



**HAL**  
open science

# Paramètres spectraux à LPC Paramètres Mapping : approches multi-linéaires et GMM (appliqué aux voyelles françaises)

Zuheng Ming

► **To cite this version:**

Zuheng Ming. Paramètres spectraux à LPC Paramètres Mapping : approches multi-linéaires et GMM (appliqué aux voyelles françaises). Autre. Université de Grenoble, 2013. Français. NNT : 2013GRENT032 . tel-00935286

**HAL Id: tel-00935286**

**<https://theses.hal.science/tel-00935286v1>**

Submitted on 23 Jan 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## THÈSE

Pour obtenir le grade de

## DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Signal, Image, Parole, Telecom (SIPT)**

Arrêté ministériel : 1 Novembre 2009

Présentée par

### Zuheng MING

Thèse dirigée par **Denis BEAUTEMPS** et codirigée par **Gang FENG**

préparée au sein du CHU de Grenoble et du Laboratoire GIPSA-lab, Département Parole et Cognition

**École Doctorale Electronique, Electrotechnique, Automatique & Traitement du Signal (EEATS)**

# Spectral Parameters to Cued Speech Parameters Mapping: Multi-linear and GMM approaches (applied to French vowels)

Thèse soutenue publiquement le **24 Juin 2013**,

devant le jury composé de :

**Mme. Régine ANDRE-OBRECHT**

Professeur, Université Toulouse III, Toulouse, *Président*

**M. Christophe D'ALESSANDRO**

Directeur de recherche CNRS, LIMSI, *Rapporteur*

**M. Yves LAPRIE**

Directeur de Recherche CNRS, LORIA, *Rapporteur (non membre du jury de soutenance)*

**M. Rafaël LABOISSIERE**

Charge de recherche, CNRS, Lyon, *Examineur*

**M. Denis BEAUTEMPS**

Chargé de Recherche CNRS, GIPSA-Lab., *Directeur de thèse*

**M. Gang FENG**

Professeur, Université de Grenoble, GIPSA-lab., *Co-directeur de thèse*

*Université Joseph Fourier / Université Pierre Mendès France /  
Université Stendhal / Université de Savoie / Grenoble INP*



# UNIVERSITÉ DE GRENOBLE



*Université Joseph Fourier / Université Pierre Mendès France /  
Université Stendhal / Université de Savoie / Grenoble INP*

# Abstract

Cued Speech (CS) is a visual communication system that uses hand shapes placed in different positions near the face, in combination with the natural speech lip-reading, to enhance speech perception from visual input for deaf people. CS is largely improving speech perception for the profound deaf people and offers to deaf people a thorough representation of the phonological system. However one of the important challenges is the question of speech communication between normal hearing people who do not practice CS but produce acoustic speech and deaf people with no residual audition who use lip-reading complemented by CS code for speech perception. One solution to this problem is based on the development of automatic translation systems. In our work, we apply the multi-linear regression approach (MLR) and Gaussian Mixture Model (GMM)-based mapping approach to map acoustic spectral parameters to the hand position in CS and the accompanying lip shape. We hence contributed to the development of automatic translation system in the framework of visual speech synthesis.

We first apply the MLR approach in order to know the limit in terms of the performance of the linear mapping approach. The performance of the MLR approach is good for estimating the lip parameters from the spectral parameters since there is strong linear correlation between the lip parameters and spectral parameters. However, the performance of MLR approach for estimating the hand position is poor since there is no relationship between the hand positions and spectral parameters. The topology structures of the hand position space and acoustic space are entirely different. In order to simulate a similar topology structure of the acoustic space for estimating the hand positions, we introduce an intermediate space in which the hand positions are translated in coherence with the distribution of vowels in the acoustic space. It proves that the similar topology structure helps to improve the performance of the MLR for estimating the hand positions. But the classification errors in the intermediate space fail to achieve the good performance in the original hand position space.

In order to release the linear constraint of the MLR approach and overcome the defect introduced by the binary property of the classification method, we apply the GMM-

based mapping approach which has both the classification-partition and regression properties. The parameters of GMM are estimated by the supervised, unsupervised and semi-supervised training methods separately in the view of the machine learning theory. The GMM-based mapping approach based on the supervised GMM benefiting from the a priori phonetic information shows high efficiency and good robustness. However, the unsupervised training method without requiring a priori knowledge can explore the latent components of GMM underlying the training data automatically. The semi-supervised training method, which intending to include advantages of the supervised and the unsupervised training method does not show improvement in comparison with the two other training methods. The Minimum Mean Square Error (MMSE) and Maximum A Posteriori Probability (MAP) are used as regression criteria separately in GMM-based mapping approach. These two mapping criteria do not perform quite differently from each other. In addition, it suggests that the MLR approach is indeed a special case of GMM-based mapping approach when the number of the Gaussians equals to one. The GMM-based mapping approach can improve the performance significantly in comparison with the MLR by increasing the number of the Gaussians.

Finally, a continuous transition achieved by the linear interpolation in the acoustic space is introduced to compare the different mapping approaches. In the case of hand position estimation, the MLR approach still produces a continuous linear transition. Owing to the class-partitioning property, the GMM-based mapping approach improves the estimation performance significantly. Unlike the classification method, the GMM-based mapping approach shows a smooth changing between the stable phases. This is due to weighting the contributions of all Gaussians. However, the local regression of GMM-based mapping approach here is meaningless since the hand positions do not relate to the spectral parameters essentially. In the case of lip parameters estimation, the different mapping approaches show the similar overall trend due to the strong linear correlation between the source and target data.

Besides, a direct prediction of lip geometry features from the natural image of mouth region-of-interest (ROI) based on the 2D Discrete Cosine Transform (DCT) combined with a Principal Component Analysis (PCA) is proposed. The results show the

possibility to estimate the geometric lip features with good accuracy using a reduced set of predictors derived from the DCT coefficients.

Keywords: Cued Speech; Acoustic speech to Cued Speech mapping; multi-linear regression (MLR); GMM; MMSE; MAP.



# Acknowledgement

First of all, I would like to express my sincere gratitude to my advisors Dr. Denis BEAUTEMPS and Prof. Gang FENG, for their support, encouragement, guidance and reviewing this thesis. Thanks to them, I had the chance to evolve during the past three years in a stimulating scientific environment while enjoying great freedom in directing my work. I thank them for their abilities to listen, their wise counsel and great availability. I also would like to thank our collaborator, Professor Sébastien Schmerber, for his welcome at the CHU of Grenoble.

Then I want to express my gratitude to the members of the jury. Thanks to Christophe D'ALESSANDRO, Yves LAPRIE and Rafaël LABOISSIERE for agreeing to report on this work and to Régine ANDRE-OBRECHT to be the president of the jury.

I am grateful to Myriam Diboui, the speaker, for having accepted the recording constraints.

I wish to thank my girl friend Xiao ZHANG sincerely for her continued encouragement and support especially in the last phase of my work to help me to finish this thesis. I would like to thank my roommates Haiyang DING and Nan YU for their interesting discussion and giving me a good mood every day. I also want to thank all the friends for their selfless helps.

Thanks to all the members of Speech and Cognition Department and GIPSA-lab who give me a happy memory and enrich my life in France.

I would sincerely like to thank my mother Zunli JI, my father Fangxin MING, my brothers Zunyi MING, who accompanied and supported me throughout my course of studies.





# Content

|   |             |
|---|-------------|
| <b>Abstract.....</b>  | <b>i</b>    |
| <b>Acknowledgement .....</b>  | <b>v</b>    |
| <b>Content.....</b>   | <b>vii</b>  |
| <b>List of Figures.....</b>   | <b>xi</b>   |
| <b>List of Tables .....</b>   | <b>xxv</b>  |
| <b>Acronyms and terms.....</b>                                      | <b>xxix</b> |
| <b>Introduction.....</b>  | <b>1</b>    |
| <b>Chapter 1. State of the art of Cued Speech .....</b>             | <b>7</b>    |
| 1.1. Introduction .....   | 7           |
| 1.2. The Cued Speech system .....                                   | 7           |
| 1.3. LPC: The French version of Cued Speech.....                    | 10          |
| 1.4. The study of the Cued Speech and LPC .....                     | 11          |
| 1.4.1 Perception studies .....                                      | 11          |
| 1.4.2 Production studies .....                                      | 15          |
| 1.4.3 Automatic systems studies .....                               | 18          |
| <b>Chapter 2. Speech and Cued Speech material .....</b>             | <b>21</b>   |
| 2.1. Introduction .....   | 21          |
| 2.2. Database recording .....                                       | 21          |
| 2.3. Extraction of lip and hand visual features.....                | 22          |
| 2.3.1. Extraction of lip visual features .....                      | 22          |
| 2.3.2. Extraction of hand visual features.....                      | 27          |
| 2.4. Extraction of spectral parameters.....                         | 29          |
| 2.5. Structure of the database.....                                 | 32          |
| <b>Chapter 3. Mapping methods .....</b>                             | <b>35</b>   |
| 3.1. Introduction .....   | 35          |
| 3.1.1. Linear mapping methods.....                                  | 35          |
| 3.1. 2. Statistical mapping methods .....                           | 37          |
| 3.2. Multi-linear mapping method based on PCA Regression (PCR)..... | 44          |
| 3.2.1. Multi-linear regression method .....                         | 44          |
| 3.2.2. Multi-linear PCA Regression model.....                       | 45          |

|  |            |
|--|------------|
| 3.3. Gaussian Mixture Model (GMM) mapping method .....   | 50         |
| 3.3.1. Gaussian Mixture Model (GMM) .....  | 50         |
| 3.3.2. GMM parameters estimation .....   | 52         |
| 3.3.3. GMM-based mapping method .....  | 60         |
| <b>Chapter 4. Estimation of lip features from natural image of mouth region-of-interest.....</b> | <b>69</b>  |
| 4.1. Introduction .....  | 69         |
| 4.2. Experimental set-up and lip material .....  | 71         |
| 4.3. Image processing and DCT.....   | 71         |
| 4.3.1. Detection of the mouth region-of-interest (ROI) .....                                     | 71         |
| 4.3.2. The 2D discrete cosine transform (DCT).....   | 72         |
| 4.3.3. Using a mask.....   | 73         |
| 4.4. Modelling.....  | 74         |
| 4.4.1. The MLR estimation modelling .....  | 74         |
| 4.4.2. The GMM-based estimation modelling .....  | 76         |
| 4.5. Evaluation.....   | 78         |
| 4.5.1. Evaluation of MLR estimation model.....   | 79         |
| 4.5.2. Evaluation of GMM-based estimation model .....  | 82         |
| 4.6. Summary.....  | 88         |
| <b>Chapter 5. Speech to Cued Speech mapping: linear approach .....</b>                           | <b>91</b>  |
| 5.1. Introduction .....  | 91         |
| 5.2. Modelling.....  | 92         |
| 5.2.1. Definition of the predictors .....  | 92         |
| 5.2.2. The selection of the predictors .....   | 92         |
| 5.2.3. The prediction .....  | 93         |
| 5.3. Evaluation of lip feature prediction .....  | 93         |
| 5.4. Evaluation of hand position prediction .....  | 98         |
| 5.5. Prediction of the hand position in the intermediate space.....                              | 101        |
| 5.5.1. Intermediate space.....   | 101        |
| 5.5.2. Prediction procedure in intermediate space.....   | 103        |
| 5.5.3. Evaluation of the hand position prediction in intermediate space.....                     | 107        |
| 5.5.4. Remapping hand position to original space .....   | 112        |
| 5.6. Summary.....  | 122        |
| <b>Chapter 6. Speech to Cued Speech mapping: GMM approach.....</b>                               | <b>125</b> |
| 6.1. Introduction .....  | 125        |
| 6.2. GMM-based mapping approach .....  | 126        |

