-defined number of key-points that are equally distanced on the interpolated manifold surface. This re-sampling procedure ensures the identification of a standard set of feature key-points as illustrated in Fig. 4.11. (Appendix E provides more examples of VSU representations.) .

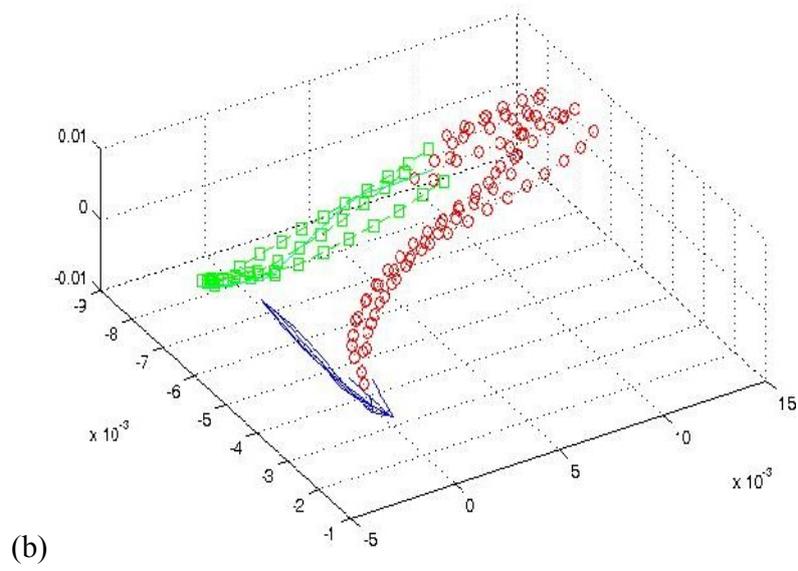
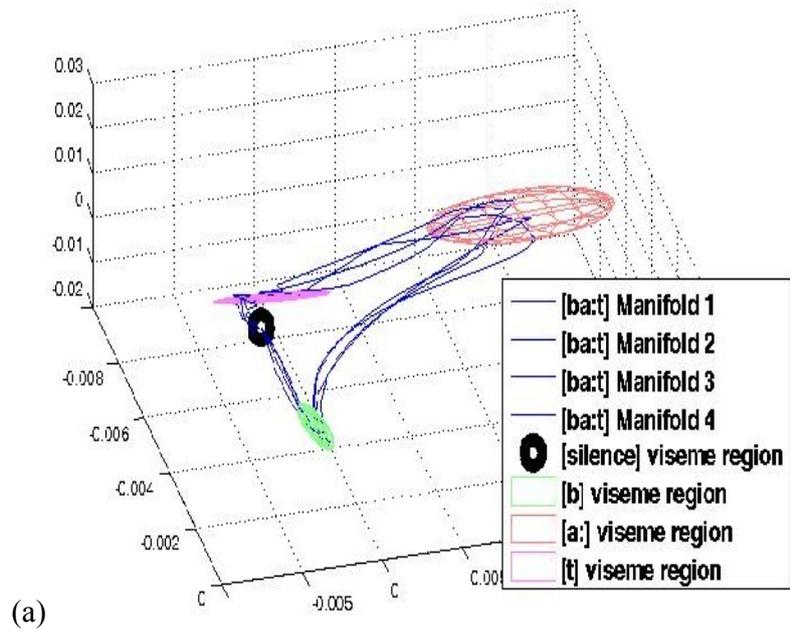


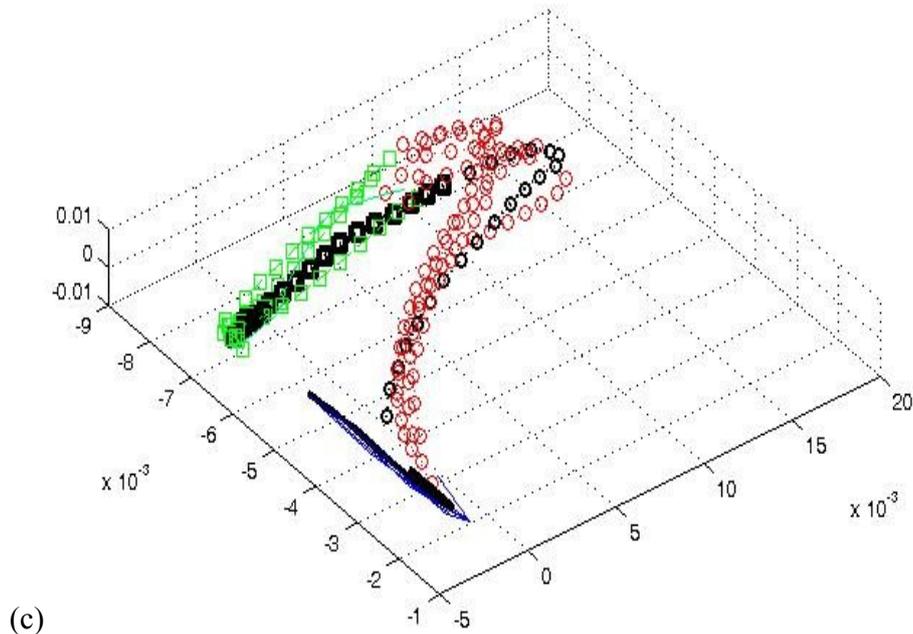


**Fig. 4.11 VSU Manifold re-sampling process.**

- (a) Two manually constructed manifolds of VSU [silence-a:] (red) and two manifolds of VSU [silence-u] (blue)
- (b) Re-sampled manifolds for all VSUs by using 20 equally distanced key-points (red and blue points)

In this way, the VSUs are obtained by manually extracting the corresponding manifold from the word manifolds. For each VSU, 5 manifolds are extracted from five instances of the same word and they are used to calculate the mean model. This manual procedure is followed by the calculation of the mean model as illustrated in Fig. 4.12 (in our implementation all VSU manifolds have been uniformly re-sampled into 20 key-points).





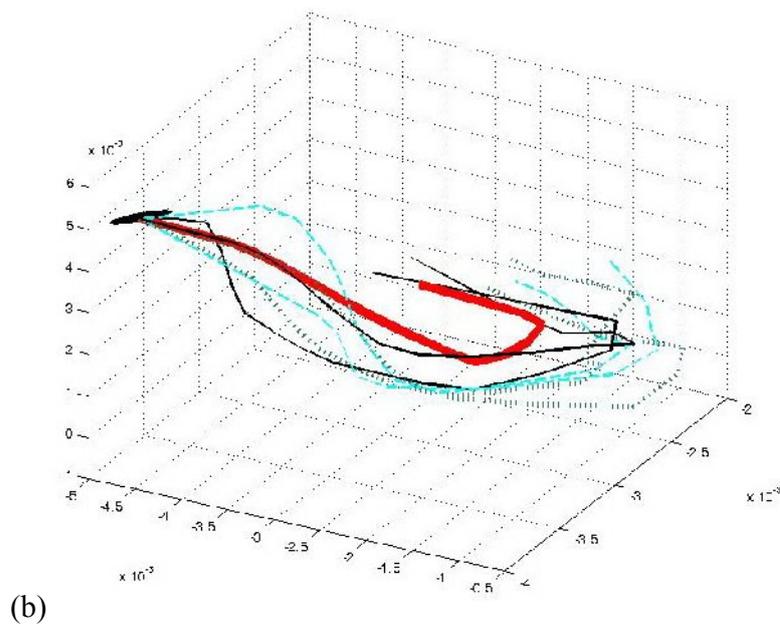
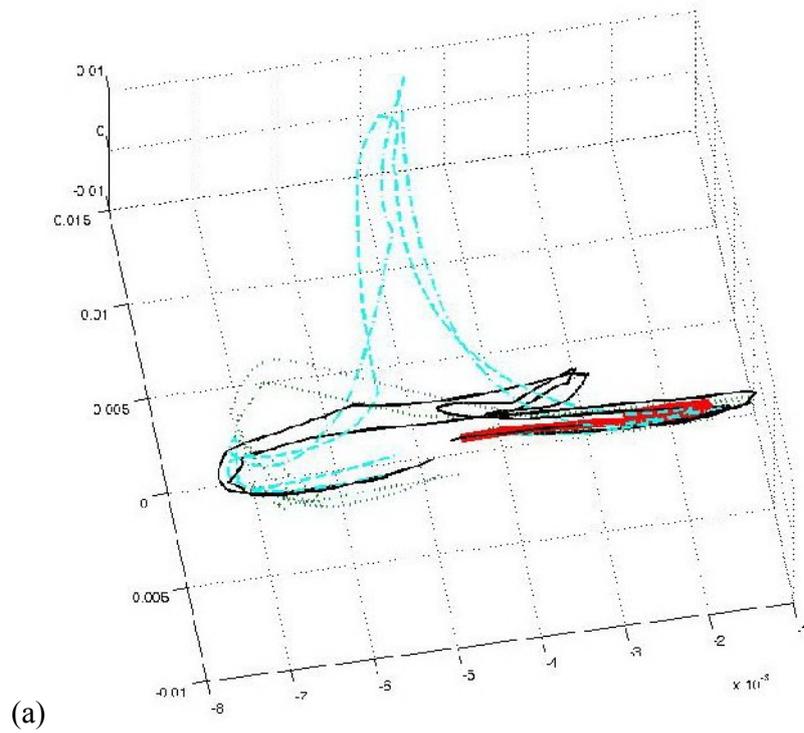
**Fig. 4.12 The calculation of VSU Mean Models.**

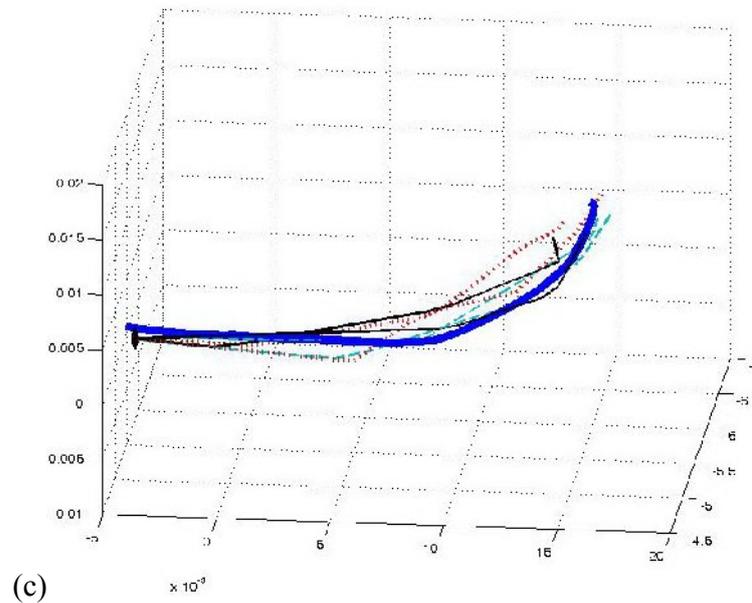
- (a) Four manifolds of the word [ba:t] displayed in blue, where the four visemes (can be observed in visual domain) are shown as follows: [silence] in black, [b] in green, [a:] in red and [t] in purple.
- (b) The VSU extracted from the re-sampled manifolds. [Silence - b] (blue points), [b-a:] (red points) and [a:-t] (green points).
- (c) The mean model for all VSUs are marked in black in the diagram ([silence-b] – black line, [b-a:] – black circles and [a:-t] - black squares).

The calculation of the VSU mean models is illustrated in Fig. 4.13. In Fig. 4.13 (a) and (b), the calculation of the mean model for VSU [silence-b] from five examples of the word [ba:t] is shown. In Fig. 4.13 (c), the calculation of the mean model for VSU [silence-a] from five examples of the word [ha:t] is illustrated.

In Fig. 4.13 (a), the mean model of VSU [silence-b] is compared against the VSUs extracted from word manifold [bu:t], [bot] and [bi:t]. Fig. 4.13 (b) shows that the VSUs extracted from word manifolds are well approximated by the [silence-b] VSU mean model. Fig. 4.13 (c) shows another example where the mean model of VSU [silence-a:] is compared with the VSUs that are extracted from the manifolds of words [ha:t], [ha:f] and

[ha:bi]. As expected, the mean model and the VSUs extracted from the word manifolds have similar shapes. The manifolds of the words shown in Fig. 4.13 are not used to calculate the VSU mean models. (Appendix F shows other 5 VSUs representation for different words)





**Fig. 4.13 The VSU Mean Models and the VSU extracted from different word manifolds.**

- (a) Mean model of VSU [silence-b] (red line) and three word manifolds (two examples each word): [bu:t] (black line), [bõ:t] (cyan line) and [bi:t] (green dot line).  
 (b) The mean model of VSU [silence-b] (red line) and the VSU samples extracted from the word manifolds displayed in (a).  
 (c) The mean model for VSU [silence-a:] (blue line) and VSU [silence-a:] samples extracted from word manifolds (two examples each word): [ha:t] (red line), [ha:lʃ] (black line) and [ha:bi] (cyan line). Note: the mean model of VSU [silence-b] is calculated from 5 examples of the word [ba:t]. The mean model of VSU [silence-a:] is calculated from 5 examples of the word [ha:t].

The VSU mean models depicted in Fig. 4.12 and Fig. 4.13 are used to train a set of HMM classifiers. In the implementation presented in this thesis, to minimize the class overlap one HMM classifier has been trained for each VSU class. In this way, the recognition is viewed as a competitive process where all VSU mean models are registered to the interpolated manifold that is calculated from the input video sequence (see Chapter 3). In other words we attempt to divide the word manifold into a number of consecutive sections, where each section is compared against the mean models of all VSUs stored in the database. To achieve this, we need to register the VSU mean models with the surface of the word manifold. In this work the registration between VSU mean

models and the surface of the word manifolds is carried out using the Dynamic Time Warping (DTW) algorithm.

### **4.3.2 Registration between VSU Model and Word Manifold**

#### **4.3.2.1 Dynamic Time Warping Review**

Dynamic Time Warping (DTW) is a classical algorithm that is applied to identify the optimal fitting (or alignment) between two time-ordered series. The warping between two time series can be used to find their corresponding regions or to determine the level of similarity between them.

Let  $X$  and  $Y$  be two time series, of lengths  $|X|$  and  $|Y|$ , where  $W = w_1, w_2, \dots, w_K$  is the warp path ( $\max(|X|, |Y|) \leq K < |X| + |Y|$ ),  $K$  is the length of the warp path,  $w_k = (i, j)$  is the  $k^{\text{th}}$  element of the path,  $i$  is the index for time series  $X$  and  $j$  is an index for time series  $Y$ . The optimal warp path is calculated by minimizing the fitting cost between the two time series as follows,

$$Dist(W) = \sum_{k=1}^{k=K} Dist(w_{ki}, w_{kj}) \quad (4.1)$$

Where  $Dist(W)$  is the distance (typically the Euclidean distance) is associated with the warp path  $W$ , and  $Dist(w_{ki}, w_{kj})$  is the distance between two data points with indexes  $i$  and  $j$ . The warp path must start at the beginning of each time series and finish at the end of both time series. This ensures that every element of each time series is used in the calculation of the warp path.

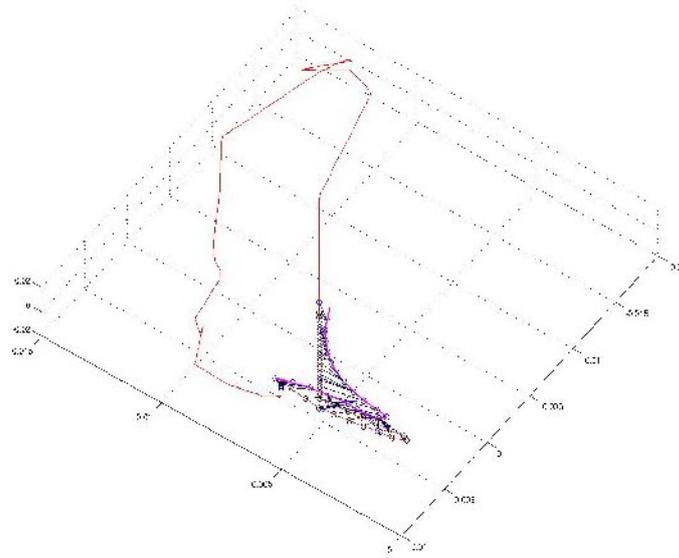
DTW is a simple solution that has been commonly used in the development of VSR systems to determine the similarity between time series and to find corresponding regions between two time series of different lengths [67-70].

#### 4.3.2.2 Registration between VSU and Word Manifold

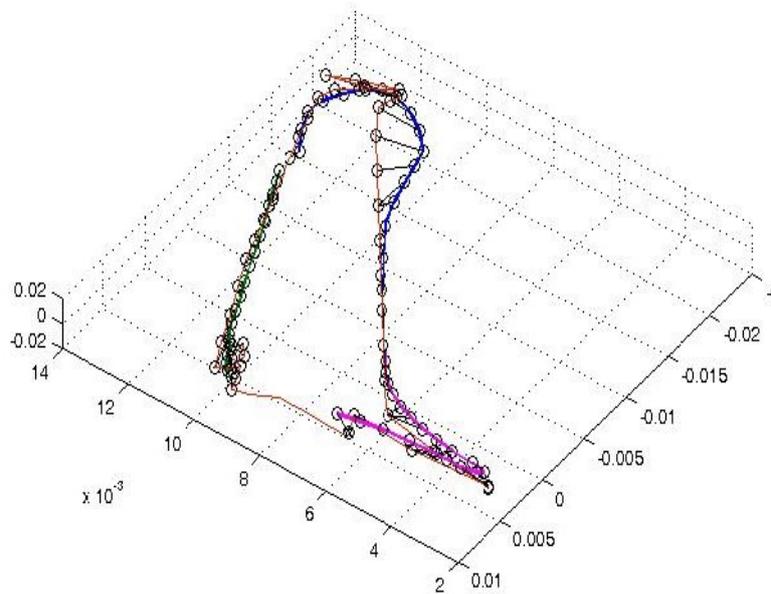
The VSU recognition process is viewed as a two-step approach. In the first step we need to register the VSU mean models to the word manifold using Dynamic Time Warping. Using this approach, the test data (word manifold) is divided into a number of consecutive sub-sections, where each sub-section is compared against the mean models of all possible VSUs. For example, the registration of the first section of the word manifold is always compared against all VSUs that start with [silence]. After the application of DTW, the registered regions are outlined based on the minimum distance between the mean models and the sub-section of word manifold. Once the best registered region is classified, the end point of the classified region is the start point of the next section of the word manifold (i.e.: after [silence-b] is classified, then the next section will be registered against all VSUs that start with [b]).

In the second step we measure the matching cost between the VSU mean models and the registered section of the manifold using HMM classification (in our implementation we have used a three-state HMM classifier (This classification topology is detailed in Chapter 5). For instance, [silence] is the start viseme of the word  $[ba:t]$  and DTW is applied to measure the local distance between the VSU mean model manifold [silence-b] and the word manifold  $[ba:t]$ . The optimal alignment (warping) between these two manifolds via point-to-point mapping is shown in Fig. 4.14.

This procedure is applied for all VSUs contained in the database and the complete registration process of the word  $[ba:t]$  is illustrated in Fig. 4.15.



**Fig. 4.14** Registration using Dynamic Time Warping between the mean model manifold of VSU [silence-b] (purple line) and the word manifold [ba:t] (red line).

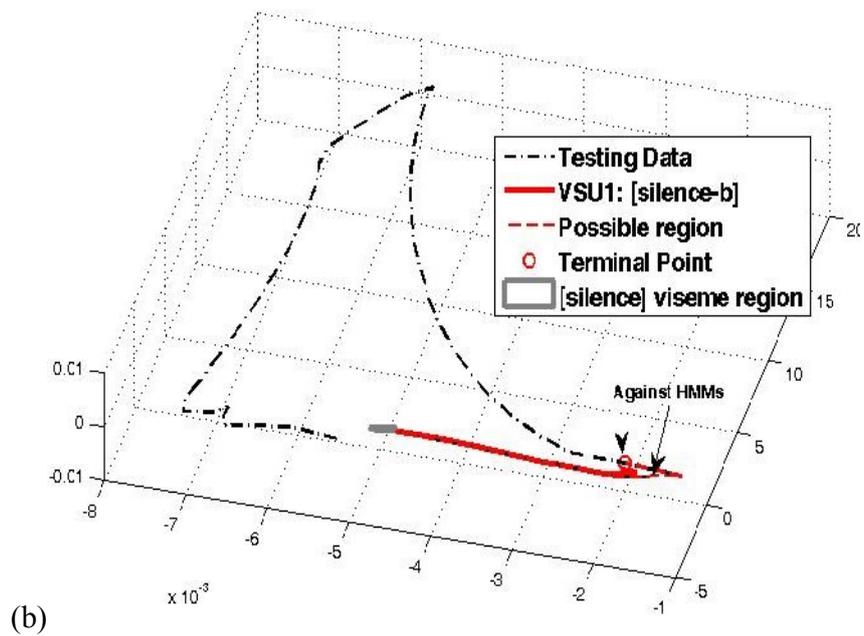
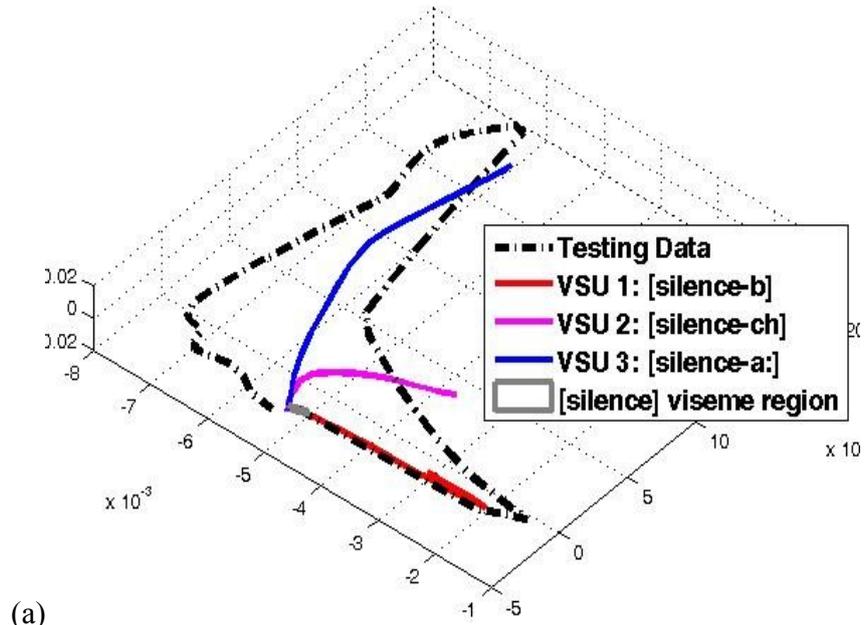


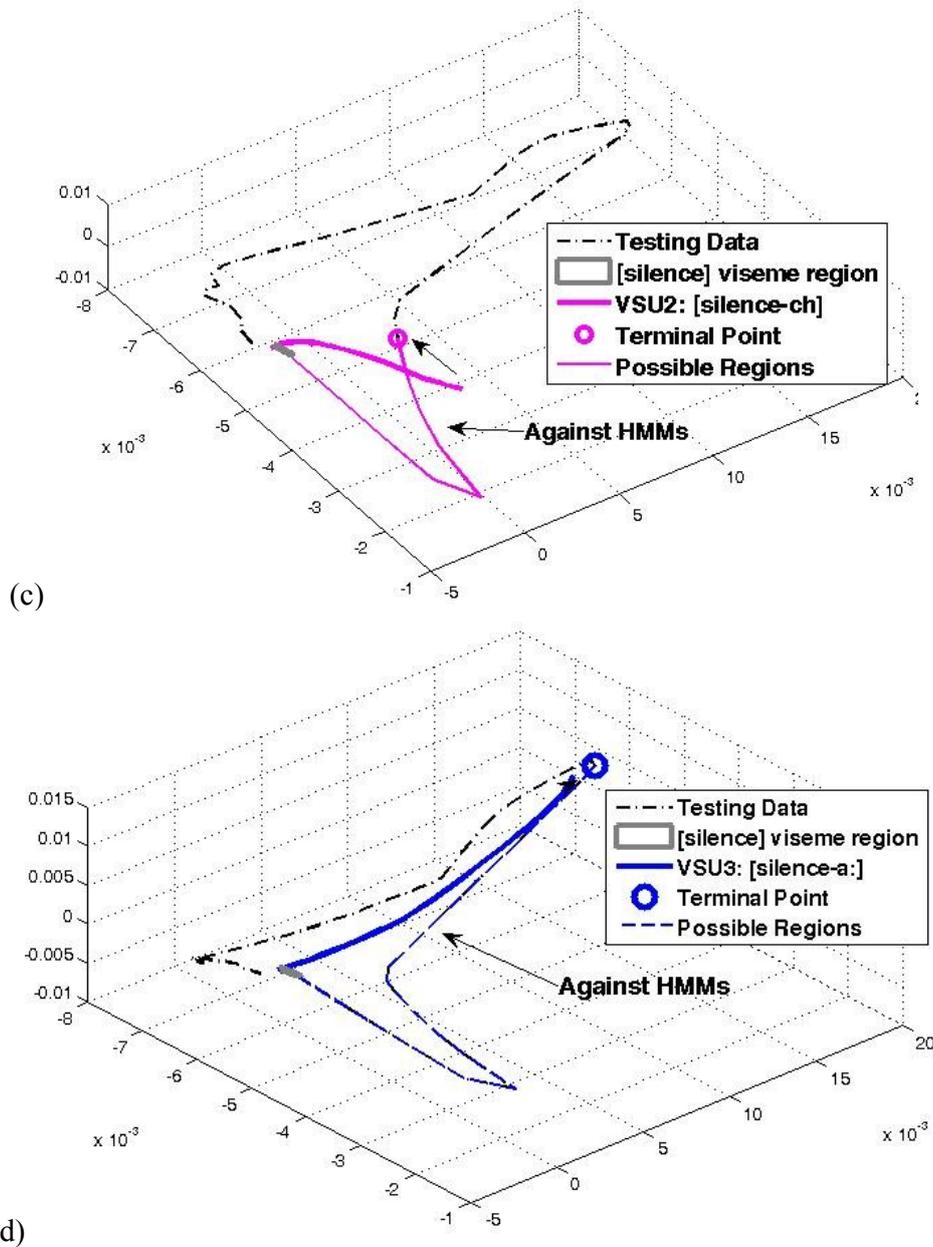
**Fig. 4.15** Complete registration using Dynamic Time Warping between the VSU mean models and the word manifold, [silence-b] (purple line), [b-a:] (blue line), [a:-t] (green line) and the word manifold [ba:t] (red line).

As illustrated in Fig. 4.15, the registration between the VSU mean models and the word manifold is applied iteratively until the last section of the manifold ends with the

state *[silence]* that is common for the beginning and the end of the word (mouth closed).

This process is illustrated step-by-step in Fig. 4.16.



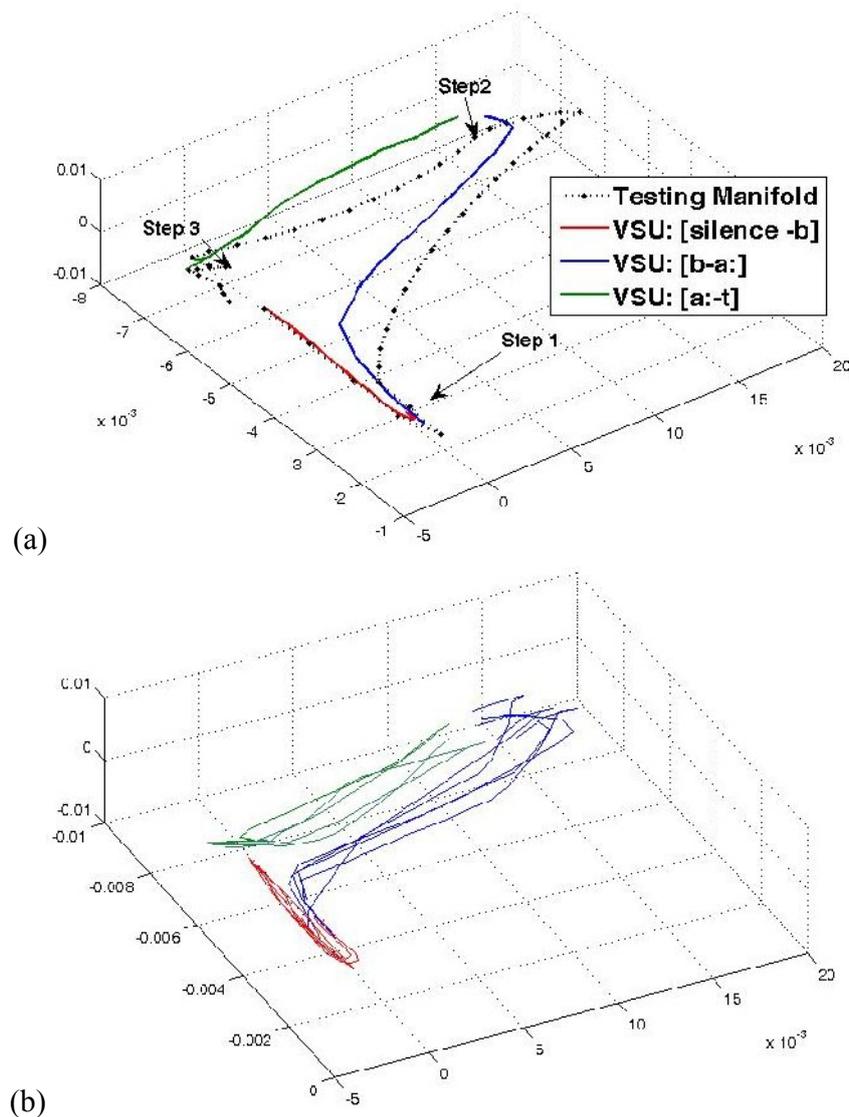


**Fig. 4.16 Step-by-Step VSU registration and classification.**

- (a) The registration of three classes of the VSU Class 1: [silence-b] (red line); Class 2: [silence-ch] (purple line); Class 3: [silence-a:] (blue line) to the word manifold (black dotted line).
- (b) Registration between the [silence-b] VSU mean model and the word manifold.
- (c) Registration between the [silence-ch] VSU mean model and the word manifold.
- (d) Registration between the [silence-a:] VSU mean model and the word manifold.

Note: the registered section of the manifold is used as input for the HMM classifier. The HMM classifier returns the match cost between the input and models contained in the database. In this example, the registration section from [silence-b] VSU mean model achieved the best matching cost (evaluated using a three-state HMM classification).

The example depicted in Fig 4.16 indicates that the VSUs that are registered with the word manifold are identified in succession. For instance in the word  $[ba:t]$ , the end point of the first VSU  $[silence-b]$  is the start point of the second VSU  $[b-a:]$ ; the start point of the VSU  $[a:-t]$  is the end point of the second VSU  $[b-a:]$ . This process is shown in Fig. 4.17.



**Fig. 4.17** The complete registration and matching between the VSU mean models contained in the database and the manifold of the word  $[ba:t]$ . (a) Registration and matching for a single word. (b) Registration and matching for five instances of the same word.

### **4.3 Summary**

Visual speech recognition is a difficult task that involves the identification of the visual speech elements based only on the visual information associated with the lips movements. The choice of the visual speech element is one of the key issues in the development of VSR systems. In this chapter, a comprehensive review of the viseme model reveals several shortcomings associated with this speech representation that can be summarized as follows:

1. There is no widely accepted consensus among researchers in regard to the optimal set of visemes.
2. Viseme representation is not able to fully characterize continuous speech (i.e. transitions between visemes are not used in this representation).
3. Visemes may be severely distorted or they may even disappear during the continuous speech process.

To address these issues, a new speech element that is referred to as a Visual Speech Unit is proposed in this thesis. VSU extends the standard viseme concept by including in this representation not only the viseme information but also the transitions between consecutive visemes. The main advantages of VSUs can be summarized as follows:

- VSUs maximize the use of information present in the word manifold.
- Transition from or to [silence] state can be used to identify the beginning and the end of the word manifold.
- VSUs are constructed using only visemes that can be observed in the visual domain.

- The VSUs are robust speech elements that show good stability when extracted from different words (i.e. VSU [silence-b] has similar characteristics when extracted from words such as [ba:t], [bi:t] or [bi:f]).

In this implementation, VSUs are manually constructed by extracting the key-points of interest from the word manifolds and they are described by the mean models that are calculated for each class of VSU. The VSU recognition process is a two-step-approach. In the first step the mean models of VSU are registered to the word manifold using the Dynamic Time Warping procedure that attempts to divide the word manifold into a number of consecutive VSUs. In the second step, the matching cost between the VSU mean model and the registered section of the word manifold is calculated using HMM classifiers.

To fully assess the discriminative power of the proposed model, we tested up to 60 VSUs that were recorded by two different speakers. In the next chapter, a large number of experiments will be conducted to evaluate the feasibility of the new speech model when applied to visual speech recognition tasks.

# Chapter 5

## Experimental Results

### 5.1. Introduction

The previous chapter discussed the proposed Visual Speech Unit (VSU) that extends the standard viseme concept by including in this new speech representation the transitions between consecutive visemes. The VSU recognition process consists of two main steps. In the first step, Dynamic Time Warping (DTW) is applied to register the mean models for each VSU class to the interpolated manifold that is calculated from the input video sequence. In the second step, the matching cost between the VSU mean models and the registered section of the manifold is calculated using Hidden Markov Model (HMM) classifiers (the HMM classification scheme that is included in the development of the proposed VSU-based VSR system is detailed in Section 5.5).

The aim of this chapter is to evaluate the accuracy of the recognition process when used in conjunction with the proposed VSU speech representation. These experiments were conducted on a set of words that are depicted in Table 5.1. The experimental tests were divided into three sets. The first set of experiments (Experiment 1) was conducted to evaluate the accuracy of the VSU models when compared with the performance attained by standard MPEG-4 visemes. The aim of the second set of experiments (Experiment 2) is to evaluate the performance of the VSU recognition with respect to the number of samples used to train the HMM classifiers. The performance of the proposed VSR system has been evaluated on data produced by two speakers (see Table 5.1). The

last set of experiments is to evaluate the performance of word recognition using VSU models that compared with using viseme models. When VSU and viseme are separately employed as the basic visual speech element for the word model, two different decision algorithms are used to identify 15 words in proposed system.

## **5.2. Description of Database**

For evaluation purposes a database generated by two Chinese speakers has been created. This database consists of 50 words where each word is spoken 10 times by speaker one and 20 words where each word is spoken 6 times by speaker two. In our database we have included simple words such as ‘boat’, ‘heart’, ‘check’, etc. and more complex words such as ‘babie’, ‘hover’, ‘bookman’, ‘chocolate’, etc (see Table 5.1). In our study we have conducted experiments to evaluate the recognition rate based on 12 classes of visemes (see Table 5.2) and 60 classes of VSUs (see Table 5.3). The video data has been captured using a SONY DCR-HC19E camera recorder with a sampling rate of 25 frames per second. The size of each image is [320\*240] and the images are captured in the standard RGB colour format. The database used to evaluate the performance of the VSR system consists of more than 40,000 colour images. Examples of images contained in the database are shown in Appendix B.

**Table 5.1: Words Database**

Speaker	Words
1	Bart, boat, beat, bet, bird, boot, barbie, book, beef, barge, birch, bookman, batch bobby, beefalo beautiful, before, heart, hot, heat, hat, hook, harpy, hobby, hoover, half, home, chard, choose, cheat, check, charge, cheap, channel, charming, chocolate, chief, wart, zart, fast, banana, January, truth, part, put, mart, mood, I, bar, card.
2	Bart, boat, beat, boot, heart, hot, heat, hook, charge, choose, heat, check, wart, zart, fat, bar, art, ill, oat, fool.

**Table 5.2: The set of MPEG-4 visemes**

Viseme Number	Phonemes	Example Words	Number of samples
1	[b], [p], [m]	<u>b</u> ut, <u>p</u> art	330
2	[s], [z]	<u>z</u> art, <u>f</u> ast	30
3	[ch], [dg]	<u>ch</u> ard, <u>ch</u> arge	174
4	[f], [v]	<u>f</u> ast, <u>h</u> alf	86
5	[I]	<u>b</u> eat, <u>h</u> eat	148
6	[A:]	<u>b</u> ut, <u>ch</u> ard,	286
7	[e]	<u>h</u> at, <u>b</u> et	136
8	[O]	<u>b</u> oat, <u>h</u> ot	112
9	[U]	<u>h</u> ook, <u>ch</u> oose	104
10	[t, d]	<u>b</u> ut, <u>b</u> ird,	268
11	[h, k, g]	<u>c</u> ard, <u>h</u> ook, <u>b</u> ug	142
12	[n]	<u>b</u> anana, <u>n</u> ight	20
13	[Th]	<u>th</u> ink, <u>th</u> at,	n/a
14	[r]	<u>r</u> ead, <u>r</u> oses	n/a

Note: This table adopts the viseme model established for facial animation applications by MPEG-4, which is an international audiovisual object-based video representation standard [41, 54].

**Table 5.3: 60 classes of Visual Speech Units**

VSU Groups	Number of classes	Example VSUs
Group 1: (Start with [silence])	9	[silence-b], [silence-ch], [silence-z], [silence-f], [silence-a:], [silence-o], [silence-i:], [silence-e], [silence-u:]
Group 2 (End with [silence])	16	[a:-silence], [o-silence], [eu-silence], [u-silence], [k-silence], [i:-silence], [ch-silence], [f-silence], [m-silence], [ng-silence], [ë-silence], [n-silence], [et-silence], [ğ-silence], [s-silence], [ə-silence]

Group 3: (Middle VSU)	35	[b-a:], [b-o:], [b-i:], [b-u:], [b-ə], [b-ë], [a:-t], [a:-b], [a:-f], [a:-ğ], [a:-ch], [o-b], [o-t], [o-k], [i:-f], [i:-p], [i:-t], [u:-t], [u:-k], [u:-f], [ë-t], [f-ə:], [f-o], [k-m], [f-a:], [w-a:], [z-a:], [ə:-t], [e-k], [ə:-ch], [n-a:], [a:-n], [ch-a:], [ch-u:], [ch-i:]
--------------------------	----	--

Note: This table displays the 60 VSU classes used in the experimental evaluation. (60 classes are generated using data produced by Speaker One and 30 classes are generated using data produced by Speaker Two).

### 5.3. Hidden Markov Models

Hidden Markov Models (HMM) are statistical pattern recognition tools that have been widely used in the development of handwriting, speech and video recognition systems. Essentially, the HMM classification performs a partition of a process into a number of discrete states [39], [71].

A Markov chain [72], [73] is a simple finite-state representation in which each state has an associated probability value where the sum of the probability values leaving a particular state is one. In this representation each state has one transition to the next state, a fact that makes the transition process stochastic. The Hidden Markov Model represents a generalization of the Markov chains since HMM is defined as a set of states (where one state is the initial state), a set of output symbols, a set of state transitions and a transition probability map for each state [72], [73], [74]. HMMs are particularly useful when applied for classification of sequential data processing via supervised learning.

As indicated in the literature survey in Chapter 2, the vast majority of vision researchers have adopted HMM classification schemes to solve the visual speech recognition task. In many proposed VSR systems the left-right HMM topology is used, where each state of the HMM is described by a set of mouth shapes and the state transitions represent the probability that a mouth shape will change to another in the visual representation of the speech elements (i.e. visemes or VSUs). The output returned

by the HMM gives information in regard to the sequence of states generated for a particular input data.

## **5.4. Hidden Markov Model Classification**

In this thesis, the viseme or VSU is represented by a time-ordered set of key-points that are obtained by re-sampling the manifolds that are manually constructed from the word manifolds (See Section 4.3). The HMM classification performs the division of the input sequence into a number of discrete states, where the observation sequence  $O_n$  is defined by the key-points of the re-sampled manifold ( $n$  represents the number of key-points calculated for each viseme or VSU manifold). This process is described in Fig. 5.1 (a) where  $O_n$  is associated with a sequence of hidden states  $S_t$ . Experimental studies on lips dynamics indicate that the lips motions associated with VSUs can be partitioned into three states using one Gaussian per state and a diagonal covariance matrix.

- **Visual Speech Unit – HMM States**

The first state describes the articulation of the first viseme of the VSU. The second state is defined by the transition to the next viseme, while the third state is the articulation of the second viseme. Fig. 5.1 (b) illustrates graphically the partition of the VSU into a sequence of three hidden states.

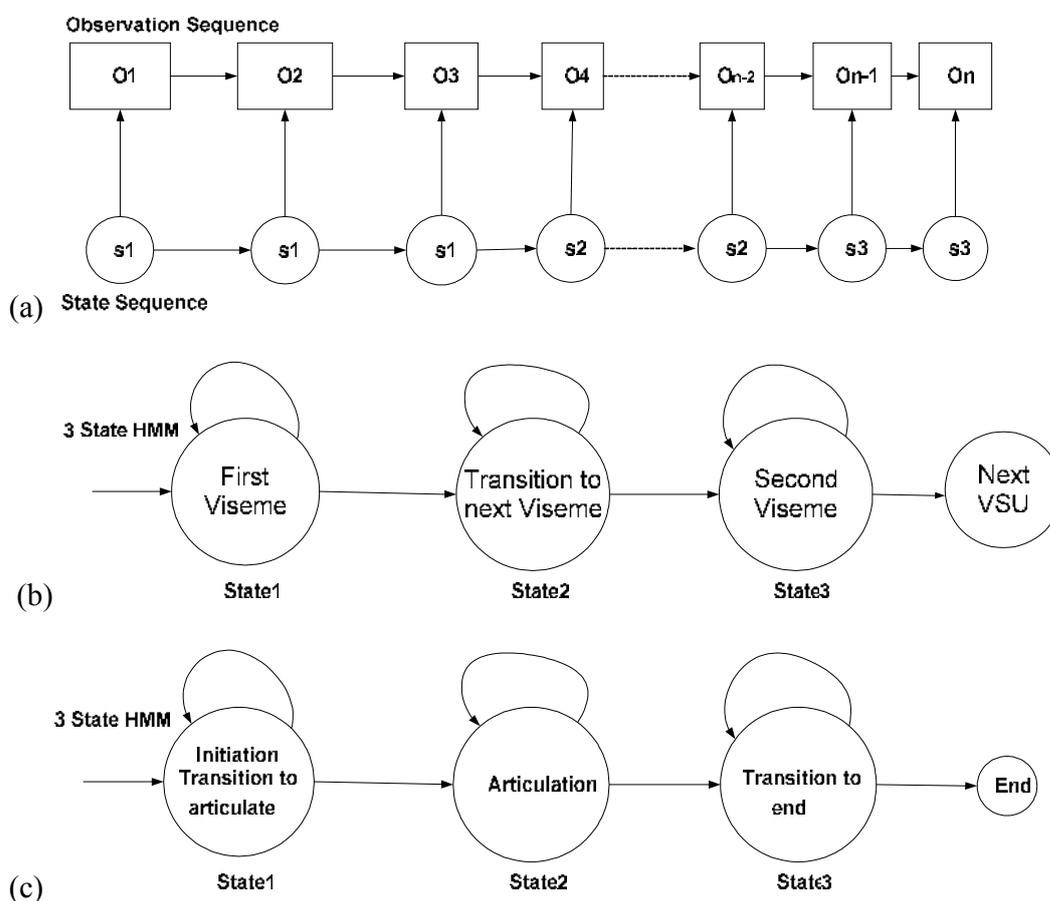
- **Viseme – HMM States**

The representation of visemes using three states HMM classifiers has been adopted by the vast majority of researchers [1, 37-40, 43]. In this work, this approach has been followed and the states generated by each viseme can be described as follows:

- The first state describes the transition from the initial state of the viseme to articulation.

- The articulation state is the part of the viseme that describes the largest variation in lips dynamics.
- The third state is the end part of the viseme when the mouth restores to the relaxed state at the end cycle of the speech process.

Among these states, the articulation provides the highest level of information in discriminating between different visemes. Fig. 5.1 (c) illustrates the partition of the viseme into a sequence of three hidden states.



**Fig. 5.1 HMM topology for VSU and viseme** (a) General observation and state sequence relationship. (b) HMM partition of the Visual Speech Unit into a sequence of three hidden states. (c) HMM partition of the viseme into a sequence of three hidden states [40].

For this implementation, the unknown HMM parameters consisting of transition probabilities and observation probabilities are estimated iteratively based on the training samples using a Baum-Welch algorithm. We have constructed one HMM classifier for each class of VSU and one HMM classifier for each viseme as well. Each trained HMM estimates the likelihood of the inputs given each of the models. The HMM classifier that returns the highest likelihood will map the input visual speech to a particular class in the database. During the training process, the number of hidden states is set to three, the length of sequence is set as the number of key-points and the maximum number of iterations is set to 30. (Appendix G shows the application of the HMM to model the VSUs mouth shapes).

## 5.5. Analysis of the Experimental Results

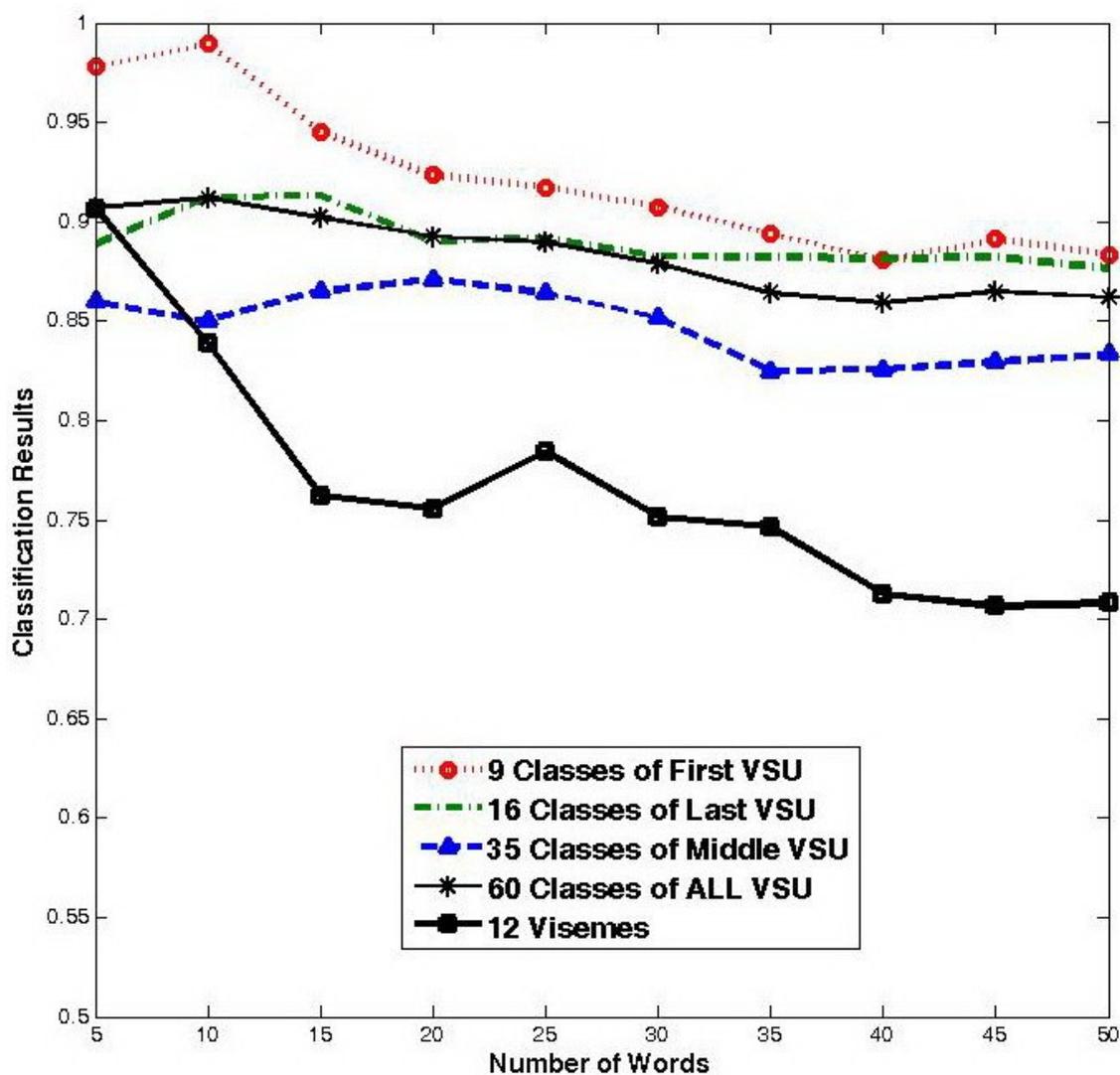
### 5.5.1. Experiment 1: Performance Evaluation for Visual Speech Units and Visemes.