---

|   |            |
|---|------------|
| 6.2.1. GMM-based mapping approach for estimating lip parameters .....   | 126        |
| 6.2.2. GMM-based mapping approach for estimating hand positions .....   | 133        |
| 6.2.3. The selection of the predictors .....  | 135        |
| 6.3. Evaluation of GMM-based mapping approach for estimating lip parameters .....   | 136        |
| 6.3.1. Evaluation of GMM-based mapping approach with supervised trained GMM for estimating lip parameters.....                | 136        |
| 6.3.2. Evaluation of GMM-based mapping approach with unsupervised trained GMM for estimating lip parameters .....             | 144        |
| 6.3.3. Evaluation of GMM-based mapping approach with semi-supervised trained GMM for estimating lip parameters .....          | 147        |
| 6.4. Evaluation of GMM-based mapping approach for estimating hand positions.....  | 151        |
| 6.4.1. Evaluation of GMM-based mapping approach with supervised trained GMM for estimating hand positions .....               | 151        |
| 6.4.2. Evaluation of GMM-based mapping approach with unsupervised trained GMM for estimating hand positions.....              | 158        |
| 6.4.3. Evaluation of GMM-based mapping approach with semi-supervised trained GMM for estimating hand positions.....           | 162        |
| 6.4.4. Discussion about the GMM-based classification method and GMM-based mapping approach for estimating hand position ..... | 164        |
| 6.5. Discussion of the approaches used for estimating the lip parameters and hand position.....                               | 168        |
| 6.6. Discussion of the residual variance obtained by the different methods .....  | 172        |
| 6.7. Summary.....   | 172        |
| <b>Chapter 7. Conclusion and perspectives.....</b>  | <b>175</b> |
| <b>Bibliography .....</b>   | <b>179</b> |



# List of Figures

|  |    |
|--|----|
| Figure 1.1 Manual cues of CS designed by Cornett for American English. (Cornett, 1988) .....   | 8  |
| Figure 1.2 Manual cues of LPC(Heracleous et al., 2009).....  | 11 |
| Figure 1.3 Average percentages obtained by Alegria et al. (1999) in its experience of perception of words and pseudo-words in the condition of lip-reading alone or with the cues of LPC.....  | 15 |
| Figure 1.4 From top to bottom: horizontal x (cm) and vertical y (cm) hand motion paths are shown in the top two frames. The two bottom frames contain the lip area ( $cm^2$ ) time course and the corresponding audio signal for the [pupøpu] sequence (Beautemps et al., 2012). ..... | 17 |
| Figure 2.1: Landmarks placed between eyebrows and the extremity of the fingers to further extraction of the coordinates used as CS hand parameters. ....   | 22 |
| Figure 2.2: The selection of the lip images is according to the instants of vowels $t_{0i}$ .  | 23 |
| Figure 2.3 The lip parameters extraction based on the points on the internal lip contour. ....   | 24 |
| Figure 2.4 The extracted lip parameter B obtained by the method of Lallouache (1991). The red line was the value of the extracted lip parameter B and the blue circles denoted the errors in the case of pronouncing the vowels [ø, u]. ....   | 25 |
| Figure 2.5 The lip parameter B was miscalculated when the lip aperture was very small such as pronouncing of the vowel [ø]......   | 26 |
| Figure 2.6 The procedure of ellipse fitting of the internal lip contour based on the internal lip ROI extracted by the points selected manually. The white region in the binary image (BW) was extracted by polygon formed by the points selected manually                             |    |

---

on the internal lip contour and then used as the ROI for ellipse fitting of the internal lip contour. ....26

Figure 2.7 The length of the major axe of the fitting ellipse was slightly shorter than the measured lip width, the distance of A1 and A2, when pronouncing the vowel [i]. .....27

Figure 2.8 The extracted lip parameter B based on the fitting ellipse of the internal lip contour. The blue circles denote the frames in which the lip height was miscalculated by the method of Lallouache (1991). .....27

Figure 2.9 The point marked in blue located in the middle of the eyebrows was the original point O. The X and Y axes were the horizontal and vertical coordinates. The relative distance of the fingertip marked in blue relative to the original point O was defined as the hand position in our experiment. ....28

Figure 2.10 The middle finger indicated by the red cross on the left image always chosen as the guiding finger. Otherwise, the index finger was chosen as the guiding finger remarked by the red cross on the right image. ....29

Figure 2.11 The spectral envelop of the vowel [ε] of a sub corpus. The spectral envelops in the red were the normal ones while the spectral envelop in blue were the abnormal ones considered as the outliers of the corpus. ....31

Figure 2.12 The speech signal, the lip parameters A,B,S and X,Y coordinates of hand position. ....33

Figure 3.1: The procedure of the GMM-based mapping method. ....61

Figure 4.1: Image of the speaker (left) and the associated mouth ROI (right). The landmark in the center of the speaker’s eyebrows is used to locate the mouth ROI. ..72

Figure 4.2: DCT coefficients matrix of the natural image of the mouth ROI (on the left) and the partial enlargement of the triangle mask area in the DCT matrix (on the right). Each pixel in the image is corresponding to a coefficient obtained by the DCT of the natural image of mouth ROI. The more darker (red) color of the pixel, the larger the coefficient. The top-left pixels indicate the largest coefficients in the DCT matrix. The triangle mask selected the most significant coefficients located at the top left of the DCT matrix.....73

Figure 4.3:  $B$  values (red crosses) and its estimation (straight line) in function of  $F_1$  (with the use of a right-angled triangle mask of side 15). .....75

Figure 4.4:  $r_1$  values (crosses) and its estimation (straight line) in function of  $F_2$  (with the right-angled triangle mask of side 15). .....76

Figure 4.5: The RVAR of the training data for lip parameter  $B$  in function of the number of the ordered predictors  $F$  in the case of a triangle mask with side length 15.  $Var_0 = 0.17 \text{ cm}^2$ . .....80

Figure 4.6: The RVAR of test data for lip parameter  $B$  in function of the number of the ordered predictors  $F$  in the case of a triangle mask with side length 15.  $Var_0 = 0.16 \text{ cm}^2$ . .....81

Figure 4.7: Estimation curve of test data for lip parameter  $B$  given by the first 20 ordered predictors, the red line is the test data, the blue line are the estimated values. ....81

Figure 4.8: Estimation curve of test data for lip parameter  $B$  in the case of images containing fingers near the mouth given by the first 20 ordered predictors, the red line is the test data, the blue line are the estimated values.....82

Figure 4.9: The RVARs of training data for lip parameter  $B$  in function of the number of the ordered predictors  $F$ . The green line corresponds to the GMM-based estimation model with 3 components and the blue dotted line is corresponding to the MLR model.  $Var_0 = 0.17 \text{ cm}^2$ . .....83



Figure 4.10: The RVARs of test data for lip parameter B in function of the number of the ordered predictors F. The green line is corresponding to the GMM-based estimation model and the blue dotted line is corresponding to the MLR model.  $Var_0 = 0.16 \text{ cm}^2$  .....84

Figure 4.11: The RMSEs (cm) of test data for lip parameter B in function of the number of the ordered predictors F. The green line is corresponding to the GMM-based estimation model and the blue dotted line is corresponding to the MLR model. ....84

Figure 4.12: Estimated value of test data for lip parameter B given by the first 20 ordered predictor. The red line denotes the test data, the green line denotes the estimated value obtained by the GMM-based estimation model and the one obtained by the MLR model.....85

Figure 4.13: The RVARs of the test data for lip features (A, B, S, Aext, Bext, Sext) in function of the number of the ordered predictors F. ....85

Figure 4.14: The RMSEs(cm) of the test data for lip features (A, B, Aext, Bext) in function of the number of the ordered predictors F. ....86

Figure 4.15: The RMSEs( $\text{cm}^2$ ) of the test data for lip features (S, Sext) in function of the number of the ordered predictors F.....86

Figure 4.16: The RMSE of lip parameter B in function of number of the ordered predictors. The red ones are the RMSE of training data and the green ones are the RMSE of test data. From left to right within one set, the number of components is 3, 4 and 5.....87

Figure 4.17: Over-fitting problem of GMM with 3 components. The RMSEs of the lip features (A, B, Aext, Bext) of test data are divergence drastically when the dimension of GMM is more than 65. ....88

Figure 5.1: The average RVAR by 5-fold cross-validation of the training data for the lip parameter B in function of the number of predictors obtained by the different audio speech spectral parameters: 2 or 4 formant (fmt), MFCCs, LSP and mixture of MFCCs-LSP.....95

Figure 5.2 : Average RVAR by 5-fold cross-validation of the test data for the lip parameter B in function of the number of predictors obtained by the different speech spectral parameters: 2 or 4 formant (fmt), MFCCs, LSP and MFCCs-LSP Mixture. .97

Figure 5.3: Average RVAR by 5-fold cross-validation of the training data for the hand coordinate x in function of the number of predictors obtained by the different audio speech spectral parameters: 2 or 4 formant (fmt), MFCCs, LSP and mixture of MFCCs-LSP.....99

Figure 5.4: Average RVAR by 5-fold cross-validation of the training data for the hand coordinate y in function of the number of predictors obtained by the different audio speech spectral parameters: 2 or 4 formant (fmt), MFCCs, LSP and mixture of MFCCs-LSP.....99

Figure 5.5: The estimated value of hand x, y coordinates by the first 16 predictors derived from the mixture spectral parameters MFCCs and LSP on training data. The upper one is corresponding to the x coordinate and the lower one is corresponding to the y coordinate. The red line denotes the measured value and the solid green line denotes the estimated value. The vertical blue dash lines separate the hand positions belonging to different vowels. .... 100

Figure 5.6: Estimation of hand position shown in the x-y space, using 1.5 standard deviation ellipses to denote the distribution of estimated values (in green) and measured data (in red) for each CS hand position, the ‘+’ are the centers of groups of the coordinates. .... 101

Figure 5.7: The left figure plots the hand position (x, y) (cm) in the original space and the right one shows the new coordinates (x', y') (cm) in the intermediate space in which the original hand positions are translated to simulate the distribution of the

formant values in the formant space. The red ellipses (std=1.5) denote the distribution of the coordinates, the '+' denotes the center of each group. .... 102

Figure 5.8: The estimation procedure in the intermediate space. The estimated values of the hand coordinates in the intermediate space are finally remapped to the original space..... 103

Figure 5.9: The new hand coordinates (red crosses) in the intermediate space. The red ellipses (std=1.5) denote the distributions of the hand coordinates corresponding to each vowel and the black '+' denotes the center of each group corresponding to each vowel..... 107

Figure 5.10: Average RVAR in the intermediate space of the training data for the hand coordinate X in function of the number of predictors obtained by the different audio speech spectral parameters: 2 or 4 formant (fmt), MFCCs, LSP and mixture of MFCCs-LSP..... 108

Figure 5.11: Average RVAR in the intermediate space of the training data for the hand coordinate Y in function of the number of predictors obtained by the different audio speech spectral parameters: 2 or 4 formant (fmt), MFCCs, LSP and mixture of MFCCs-LSP..... 108

Figure 5.12: The hand coordinates (red crosses) and the estimated values (green crosses) in the intermediate space on the training data. The coordinates were estimated by 16 predictors from the mixture of MFCCs and LSP. The red ellipses (std=1.5) denote the distributions of data and the '+' denotes the center of each ellipse. .... 109

Figure 5.13: The estimated value of hand x, y coordinates by the first 16 predictors derived from the mixture spectral parameters MFCCs and LSP on training data in the intermediate spaces. The upper one is corresponding to the x coordinate and the lower one is corresponding to the y coordinate. The red solid line denotes the measured value and the green solid line denotes the estimated value. The vertical blue dash lines separate the groups of coordinates belonging to the different vowels. .... 110

Figure 5.14: The classification results of estimated hand coordinates by LDA in intermediate space on training data. The symbol ‘square’ shows the misclassification. The coordinates were estimated by 16 predictors from the mixture of MFCCs and LSP. The green ellipses (std=1.5) denote the distributions of the estimated hand coordinates. Each class of estimated hand coordinates is labeled by different symbols.  
 ..... 114

Figure 5.15: The classification results of estimated hand coordinates by QDA in intermediate space on training data. The symbol ‘square’ shows the misclassification. The coordinates were estimated by 16 predictors from the mixture of MFCCs and LSP. The green ellipses (std=1.5) denote the distributions of the estimated hand coordinates. Each class of estimated hand coordinates is labeled by different symbols.  
 ..... 115

Figure 5.16: Remapping estimated hand position from the intermediate space to the original space. Using 1.5 standard deviation ellipses to denote the distribution of estimated values (in green) and measured data (in red) for each CS hand location. The ‘+’ is the center of each group of the coordinates..... 117

Figure 5.17: The remapping results of estimated hand x, y coordinates on training data in original space. The upper one is corresponding to the x coordinate and the lower one is corresponding to the y coordinate. The red line denotes the measured value and the solid green line denotes the estimated value. The vertical blue dash lines separate the groups of coordinates belonging to the different vowels. .... 118

Figure 5.18: The misclassification individuals (obtained by QDA) in intermediate space. Individuals belonging to the different vowels are labeled with the different symbols. The green ellipse (std=1.5) denote the distribution of estimated hand positions. .... 119

Figure 5.19: The misclassification individuals (obtained by QDA) in original space. Individuals belonging to the different vowels are labeled with the different symbols. The green ellipse (std=1.5) denote the distribution of estimated hand positions..... 119

Figure 5.20: The remapping results of estimated  $x$ ,  $y$  coordinates on test data in original space. The upper one is corresponding to the  $x$  coordinate and the lower one is corresponding to the  $y$  coordinate. The red line denotes the measured value and the solid green line denotes the estimated value. The vertical blue dash lines separate the groups of coordinates belonging to the different vowels..... 121

Figure 6.1: The 3 different lip visemes are corresponding to the three groups of vowels highlighted by the three underlines with different colors. The measured values of lip width (A) and lip height (B) are denoted by the red dotted line. The vertical blue dash lines separate the groups of coordinates belonging to the different vowels..... 127

Figure 6.2: The three groups of the lip parameters in the A-B distribution plan are corresponding to the three lip visemes. The points in the plan are denoted by the red crosses, triangles and squares respectively corresponding to the different groups. The ellipses (std=2) in three different colors denote the distributions of the three groups of lip parameters and the red '+' denote the centers of the ellipses..... 128

Figure 6.3: The semi-supervised method for training GMM. The three groups trained by the supervised method are denoted by the three ellipses (std=2) with different colors. The Gaussians inside each group are trained by the unsupervised method denoted by the different colors. The color bar in the bottom of the figure indicates the colors corresponding to the Gaussians. The centers of the Gaussians are indicated by the labels 'c1', 'c2', ..., 'c10'..... 132

Figure 6.4: The target data, i.e. the hand coordinates, gather in five groups highlighted by the five underlines with different colors by the CS rule. The vertical blue dash lines separate the groups of coordinates belonging to the different vowels..... 134

Figure 6.5: The five groups of the hand position shown in X-Y plan. The points in the plan are denoted in the red crosses, triangles, squares, circle, plus sign corresponding to the different five groups. The ellipses (std=2) in five different colors denote the distributions of the five groups of hand position and the red '+' denote the centers of the ellipses..... 135

Figure 6.6: Average RVARs of the training data based on the supervised trained GMM with 10 Gaussians for estimating the lip parameters A,B and S in function of the number of predictors. MMSE is used as the regression criterion. .... 137

Figure 6.7: Average RMSEs of the training data based on the supervised trained GMM with 10 Gaussians for estimating the lip parameters A, B and S in function of the number of predictors. MMSE is used as the regression criterion. .... 138

Figure 6.8: Average RVARs and RMSEs of the training data for estimating the lip parameter B in function of the number of predictors based on the MLR approach and GMM-based mapping approach with 10 supervised trained Gaussians corresponding to the 10 vowels in the regression criteria of MMSE and MAP separately..... 138

Figure 6.9: Average RVARs of the test data based on the supervised trained GMM with 10 Gaussians for estimating the lip parameters A, B and S in function of the number of predictors. MMSE is used as the regression criterion. .... 139

Figure 6.10: Average RVARs and RMSEs of the test data for estimating the lip parameter B in function of the number of predictors based on the MLR approach and GMM-based mapping approach with 10 supervised trained Gaussians corresponding to the 10 vowels in the regression criteria of MMSE and MAP separately..... 140

Figure 6.11: The estimated value of the lip parameter B obtained by the GMM-based mapping approach with 10 supervised trained Gaussians. MMSE is used as the regression criterion. The red dotted line denotes the measured value of test data and the solid green line denotes the estimated value. The red circles indicate the misclassified individuals..... 140

Figure 6.12: Average RVARs of the training data based on the supervised trained GMM with 3 Gaussians for estimating the lip parameters A, B and S in function of the number of predictors. MMSE is used as the regression criterion. .... 141

Figure 6.13: Average RVARs and RMSEs of the training data for estimating the lip parameter B in function of the number of predictors based on the MLR approach and GMM-based mapping approach with 3 supervised trained Gaussians corresponding to the 3 lip visemes in the regression criteria of MMSE and MAP separately. .... 142

- 
- Figure 6.14: Average RVARs of the test data based on the supervised trained GMM with 3 Gaussians for estimating the lip parameters A, B and S in function of the number of predictors. MMSE is used as the regression criterion. .... 142
- Figure 6.15: Average RVARs and RMSEs of the test data for estimating the lip parameter B in function of the number of predictors based on the MLR approach and GMM-based mapping approach with 3 supervised trained Gaussians corresponding to the 3 lip visemes in the regression criteria of MMSE and MAP separately. .... 143
- Figure 6.16: The estimated value of the lip parameter B obtained by GMM-based mapping approach with 3 supervised trained Gaussians in the regression criterion of MMSE. The red dotted line denotes the measured value and the solid line denotes the estimated value. .... 143
- Figure 6.17: Average RVARs of the training data based on the unsupervised trained GMM for estimating the lip parameters in function of the number of the Gaussians/components. MMSE is used as the regression criterion. .... 144
- Figure 6.18: Average RMSEs of the training data based on the unsupervised trained GMM for estimating the lip parameters in function of the number of the Gaussians/components. MMSE is used as the regression criterion. .... 145
- Figure 6.19: Average RVARs and RMSEs of the training data for estimating the lip parameter B in function of the number of Gaussians based on the MLR approach, GMM-based mapping approach with 3 supervised Gaussians and unsupervised Gaussians in the regression criteria MMSE and MAP respectively. .... 146
- Figure 6.20: Average RVARs and RMSEs of the test data for estimating the lip parameter B in function of the number of Gaussians based on the MLR approach, GMM-based mapping approach with 3 supervised Gaussians and unsupervised Gaussians in the regression criteria MMSE and MAP respectively. .... 146
- Figure 6.21: Average RVARs of the training data based on the semi-supervised trained GMM for estimating the lip parameters in function of the number of the Gaussians. MMSE is used as the regression criterion. .... 147

Figure 6.22: Average RVARs of the test data based on the semi-supervised trained GMM for estimating the lip parameters in function of the number of the Gaussians. MMSE is used as the regression criterion..... 149

Figure 6.23: Average RVARs of the test data based on the semi-supervised trained GMM for estimating the lip parameters in function of the number of the Gaussians. The arbitrary covariance matrices instead of the common matrix are used in the MMSE regression method. .... 149

Figure 6.24: Average RVARs (on left) and RMSEs (on right) of the training data based on the supervised trained GMM with 10 Gaussians for estimating the hand position in function of the number of predictors. MMSE is used as the regression criterion. .... 151

Figure 6.25: Average RVARs (on left) and RMSEs (on right) of the test data based on the supervised trained GMM with 10 Gaussians for estimating the hand position in function of the number of predictors. MMSE is used as the regression criterion. .... 152

Figure 6.26: Average RVARs of the training data for estimating X (on left) and Y (on right) coordinates in function of the number of predictors based on the direct and indirect MLR approaches and the GMM-based mapping approach with 10 supervised trained Gaussians in the regression criterion of MMSE. .... 152