The 60 classes of VSUs listed in Table 5.3 are divided into three distinct groups. The first group is defined by the VSUs that start from the *[silence]* state. The second group is formed by the VSUs whose last state is *[silence]*. The third group consists of “middle” VSUs, which are defined by the articulation of two consecutive visemes and the transitory information between them. The reason to adopt this database segregation is to speed up the recognition process by using the knowledge that the VSUs that contain the state *[silence]* are located either at the beginning or at the end of the word manifold.

This experiment is conducted to evaluate the classification accuracy when visemes and VSUs are employed as speech elements and the number of words in the database is increased. For each VSU, 5 samples are used for training and the others for testing. For

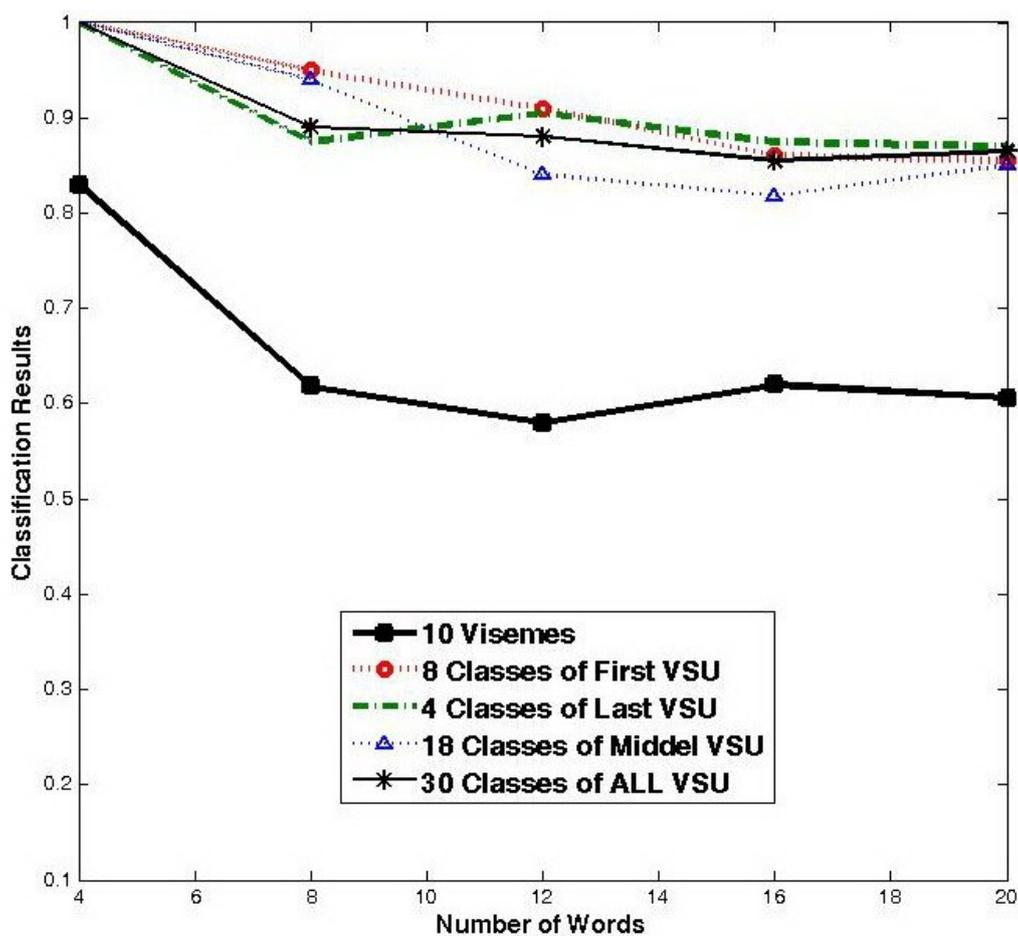
each viseme, half of the samples are used for training and the other half for testing. The classification results for speaker one is depicted in Fig. 5.2 (60 classes of VSUs and 12 classes of visemes). The classification results for speaker two are depicted in Fig. 5.3 (30 classes of VSUs and 10 classes of visemes). Based on the experimental results, it is noticed that the correct identification of the visemes in the input video sequence drops significantly with the increase in the number of words in the database. Conversely, the recognition rate for VSUs suffers a minor reduction with the increase in the size of the database. This drop in recognition accuracy when visemes have been used as speech elements was expected due to the viseme distortion and the occurrence of silent visemes. For example, in the EM-PCA manifold of the word ‘Barbie’  $[ba:bi]$  we can observe that the second viseme  $[b]$  is severely distorted when compared to the first viseme  $[b]$ . In the manifold of the word ‘beat’  $[bi:t]$ , the viseme  $[t]$  is invisible because the mouth is closing fast and in the manifold of the word ‘fast’  $[fa:st]$ , the transition between visemes  $[s]$  and  $[t]$  reveals more information than either of the visemes (see Appendix C for more examples).

During the classification process, it has been discovered that some un-expected registration results occurred when the word manifold is generated under the complex conditions (e.g.: speaker is tired). In these situations, the DTW-based registration failed to track multiple occurrences of the same VSU in complex words (e.g.: ‘banana’). This problem generates most of the classification errors. Fig. 5.4 depicts examples when the DTW-based registration produces correct and incorrect VSU registration results when applied to different word manifolds.



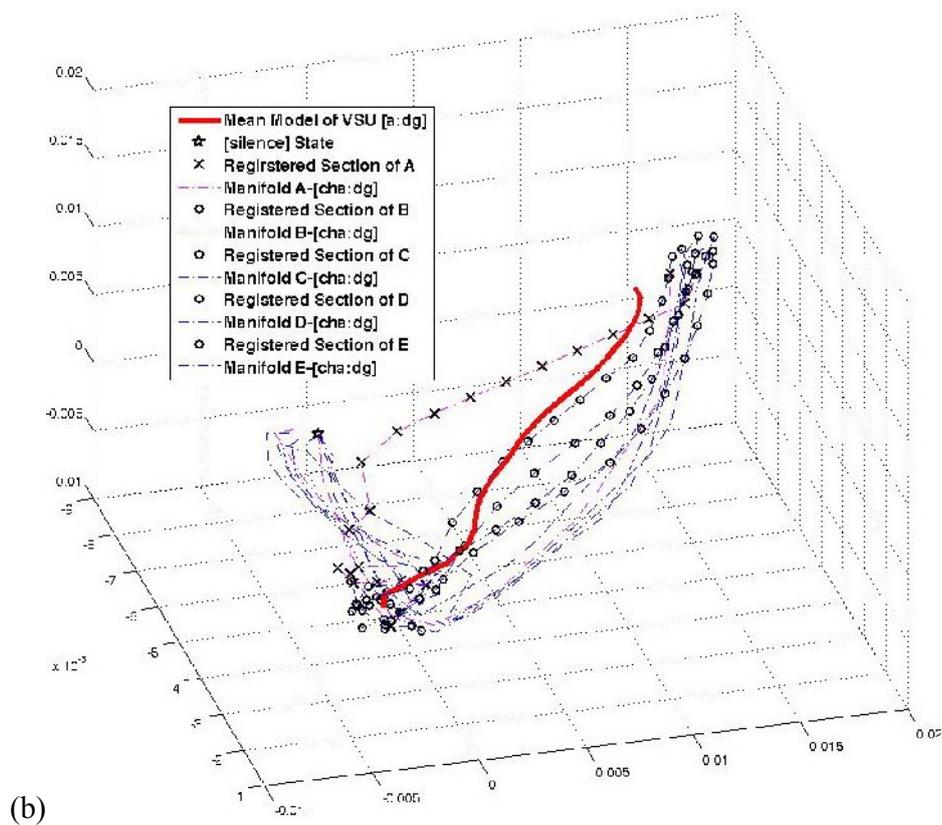
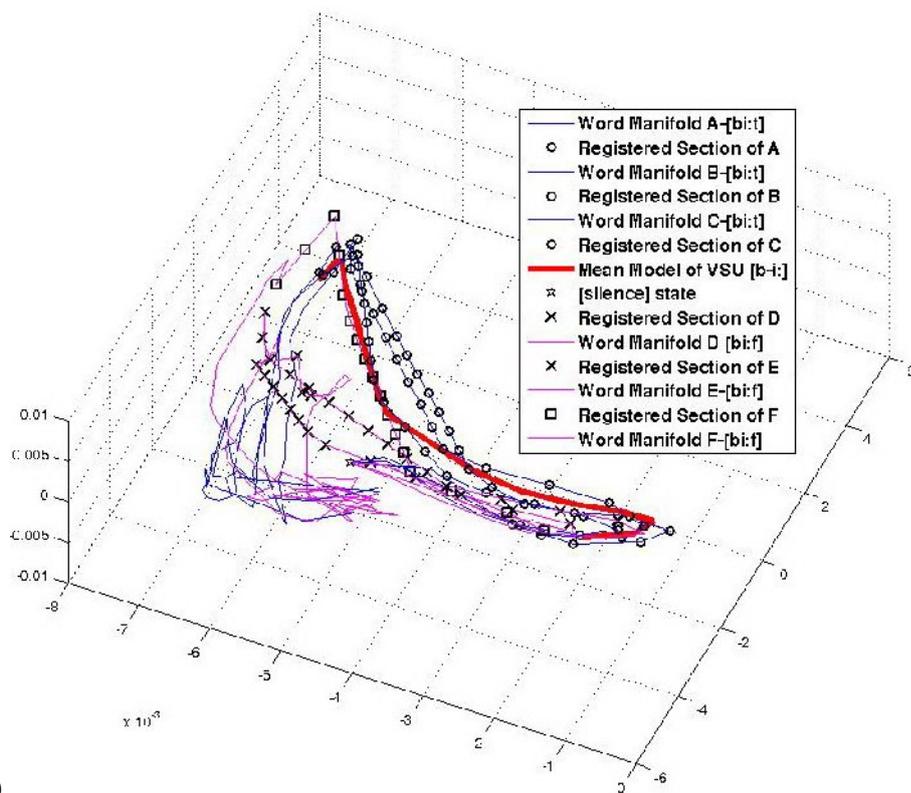
Viseme	[b,p,m]	[s,z]	[ch]	[f,v]	[l]	[a:]	[e,ə]	[o]	[u]	[t,d]	[k,g]	[n]
Average Rate	95%	33%	62%	85%	56%	82%	33%	90%	52%	81%	28%	80%

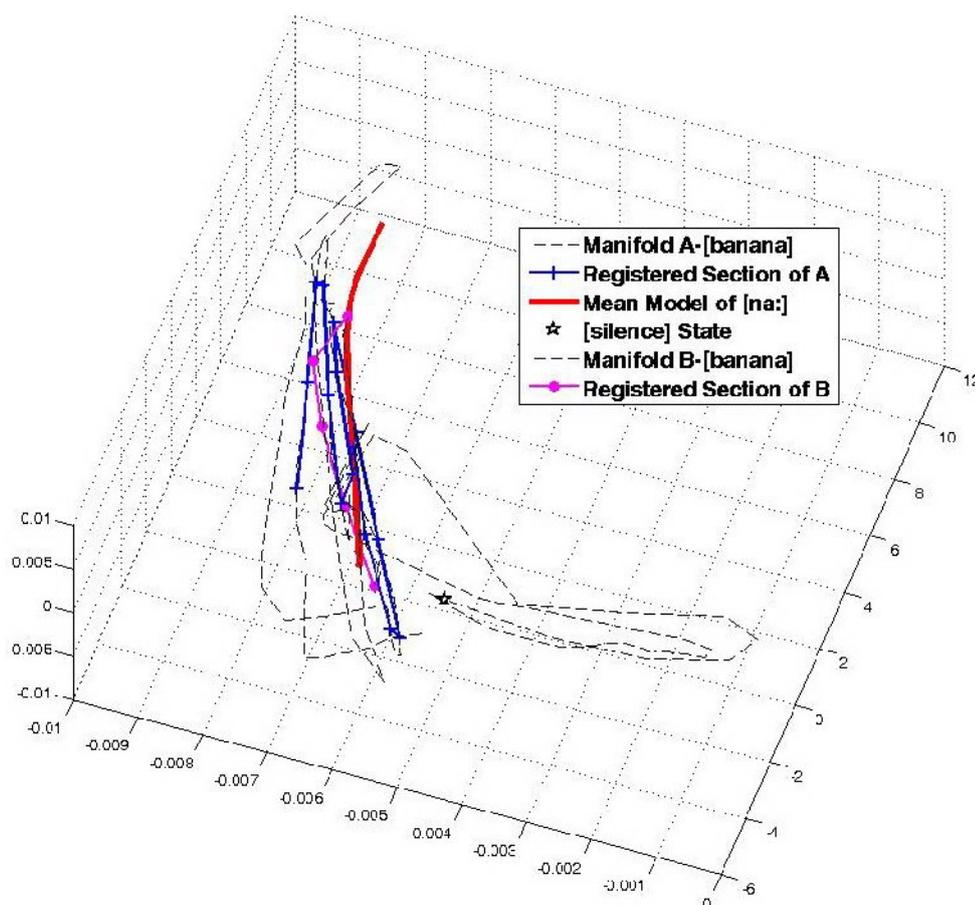
Fig. 5.2 Viseme vs. VSU classification for speaker one. NOTE: The average recognition rate for 12 visemes is 71% while the average recognition rate for 60 VSUs is 88%.



Viseme	[b,p,m]	[ch]	[f,v]	[l]	[a:]	[e,ə]	[o]	[u]	[t,d]	[k,g]
Average Rate	80%	70%	75%	85%	85%	55%	36%	90%	43%	90%

Fig. 5.3. Viseme vs. VSU classification for speaker two. NOTE: The average recognition rate for 10 visemes is 61% while the average recognition rate for 30 VSUs is 86%.





(c)

**Fig. 5.4. Correct and incorrect VSU registration.**

NOTE: All displayed word manifolds are not used to calculate the mean model of VSU.

In the Fig. 5.4 (a), the mean model VSU [b-i:] (red line) is used to register the corresponding sections in three examples of the word [bi:t] (blue line) and three examples of the word [bi:f] (pink line). The registered sections for examples of the word [bi:t] (black cycle line) and one registered section from word [bi:f] (black square line) are correct, while two registered sections for word [bi:f] (black cross on pink line) are incorrectly identified. These miss-registrations are caused by the incorrect articulations for visemes [b] and [i:].

In Fig. 5.4 (b), the VSU mean model [a:-dg] (red line) is used to register the corresponding sections for five examples of the word [cha:dg] (dash-dot line). The classification results for four registered sections (black cycle on blue line) - manifold B to E are correct while one registered section (black cross on pink line) - manifold A is incorrectly classified.

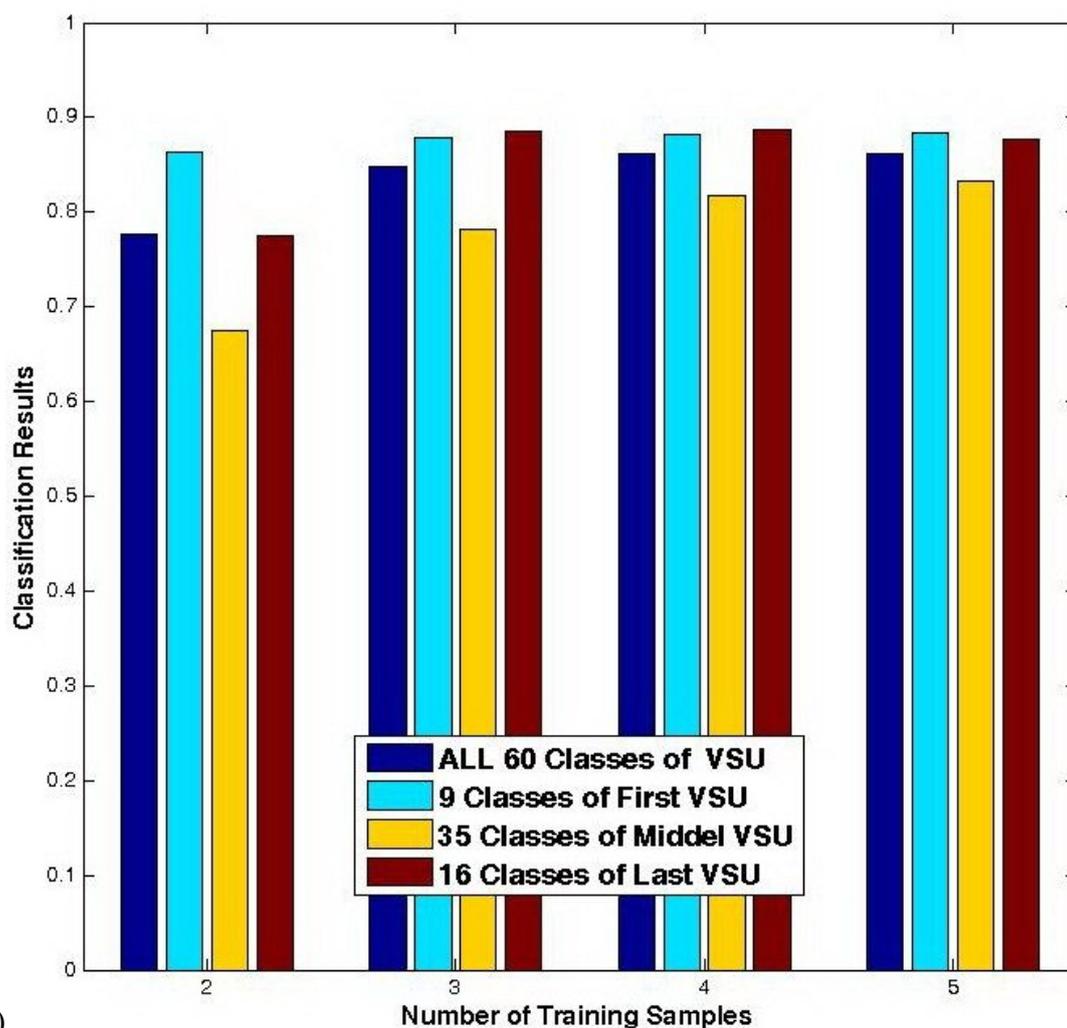
In Fig. 5.4 (c), the mean model of VSU [n-a:] (red line) is used to register the corresponding sections from two examples of the word [banana] (black line). The classification result for registered section (pink point line) of manifold B is correct; the registered section (blue cross line) of the manifold A is incorrectly classified. The reason that caused the registration failure for manifold A is those two sections of the VSU [n-a:] are too closely positioned in the EM-PCA space to allow precise identification. (Appendix F shows more registration examples for VSU [n-a:].)

### **5.5.2. Experiment 2: Performance Evaluation for Visual Speech Units with the Variation in the Number of Training Examples.**

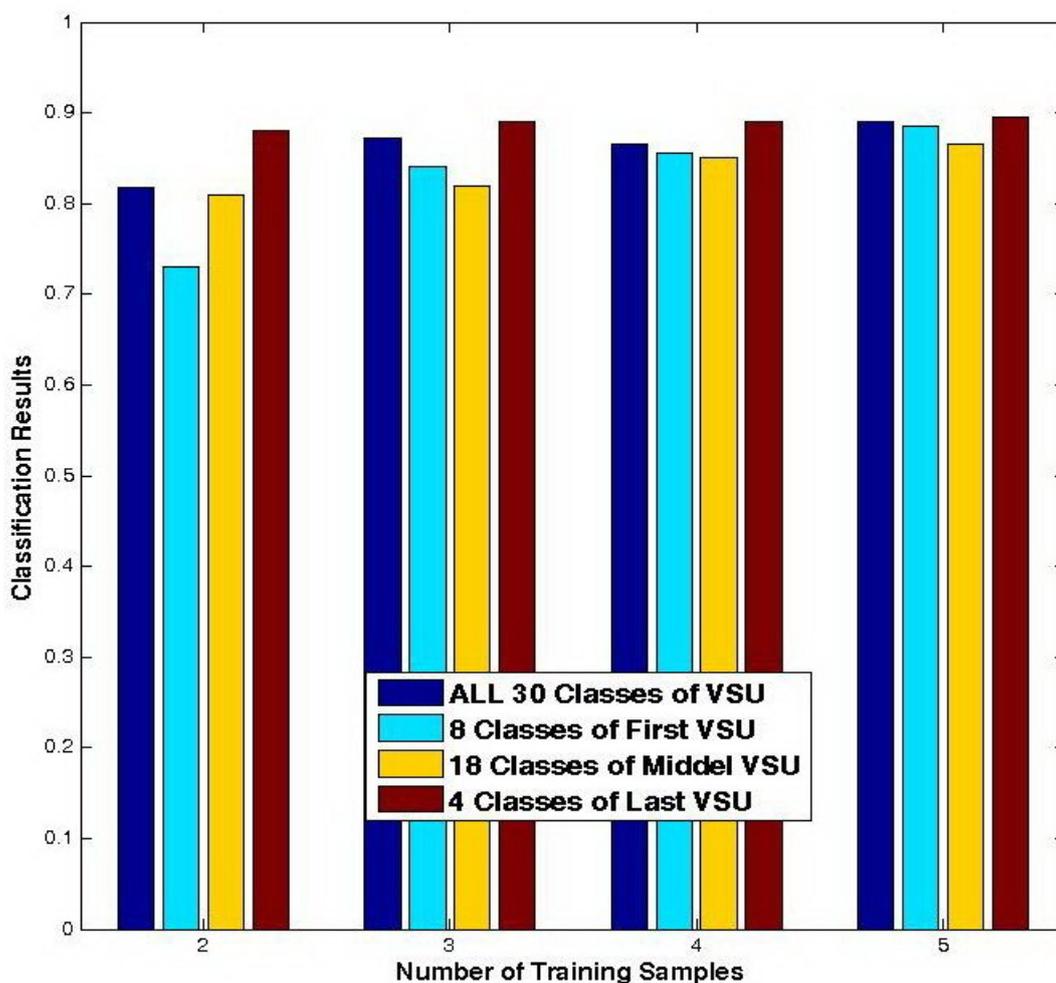
In this experiment we evaluate the recognition rate for each class of VSU when the number of samples employed to train the HMM classifiers is varied. In this experiment, 2, 3, 4 and 5 samples generated by Speaker One are used to train the HMM classifiers for each VSU class and the experimental results of 60 VSUs are shown in Fig. 5.5 (a). For Speaker Two data, 2, 3, 4 and 5 samples are used to train the HMM classifiers for each VSU class and the results of 30 VSUs are illustrated in Fig. 5.5(b).

As expected, the recognition rate is higher when the number of samples used in the training stage is increased. In Fig. 5.5 it can be also observed that the recognition rate for Group 3 (middle VSUs) is lower than the recognition rate for Groups 1 and 2. This is

explained by the fact that the VSUs contained in Groups 1 and 2 starts or ends with *[silence]* and this state can be precisely located in the word manifold.