Figure 6.27: Average RVARs of the test data for estimating X (on left) and Y (on right) coordinates in function of the number of predictors based on the direct and indirect MLR approaches and the GMM-based mapping approach with 10 supervised trained Gaussians in the regression criterion of MMSE. .... 153

Figure 6.28: Average RVARs (on left) and RMSEs (on right) of the training data based on the supervised trained GMM with 5 Gaussians for estimating the hand position in function of the number of predictors. MMSE is used as the regression criterion. .... 153

Figure 6.29: Average RVARs (on left) and RMSEs (on right) of the test data based on the supervised trained GMM with 5 Gaussians for estimating the hand position in function of the number of predictors. MMSE is used as the regression criterion. .... 154



Figure 6.30: Average RVARs of the training data for estimating X (on left) and Y (on right) coordinates in function of the number of predictors based on supervised trained GMM with 5 Gaussians and supervised trained GMM with 10 Gaussians. .... 154

Figure 6.31: Average RVARs of the training data for estimating X (on left) and Y (on right) coordinates in function of the number of predictors based on supervised trained GMM with 5 Gaussians and supervised trained GMM with 10 Gaussians. .... 155

Figure 6.32: The estimated value of X coordinates of hand positions obtained by the supervised trained GMM with 5 Gaussians (the upper plan) and 10 Gaussians (the lower plan). The red dotted line denotes the measured value and the green solid line denotes the estimated value. The red circles denote the misclassified individuals of the supervised trained GMM with 10 Gaussians. .... 156

Figure 6.33: Average RVARs of the training data for estimating X (on left) and Y (on right) coordinates in function of the number of predictors based on supervised trained GMM with 5 Gaussians in the regression criteria MMSE and MAP respectively. .... 157

Figure 6.34: Average RVARs of the test data for estimating X (on left) and Y (on right) coordinates in function of the number of predictors based on supervised trained GMM with 5 Gaussians in the regression criteria MMSE and MAP respectively. .... 157

Figure 6.35: The estimated value of X coordinates of hand position obtained by the supervised trained GMM with 5 Gaussians under the criteria of MMSE (the upper plan) and MAP (the lower plan). The red dotted line denotes the measured value and the green solid line denotes the estimated value. The red circles denote the misclassification which is eliminated in the MAP method while the blue circles denote the misclassification which is aggravated in the MAP method. .... 158

Figure 6.36: Average RVARs (on left) and RMSEs (on right) of the training data based on unsupervised trained GMM for estimating the hand position in function of the number of Gaussians. MMSE is used as the regression criterion. .... 159

Figure 6.37: Average RVARs (on left) and RMSEs (on right) of the test data based on unsupervised trained GMM for estimating the hand position in function of the number of Gaussians. MMSE is used as the regression criterion. .... 159

Figure 6.38: Average RVARs of the training data for estimating X (on left) and Y (on right) coordinates in function of the number of Gaussians based on supervised trained GMM with 5 Gaussians and unsupervised trained GMM.. ..... 160

Figure 6.39: Average RVARs of the test data for estimating X (on left) and Y (on right) coordinates in function of the number of Gaussians based on supervised trained GMM with 5 Gaussians and unsupervised trained GMM.. ..... 160

Figure 6.40: The supervised trained GMM with 5 Gaussians projects on the (X, Y) coordinates plan. The five different colors indicated the different Gaussians corresponding to the five hand positions defined in the CS: side, cheek, mouth, chin and throat. .... 161

Figure 6.41: The unsupervised trained GMM with 5 Gaussians projects on the (X, Y) coordinates plan. The five different colors indicate the five different Gaussians trained automatically by EM algorithm. .... 162

Figure 6.42: Average RVARs (on left) and RMSEs (on right) of the training data based on the semi-supervised trained GMM for estimating the hand position in function of the number of Gaussians. MMSE is used as the regression criterion. .... 164

Figure 6.43: Average RVARs (on left) and RMSEs (on right) of the test data based on the semi-supervised trained GMM for estimating the hand position in function of the number of Gaussians. MMSE is used as the regression criterion..... 164

Figure 6.44: Average RVARs of X (on left) and Y (on right) of the training data in function of the number of predictors based on GMM-based classification method and the GMM-based mapping approach with 5 supervised Gaussians in the regression criteria of MMSE and MAP respectively. .... 165

Figure 6.45: Average RVARs of X (on left) and Y (on right) of the test data in function of the number of predictors based on GMM-based classification method and the GMM-based mapping approach with 5 supervised Gaussians in the regression criteria of MMSE and MAP respectively. .... 165

Figure 6.46: The estimated value of X coordinates of hand positions based on the GMM-based mapping approach (the upper plan) and the GMM-based classification method (the lower plan). The red dotted line denotes the measured value and the green solid line denotes the estimated value. The blue circles indicate the errors obtained by GMM-based mapping approach and classification method respectively. ....166

Figure 6.47: The estimated hand positions obtained by the GMM-based mapping approach. The reference positions are denoted in the red crosses and the estimated values are denoted in the green crosses. The blue arrow and the blue square with dotted line denote the deviation of a estimated hand position corresponding to the vowel [i]. ....167

Figure 6.48: The estimated hand positions obtained by the GMM-based classification method. The reference positions are denoted in the red crosses and the estimated values are denoted in the green crosses. The blue arrow and the blue square with dotted line denote that a hand position corresponding to the vowel [i] is misclassified as a hand position corresponding to the vowel [e]. ....167

Figure 6.49 The linear interpolation in the acoustic space between vowels [a] and [i]. ....170

Figure 6.50 The continuous transition of X coordinates of hand position achieved by linear interpolation between the vowels [a] and [i]. ....171

Figure 6.51 The continuous transition of lip parameter B achieved by linear interpolation between the vowels [a] and [i]. ....172

# List of Tables

|  |     |
|--|-----|
| Table 1.1 Percentages of correct reception of syllables and keywords obtained by Nicholls et al. (1982) in each of the presentation conditions.....  | 13  |
| Table 2.1 List of the ten French vowels with their occurrence.....   | 32  |
| Table 4.1: Performance of different masks. The line corresponding to each mask means the number of the predictors that explain 95% of the variance. If using all of the predictors still cannot explain 95% of the variance, the asymptotic value of explained variance is shown in the table..... | 79  |
| Table 5.1: The RVARs of the training data for the lip parameters with the predictors obtained by 2 Formant (Formant1, Formant2), 4 Formant (Formant1-Formant4), MFCCs, LSP and mixture of MFCCs and LSP respectively.....  | 96  |
| Table 5.2: The RMSEs of the training data for the lip parameters with the predictors obtained by 2 Formant (Formant1, Formant2), 4 Formant (Formant1-Formant4), MFCCs, LSP and mixture of MFCCs and LSP respectively.....  | 96  |
| Table 5.3: The RVARs of the test data for the lip parameters with the predictors obtained by 2 Formant (Formant1, Formant2), 4 Formant (Formant1-Formant4), MFCCs, LSP and mixture of MFCCs and LSP respectively.....  | 97  |
| Table 5.4: The RMSEs of the test data for the lip parameters with the predictors obtained by 2 Formant (Formant1, Formant2), 4 Formant (Formant1-Formant4), MFCCs, LSP and mixture of MFCCs and LSP respectively.....  | 98  |
| Table 5.5: Shift vector $\mathbf{v} = (\Delta x, \Delta y)^T$ (cm) which is varied depending on the different vowels. ....   | 102 |

|  |     |
|--|-----|
| Table 5.6: The average RVARs and RMSEs of the X and Y coordinates estimated by the different predictors in intermediate space on training data.....  | 111 |
| Table 5.7: The average RVARs and RMSEs of the X and Y coordinates estimated by the different predictors in intermediate space on test data.....  | 111 |
| Table 5.8: The average score of the LDA classification on the training and test data with the 5-fold cross-validation. ....  | 114 |
| Table 5.9: The average score of the QDA classification on the training and test data with the 5-fold cross-validation. ....  | 116 |
| Table 5.10: The average RMSE (cm) and RVAR of the remapping coordinates on the training and test data of the 5-fold cross-validation.....  | 120 |
| Table 5.11: The average $\overline{RMSE}_0$ (cm) of the direct MLR approach. The average $\overline{RMSE}_1$ (cm) of MLR approach obtained in the intermediate space. The average $\overline{RMSE}_2$ (cm) of the indirect MLR approach..... | 121 |
| Table 5.12: The average $\overline{RVAR}_0$ of the direct MLR approach. The average $\overline{RVAR}_1$ of MLR approach obtained in the intermediate space. The average $\overline{RVAR}_2$ of the indirect MLR approach. ....               | 122 |
| Table 6.1: The number of the components within each group in the training of the semi-supervised GMM.....  | 148 |
| Table 6.2: The best RVARs obtained by the GMM-based mapping approach with different GMMs and MLR approach. ....  | 150 |
| Table 6.3: The best RMSEs obtained by the GMM-based mapping approach with different GMMs and MLR approach. ....  | 150 |
| Table 6.4: The configuration of the number of the Gaussians belonging to the groups in the semi-supervised training process. ....  | 163 |

Table 6.5: The best RVARs obtained by the GMM-based mapping approach with different GMMs and MLR approach. ....168

Table 6.6: The best RMSEs obtained by the GMM-based mapping approach with different GMMs and MLR approach. ....168



# Acronyms and terms

GMM: Gaussian Mixture Model

HMM: Hidden Markov Model

MFCCs: Mel-Frequency Cepstral coefficients

LPC: Linear Predictive Coding

MLR: Multi-Linear Regression

LDA: Linear Discriminant Analysis

QDA: Quadratic Discriminant Analysis

MLE: Maximum Likelihood Estimation

MAP: Maximum A Posteriori Probability

MMSE: Minimum Mean-Square Error

PCA: Principal Analysis Component

PDF: Probability Density Function

RMSE: Root-Mean-Square Error

RVAR: Residual Variance

EM: Expectation-Maximization





# Introduction

The framework of this thesis is speech communication for orally educated deaf people. Speech is concerned here in its multimodal dimensions and in the context of automatic processing. Indeed, the benefit of visual information for speech perception (called “lip-reading”) is widely admitted. From the precursory works of Sumby et al. (1954), then those of Summerfield (1979) to those of Benoit et al. (1991) as far as the French language is concerned, it is well established that the visual information from the speaker’s face is used to enhance speech perception under noisy environment. Moreover, even in the context of clear auditory speech, vision remains important: shadowing experiments have shown, for instance, that performances are improved by an average factor of 7.5% in case of audiovisual stimuli in comparison with the simple audio presentation (Reisberg et al., 1987). The well known “McGurk effect” demonstrates the ability to integrate auditory and visual information even if the two modalities are not congruent (McGurk et al., 1976). As we have exposed here, normal-hearing people have competences in lip-reading (Cotton, 1935; Dodd, 1977). However, it has been shown that the initial performances – i.e. without specific training - vary greatly from one individual to another. Bernstein et al. (2000) have compared the performances of 96 normal-hearing people with 72 profoundly deaf people. The authors have observed very variable performances between the individuals of both groups, but have clearly showed that the best lip readers were deaf people.

However, even with high lip-reading performances, without knowledge about the semantic context, speech cannot be thoroughly perceived. The best lip readers scarcely reach perfection. Only 41.9% to 59.1% of the 10 different vowels are recognized in a [hVg] context (Montgomery et al., 1983) and 32% when relating to low predicted words (Nicholls et al., 1982). The main reason for this lies in the ambiguity of the visual pattern. However, as far as the orally educated deaf people are concerned, the act of lip-reading remains the main modality of perceiving speech. This led Cornett (1967) to develop the Cued Speech (CS) system as a complement to lip information.

CS is a visual communication system that uses of hand shapes placed in different positions near the face in combination with the natural speech lip-reading to enhance speech perception from visual input. This is a system where the speaker, facing the perceiver, moves his hand in close relation with speech (See also (Attina et al., 2004) for a detailed study on CS temporal organization in French language). CS is largely improving speech perception for hearing-impairing people (Nicholls et al., 1982). Moreover, CS offers to deaf people a thorough representation of the phonological system, inasmuch as they have been exposed to this method since their youth, and therefore it has a positive impact on the language development (Leybaert, 2000).

As we have seen from this short review, the CS method offers a real advantage for completing speech perception. Nowadays, one of the important challenges is the question of speech communication between normal hearing people who do not practice CS but produce acoustic speech and deaf people with no residual audition who use lip-reading completed by CS code for speech perception. To solve this question, one can use a human translator. Another solution is based on the development of automatic translation systems. For this, and in a more general framework, two sources of information could contribute to this translation operation: (i) *a priori* knowledge of the phonetic, phonologic and linguistic constraints; (ii) *a priori* knowledge of the correlations between the different vocal activity: neuronal and neuromuscular activities, articulator movements, aerodynamic parameters, vocal tract geometry, face deformation and acoustic sound. Different methods allow their modelling and their optimal merging with the input signals and the output ones. On an axis ordering the methods in function of their dependence upon the used language, one can find at the two extremities: (i) the method using the phonetic level of interface, combining speech recognition and speech synthesis to take into account speech phonology organization. Note that the recognition and synthesis processing can call on very various modelling techniques. If the phonetic models based on Hidden Markov Models (HMM) are the basis of the main recognition systems, synthesis based on concatenation of multi-parametered units of various lengths is still very popular. Note the increasing interest of synthesis by trajectory models based on HMM allowing the jointed learning of the recognition and synthesis systems (Tokuda et al., 2000; Zen et al., 2004; Zen et al., 2011); (ii) The methods using the correlation

between signals without the help of the phonetic level but using various mapping techniques. These techniques capture the correlations between input and output samples using Vector Quantification (VQ) or Gaussian Mixture Model (GMM) (Toda et al., 2004; Uto et al., 2006).

The classic method to convert audio speech to CS components consists of coupling a recognition system to a text-to-visual speech synthesizer (Duchnowski et al., 2000; Attina et al., 2004; Gibert et al., 2005; Beauteemps et al., 2007). In the classic method, the audio speech is usually recognized as the sequences of phonemes by the recognition system. Then the recognized phonemes are converted into sequences of codes for cues via finite-state grammar. Finally, the codes are sent to the display system to form the CS components and overlay the appropriate cues on the image. The link between the audio and CS in this method requires at least the phonetic level.

Before this work, no studies aimed at using the very low signal level. This thesis is a contribution to this challenge in the case of French vowels. A new approach based on the mapping from speech spectral parameters to the visual components made of CS and lip parameters is proposed. This will be detailed in Chapters 3, 5 and 6. In this context the objective of the mapping process is to deliver visual parameters that can be used as target parameters for visual speech synthesis.

Another issue of this thesis concerns the mapping (or estimation) of the geometric lip parameters by the appearance features obtained by the natural image of the mouth region-of-interest (ROI) without using the artifices of lips (e.g. using the blue artifices for extracting the lip contour). Various sets of visual features for automatic lip-reading have been proposed in the literature. In general, they can be grouped into three categories as suggested by Potamianos et al. (2012): shape based features, appearance based features and features that are a combination of both shape and appearance. In the first one, the inner and outer lip contours are extracted from the image view of the face. A lip contour model can be obtained statistically (Luettin et al., 1996; Dupont et al., 2000) or parametrically (Hennecke et al., 1996; Chiou et al., 1997). In the second one, visual feature vectors are obtained by appropriate transformations, such as Discrete Cosine Transform (DCT) or Principal Component Analysis (PCA). The transformations are applied to the pixels of the images

corresponding to the speaker's mouth ROI (Matthews et al., 1996; Gray et al., 1997). Lastly, the combinations of both shape and appearance based features have been used, e.g. Matthews (1998) built active appearance models based on both shape and appearance features. In our work, the innovative methods for mapping the lip appearance based features to the shape based features are presented.

The thesis is organized as follows:

**Chapter 1** presents the state of art of CS. In this chapter, the development of CS as well as the principles of its construction are presented firstly. Then, we report the studies proposed in the literature in the field of CS perception, CS production and the CS automatic synthesis system.

**Chapter 2** presents the experimental set-up and the material used in this work. The lip width, height and area of inner lip contour are extracted as the lip features. The hand positions related to the point located at the center of the eyebrows are extracted as the hand features. The formant, MFCCs, LSP and the concatenate parameters of the MFCCs and LSP are extracted as the acoustic speech features. Finally the synchronized lip parameters, hand positions and the acoustic spectral parameters constitute the corpus used in this work.

**Chapter 3** provides the detailed theoretical description of the approaches used in this work. The chapter starts with the MLR method based on the preliminary PCA processing, which is the most researched and direct approach in the regression or estimation problem. Then we present the more complicated stochastic mapping approach based on GMM. The Expectation–Maximization (EM) training method for GMM is elaborated in the following section. The Minimum Mean Square Error (MMSE) and Maximum A Posteriori Probability (MAP) used as the regression criteria in the GMM-based mapping approach are presented at last.

**Chapter 4** focuses on the estimation of the lip geometric parameters by the lip appearance based features obtained by the PCA on the DCT of the natural image of mouth ROI. Different DCT masks are used to select the DCT coefficients to optimize the estimation performance. Then the evaluation of two different methods,

i.e. the MLR and GMM-based regression methods are presented. Finally a summary of the comparison of the two methods is drawn.

**Chapter 5** presents the MLR for estimating the lip parameters and hand positions of CS by the acoustic speech features. Firstly the evaluation of MLR mapping approach with the different predictors derived from the PCA of the different spectral parameters (formant, MFCCs, LSP and mixture of the MFCCs and LSP) is presented. Then an intermediate space is introduced for estimating the hand position. The evaluation of the indirect MLR mapping approach is presented. Lastly, a summary of the MLR mapping approach is drawn.

**Chapter 6** presents the GMM-based mapping approach for estimating the lip and hand visual features of CS by the acoustic speech feature. We describe three different training methods for estimating the parameters of GMM firstly. In the following sections, the evaluation of GMM-based mapping approach with GMM trained in three different ways is presented. Then a discussion of approaches used for estimating the lip parameters and hand positions is made. Finally a summary is drawn in the end of the chapter.

Finally, the **Conclusion** summarizes the main results of this thesis and discusses suggestions for future work.



# Chapter 1. State of the art of Cued Speech

## 1.1. Introduction

Lip-reading which is the main modality to allow the access to speech for deaf people does not allow to recover some phonetic features of nasality, for example [p] vs [m], and voicing such as [p] vs [b]. This problem is due to the similarity of labial shapes. Thus, due to insufficient information provided by lip-reading, deaf children can not acquire and master speech by relying only on the traditional oral education. Several techniques have emerged to overcome this lack of information. In general, these techniques are based on visual cues, often coded with the hand, to provide additional information. In this section we present firstly the CS with a brief history of its development, and secondly, we will present a description of the principles of its construction. In addition, we focus on the adaptation of the CS to the French language. Finally, we report some studies in the field of production, perception and automatic synthesis system of CS, which show the effectiveness of this system in face to face communications for hearing-impaired people.







## 1.2. The Cued Speech system

The Cued Speech is a system using the hand codes as a complement to lip-reading, invented by Dr. Cornett and aims to enable hearing-impaired people access to spoken language. In fact, being the vice-president of the first university for the hearing-impaired people in the USA, Gallaudet College, Cornett discovered that prelinguistic children with profound hearing impairments have low reading comprehension. He understood that the access to spoken language is limited by their hearing impairment and the insufficient oralist re-education. Furthermore, existing system such as gestural sign language which is language-oriented is not efficient to improve reading performance. Thus Cornett developed a system aiming to complement lip-reading for speech perception.



In the system of CS, the hand placing at specific location around the face is used for coding the vowels, while the forms are used to encode the consonants. Cornett had built the system by setting two major criteria: the provided visual contrast should be maximum with the minimum effort for encoding. Thus, the number of configurations and positions is limited in order to save the effort for encoding and each position of the hand identifies a group of vowels meanwhile each configuration of the hand identifies a group of consonants. The final version defined by Cornett for American English is based on four hand positions and eight configurations (four groups of vowels and eight groups of consonants showing in Figure 1.1). The phonemes of each of these groups can be differentiated by the lip shape while the phonemes with the similar labial shape will be distinguished by the use of different hand position or configuration. For example, the consonants [p], [b] and [m] (which have the same lip shape) are encoded by the respective configurations 1, 4 and 5. Thus, the information of the hand (position or configuration) and lip shape allow the identification of a single percept.