(a)



(b)

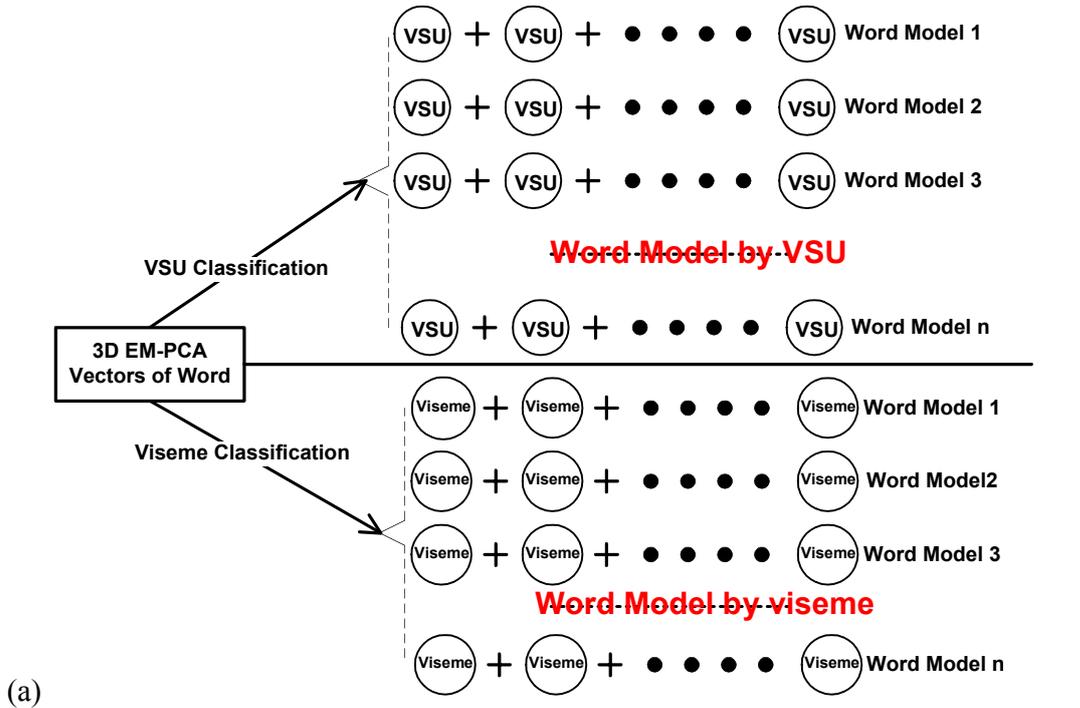
**Fig. 5.5 Visual Speech Unit classification with respect to the number of training examples.**

(a) Speaker One. (b) Speaker Two. In blue the overall recognition rate for all groups is depicted. In light blue the recognition rate for Group 1-First VSUs, in yellow the recognition rate for Group 3-Middle VSUs and in dark red the recognition rate for Group 2 - VSUs are depicted.

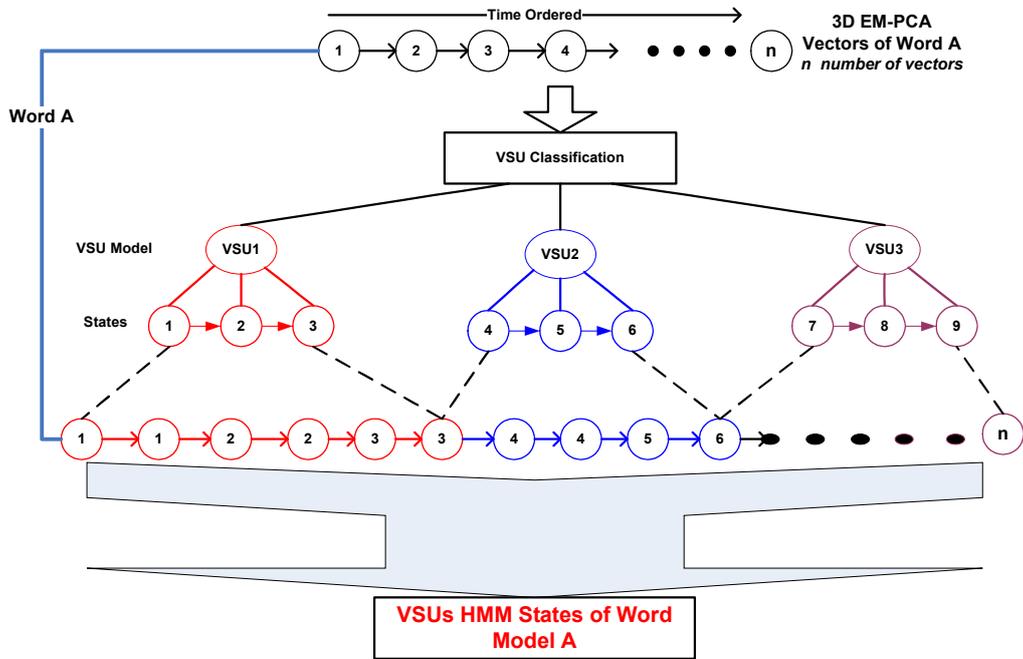
### 5.5.3. Experiment 3: Performance Evaluation for Visual Speech Units and Visemes in the Context of Word Recognition

This experiment is conducted to evaluate the word classification accuracy when the VSUs and visemes are employed as speech elements. The lips dynamics associated with VSUs and visemes are partitioned using three HMM states (see Section 5.4) and each word is modeled as sequence of time-ordered VSUs or visemes. This process is displayed

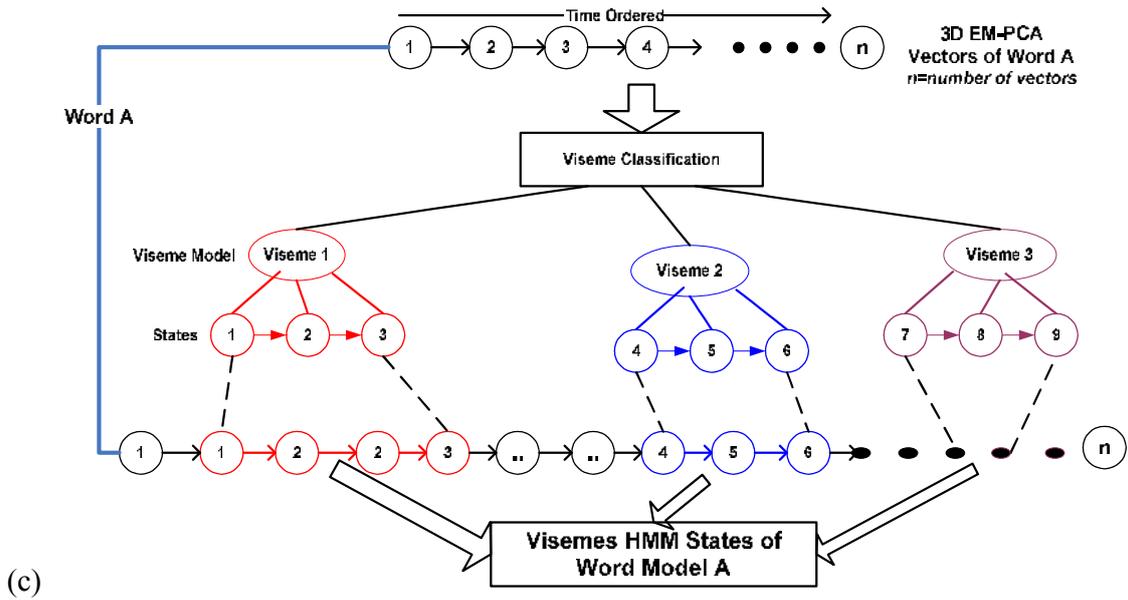
in Figure 5.6. In this diagram the 3-dimensional (3D) EM-PCA vectors that describes visually the spoken word (referred as “manifold” as detailed in Section 3.4) is partitioned into a set of time-ordered VSUs or visemes.



(a)

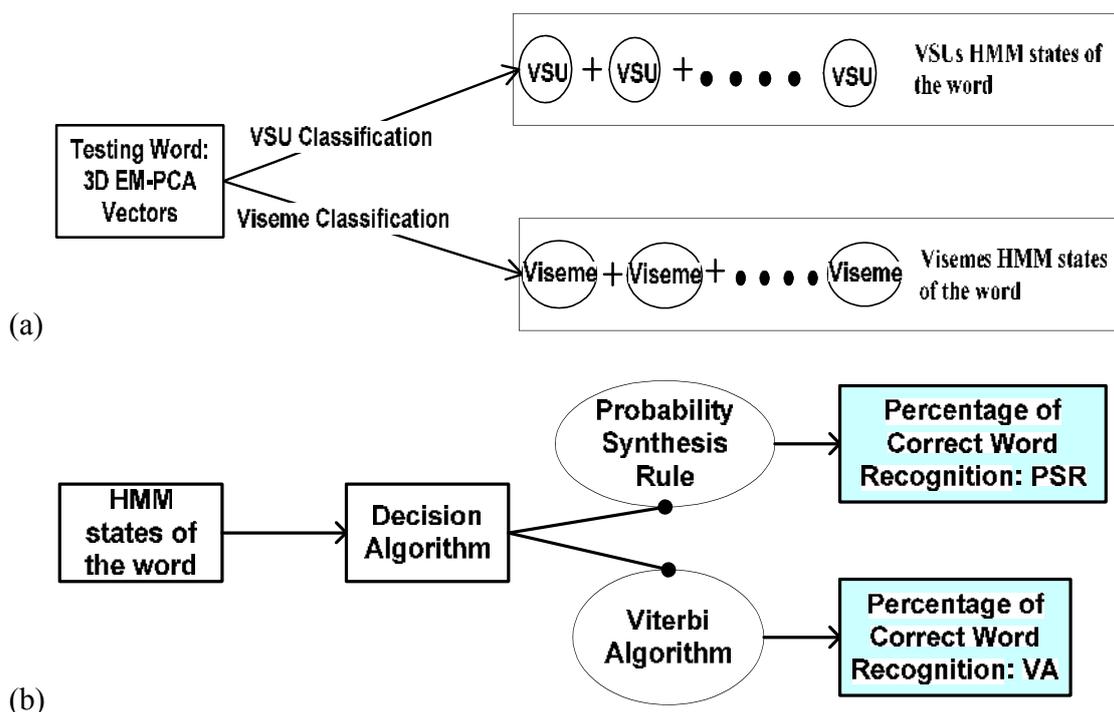


(b)



**Fig. 5.6. Word-based recognition when the VSUs and visemes are used to model the visual speech.** (a) Word recognition process. (b-c) The classification process when the VSUs (b) and visemes (c) are applied for word recognition.

The word recognition process consists of two stages (see Figure 5.7). In the first stage the 3D EM-PCA vectors associated with the input word is partitioned into a set of basic visual speech elements (VSU or viseme) and the resulting sequence is described by a set of ordered HMM states. In this representation each video frame (vector) is labeled to a particular HMM state. In the second stage, the HMM state sequence resulting from stage one is recognized using a decision algorithm that is based either on a probability synthesis rule approach or on a Viterbi algorithm.



**Fig. 5.7. Word Recognition Process.** (a) Stage 1: Generation of the HMM state sequence. (b) Stage 2: Word recognition process.

As indicated earlier, in this implementation the word recognition is evaluated using two decision algorithms:

(a) **Probability Synthesis Rule (PSR).** This approach evaluates the recognition of each independent speech element (VSU or viseme) in the HMM sequence associated with the input word. (For more details refer Dong et al [86] and Alaa EI. Sagheer et al [87]). For instance, the word [ba:bi] will generate the following VSUs ([silence-b] + [b-a:] + [a:-b] + [b-i:]) and visemes ([b] + [a:] + [b] + [i]) sequences. The testing sequence will be classified as [ba:bi] only if all VSUs ([silence-b] + [b-a:] + [a:-b] + [b-i:]) are correct classified or visemes ([b] + [a:] + [b] + [i]) are correct classified. This result is denoted as PSR in Table 5.4 for VSU based recognition and viseme based recognition.

(b) **Viterbi algorithm (VA)**. Using this approach, the transition and emission probabilities matrix between the HMM states for all words in the database (in this experiment a database containing 15 words is used. 5 instances for each word) are re-estimated using the Baum-Welch algorithm. Given a HMM state sequence calculated from EM-PCA vectors associated with the input word (Stage 1), the most likely state path specified by transition and emission probabilities matrix between the hidden states associated with the input word and the words contained in the database is calculated using the viterbi algorithm. (For more details of this procedure refer Durbin et al [88]). Based on above results, the percentages of the most likely probable HMM states of input word that agrees with the training HMM state sequences contained in the word database are calculated.

For instance, given most likely probable HMM states of testing word *likelystates* and one training HMM states sequence of word model **A**, the length (**len**) are normalized by the length of the HMM state sequence that calculated from the testing word. Based on time-ordered of both sequences, each state of *likelystates* with each state of word model **A** will be compared one by one if they are same or not. (e.g.: If the fourth state of word model **A** is **1** and the fourth state of *likelystates* is **1**, then they are same). The percentage of *likelystates* that agrees with word model **A** is calculated as the number of same states in the total number of states (**len**). This calculation is shown as follow:

$$\text{Percentage 1: } \text{sum}(\mathbf{A}==\text{likelystates}) \rightarrow \frac{\text{Number of same states}}{\text{len}} = \underline{\underline{\mathbf{0.8200}}}$$

In order to find the best accuracy among all word models, the percentage of *likelystates* that agrees with other word models are also calculated as follow:

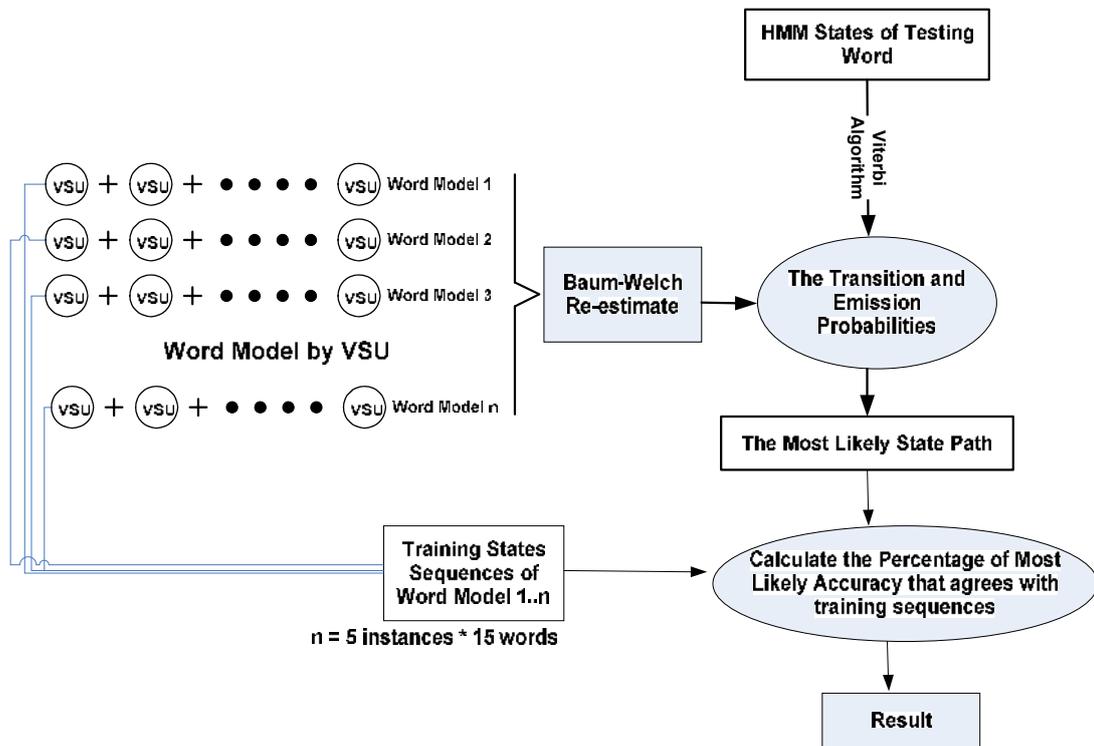
Percentage 2:  $\text{sum}(\mathbf{B}==\text{likelystates}) \rightarrow \frac{\text{Number of same states}}{\text{len}} = 0.4100$

...

Percentage n:  $\text{sum}(\mathbf{N}==\text{likelystates}) \rightarrow \frac{\text{Number of same states}}{\text{len}} = 0.3430$

**Note:** `sum()` is the matlab function to compute total number of same states between word model (A, B, ..., N) and testing data (**likelystates**). In the example provided above the best accuracy is achieved for the word model **A**.

This classification algorithm is implemented using the HMM functions of the MATLAB Statistics Toolbox [<http://www.mathworks.com/>].



**Fig.5.8.** The application of the Viterbi algorithm for word recognition. Note: this procedure is also applied when the visemes are applied for word recognition.

The experimental results depicted in Table 5.4 are obtained when the PSR and VA word-based classification schemes were applied to a database of 15 words generated by the speaker one where each word consists of at least 3 visemes or VSUs. For each word 5

samples are used for training and 5 samples are used for testing. The classification rate is calculated as follow equation:

$$\text{Classification Rate} = \frac{\text{Number of Correct Recognition}}{\text{Total Number of Testing}} \times 100\%$$

**Table 5.4. Word Correct Recognition Rate**

<b>Word</b>	<b>VSU-based Classification</b>		<b>Viseme-based Classification</b>	
	<b>PSR</b>	<b>VA</b>	<b>PSR</b>	<b>VA</b>
<b>Bart</b>	80%	100%	100%	100%
<b>Boat</b>	80%	100%	100%	100%
<b>Boot</b>	100%	100%	60%	100%
<b>Barbie</b>	80%	80%	40%	40%
<b>Beef</b>	60%	100%	100%	80%
<b>Birch</b>	80%	100%	60%	100%
<b>Bobby</b>	100%	40%	20%	60%
<b>Heart</b>	100%	100%	100%	80%
<b>Hot</b>	100%	100%	80%	40%
<b>Harpy</b>	100%	80%	80%	100%
<b>Hobby</b>	60%	60%	40%	60%
<b>Charge</b>	80%	100%	40%	80%
<b>Zart</b>	100%	80%	100%	100%
<b>Fast</b>	100%	100%	40%	100%
<b>Banana</b>	20%	20%	0%	40%
<b><u>Average Rate</u></b>	<b>82%</b>	<b>84%</b>	<b>64%</b>	<b>77%</b>

*PSA: Probability Synthesis Rule. VA: Viterbi Algorithm*

The experimental results indicate that the correct word recognition based on VSUs classification is 7%-12% higher than the correct word recognition based on visemes classification. It can be also observed that the recognition rate obtained when the Viterbi algorithm (VA) is applied for classification is higher than that attained by the probability synthesis rule (PSR). This is motivated by the fact that the Viterbi algorithm attempts finding the most likely sequence of hidden states between the HMM sequences calculated for the input word and those calculated for the words used for training. For instance, let's assume that the viseme [a:] is not recognized as part of the word 'charge' [cha:dg]. If viseme [ch] and [dg] are correctly classified, the word 'charge' can still be recognized since the vast majority of hidden states are correctly identified within the Viterbi path.

Although the experimental results depicted in Table 5.4 are only indicative since they are produced on a small database, they strengthen the conclusion that the VSUs provide a more elaborate visual speech representation than the standard visemes.

## **5.6. Summary**

In this chapter, three experiments were conducted to assess the performance of the proposed Visual Speech Unit representation. In this approach the VSUs are constructed from the re-sampled word manifolds and the recognition between the VSUs extracted from the input data and the model VSU stored in the database is carried out using HMM classifiers.

In this thesis three distinct experiments were conducted.

- Experiment 1 investigates the performance of VSU and standard viseme representations when applied to the recognition of a set of words. It is

observed that the recognition rate for VSUs generated by both speakers (80-90%) is higher than the recognition rate of visemes (62-72%).

- Experiment 2 evaluates the classification accuracy attained for VSUs when the number of samples applied to train the HMM classifiers is varied. The experimental data indicates that the recognition rate is higher when the number of training samples is increased. Another important finding resulting from this investigation is the fact that the classification accuracy for Group 3 (middle VSU) is slightly lower than the recognition rate obtained for Group 1 and Group 2 VSU categories.
- Experiment 3 presents the recognition accuracy attained for word recognition based on VSU model concept or viseme model. Based on either of two different decision algorithms, the result shows better accurate rate when VSU is used as the basic visual speech element.

The experimental results presented in this chapter indicate that the Visual Speech Unit is an accurate representation for word based visual speech recognition.

# Chapter 6

## Conclusions and Future Work

### 6.1. Conclusions

#### 6.1.1 Thesis Summary

Visual Speech Recognition (VSR) is a very challenging task that involves collaborative research efforts in multiple areas such as computer vision, pattern recognition, image processing and human actions modeling. In general, five tasks are required in any VSR system. First, the human face has to be located and tracked in each frame of the video sequence. Second, the region-of-interest (ROI) surrounding the lips have to be extracted from input video data. Third, the optimal visual features have to be calculated in order to produce a representation that describes the shape of lips in each image. Forth, an accurate visual speech model has to be generated to encode the lip motions during the speech process. Finally, the last task is to recognize the visual speech models in the input video data.

In many multimedia systems such as audio-visual speech recognition (AVSR) [18, 19], mobile phone applications, human-computer interaction [58] and sign language recognition [22, 82], VSR provides useful cues since the visual information may improve the overall accuracy of audio and hand recognition systems when they are operated in environments characterized by a high level of noise. VSR techniques have also been applied in the development of systems for person identification [77], machine control or game animation.

## ***APPENDIX A: VISEME MODEL IN LITERATURES***

---

To be successful VSR has to address complex issues such as feature extraction techniques, classification algorithms and recognition tasks. In this thesis, several new techniques have been applied to address issues that arise in the development of VSR applications and they can be summarized as follows:

- **Intensity-based Lip Segmentation**

The pseudo-hue based on the RGB data is calculated and the lips are segmented by applying a histogram-based thresholding scheme. The image area describing the lips is extracted for each frame from the input video sequence.

- **Manifold Generation**

The grayscale data around the lips region is extracted and this information is used to generate the low-dimensional space that is calculated using the EM-PCA procedure. This grayscale data is projected onto the low-dimensional space and for each frame will be calculated a low dimensional point (vector). The feature points obtained after data projection on the low-dimensional EM-PCA space are joined by a poly-line by ordering the frames in ascending order with respect to time. The aim of this procedure is to obtain a discrete manifold where for each mouth shape a low dimensional vector is assigned. To obtain a continuous representation, the manifold is interpolated using cubic-spline.

- **Visual Speech Unit Modeling**

The proposed VSU model extends the standard viseme model by including in the new representation the transition between consecutive visemes. In this manner, the manifold representation generated from the input image sequence describing visually the spoken word is broken into an ordered sequence of VSUs. In the training process, the

## ***APPENDIX A: VISEME MODEL IN LITERATURES***

---

VSUs are constructed from training data and for each class of VSU a mean model is generated based on the re-sampled EM-PCA manifold representation.

- **Visual Speech Unit Registration and Classification**

Finally, the registration process between the VSU mean models and the continuous manifold calculated from the input video sequence is carried out using Dynamic Time Warping (DTW). In this way, a two-step approach is adopted in the VSU recognition process. In the first step, DTW is applied to register the VSU models and the input continuous manifold. In the second step, HMM is employed to calculate the matching cost between the registered section of input manifolds and VSUs contained in the database. The classification result is based on the best matching cost of the registered section of manifold and the VSU model which is contained in the database.

- **Experimental results**

The developed VSR system has been evaluated on real data generated by two speakers and the experimental data indicates that the VSU recognition rate (80-90%) is significantly higher than the recognition rate obtained for MPEG-4 visemes (62%-72%). It is useful to note two facts that might cause the lack of accuracy for standard viseme recognition (10-20% lower than VSU recognition rate). First, during the training section, the viseme samples are difficult to construct because they are presented by a small number of mouth shapes. Second, a large variation was noticed even within the same class of visemes. For example, in the word [ba:bi:], the first viseme [b] shows different characteristics when compared with the second viseme [b] in the visual speech representation. In another fact, the VSU provides a more accurate representation for speech modeling into the word recognition test than the standard viseme representation

and the reported results confirm the superiority of the VSU representation when applied to continuous visual speech.