**Cues for vowels and diphthongs**

|  |   |  |  |  |  |
|--|---|--|--|--|--|
| <br><b>Side (*)</b><br>ɑ: (father)<br>ʌ (but)<br>əʊ (home)<br>ə (the) | <br><b>Mouth</b><br>i: (see)<br>ɜ: (her) | <br><b>Chin</b><br>ɔ: (tall)<br>e (tent)<br>u: (blue) | <br><b>Throat</b><br>æ (that)<br>ɪ (is)<br>ʊ (book) | <br><b>Side-throat</b><br>eɪ (day)<br>ɔɪ (boy) | <br><b>Chin-throat</b><br>aɪ (my)<br>aʊ (how) |
|--|---|--|--|--|--|

**Cues for consonants**









|  |   |  |  |
|--|---|--|--|
| <br><b>Configuration 1</b><br>p (picture)<br>d (deep)<br>ʒ (treasure)         | <br><b>Configuration 2</b><br>k (caves)<br>v (visual)<br>ð (the)<br>z (cues) | <br><b>Configuration 3</b><br>s (sea)<br>r (rate)<br>h (horse)    | <br><b>Configuration 4</b><br>b (both)<br>n (name)<br>w (white)   |
| <br><b>Configuration 5</b><br>t (training)<br>m (mother)<br>f (father)<br>(*) | <br><b>Configuration 6</b><br>l (look)<br>ʃ (shell)<br>w (wet)               | <br><b>Configuration 7</b><br>g (give)<br>θ (thin)<br>dʒ (jogger) | <br><b>Configuration 8</b><br>j (you)<br>ɪŋ (young)<br>tʃ (child) |

Figure 1.1 Manual cues of CS designed by Cornett for American English. (Cornett, 1988)

On the other hand, maximizing the contrast in each group of phonemes requires a proper assembly. Cornett used visemes established previously by Woodward et al. (1960) to group together visually contrasting consonants. Untreated cases by Woodward et al. (1960) were classified by empirical choices. The frequency tables established by Denes (1963) were also used to group phonemes. The use of these tables is intended to minimize the energy during coding and to facilitate the movement of the hand for the combinations of the most common consonants in the language. Thus, the most common consonants like [t], [m] and [f] are encoded by the configurations easier to perform and require the least energy such as configuration 5. In terms of vowels, since the grouping visemes is difficult (especially at the time of the creation of CS), Cornett relied on visual features such as opening, rounding and stretching to distribute vowels in each position. The diphthongs are encoded by sliding of the hand between two positions of vowels. In some cases as to encode the word "papa" when the code is repeated, hand-arm together performs a front-to-back motion to indicate a repetition code. Finally, movements for coding the CS can be summarized in four types:

- The movement of the hand from one position to the other;
- The change in configuration of the hand;
- The shift between two positions to encode diphthongs;
- A slight front-to-back movement of hand-arm for indicating a code repetition.

The CS system allows the deaf to have a complete phonological representation of language and allows them to develop skills in reading and writing comparable to those with normal hearing (Leybaert et al., 1996). Thus, Cornett hoped to give opportunity to the deaf child “to acquire a complete and accurate model of spoken language by the visual channel” (Destombes, 1982). In addition, Cornett also wished to improve the social life of the hearing-impaired or deaf children. He wanted to reduce the barriers that complicate the initial communication between deaf children and their parents with normal hearing (Périer et al., 1987).

The CS system for the English language uses twelve keys (or "cues") so that they provide enough additional information to lip-reading to allow accurate identification of phonemes. However these keys used alone without lip-reading are confused.

The CS is defined by a syllabic system where the CV syllable is considered as the basic unit. Thus, the hand moves to a position simultaneous with a specific shape providing the code for the consonant C and the vowel V of the syllable CV. The combination of the position and configuration of the hand allows the speaker (receiver), for example, to differentiate the syllable [ma] with [pa] or [mi]. In special cases of an isolated consonant or an isolated vowel, a neutral configuration or position has been provided. Indeed, an isolated consonant is encoded by a single hand with the corresponding configuration pointing to position 'side' which is the "neutral" position (marked as '\*' in the Cues for vowels and diphthongs in Figure 1.1). Similarly, an isolated vowel is encoded by a single hand with the corresponding position presenting the No. 5 configuration which is the "neutral" configuration (marked as '\*' in the Cues for consonants in Figure 1.1).

### **1.3. LPC: The French version of CS**

The CS has been adapted to more than 60 languages and dialects around the world including French. The CS was imported to France in 1977. The adaptation has been called firstly *Langage Codé Cornett* (L.C.C.). Then, the adaptation has evolved into *Langage Parlé Complété* (LPC) with the effort of the association for the promotion and development of LPC (l'A.L.P.C.). And recently the name was changed to *Langue Française Parlée Complétée* (L.P.C.) to emphasize the fact that the LPC is based entirely on the French language.

Inherited from CS, the LPC has maintained the main criteria for its construction, namely the maximization of visual contrast in each group of the cues of LPC and minimizing of the effort of coding. The functioning principle is still the same. The coding unit is always the CV syllable. However, the adaptation to the French language raises some changes from the grouping of phonemes and number of cues (see Figure 1.2). Specifically, five hand positions are used to encode the vowels and eight hand configurations are used to encode the consonants. There is also no change

in respect to the rules for the isolated phonemes: the position 'side' is still the "neutral" position and configuration No.5 (hand fully extended) still remains as "neutral". All movements mentioned for CS therefore keep the same functions except for the sliding between two positions to encode diphthongs which are nonexistent in French.

Finally, it is very important to mention that the LPC relies on the phonetic transcription of what is spoken but not on the spelling. Thus a speaker communicating with deaf people codes everything what he utters (the syllable of the words) while the liaison between the words also has been taken into account. This makes it possible to transmit all phonological contrasts without ambiguity or confusion.

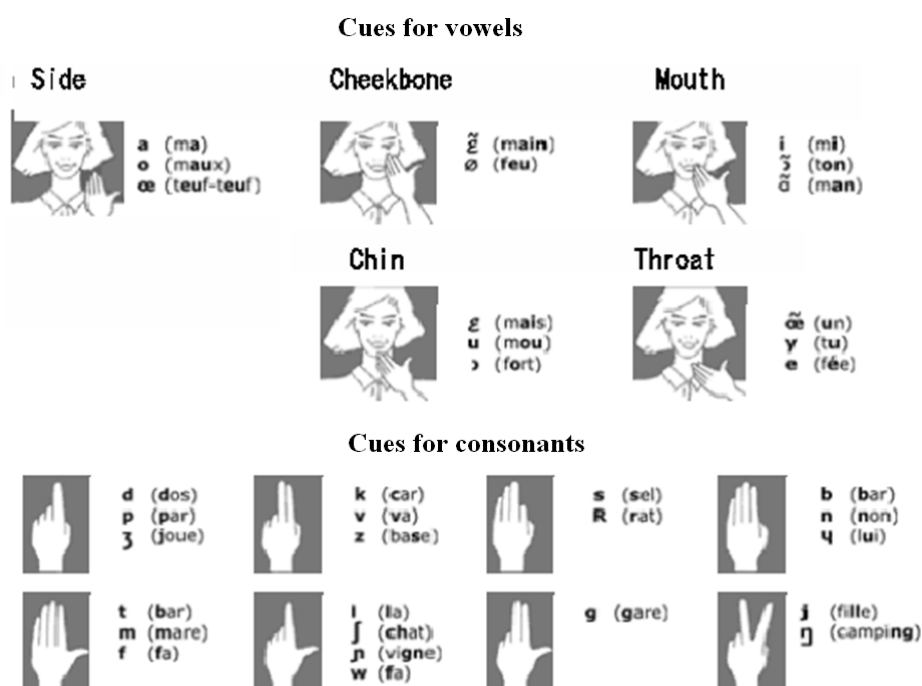


Figure 1.2 Manual cues of LPC (Heracleous et al., 2009).

## 1.4. The study of the Cued Speech and LPC

### 1.4.1 Perception studies

The value of LPC lies in its effectiveness in improving the speech perception. This is the main reason why this system was invented. The expansion of CS in the world by its adaptations to different languages and its increasing use in several environments

(at home or at school), demonstrate the effectiveness of the system for good reception of speech. The website of l'ALPC has presented the concrete evidence of parents using the LPC code to communicate with their deaf children and demonstrates the contribution of perceptual LPC.

On the experimental side, several studies have been conducted on different versions of CS in the world. For CS, in its original version we can cite the work of Ling et al. (1975), Clarke et al. (1976), Nicholls et al. (1982) and Uchanski et al. (1994). For the French version of the CS, we can find such studies of Charlier et al. (1990), Alegria et al. (1999), Attina et al. (2002), Attina et al. (2004), Attina (2005), Aboutabit (2007) and Heracleous et al. (2009). We present some of their work in the following section.

As the first truly systematic study on the perception of CS, Nicholls et al. (1982) tested the reception in seven different conditions of multiple messages by 18 hearing-impaired children aged 9-16 years. The participating hearing-impaired children chosen from an Australian school where the CS was practiced for ten years were exposed to the CS at least four years before this experience. Messages used for the experiment were categorized into two types. They were either syllable type (consonant-vowel (CV) or vowel-consonant (VC)) or keywords in sentences. Syllabic stimuli were constructed by combining 28 consonants and three vowels [a], [i] and [u]. Keywords were inserted at the end of sentences in two semantic contexts: either in a context that may help to predict the keyword (High-Predictability, HP) or not (Low-Predictability, LP). For example, the prediction of the word "purse" can be considered easy in the context of "Mum's money is in her purse." However, the task seems less easy for the word "room" in the context "Go in that room". The authors presented two types of stimuli in seven different experimental conditions: either with the audio, lip-reading or only keywords of the CS (respectively denoted as A, L and C) or by combining two of the above three conditions (AL, LC or AC) or the three conditions together (ALC). Table 1.1 summarizes the mean scores of perception obtained in this experiment.

Table 1.1 Percentages of correct reception of syllables and keywords obtained by Nicholls *et al.* (1982) in each of the presentation conditions.

| Conditions                          | A   | L    | C    | AL   | AC   | LC   | ALC  |
|-------------------------------------|-----|------|------|------|------|------|------|
| Results of syllables (%)            | 2,3 | 30   | 36   | 35   | 39   | 83,5 | 80,4 |
| Results of words in sentences LP(%) | 0,9 | 25,5 | 42,9 | 42   | 59,2 | 96,6 | 95   |
| Results of words in sentences HP(%) | 2,5 | 32   | 50   | 47,8 | 68,8 | 96,2 | 96   |

First, as regards the receipt of syllables, the first thing to note is that the scores of the condition ‘audio only’ are very low. This is quite reasonable since the subjects participating in this experiment are profoundly hearing-impaired. Second, it is the effect of vocalic context on the identification of labial vowels. The authors that in the conditions L, AL, LC and ALC where lip-reading is used the results in context of vowel [a] and [i] are better than rounded vowel [u]. Finally, the most important result drawn from this experience is that the performances of perception are significantly best when lip-reading is associated with coding manual of CS. Indeed, the percentage of correctly identified syllables in the condition LC and ALC is much higher (over 40% at least) than in the other conditions.

Moreover, the same remarks can be made in the case of receipt of keywords. The performances of the condition AL are significantly superior to the ones in conditions A and L only. Similarly, the performances of condition AC are significantly higher than in condition A and C only. The average scores of condition LC and ALC are not very different and still are superior to all other scores. The ambiguity or non receipt of the visual modality may explain the differences between these results. If the stimuli are presented in visual condition in which the reception of phonemes is ambiguous (condition L or C), the sound integrated with gesture is limited for the perception by the profoundly hearing-impaired people. However, if the visual message has no ambiguity (in condition LC) the sound brings no benefit to the perception.

On the other hand, the authors note an interaction between levels of contextual predictability (HP or LP) and conditions L, AL, AC and C (conditions where the visual information is ambiguous). Under these conditions, the scores of the level HP are indeed superior to those of the level LP. The differences in predictability are

however not significant for conditions LC and ALC. This suggests that the visual ambiguity of the message is mitigated by the semantic context.

Finally, the authors can conclude from this study that the word can be received clearly and accurately through vision only, without making any sound. The context semantics can contribute to the understanding of the message especially when it involves visual ambiguities. In this study the effectiveness of CS to improve speech perception for the hearing-impaired children is significant. The authors can therefore say that a hearing-impaired person can perceive the coded speech as well as a hearing normal person perceives the oral speech.

The work of Alegria et al. (1999) had two objectives: to deepen the notion that LPC enhances the lip-reading and to explore the way in which the information from these two streams are combined. For this purpose, words and pseudo-words were presented to 31 deaf children with different characteristics according to the age, which are the duration and the beginning of exposure to LPC. The participants were divided into two groups. The group "LPC-early" consisted of 7 children aged 8 to 12 years who were exposed to CS before the age of 2 years with an average duration of 9 years and 5 months. The rest of the children (24 children) were the second group called "LPC-late." Children in this group aged 11-19 years were exposed to LPC for an average duration of 6 years and 5 months after the age of 2 years. Words and pseudo-words were presented consisting of four phonemes following four structures: CV-CV, VC-CV, V-CVC, V-CCV. The authors used 8 words and 8 pseudo-words to constitute 64 combinations by the structures. Each combination was presented twice with or without using the LPC. The results obtained in this experiment are shown in Figure 1.3.

Firstly, the authors note that the use of CS in the two conditions (early or late exposure) significantly improves the reception performance of words and pseudo-words. However, the authors also notice some significant differences between the receipts of two types of combinations. Receipt of pseudo-words is indeed lower than that of words for both groups of participants. This can be explained by the fact that subjects rely only on the phonological information of the lip-reading and CS to identify a pseudo-word while they use also their lexical knowledge to determine a

word. The "lexical" effect may explain partly the inferiority of the percentages of pseudo-words obtained by the "LPC-late" compared to the group "LPC-early." In general, the performances of the children exposed early to LPC are superior to those exposed later. This result is confirmed by other studies showing the importance of the duration of exposure to LPC (Nicholls et al., 1982; Périer et al., 1987).

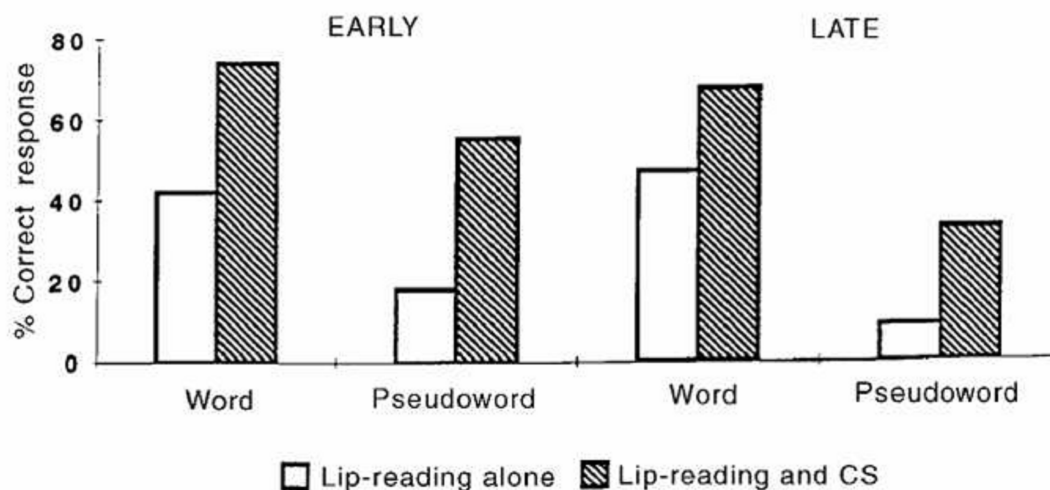


Figure 1.3 Average percentages obtained by Alegria et al. (1999) in its experience of perception of words and pseudo-words in the condition of lip-reading alone or with the cues of LPC.

In conclusion, lip-reading transmits only part of the information. The other part may also be transmitted visually by the cues of LPC. The ambiguity of lip-reading can be reduced by the LPC. This indicates that deaf children can benefit from this code to develop their oral language with the similar performance to the hearing normal people.

## 1.4.2 Production studies

Since the invention of the CS, no fundamental study has been devoted to the analysis of the skilled production of CS gestures, i.e. the temporal organization existing between lip movements and hand gestures, until the studies of Attina et al. (2002). Except for Cornett pointing out some consonant clusters where speech should be delayed to leave the hand enough time to reach the correct position (Cornett, 1967), the problems of cues presentation timing are only incidentally touched on in the



course of technological investigations. In the Cornett's Autocuer system (Cornett, 1988), cues are defined from the sound recognition of the pronounced word and are displayed in groups of LEDs on glasses worn by the speech-reader. The whole process involves a delay of 150 to 200 ms for the cue display compared to the production time of the corresponding sound. This system, designed for isolated words, attained 82% correct identification.

The pioneering work in this field was conducted at *l'Institut de la Communication Parlée* (ICP) to show how the hand movement co-produces on the consonant and the vowel in LPC. Firstly, Attina et al. (2002, 2004) have focused on the temporal organization of manual cues in conjunction with the movement of the lips and the corresponding acoustic signal. In these studies, Attina et al. (2004) shows an advance of the start of the movement of the hand with an average 200ms to the realization of acoustic syllable CV from the analysis of a LPC coder. In their experiments, the authors track the two-dimensional (2D) position of a point (marked by an easily identifiable tablet) placed in the middle of the back of the hand from a video recording of a LPC coder as well as the internal lips area. The hand movements are characterized by slow transitions according to the x and y trajectories in function of the time of the 2D position. The authors define the interval by the peak acceleration (M2) as the start and the peak deceleration (M3) as the end which is also the beginning of the gesture transiting towards the following position (Figure 1.4).

In this figure, we can observe some events selected by the authors to describe the different signals. We introduce the nomenclature used by the authors in the experiments:

- M1: the beginning of the transition from one hand position to another;
- M2: the start of holding the hand position;
- M3: the end of holding the hand position;
- L1: the beginning of the lip transition;
- L2: the end of the lip transition;

- A1: the start of the acoustic realization of the consonant;

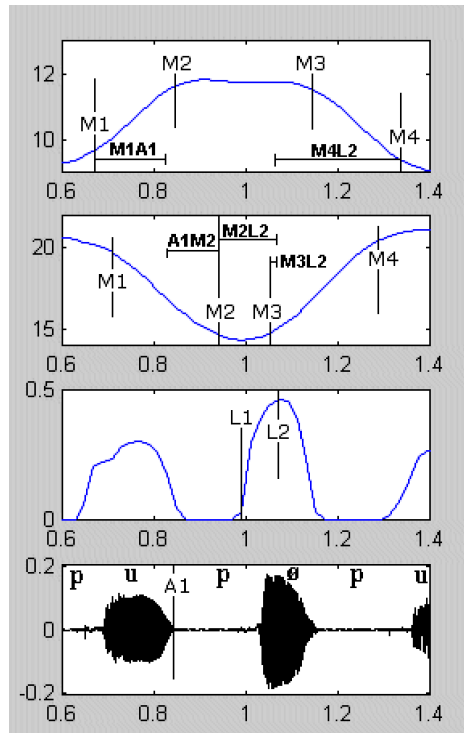


Figure 1.4 From top to bottom: horizontal  $x$  (cm) and vertical  $y$  (cm) hand motion paths are shown in the top two frames. The two bottom frames contain the lip area ( $\text{cm}^2$ ) time course and the corresponding audio signal for the [pupøpu] sequence (Beautemps et al., 2012).

The target of the hand is thus assumed to be reached when  $x$  and  $y$  simultaneously reach their plateau. Using this kinematic analysis of hand gesture, the authors show that the hand reaches the target position in quasi-synchronous with the beginning of the acoustic realization of the syllable CV, thus at the beginning of consonant before the acoustic realization of the vowel. The hand left its position to a new target before the peak of the lip area for the vowel.

The results of the advance of the hand have been confirmed by the analysis of the production of three additional coders (Attina, 2005). Each temporal pattern of hand-lips coordination of the three encoders is similar to that found in the first one. Attina (2005) studied with the experiment *Gating*, researching the advance of the hand before the labial gesture in the perception of LPC by 16 deep deaf subjects. The experiment consisted of cutting temporally each stimulus into several key points and presenting them gradually from the beginning to the different truncation points. If a

stimulus has  $p$  points, it is divided into a series with  $p+1$  presentations from the beginning ( $P_0$ ) to each key points ( $P_i$ ,  $i$  from 1 to  $p$ ):  $P_0$  to  $P_1$ ,  $P_0$  to  $P_2$ , ...,  $P_0$  to  $P_p$  and  $P_0$  to the end. In this experiment, participants practiced LPC daily except two of them. Stimuli tested were nonsense structures composed of five syllables with the type [mytymaCVma]. The fourth syllable (denoted CV) of the structure can vary and this is what subjects should identify. Each stimuli was truncated to 6 points around the CV syllable to be identified: The first point corresponds to the instant M1, that is to say it is the moment that the hand leaves his position beside the [ma] and passes to the position of the syllable [CV], and the second point corresponds to the configuration of the hand for the consonant C and the third corresponds to the instant M2, that is to say at the beginning of reaching the target position of the syllable CV; the fourth point corresponds to the instant when the configuration and position of the hand are identifiable; the fifth instant at the beginning of the movement of the lips to achieve the target vowel (L1); the sixth instant at the realization of the lips of the vowel (L2). The participants' task was to identify the coded syllable at each truncation point. It was shown that point No.4 is the point where the participants correctly identified over 80% of the manual groups of LPC. In fact this result is predictable because at this moment the manual information (on the configuration and position) is almost fully visible meanwhile the lip information is not prepared yet according to the temporal pattern observed in the production. This seems to confirm that the advance of the hand to the lips observed in the production is recovered in perception. This allows authors Attina et al. (2004) to propose the following hypothesis: the position of the hand could give a subset of vowels firstly and then the corresponding lips which would disambiguate information from the hand. These results were used as reference for the development of audio-visual synthesizer 2D first (Attina et al., 2004) and 3D of LPC (Gibert, 2006). Finally, note that these results were obtained on a corpus of logatomes by technical means and manuals.