### **6.1.2 Contributions**

As indicated in the literature review provided in Chapter 2, the most difficult problems that have to be addressed by VSR are the feature extraction, the development of accurate speech models and classification. The first task of feature extraction involves the extraction of the lips in the image data. In practice, various approaches have been proposed where the most simplistic highlight the lips in image data by applying lipstick. Although this approach is effective, it is not comfortable for users and such systems can be operated only in constrained environments. Thus, the main research efforts have been concentrated in the development of vision-based lip segmentation algorithms. In this manner, approaches based on the evaluation of the shape and colour skin models proved to be the most promising. The shape-based approaches require complex initialization procedures and proved to be cumbersome when applied to continuous data, thus in this work has been developed an intensity-based approach that identifies the lips in data converted to the pseudo-hue representation. The development of the lips segmentation algorithm represents a minor contribution of this research work.

Feature extraction was another major topic of interest for this research. In this thesis it has been detailed the application of the EM-PCA manifolds to generate a compact representation that is able to encode the lips motions in the visual domain. While the words are defined by image sequences of different lengths, in this work the discrete manifolds were interpolated to generate a continuous representation. The developed feature extraction scheme represents an important contribution of this work.

## ***APPENDIX A: VISEME MODEL IN LITERATURES***

---

The appropriate selection of the visual speech model is the key issue in the implementation of VSR systems. The vast majority of the proposed VSR systems employed visemes to model the visual speech where continuous speech is viewed as a simple combination of standard visemes. In this investigation, we noted that visemes offer only a partial representation when applied to the representation of the words in continuous speech, since the transitions between visemes are not used in the recognition process. To address this problem, in this thesis a new speech model referred to as Visual Speech Unit (VSU) is proposed and represents the major contribution of this work. Other minor contributions are located in the development of HMM classification schemes.

### **6.2. Future Work**

A detailed analysis of the experimental results indicates that two factors contribute to errors in the recognition process. These two factors can be summarized as follows:

1. The errors in classification are mostly generated by the errors in registration between the VSU models and continuous manifold.
2. The image data are generated by two speakers and the database is defined only by a limited number of VSUs.

In order to overcome the abovementioned issues and improve the recognition accuracy of the proposed VSR technique, future investigations need to be focused on the following areas:

- Improve the DTW technique that performs the registration between VSU models and continuous manifold. In the implementation detailed in this thesis, VSUs that start or end with [silence] can be precisely located in the words manifold, but the registration of “middle” VSUs can be improved especially when dealing with

## ***APPENDIX A: VISEME MODEL IN LITERATURES***

---

complex words (e.g.: like ‘banana’, ‘January’, etc) that consist of multiple “middle” VSUs.

- Evaluate the proposed approach on a larger number of VSUs that are generated by multiple speakers. Based on the standard MPEG-4 viseme category, the total number of VSUs that can be theoretically constructed is 196. Thus, the performance of the proposed VSR system needs to be evaluated on more comprehensive databases defined by an increased number of VSU models and a larger vocabulary.
- Evaluate the proposed VSR system when applied to identify the words in larger video sequences where multiple words are spoken by the speaker.
- The data evaluated in this thesis did not include images showing 3D rotations of the speaker’s head. In order to deploy the proposed VSR system in real world applications, additional work is required to extend the proposed VSU representation to cover 3D rotations. The VSU model will be extended to cover the front and side face of multiple degrees of lips. This investigation can help in the discrimination of visual speech in very complex environment and also be used for 3D human speech animation modeling.
- Future research will be also concerned with the inclusion of the VSU based visual speech recognition in the implementation of a robust sign language gesture recognition system in order to increase its overall performance.
- The proposed VSR system can also be deployed into the development of systems for vehicle control and interaction with industrial robots.

## Appendix A

### Viseme Models

Table A.1 to A.3 show three different viseme categories that have been proposed in the literature on VSR. The viseme category displayed in Table A.1 is the adopted viseme standard in this thesis, this viseme table is introduced by I.S. Pandzic and R. Forchheimer which is an international audio-visual object-based video representation standard.

Viseme Number	Phonemes	Example Words	Vowels or Consonants	Image Example in Database
1	[b], [p], [m]	put, <u>bed</u> , <u>me</u>	consonants	
2	[s], [z]	<u>Z</u> eal, <u>s</u> it	consonants	
3	[ch], [dZ]	<u>ch</u> ard, <u>jo</u> in	consonants	
4	[f], [v]	<u>f</u> ar, <u>vo</u> ice	consonants	
5	[t, d]	<u>t</u> ick, <u>do</u> or,	consonants	
6	[k, g]	<u>g</u> ate, <u>ki</u> ck	consonants	
7	[n, l]	<u>N</u> eed, <u>l</u> ead	consonants	
8	[Th]	<u>th</u> ink, <u>th</u> at,	consonants	n/a in database
9	[r]	<u>r</u> ead	consonants	n/a in database
10	[I]	<u>be</u> at, <u>he</u> at	vowel	
11	[A:]	<u>bu</u> t, <u>cha</u> rd, <u>bar</u> bie	vowel	
12	[e]	<u>h</u> at, <u>be</u> t	vowel	
13	[O]	<u>bo</u> at, <u>ho</u> t	vowel	
14	[U]	<u>ho</u> ok, <u>choo</u> se	vowel	

**Table A.1 Viseme Model of MPEG-4 standard for English [27, 39-40, 51]**

## **APPENDIX A: VISEME MODEL IN LITERATURES**

<b>viseme Class</b>	<b>Phonemes in cluster</b>
silence	[silence], [sp]
Lip-rounding based vowels	[ao], [ah], [aa], [er], [oy], [aw], [hh], [uw], [uh], [ow], [ae], [eh], [ey], [ay], [ih], [iy], [ax]
Alveolar-semivowels Alveolar-fricative Alveolar Palato-alveolar Bilabial Dental Labio-dental Velar	[l], [el], [r], [y] [s], [z] [t], [d], [n], [en] [sh], [zh], [ch], [jh] [p], [b], [m] [th], [dh] [f], [v] [ng], [k], [g], [w]

**Table A.2 44 Phoneme to 13 Viseme Mapping using the HTK phone set [1, 76]**

<b>Viseme number</b>	<b>Viseme description</b>
1	Mouth close
2	Slightly open in small degree of mouth opening
3	Medium degree of mouth opening
4	High degree of mouth opening
5	Mouth open when teeth are observable
6	Mouth open when teeth are not observable.

**Table A.3 Representation of six major viseme classes [57].**

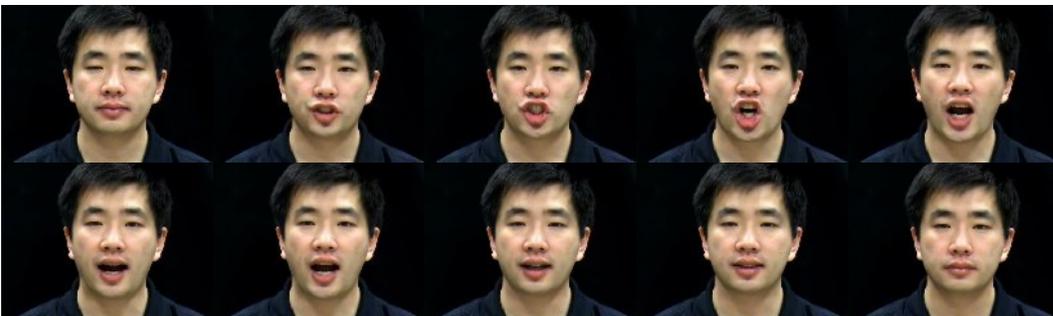
## Appendix B

### Original Images Dataset

Figures B.1 to B.4 show the original images generated by two speakers.



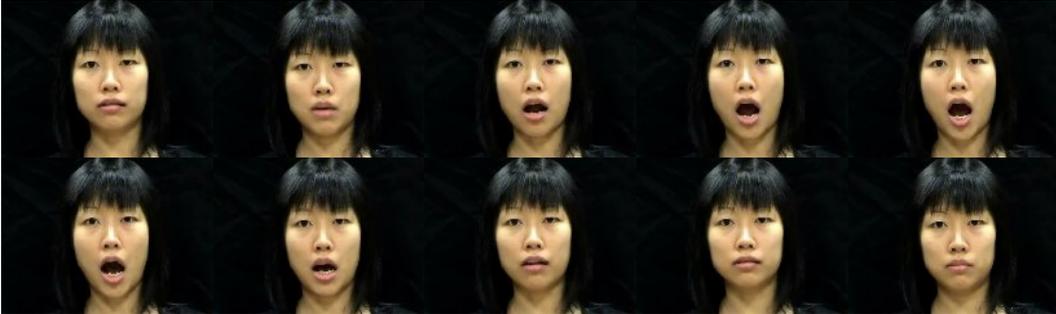
**Figure B.1: Samples of original frames from video sequence 1(Speaker One)**



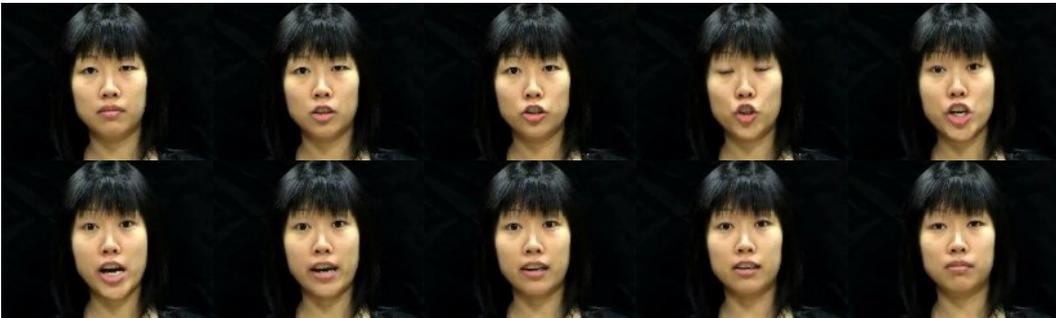
**Figure B.2: Samples of original frames from video sequence 2 (Speaker One)**

*APPENDIX B: ORIGINAL IMAGES DATASET*

---



**Figure B.3: Samples of original frames from video sequence 3(Speaker Two)**

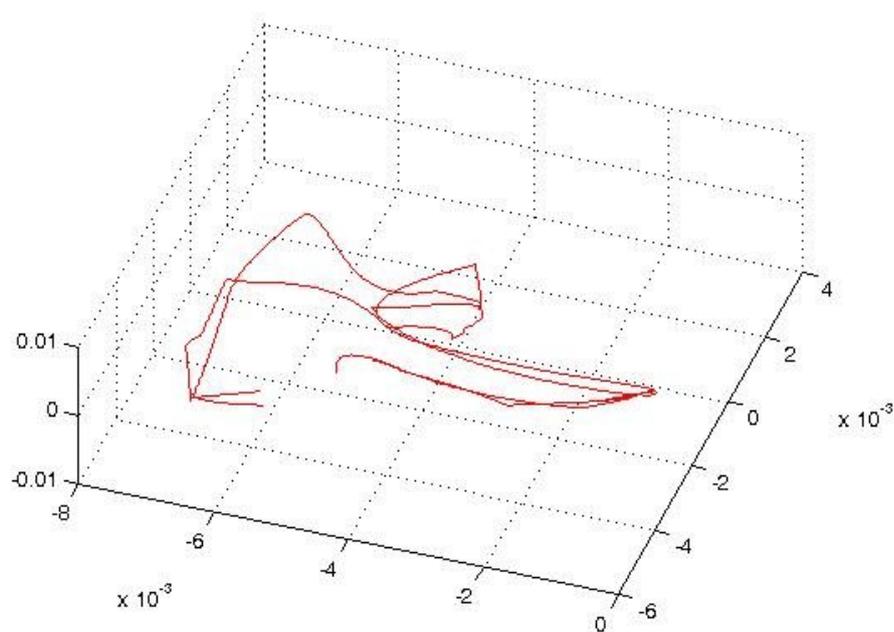


**Figure B.4: Samples of original frames from video sequence 4(Speaker Two)**

## Appendix C

### Continuous Manifold Representation

Figures C.1 to C.6 depict the manifolds of several words analyzed in this thesis. Each figure contains 2 examples of each word.



**Figure C.1: Two continuous manifolds of the word [bu:t]**

## APPENDIX C: CONTINUOUS MANIFOLD REPRESENTATION

---

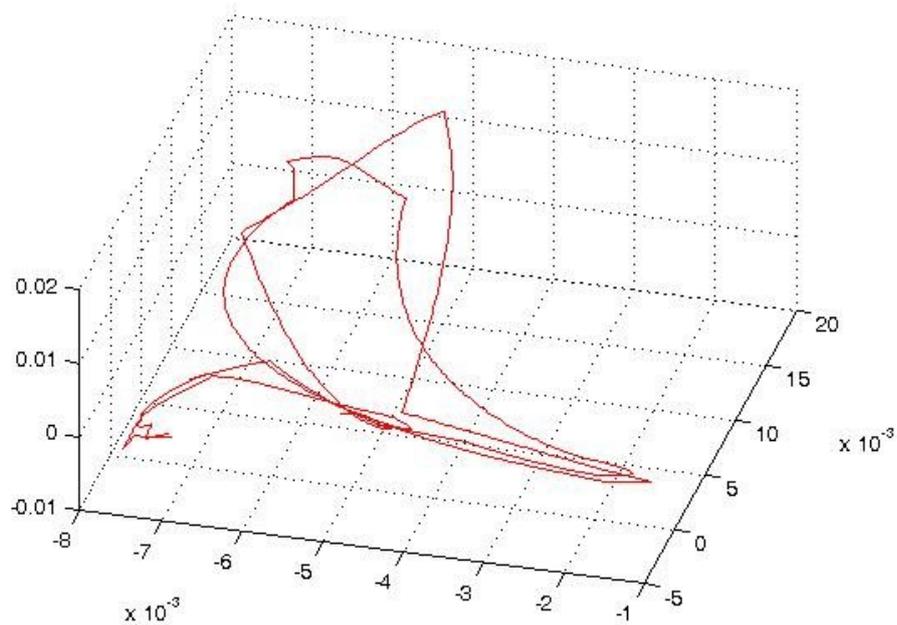


Figure C.2: Two continuous manifolds of the word [ba:bi]

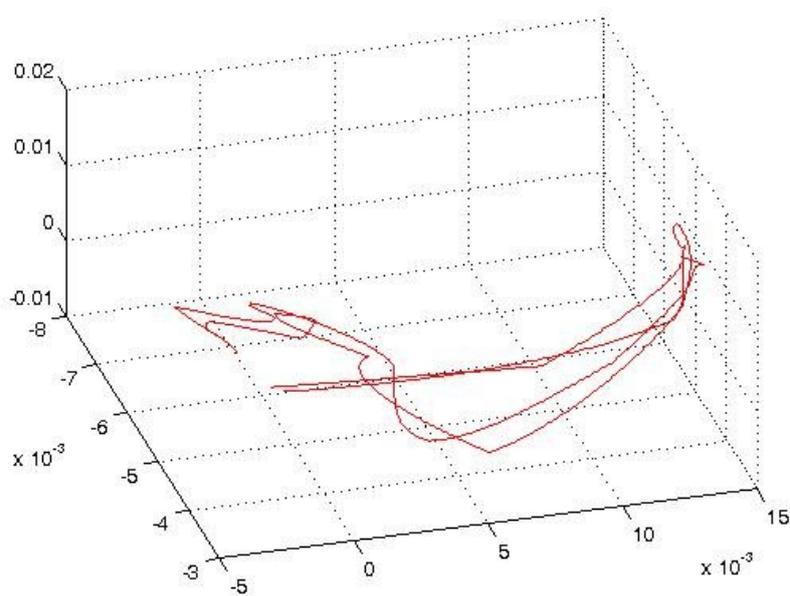


Figure C.3: Two continuous manifolds of word [chu:s]

## APPENDIX C: CONTINUOUS MANIFOLD REPRESENTATION

---

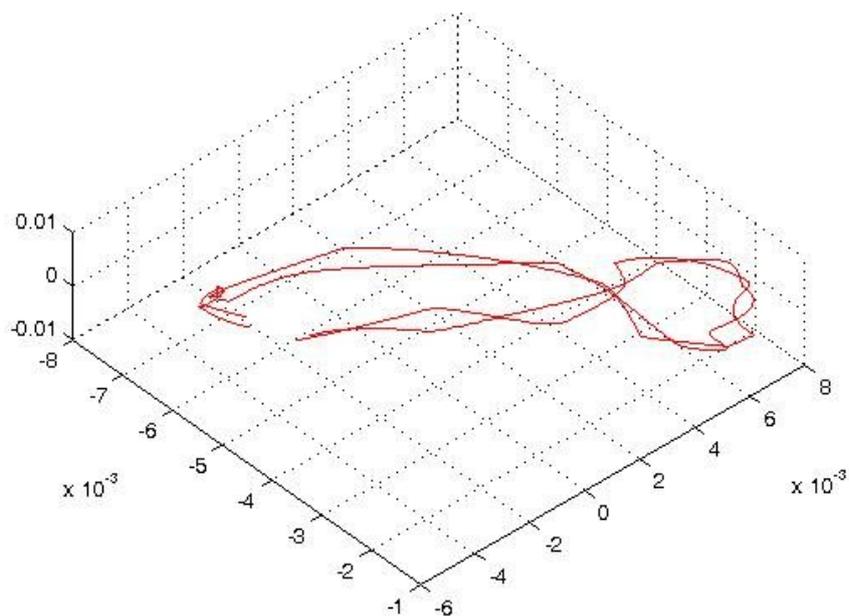


Figure C.4: Two continuous manifolds of the word [hot]

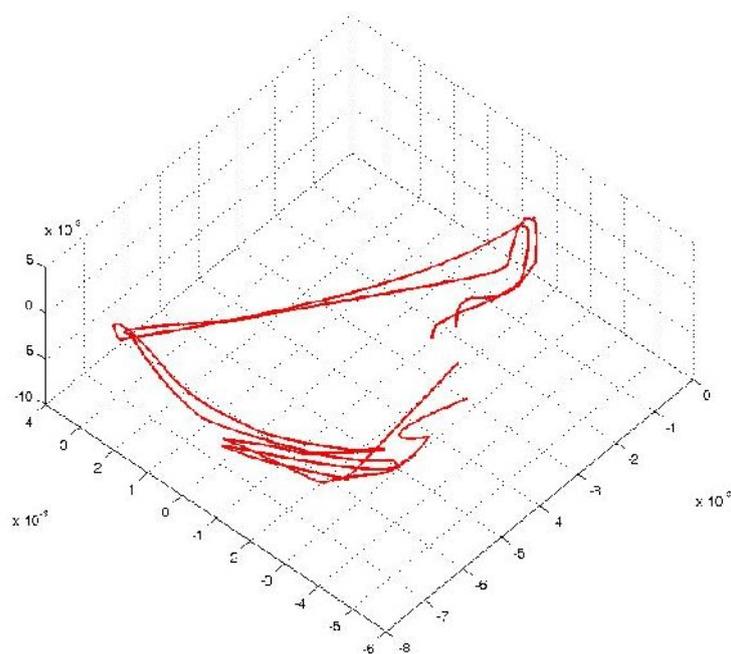


Figure C.5: Two continuous manifolds of the word [bi:t]

## APPENDIX C: CONTINUOUS MANIFOLD REPRESENTATION

---

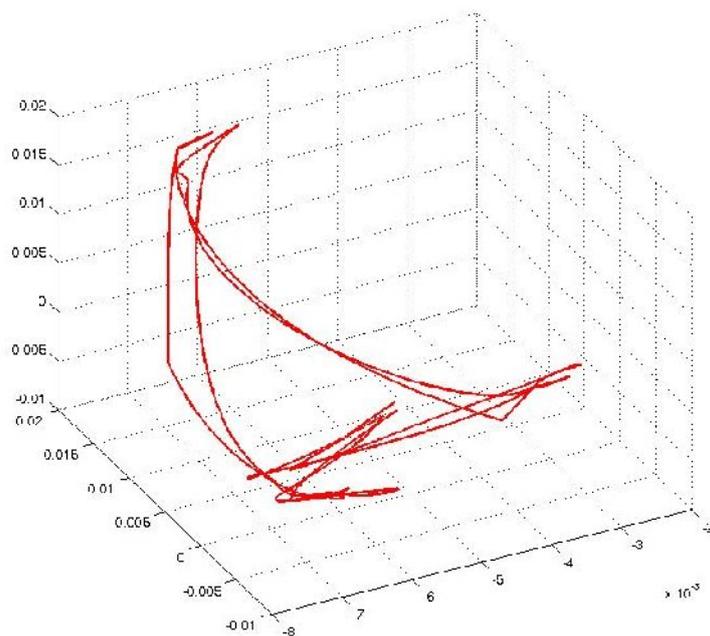


Figure C.6: Two continuous manifolds of the word [fäst].

## Appendix D

### Viseme Representation

Figures D.1 to D.3 show the viseme mapping in single word manifolds. These figures demonstrate that the visemes cover only a small part of the word manifold.

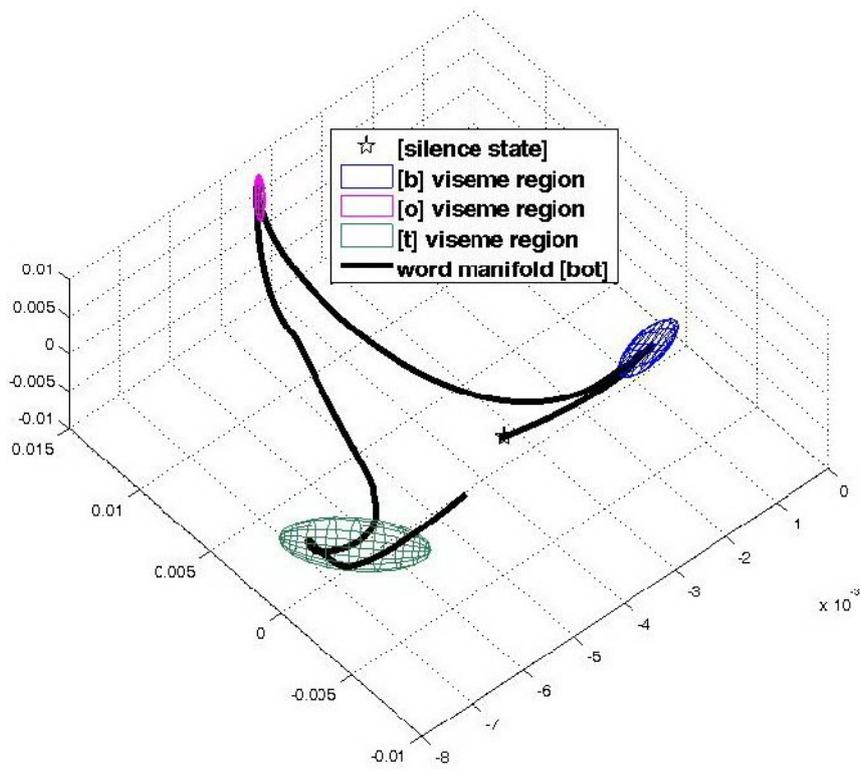


Figure D.1: Viseme [b], [o] and [t] - word manifold- [bot].

## APPENDIX D: VISEME REPRESENTATION

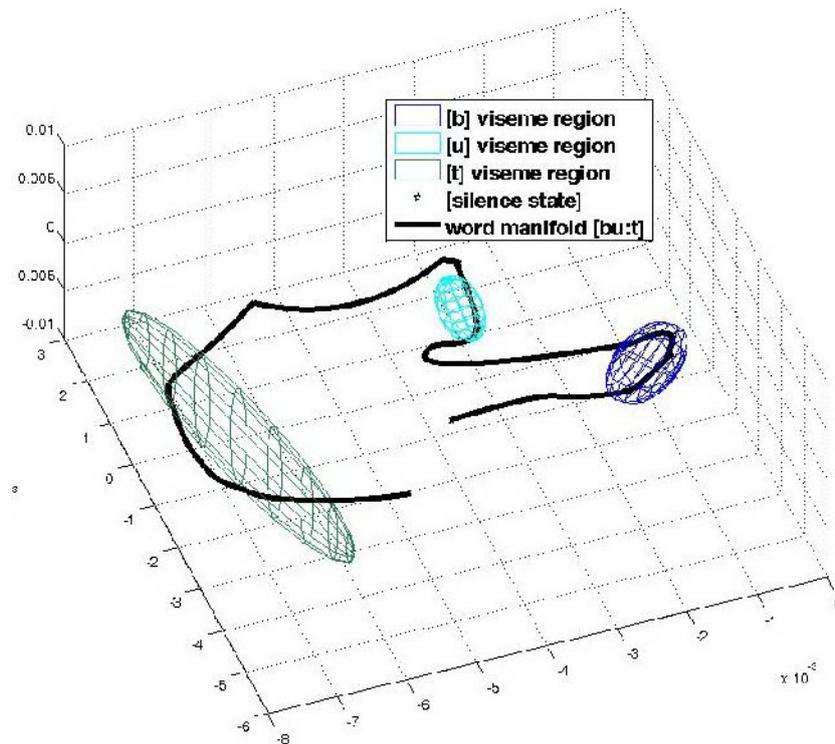


Figure D.2: Viseme [b], [u] and [t] - word manifold-[bu:t].

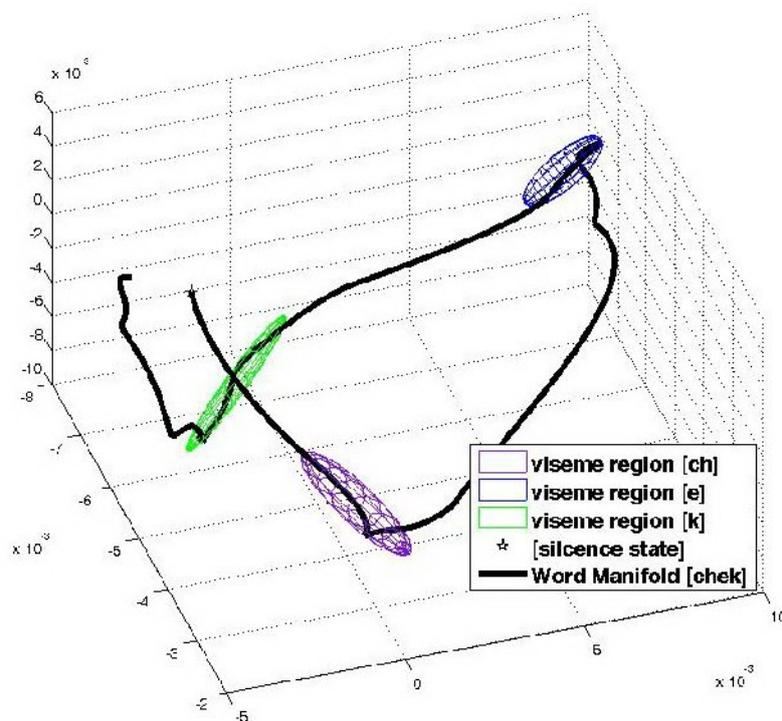


Figure D.3: Viseme [ch], [e] and [k] - manifold-[chek].

## APPENDIX D: VISEME REPRESENTATION

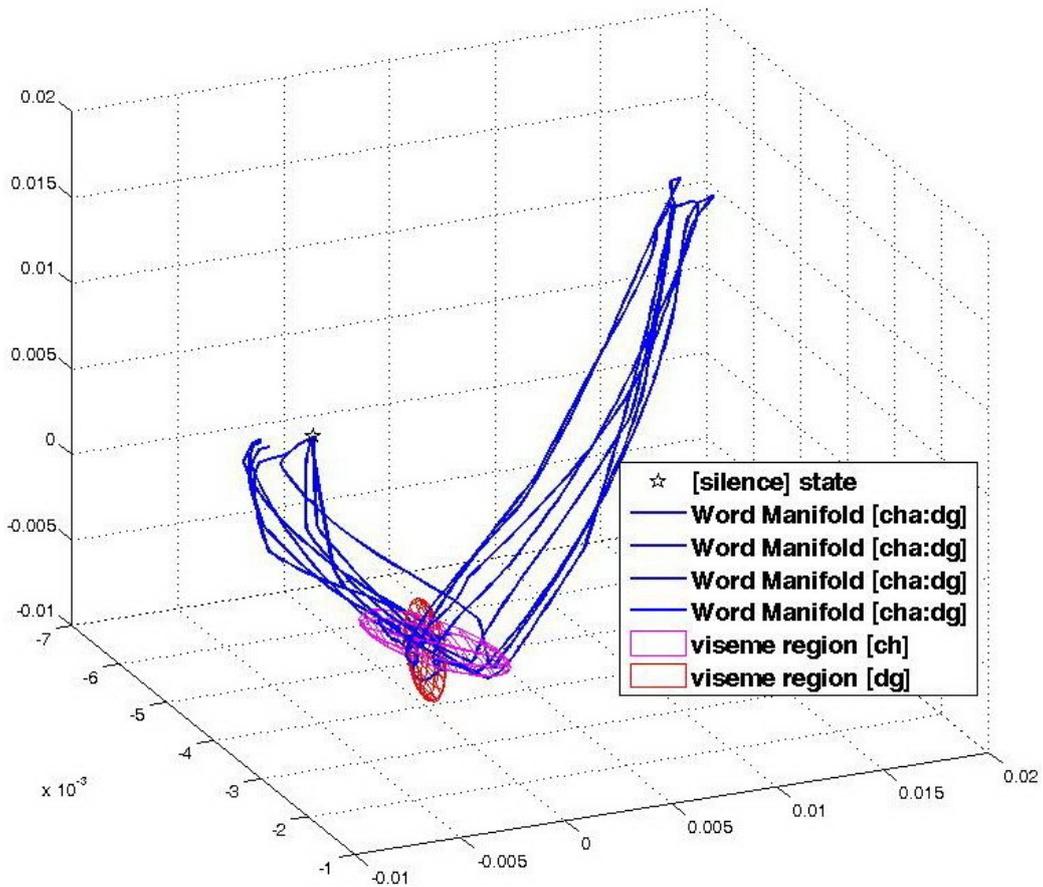
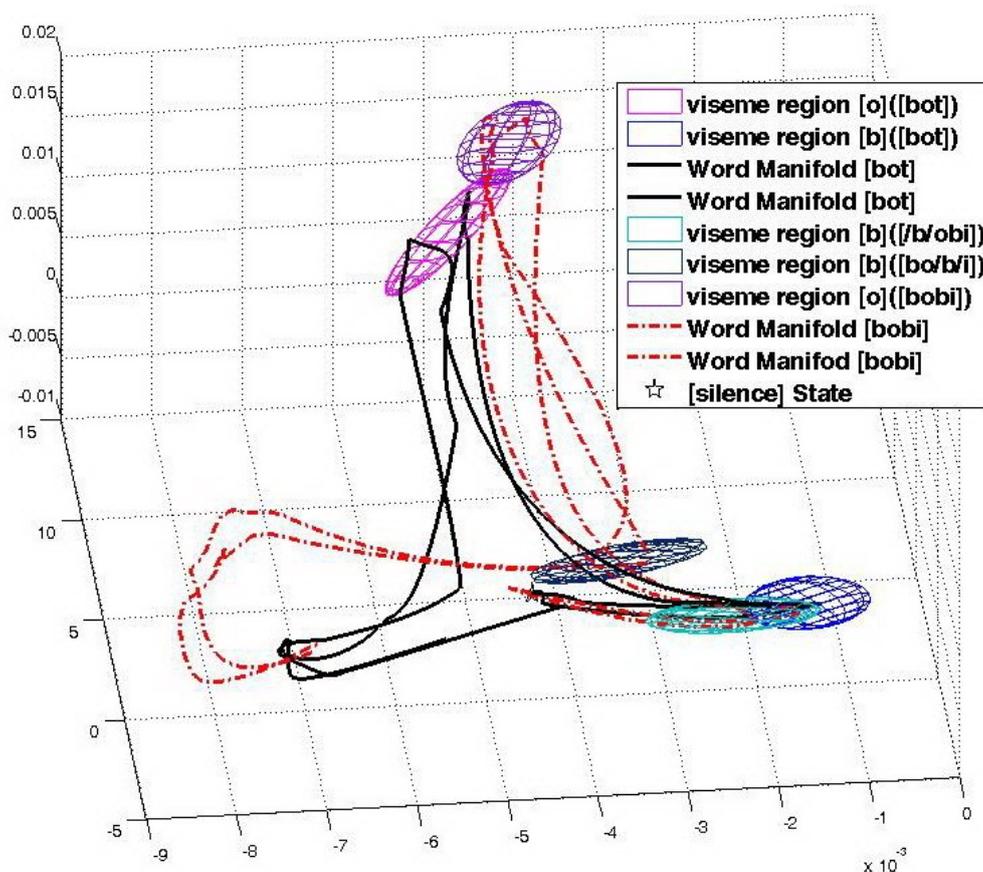


Figure D.4: Representation of same class of viseme [ch] and [dg] extracted from the word manifolds [cha:dg] (four examples).

Note the overlap between the viseme [ch] and viseme [dg] in the EM-PCA space.

## APPENDIX D: VISEME REPRESENTATION



**Figure D.5: Representation of viseme [b] and viseme [o] extracted from different words manifolds - [bot] and [bobi] (2 examples each word).**

The viseme [o] is displayed in the pink region when extracted from the manifold of the word [bot] and in the purple region when extracted from the manifold of the word [bobi]. Note a large region required to map the viseme [o] in the feature space. The viseme [b] is displayed in the cyan region when extracted from the manifold of the word [babi]. We can observe the large variation between the first and the second viseme [b] in the manifold of the word [babi].

## APPENDIX D: VISEME REPRESENTATION

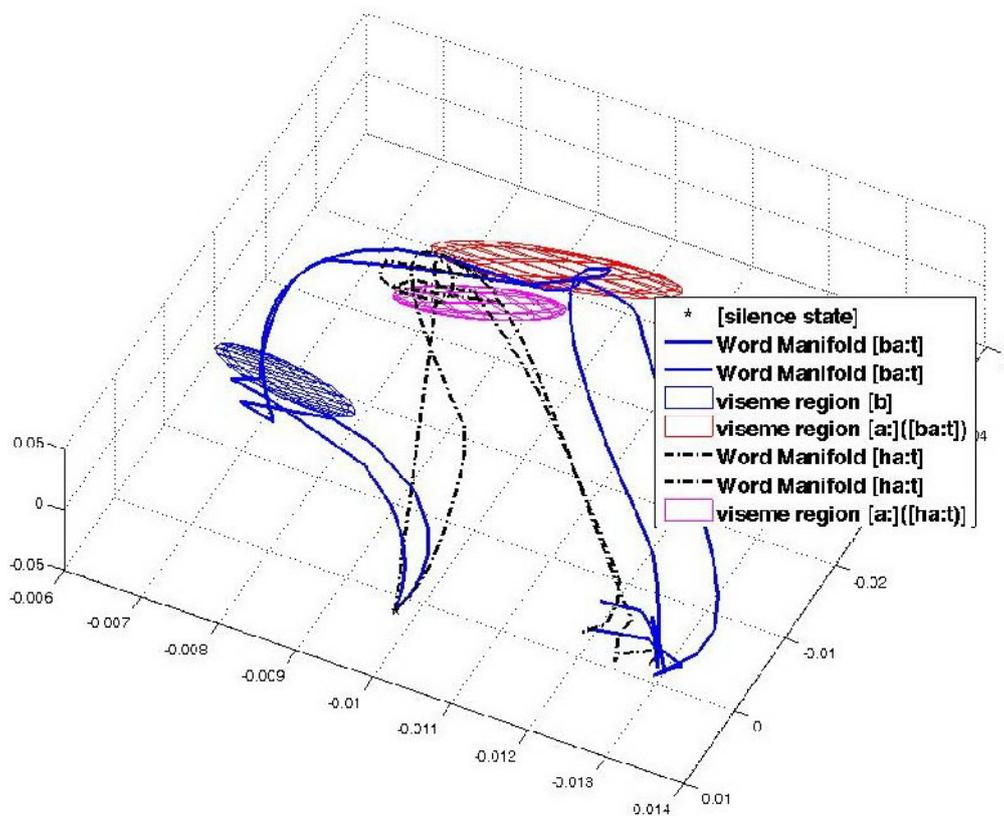
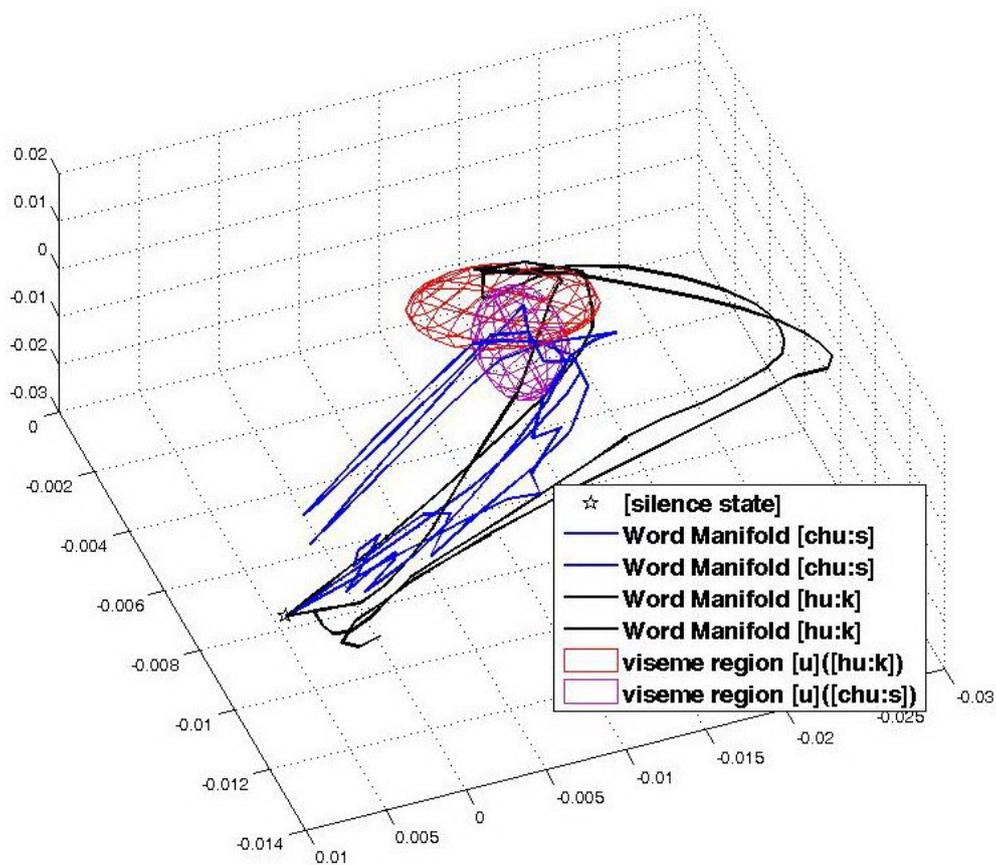


Figure D.6: Representation of viseme [b] and viseme [a:] extracted from different word manifolds- [ba:t] and [ha:t] (2 examples each word).

The viseme [a:] is displayed in the red region when is extracted from the manifold of the word [ba:t] and in the pink region when extracted from the manifold of the word [ha:t].

## APPENDIX D: VISEME REPRESENTATION



**Figure D.7: Representation of viseme [u:] extracted from different word manifolds- [chu:s] and [hu:k] (2 examples each word).**

The viseme [u:] is displayed in the red region when is extracted from the manifold of the word [hu:k] and in the purple region when is extracted from the word [chu:s]. Again we can notice that a large region is required to map the viseme [u:] in the feature space.

## Appendix E

### Visual Speech Unit Representation 1

Figures E.1 to E.3 show the re-sampled manifolds of several Visual Speech Units. In each figure, there are two samples for each Visual Speech Unit (VSU).

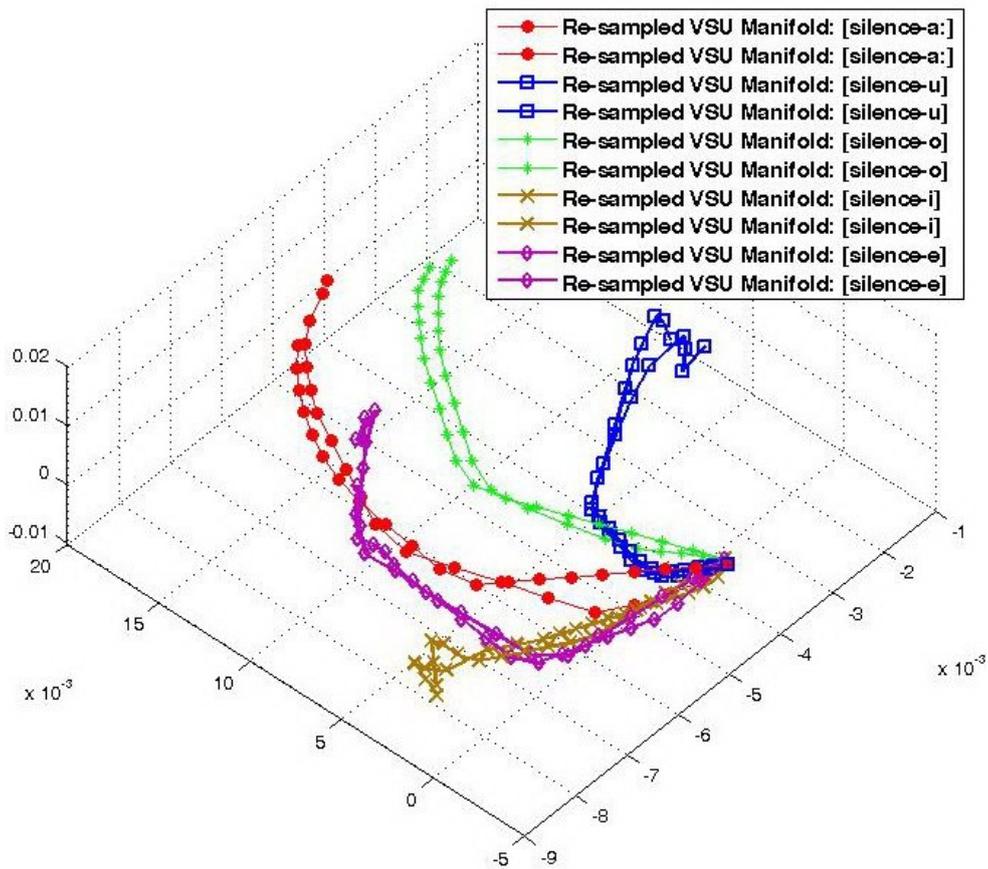


Figure E.1: Re-sampled VSU Manifolds. Five VSUs which all start with viseme [silence] (two samples for each VSU).

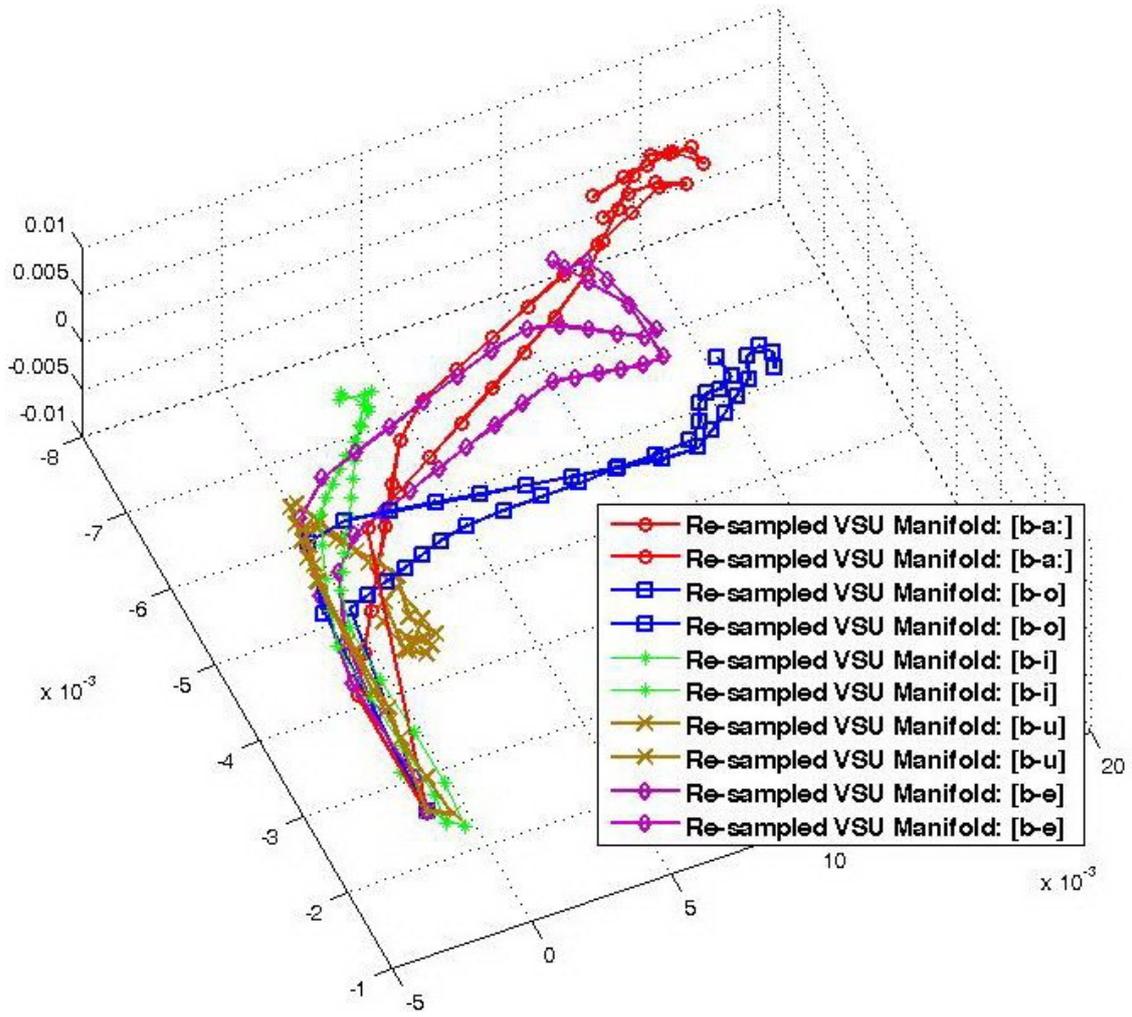


Figure E.2: Re-sampled VSU Manifolds. Five VSUs which all start with viseme [b] (two samples for each VSU).

*APPENDIX E: VISUAL SPEECH UNIT REPRESENTATION 1*

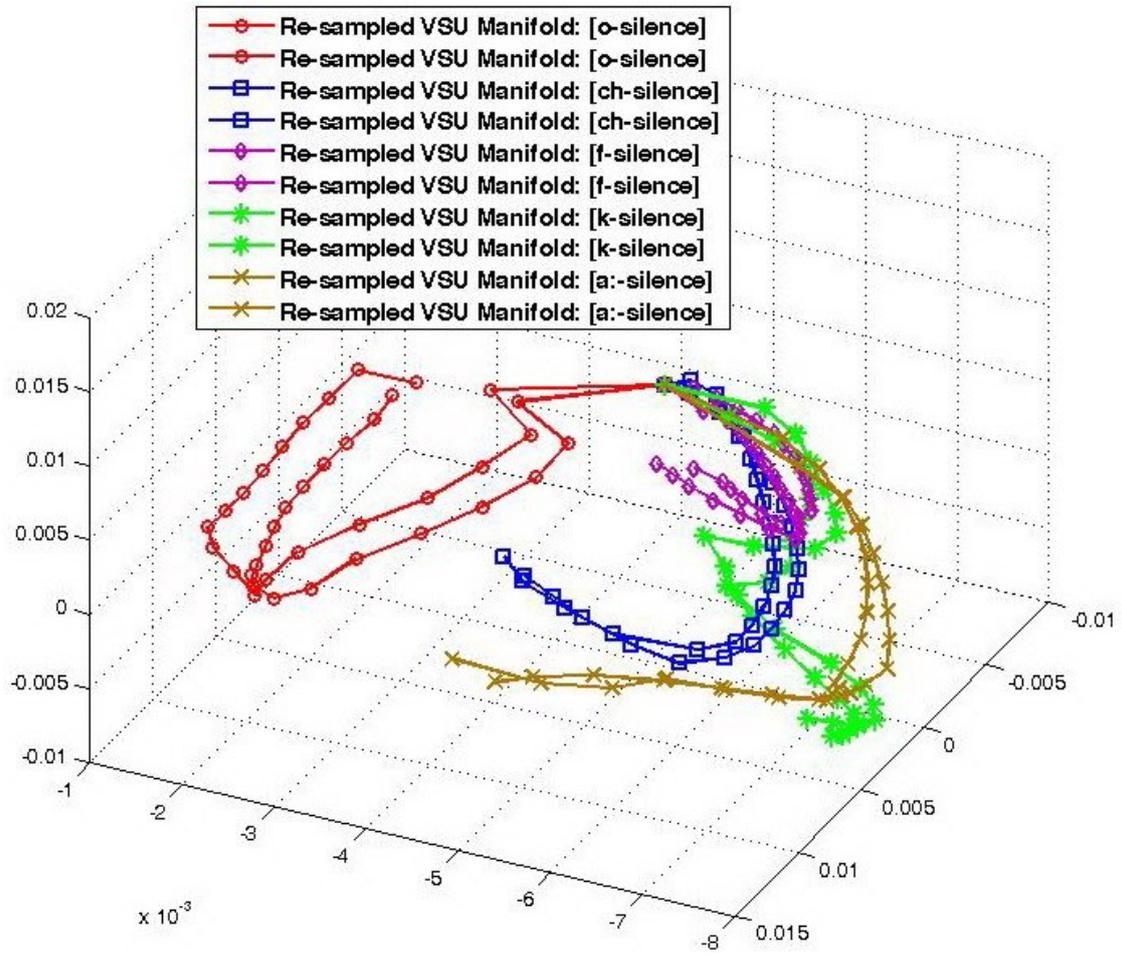


Figure E.3: Re-sampled VSU Manifolds. Five VSUs which all end with viseme [silence] (two samples for each VSU).

## Appendix F

### Visual Speech Unit Representation 2

Figures F.1 to F.5 show the mean models for different VSUs and VSU samples extracted from different word manifolds. In these examples, each figure has two parts: (a) illustrates the VSU mean models and the word manifolds; (b) illustrates the similarity between the VSU mean models and the VSU samples extracted from the words manifolds displayed in (a).

Figure F. 1 to F. 5 indicate that:

1. The part of word manifolds show similar characteristics when they contain the same VSU.
2. There can be noticed some variation for the same VSU such as [na:] when appears in succession in a complex words such as ‘banana’ (Figure F.4).

The words manifolds displayed in Figures F.1 to F.5 are not used to calculate the VSU mean model.

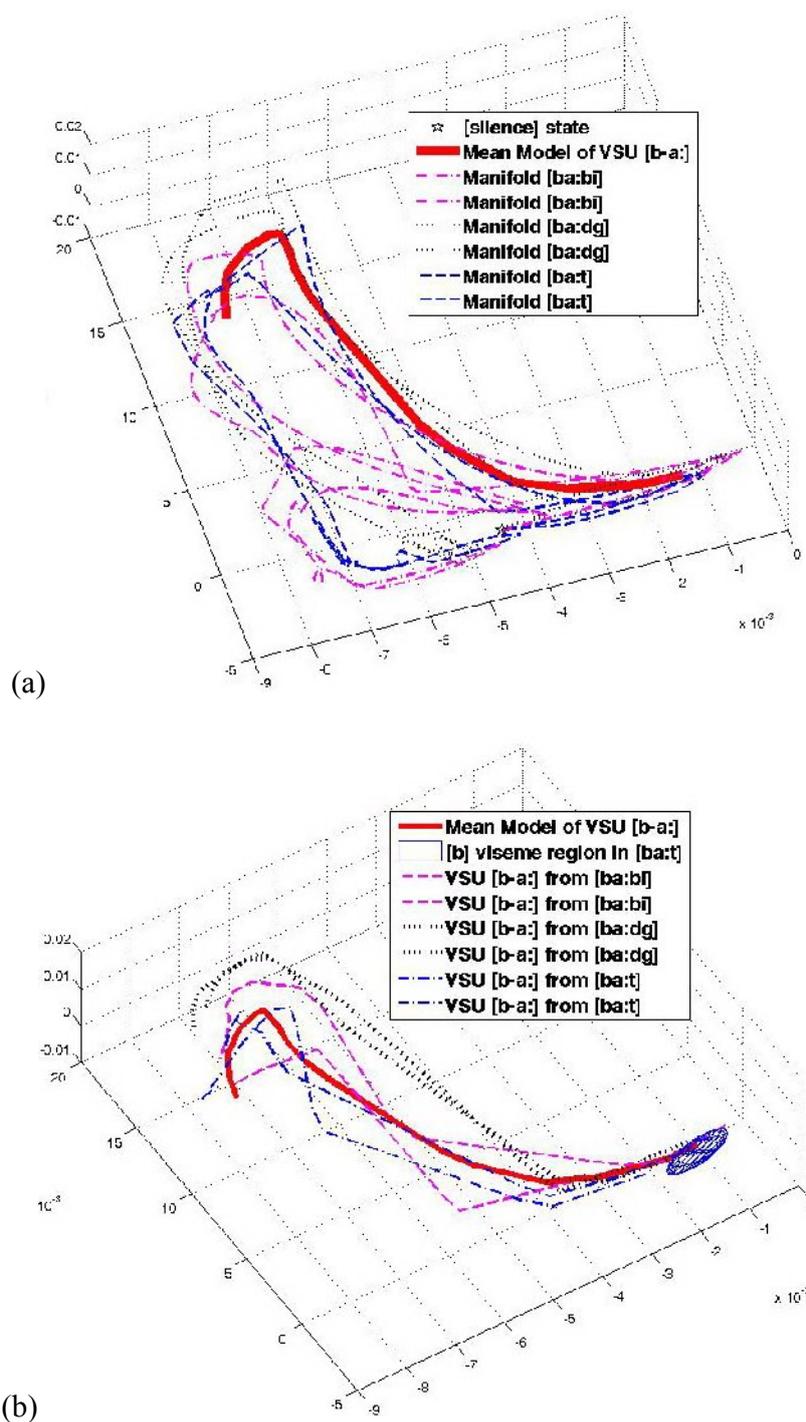
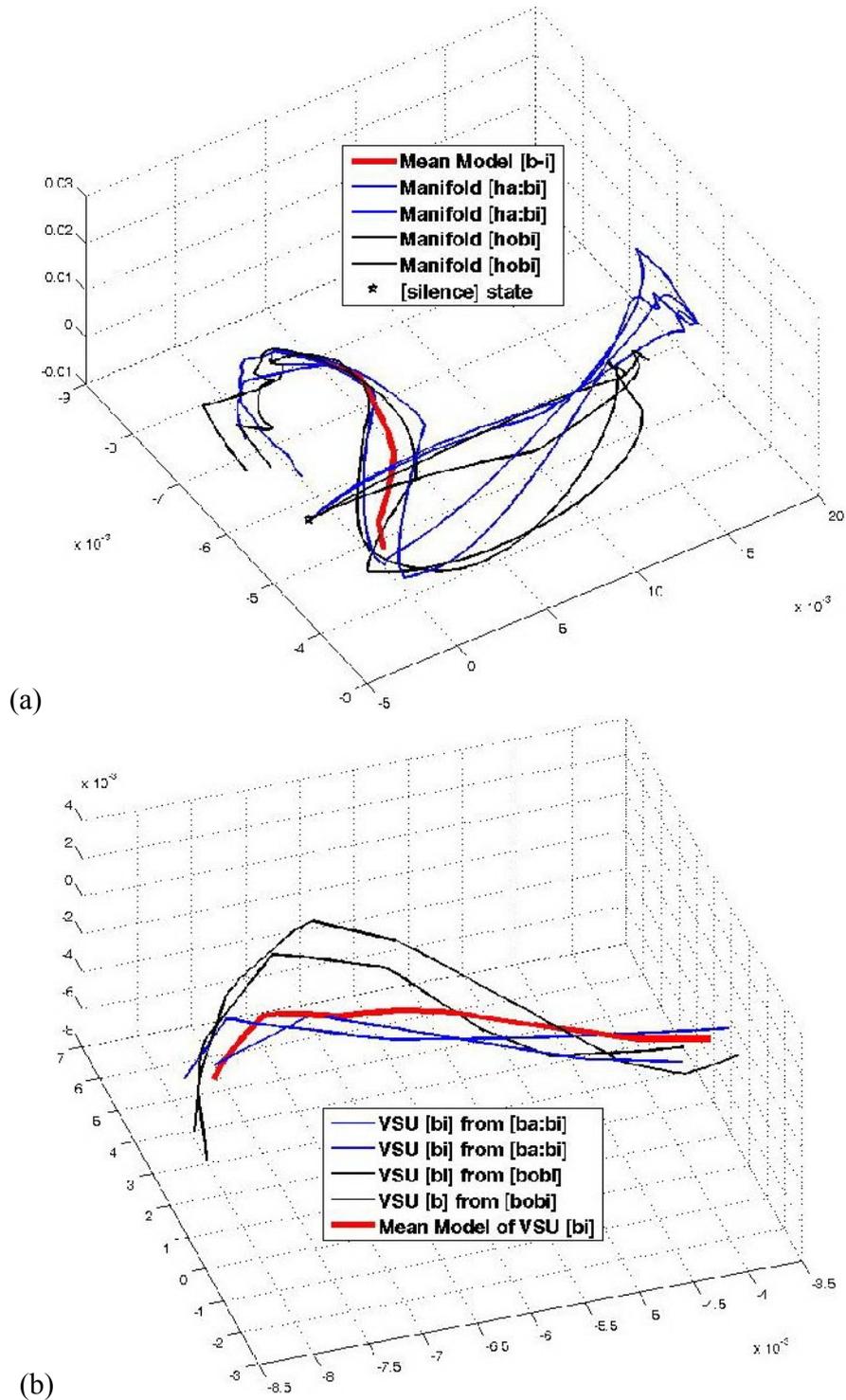


Figure F.1: The mean model of VSU [b-a:] and the VSUs extracted from different words (two examples each word). (a) Mean model of VSU [b-a:] (red), word manifolds: [ba:bi] (pink), [ba:dg] (black) and [ba:t] (blue). (b) Mean model of VSU [b-a:] (red), test VSUs [b-a:] extracted from [ba:bi] (pink), [ba:dg] (black) and [ba:t] (blue).

**APPENDIX F: VISUAL SPEECH UNIT REPRESENTATION 2**



**Figure F.2: The mean model of VSU [b-i] and the VSUs extracted from different words (two examples each word).** (a) Mean model of VSU [b-i] (red), word manifolds: [ba:bi] (blue) and [bobi] (black). (b) Mean model of VSU [b-i] (red), test VSUs [b-i] extracted from [ba:bi] (blue) and [bobi] (black).

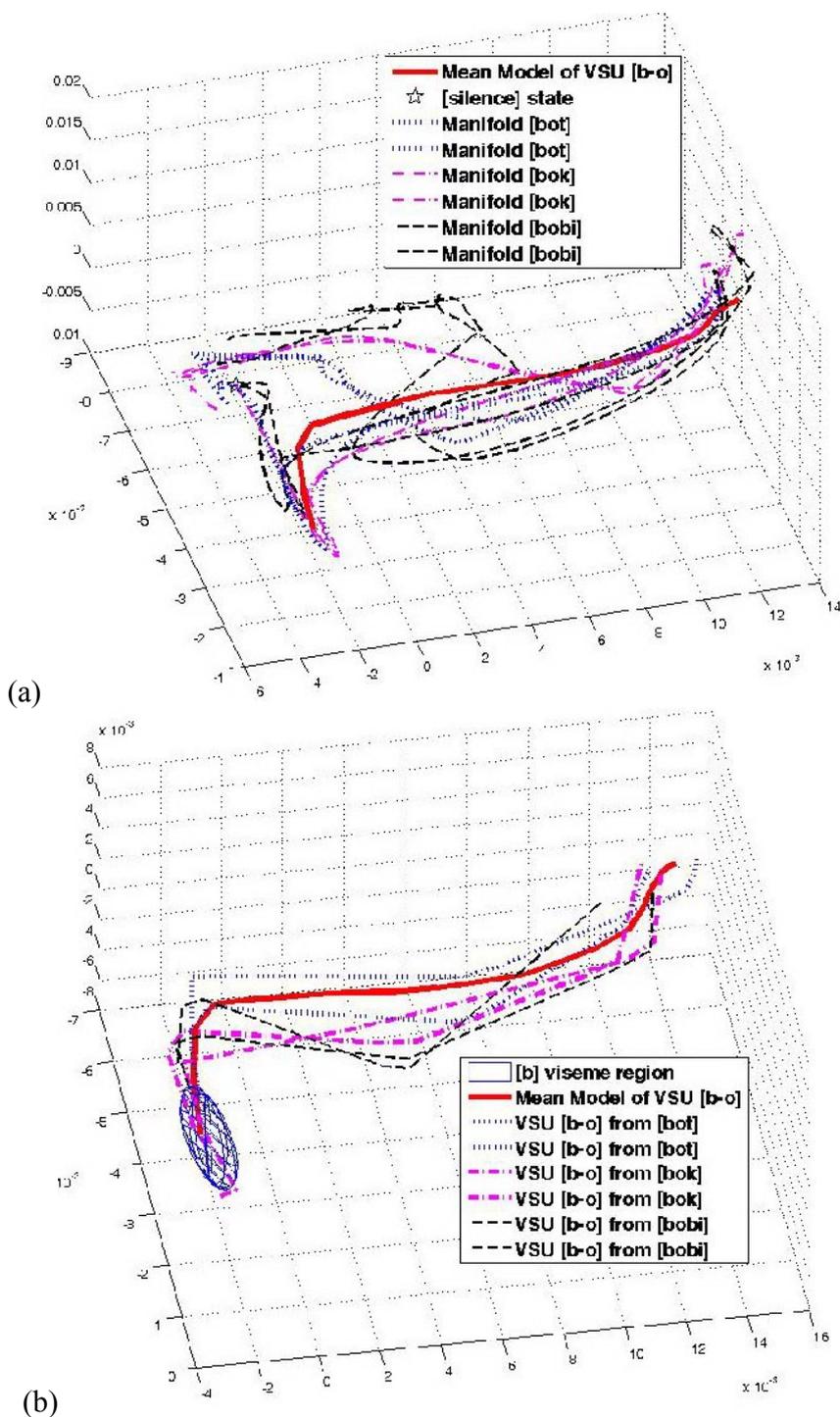
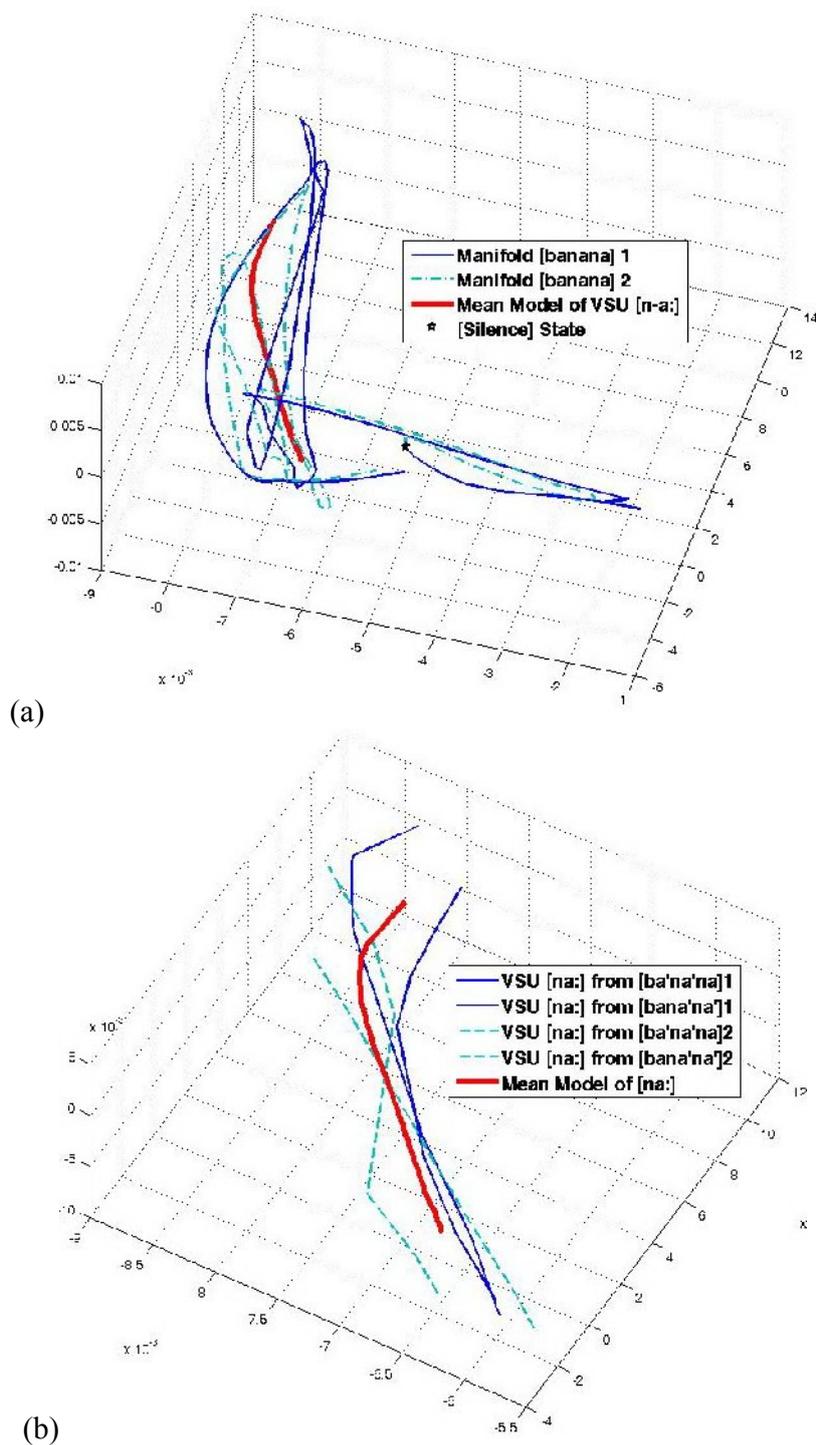
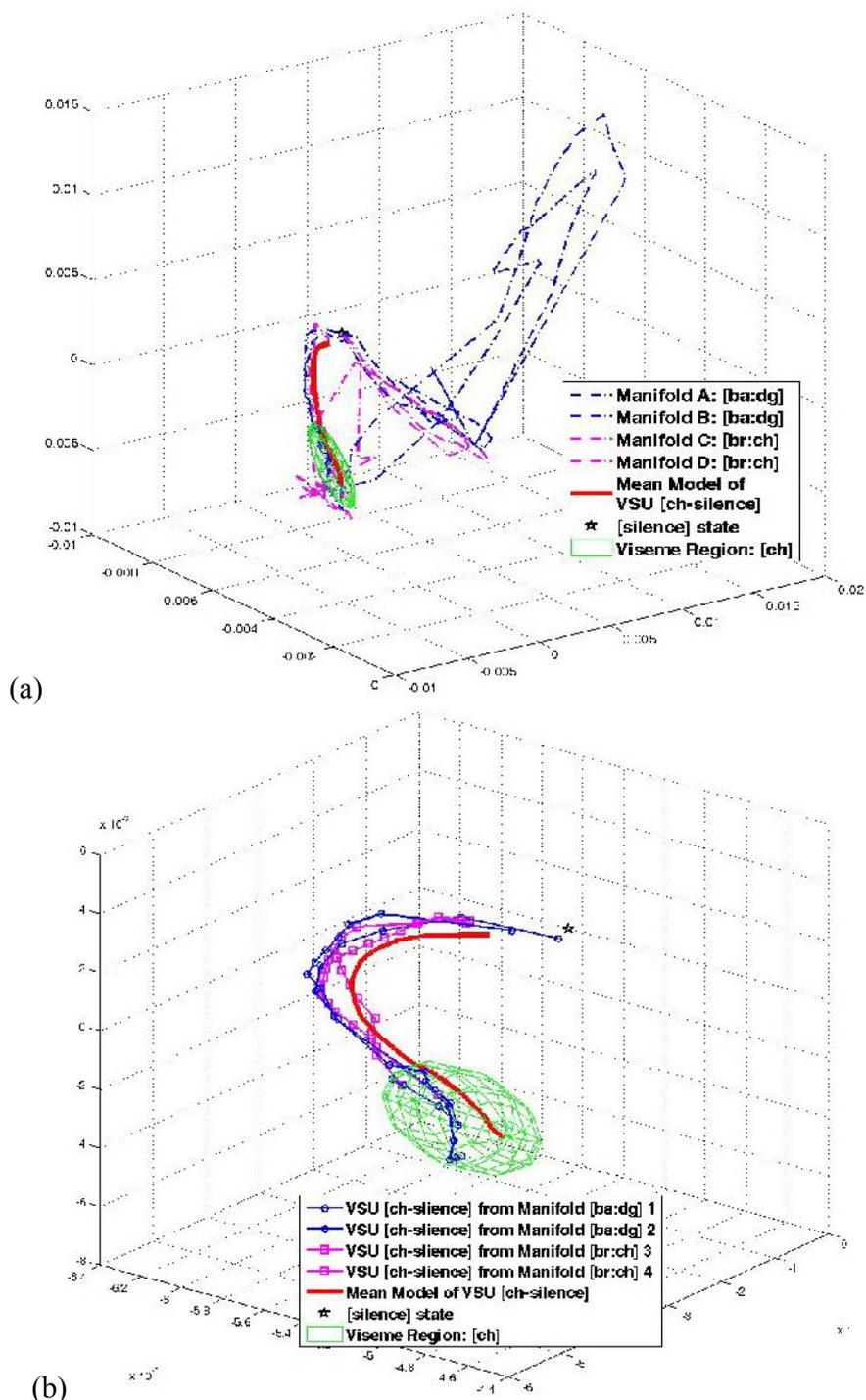


Figure F.3: The mean model of VSU [b-o] and the VSUs extracted from different words (two examples each word). (a) Mean model of VSU [b-o] (red), word manifolds: [bot] (blue dot), [bok] (pink) and [bobi] (black dash). (b) Mean model of VSU [b-o] (red), test VSUs [b-o] extracted from [bot] (blue dot), [bok] (pink) and [bobi] (black dash).

## APPENDIX F: VISUAL SPEECH UNIT REPRESENTATION 2



**Figure F.4: The mean model of VSU [n-a:] and the VSUs extracted from the word-‘banana’ (two examples).** (a) Mean model of VSU [n-a:] (red), word manifolds-‘banana’: example 1 (blue) and example 2 (cyan dash). (b) Mean model of VSU [n-a:] (red), test VSUs [n-a:] extracted from example 1 (blue) and example 2 (cyan dash).

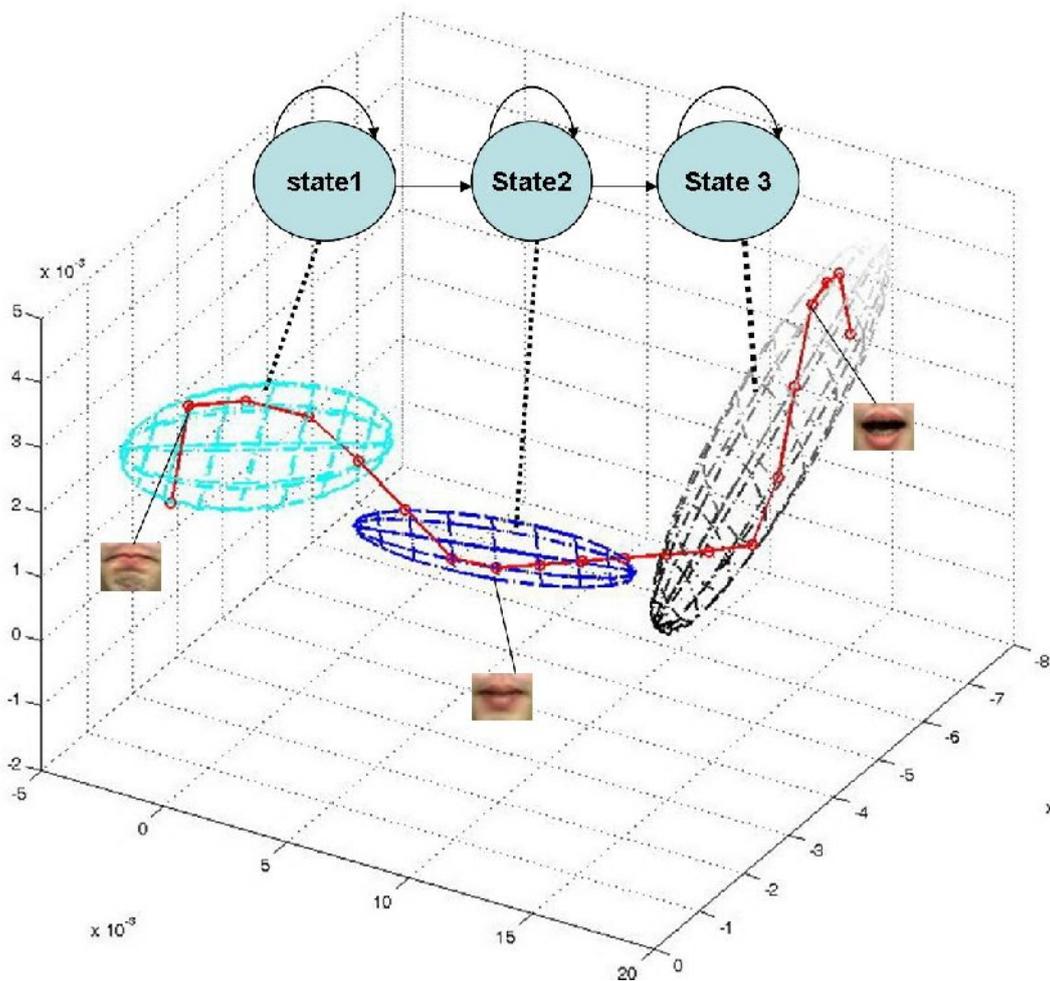


**Figure F.5: The mean model of VSU [ch-silence] and the VSUs extracted from different words (two examples each word).** (a) Mean model of VSU [ch-silence] (red), word manifolds: “barge” [ba:dg] (blue dash) and “birch” [bɜ:ch] (pink dash). (b) Mean model of VSU [ch-silence] (red), test VSUs [ch-silence] extracted from [ba:dg] (blue cycle), and [bɜ:ch] (pink square). Note: [ch] and [dg] are in the same class of viseme (green region).

## Appendix G

### Visual Speech Unit Representation 3

Figure G.1 shows one example where the mouth shapes associated with a VSU are modeled using three-state HMMs.



**Figure G.1: VSU modeling using three state HMMs.** The manifold of the [b-a:] is plotted with a red line.

## **Publications Resulting from this Research**

1. Dahai Yu, Ovidiu Ghita, Alistair Sutherland, Paul F. Whelan, “Dictionary Based Lip Reading Classification”, Irish Machine Vision & Image Processing Conference (IMVIP 2006) (Poster), Dublin City University, Ireland, 2006. **(Poster Presentation)**
2. Dahai Yu, Ovidiu Ghita, Alistair Sutherland, Paul F. Whelan, “Dictionary Based Lip Reading Classification”, China-Ireland International Conference on Information and Communications Technologies (CIICT 2006), Hangzhou, China, 2006. **(Oral Presentation)**
3. Dahai Yu, Ovidiu Ghita, Alistair Sutherland, Paul F. Whelan (2007), “A PCA based Manifold Representation for Visual Speech Recognition”, China-Ireland International Conference on Information and Communications Technologies (CIICT2007), August 28th - 29th, 2007 **(Oral Presentation, Microsoft Best Paper Award)**.
4. Dahai Yu, Ovidiu Ghita, Alistair Sutherland, Paul F. Whelan (2007), “A New Manifold Representation for Visual Speech Recognition”, 12th International Conference on Computer Analysis of Images and Patterns (CAIP 2007), Vienna, Austria. August 27th - 29th, 2007. **(Poster Presentation)**
5. Dahai Yu, Ovidiu Ghita, Alistair Sutherland, Paul F. Whelan (2007), “A New Manifold Representation for Visual Speech Recognition”, Proceedings of the International Machine Vision & Image Processing Conference 2007 (IMVIP

## ***LIST OF PUBLICATIONS***

---

- 2007), 5th-7th September, National University of Ireland, Maynooth (NUIM), IEEE Computer Society Press. **(Poster Presentation, Best Poster Award)**
6. Dahai Yu, Ovidiu Ghita, Alistair Sutherland, Paul F. Whelan (2008), “The Application of Visual Speech Units for Visual Speech Recognition”, Image and Vision Computing, ELSEVIER, 2008. (Under Review, Submission Date: 30/03/2008)
  7. Dahai Yu, Ovidiu Ghita, Alistair Sutherland, Paul F. Whelan (2008), “A Novel Visual Speech Representation and HMM Classification for Visual Speech Recognition”, the 3<sup>rd</sup> Pacific-Rim Symposium on Image and Video Technology, PSIVT, 2009. (Under Review, Submitted on 18<sup>th</sup> Aug)

## References

- [1] G. Potamianos, C. Neti, G. Gravier, A. Garg and A.W. Senior, "Recent Advances in the automatic recognition of Audio-Visual Speech", Proceedings of the IEEE, vol, 91, no. 9, pp. 1306-1326, 2003.
- [2] N. Eveno, A. Caplier, P.Y. Coulon. "A new color transformation for lips segmentation", Proceedings of IEEE Fourth Workshop on Multimedia Signal, pp. 3-8, Cannes, France, 2001.
- [3] N. Eveno, A. Caplier and P. Coulon, "Accurate and Quasi-Automatic Lip Tracking", Proceedings of IEEE Trans. Circuits Syst. Video Techn. 14(5): 706-715., 2004
- [4] X. Zhang and R.M. Mersereau, "Lip Features extraction Towards an Automatic Speech-reading System", Proceedings of ICIP00, Wa07.05, Vancouver, Canada, 2000.
- [5] T. Coianiz, L. Torresani and B. Caprile, "2D Deformable Models for Visual Speech Analysis", Proceedings of Springer, Speechreading by Humans and Machines, D.G. stork & M. E. Hennecke Eds., NY, 1996.
- [6] A. Hulbert and T. Poggio, "Synthesizing a colour algorithm from examples," Proceedings of Science, vol. 239, pp. 482-485, 1998.
- [7] L.G.D. Silveira, J. Facon and D.L. Borges, "Visual speech recognition: a solution from feature extraction to words classification", Proceedings of Computer Graphics and Image, SIBGRAPI03, pp. 399 - 405 ;XVI Brazilian Symposium on 12-15 Oct. 2003
- [8] J. Luettin, N.A. Thacker, and S.W. Beet, "Active Shape Models for Visual Speech Feature Extraction," Univ. of Sheffield, U.K., Electronic System Group Rep. 95/44, 1995.
- [9] Zhaorong Li; Haizhou Ai, "Texture-Constrained Shape Prediction for Mouth Contour Extraction and its State Estimation", Proceedings of Pattern Recognition, ICPR, 18th International Conference on, vol.2, pp. 88 - 91 , 20-24 Aug. 2006
- [10] Y. Tian, T. Kanade and J. Cohn, "Robust Lip Tracking by Combining Shape, Color and Motion", Proceedings of the 4th Asian Conference on Computer Vision (ACCV'00), January, 2000.
- [11] M. Heckmann, F. Berthommier and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition", Proceedings of EURASIP J. Appl. Signal, vol. 2002, pp. 1260-1273, Nov. 2002.
- [12] J. Chaloupka, "Automatic Lips Reading for Audio-Visual Speech Processing and Recognition", Proceedings of Of ICSLP, pp. 2505-2508, Jeju Island, Korea, Oct. 2004
- [13] N. Tsapatsoulis, Y. Avrithis and S. Kollias, "Effeicient Face Detection for Multimedia Applications", Proceedings of ICIP00, TA07.11, Vancouver, Canada, Sept. 2000.

- 
- [14] A.W.C. Liew, S.H. Leung, and W.H. Lau, "Segmentation of color lip images by spatial fuzzy clustering" Proceedings of IEEE Trans. Fuzzy Syst. vol. 11 no. 4, pp. 542-549, Aug. 2003.
- [15] M. Hennecke, V. Prasad and D. Stork, "Using deformable template to infer visual speech dynamics", Proceedings of 28<sup>th</sup> Annual Asimolar Conference on Signals, Systems and Computer, vol. 2, pp. 576-582, 1994.
- [16] T. F. Cootes, "Statistical models of appearance for computer vision", Online Technical Report Available from <http://www.isbe.man.ac.uk/bim/refs.html>, 2001.
- [17] Y. P. Guan, "Automatic Extraction of Lip Based on Wavelet Edge Detection", Proceedings of the Eighth international Symposium on Symbolic and Numeric Algorithms, Scientific Computing SYNASC. IEEE Computer Society, pp. 125-132. Sept. 2006
- [18] N. Chalapathy, P. Gerasimos, et al., "Audio-Visual Speech Recognition" IBM work shop.
- [19] A.V. Nefian, L.H. Liang, X. Liu, X. Pi, "Audio-Visual Speech Recognition", Intel Technology&Research.<http://www.intel.com/technology/computing/applications/avcsr.htm>
- [20] T.J. Hazen, "Visual Model Structures and Synchrony Constraints for Audio-Visual Speech Recognition", Proceedings of IEEE Transactions on Speech and Audio, 2006.
- [21] G. Potamianos, C. Neti, J.Huang, J.H. Connell, S. Chu, V. Libal, E. Marcheret, N. Haas and J. Jiang, "Towards Practical Deployment of Audio-Visual Speech Recognition", Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 3, pp. 777-780, 2004.
- [22] A. Shamaie and A. Sutherland, "Accurate Recognition of Large Number of Hand Gestures", Proceedings of Iranian Conference on Machine Vision and Image, University of Technology, Tehran, 2003.
- [23] E. D. Petajan, "Automatic lipreading to enhance speech recognition", Ph.D. dissertation, Univ. Illinois, Urbana-Champaign, 1984.
- [24] E. D. Petajan, B. Bischoff, D. Bodoff and N. M. Brooke, "An improved automatic lipreading system to enhance speech recognition", Proceedings of the SIGCHI Conference of Human Factors in Computing Systems, pp. 19-25, 1988.
- [25] C. Bregler and S.M. Omohundro, "Non-linear manifold learning for visual speech recognition", Proceedings of the International Conference on Computer Vision, 1995, pp. 494-499.
- [26] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox and R. Harvey, "Extraction of Visual Features for Lip-reading", Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, vol. 24, no. 2, pp. 198-213.
- [27] W. Yau, K. D. Kant and A. S. Poosapadi. "Visual recognition of speech consonants using facial movement features", Proceedings of Integrated Computer-Aided Engineering, 2007, vol. 14, no. 1, pp. 49-61.
- [28] R. Harvey, I. Matthews and J.A. Bangham and S. Cox, "Lip reading from scale-space measurements", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1997, pp. 582-587.

- 
- [29] X.P. Hong, H.X. Yao, Y.Q. Wan and R. Chen, "A PCA Based Visual DCT Feature Extraction Method for Lip-Reading", Proceedings of Intelligent Information Hiding and Multimedia Signal Processing, 2006, pp. 321-326.
- [30] Y. Fu, X. Zhou, M. Liu, M. Hasegawa-Johnson and T. S. Huang, "Lip-reading by Locality Discriminate Graph", Proceedings of the IEEE International Conference on Image Processing (ICIP), 2007.
- [31] K. D. Lee, M. J. Lee and S. Y. Lee, "Extraction of Frame-difference features based on PCA and ICA for Lip-reading", Proceedings of International Joint Conference on Neural Network, vol. 1, pp. 232-237, Aug 2005.
- [32] M. Gordan, C. Kotropoulos and I. Pitas, "Application of support vector machines classifiers to visual speech recognition", Proceedings of the 2002 Int. Conf. on Image, 2002.
- [33] Y. Freund and R. Schapire. "A short introduction to boosting", Journal of the Japanese Society for Artificial Intelligence, vol. 14, pp. 771-780, 1999.
- [34] A. J. Goldschen, "continuous automatic speech recognition by lipreading", Ph. D dissertation, George Washington Univ., Washington, DC, 1993.
- [35] G. Potaminanos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading", in Pro. Proceedings of Image Processing, vol. I, Chicago, IL, pp. 173-177, Oct. 1998
- [36] K.Yu, X. J, H. B, "Sentence lip-reading using hidden Markov Model with integrated grammar", Proceedings of World Scientific Series in Machine Perception and Artificial Intelligence Series, Hidden Markov models: applications in computer vision, 981-02-4564-5, Page 161-176, 2001
- [37] M.T. Chan. "HMM-Based Audio-Visual Speech Recognition Integrating Geometric- and Appearance-Based Visual Features". In J.-L. Dugelay and K. Rose, editors, Proceedings of the 2001 IEEE Fourth Workshop on Multimedia Signal Processing, pages 9--14, Cannes, France, October 2001.
- [38] S. W. Foo and L. Dong , "A boosted multi-HMM classifier for recognition of visual speech elements", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 6-10 April, 2003.
- [39] S.W. Foo and Y. Lian, "Recognition of visual speech elements using adaptively boosted HMM", Proceedings of IEEE Transactions on Circuits and Systems For Video Technology, 14(5), pp. 693-705, 2004.
- [40] L. Dong, S.W. Foo, Y. Lian, "A two-channel training algorithm for Hidden Markov Model and its application to lip reading," Proceedings of EURASIP Journal on Applied Signal, pp. 1382-1399, 2005.
- [41] W.C. Yau, D.K. Kumar, H. Weghorn, "Visual Speech Recognition Using Motion Featrues and HMM", Proceedings of CAIP, LNCS 4673, pp. 832-839 2007.
- [42] S. Werda, W. Mahdi, A.B. Hamadou, "Lip Localization and Viseme Classification for visual speech recognition", Proceedings of International Journal of Computing & Information Sciences, Vol. 5, pp. 62-75, No. 1, April 2007.
- [43] S. Lee and D. Yook, "Viseme recognition experiment using context dependent Hidden Markov Models", Proceedings of Third Intl. Conf. on Intelligent Data Engineering and Automated learning, Vol. 2412 pp. 557-561, 2002

- 
- [44] R.L. Hsu, M. Abdel-Mottaleb and A.K. Jain, "Face Detection in Colour Images," Proceedings of IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 696- 707, May 2002.
- [45] J. Yang and A. Waibel, "A Real-Time Face Tracker," Proceedings of the IEEE Workshop Applications of Computer Vision, pp. 142-147, Dec. 1996.
- [46] M.H. Yang and N. Ahuja. "Detecting human faces in colour images2", Proceedings of International Conference on Image Processing ICIP, vol.1, pp. 127-130.
- [47] B.D. Zarit, B.J. Super and F.K.H. Quek, "Comparison of five colour models in skin pixel classification". Proceedings of the ICCV Int'l Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems, pp. 58-63, 1999.
- [48] J.Y. Lee and S.I. Yoo, "An elliptical boundary model for skin colour detection", Proceedings of the International Conference on Imaging Science, Systems and Technology, 2002.
- [49] J. Y. Kim, S. Y. Na and R. Cole "Lip Detection using Confidence-based Adaptive Thresholding", Proceedings of Springer, Advance in Visual Computing, vol. 4291, pp. 731-740, 2006.
- [50] A. Abutaleb, "Automatic thresholding of gray level pictures using two-dimensional entropy", Proceedings of Computer Graphics and Image Understanding, 41(1), pp. 22-32, 1989.
- [51] S. Roweis, "EM Algorithms for PCA and SPCA", Proceedings of Advances in Neural Information Processing Systems, vol. 10, pp. 626-632, 1998.
- [52] T.K. Moon, "The EM Algorithm", Proceedings of IEEE Signal Processing Magazine, November 1996.
- [53] J. Bilmes, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", Proceedings of International Computer Science Institute and Department of Electrical Engineering and Computer Science U.C. Berkeley, TR-97-021, April 1998.
- [54] I.S. Pandzic, R. Forchheimer, Editors, "MPEG-4 Facial Animation – The standard, Implementation and Applications", John Wiley & Sons Ltd, ISBN 0-470-84465-5, 2002.
- [55] Say Wei Foo, Liang Dong, "Recognition of Visual Speech Elements Using Hidden Markov Models", Proceedings of IEEE Pacific Rim Conference on Multimedia, pp. 607-614, 2002.
- [56] W.C. Yau, D. K. Kumar, S. P. Arjunan, S. Kumar, "Visual Speech Recognition Using Image Moments and Multi-resolution Wavelet Images", Proceedings of Computer Graphics, Imaging and Visualisation, pp. 194-199, 2006
- [57] M. Leszczynski and W. Skarberk, "Viseme Recognition – a comparative study", Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, , pp. 287-292, 2005.
- [58] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu and T. Huang, "AVICAR: Audio-Visual speech Corpus in a Car Environment", Proceedings of Inter-speech - International Conference on Spoken Language Processing, Jeju, Korea, 2004.

- 
- [59] Y. Chang, C. Hu, and M. Turk. "Manifold of facial expression". Proceedings of IEEE International Workshop on Analysis and Modeling of Faces and Gestures, pp. 28-35, 2003.
- [60] G. A. Kalberer, P. Muller and L.V. Gool, "Visual Speech, a Trajectory in Viseme Space", Proceedings of International journal of imaging systems and technology vol.13, issue:1, pp.74-84, 2003.
- [61] M. Visser, M. Poel and A. Nijholt "Classifying Visemes for Automatic Lipreading", Proceedings of TSD' 99, LNAI 1692, pp. 349-352, 1999.
- [62] E. Bozkurt, C. E. Erdem, E. Erzin, T. Erdem and M. Ozkan, " Comparison of Phoneme and Viseme Based Acoustic Units for Speech Driven Realistic Lip Animation", Proceedings of Signal Processing and Communications Application, pp. 1-4, June, 2007.
- [63] K. Saenko, T. Darrel and J. Glass, "Articulatory features for Robust Visual Speech Recognition", Proceedings of ICMI, 2004.
- [64] K.C. Scott, D.S. Kagels, S.H. Watson, H. Rom, J.R. Wright, M. Lee and K.J. Hussey, "Synthesis of speaker facial movement to match selected speech sequences", Proceedings of 5th Australian Conf. On Speech Science and Technology, 1994
- [65] B. Rauch, "The use of Visual Information in Automatic Speech Recognition", Proceedings of Speech Signal Processing Group, IGK Annual Meeting, Saarland University, 14 July, 2005.
- [66] J. Yang, J. Xiao and M. Ritter, "Automatic Selection of Visemes for Image-based Visual Speech Synthesis", Proceedings of the IEEE International Conference on Multimedia and Expo, vol. 2, pp. 1081-1084, 2000.
- [67] C. A. Ratanamahatana and E. Keogh, "Everything you know about Dynamic Time Warping is Wrong", Proceedings of the 3<sup>rd</sup> SIGKDD Workshop on Mining Temporal and Sequential Data, 2004
- [68] S. Salvador and P. Chan. "Fast DTW: Toward Accurate Dynamic Time Warping in Linear Time and Space", Proceedings of the KDD Workshop on Mining Temporal and Sequential Data, pp. 70-80, 2004.
- [69] T. Oates, L. Firoiu and P. R. Cohen, "Clustering Time Series with Hidden Markov Models and Dynamic Time warping", IJCAI-99 workshop on sequence learning. 1999.
- [70] L. Gu, J.G. Harris, R. Shrivastav and C. Sapienza," Disordered Speech Evaluation Using Objective Quality Measures" Proceedings of Acoustics, Speech, and Signal Processing, 2005. IEEE International Conference on Volume 1, March 18-23, 2005.
- [71] D Jong, A. Kenneth, W M. Spear and D.F. Gordon, "Using Markov Chains to Analyze GAFOs". Proceedings of Foundations of Genetic Algorithms, pp. 115-137, 1994.
- [72] J. Yang and Y. Xu, "Hidden Markov Model for Gesture Recognition", Proceedings of Technical report CMU-RI-TR-94-10, Robotics Institute, Carnegie Mellon University, May, 1994.
- [73] L.R. Rabiner. "A tutorial on Hidden Markov Models and selected applications in speech recognition", Proceedings of the IEEE, 77(2), pp.257-286, 1989.

- 
- [74] L. Rabiner and B.H. Juang, "An Introduction to Hidden Markov Models", Proceedings of IEEE Signal Processing Magazine, January 1986.
- [75] Z. Ghahramani, Machine Learning Toolbox, version 1.0, 01-04-1996, University of Toronto.
- [76] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland, "The HTK Book", Proceedings of United Kingdom: Entropic Ltd, 1999.
- [77] Luettin, J. and Thacker, N. A. "Speech Reading using Probabilistic Models". Proceedings of Computer Vision Image Understanding, vol. 65, issue. 2, pp. 163-178, Feb. 1999.
- [78] T. F. Cootes, G. J. Edwards, and C. J. Taylor. "Active appearance models". Proceedings of IEEE TPAMI, 23(6):681–685, 2001.
- [79] F. Dornaika, J. Ahlberg, "Efficient Active Appearance Model for Real-time Head and Facial Features Tracking", Proceedings of IEEE international workshop on Analysis and Modeling of Faces and Gestures (AMFG), 2003.
- [80] P. Kakumanu, S. Makrogiannis, N. Bourbakis, "A Survey of Skin-color Modeling and Detection Methods", Proceedings of Pattern Recognition, Vol. 40, issue. 3, pp. 1106-1122, March, 2007.
- [81] H. Jun and Z. Hua, "A Real Time Lip Detection Method in Lip-reading", Proceedings of 26<sup>th</sup> Chinese Control Conference (CCC2007), pp. 516-520, July, 2007.
- [82] A. George, C. Tommy, H. Jef, S. Alistair, "Real-Time Hand Gesture Segmentation, Tracking and Recognition", Proceedings of 9th European Conference on Computer Vision (ECCV2006), Graz, Austria, May 7 - 13, 2006.
- [83] Z. M. Wang, L. H. Cai, H. Z. Ai, "A dynamic viseme model for personalizing a talking head," Proceedings of Signal Processing, 6th International Conference on, vol.2, no., pp. 1015-1018 vol.2, 26-30 Aug. 2002.
- [84] I. Ravyse, D. Jiang, X. Jiang, G. Lv, Y. Hou, M. Sahli and R. Zhao, "Proceedings of DBN Based Models for Audio-Visual Speech Analysis and Recognition", LNCS 4261, Advances in Multimedia Info processing (PCM) 2006.
- [85] A J. Goldschen, O. N. Garcia, E. Petajan, "Continuous Optical Automatic Speech Recognition by lipreading", Proceedings of 28<sup>th</sup> Asilomar Conference on Signals, Systems and Computers, 1994.
- [86] L. Dong, S. W. Foo and Y. Lian, "Modeling Continuous Visual Speech Using Boosted Viseme Model", Proceedings of IEEE, ICICS-PCM, Dec. 2003
- [87] Alaa El. Sagheer, N. Tsuruta, Rin. Taniguchi and S. Maeda, "Combination of Hypercolumn Neural Networks Model with Hidden Markov Model Based Lip-reading for Arabic Language", Proceedings of Introduction for Preparing FIC2004 Manuscript, 2004.
- [88] Durbin, R., S. Eddy, A. Krogh, and G. Mitchison, Proceedings of Biological Sequence Analysis, Cambridge University, 1998.
- [89] C. Bregler AND S.Omohundro,"Nonlinear image interpolation using manifold learning", Proceedings of Advances in Neural Information Processing Systems 7. 973—980, 1995.

## **Bibliography**

- [B1] P.F. Whelan and D. Molloy, “Machine Vision Algorithms in Java: Techniques and implementations”, ISBN 1-85233-218-2, TA1634, W54, 2000.
- [B2] S. Gong, S. J. McKenna and A. Psarrou, “Dynamic Vision from Images to Face Recognition”, ISBN 1-86094-181-8, 2000.