### **1.4.3 Automatic systems studies**

In the Cornett's Autocuer system (Cornett, 1988), cues are defined from the sound recognition of the pronounced word and are displayed in groups of LEDs on glasses worn by the speech-reader. The whole process involves a delay of 150 to 200 ms for

the cue display, compared to the production time of the corresponding sound. This system, designed for isolated words, attained 82% correct identification.

In the system for the automatic generation of CS developed by Duchnowski et al. (2000) for American English, the cues are presented with the help of pre-recorded hands, and rules for temporal coordination with sound are proposed. This system uses a phonetic recognizer of audio speech to obtain a list of phones which are then converted to a time-marked stream of cue codes. The appropriate cues are visually displayed by superimposing hand shapes on the video signal of the speaker's face. The display is presented with a delay of two seconds, a delay that is necessary to correctly identify the cue (since the cue can only be determined at the end of each CV syllable). The superimposed hand shapes are always digitized images of a real hand. Using the system reported here, experienced cue receivers can recognize roughly 66% of the keywords in cued low-context sentences correctly, compared to roughly 33% by speech reading alone but they were still under the 90% level obtained with manual CS. In particular, human cuers often begin to form cues well before producing audible sound. To achieve a similar effect, in the “synchronous” display, the time at which cues were displayed was advanced by 100 ms relative to the start time determined by the recognizer. In addition, 150 ms was allocated to the transition provided the hand could pause at the target position for at least 100 ms.

Many works focus on the automatic recognition of CS. In automatic recognition of CS, lip shape and hand gesture recognition are required. In addition, the integration of the two modalities is of the greatest importance. In the study of Heracleous et al. (2009) which will be elaborated later, lip shape component is fused with hand gestures components to realize CS recognition. Using concatenative feature fusion and multi-stream HMM decision fusion to achieve the CS vowel and consonant recognition. With the automatic recognition of CS, people can translate automatically the CS to audio speech. The classic method to convert audio speech to visual speech consists of coupling a speech recognition system to a text-to-visual speech synthesizer (Duchnowski et al., 2000; Attina et al., 2004; Gibert et al., 2005; Beautemps et al., 2007). In the classic method, the audio speech is usually recognized as the sequences of phonemes by the recognition system. Then the recognized phonemes are converted

into sequences of codes for cues via finite-state grammar. Finally, the codes are sent to the display system to form the CS components and overlay the appropriate cues on the image. The link between the audio and CS in this method requires at least the phonetic level.

Heracleous et al. (2009) researched the automatic vowel recognition as used in the CS for French, combined with lip-reading and based on HMM. In automatic recognition of CS, lip shape and hand gesture recognition are required and the integration of the two modalities is importance. In this study, lip shape component is fused with hand position component to realize CS recognition. A 3-state, left-to-right with no skip HMM topology was used. Each state was modeled with 32 Gaussian mixtures. In addition to the static lip and hand parameters, the first and second derivatives were used as well. In automatic speech recognition, a diagonal covariance matrix is often used because of the assumption that the parameters are uncorrelated. In lip-reading, however, parameters show a strong correlation. In this study, PCA was applied to decorrelate the lip shape parameters. All 24 PCA lip shape components were used for HMM training and 3 visemes ( $V1:[\bar{\text{ɔ}}, \gamma, \text{o}, \emptyset, \text{u}]$ ;  $V2:[\text{a}, \bar{\text{e}}, \text{i}, \bar{\text{e}}, \text{e}, \varepsilon]$ ;  $V3:[\bar{\text{a}}, \text{o}, \bar{\text{e}}]$ ) were defined according to the same manner of the phonemes articulated in the visual domain. By using the multi-stream HMM decision fusion with 32 Gaussian mixtures per state, the vowel recognition accuracy obtained 87.6% showing a 19.6% relative improvement compared to the sole use of lip shape parameters.

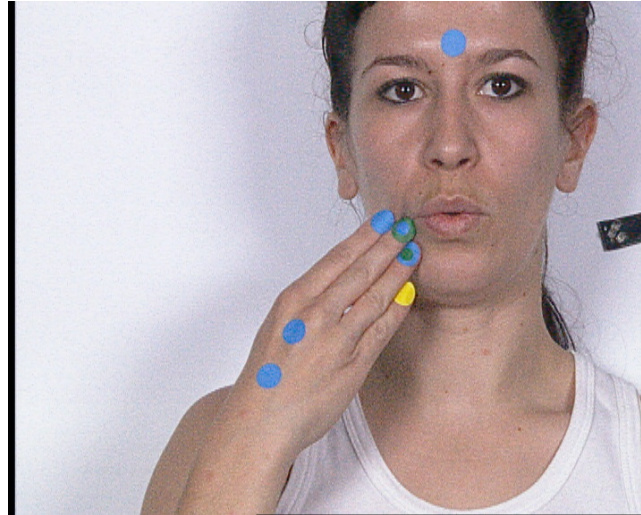
# Chapter 2. Speech and Cued Speech material

## 2.1. Introduction

Data base is indispensable to any statistical analysis. In our work, either the MLR approach or the GMM-based mapping approach is dependent on the statistical property underlying the data set. In our database, we tried to cover the ten French vowels used in the Cued Speech (CS). Each French word used in our experiment were repeated for 10 times to assure the statistical characteristic can be found in the experiment. Then the experimental set-up will be presented in the next section.

## 2.2. Database recording

The data derived from a video recording of a speaker pronouncing and coding in CS a set of 50 isolated French words. The words were made of 32 digits (from 0 to 31), 12 months and 6 more words that are ordinary. Each word was presented once on a monitor placed in front of the speaker, in a random order. The corpus has been uttered 10 times. The speaker is a female native speaker of French graduated in CS. The recording has been made in a soundproof booth and the image video recording rate was set on 25 images per second. The speaker seated in front of a microphone and a camera connected to a Betacam recorder. Landmarks were placed between eyebrows and at the extremity of the fingers to further extraction of the coordinates used as CS hand parameters (see Figure 2.1). In addition, a square paper was recorded for pixel-to-centimeter conversion.



*Figure 2.1: Landmarks placed between eyebrows and the extremity of the fingers to further extraction of the coordinates used as CS hand parameters.*

The video recording has been done with the PAL format, thus saved as numerical Bitmap RGB images made of the interlaced half-frames of the video (respectively even and odd lines). Each image was de-interlaced into two half-frames and the missing lines of the each half-frame were filled by linear interpolation, as to obtain two de-interlaced full frames corresponding to two recordings separated with 20 ms.

## **2.3. Extraction of lip and hand visual features**

These frames constitute the set of images at the rate of 50 Hz that we will refer to in the following. For each word, the coordinates of the inner contour of the lips have been manually selected on the corresponding image. Then the coordinates were converted into centimeters with the use of the pixel-to-centimeter conversion formula.

### **2.3.1. Extraction of lip visual features**

The lip features were extracted from the points on the inner lip contour. Before selecting the points on the lip images, the audio signal was used to locate the instants of the vowels. This was done with a frequency not higher than 50 Hz in order to synchronize it with the video image rate inside the isolated word. When the instants of vowels  $t_{0i}$  were fixed, the corresponding lip images can be selected from the video. Then the points on the internal lip contour can be selected manually from the lip images.

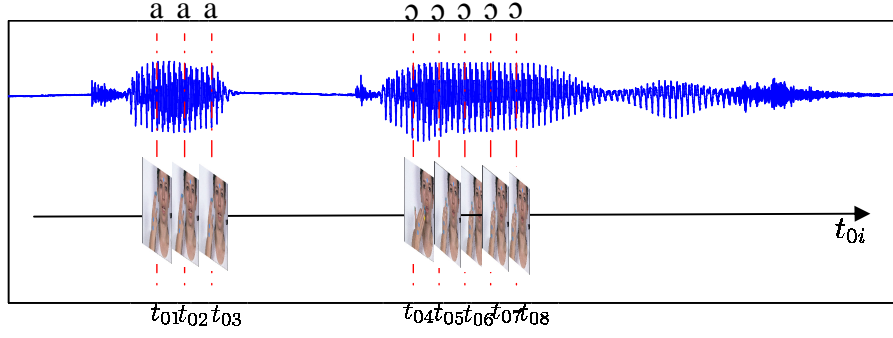


Figure 2.2: The selection of the lip images is according to the instants of vowels  $t_{0i}$ .

Following Gibbon et al. (2000), only a few parameters are considered to define lip shapes as lip features. These parameters are: the horizontal width of the lip, the vertical height and area of the internal lip contour, the lip protrusion and the lip contact point (Benoit et al., 1991). In the audio-visual speech or lip-reading research area the lip width, height and area were normally used as the visual features of the lips (Chen et al., 1997; Liew et al., 2002). In our work, we also used the lip width, height and area of the internal lips denoted as  $A$ ,  $B$  and  $S$  as the lip parameters as the visual features. According to Lallouache (1991), the lip parameters are calculated based on the points selected on the internal lip contour. The lip area was calculated directly from the points on the lip contour by the equation (2. 1):

$$S = \frac{1}{2} \sum_k (X_k Y_{k+1} - X_{k+1} Y_k) \quad (2. 1)$$

, the  $X_k, Y_k$  were the coordinates of the point on the lip contour. The lip width was calculated by the two points  $A1$  and  $A2$  which are the two extremes in the horizontal direction of the internal lip contour. The distance of the  $A1$  and  $A2$  was calculated as the lip parameter  $A$ . For calculating the lip height  $B$ , we defined two points  $B1$  and  $B2$  which are the intersections of the principal inertial axe and the internal lip contour. The distance of the  $B1$  and  $B2$  was calculated as the lip height  $B$ . The principal inertial axe passed through the barycentre  $(X_G, Y_G)$  of the internal lip contour which were calculated by the equations (2. 2) and (2. 3):

$$X_G = \frac{1}{2} \sum_k \frac{X_k^2 (Y_{k+1} - Y_k)}{S} \quad (2. 2)$$



$$Y_G = \frac{1}{2} \sum_k \frac{Y_k^2 (X_{k+1} - X_k)}{S} \quad (2.3)$$

The inertial moments  $J_x, J_y$  and  $J_{xy}$  were calculated by the equations (2.4), (2.5) and (2.6):

$$J_x = \frac{1}{3} \sum_k Y_k^3 (X_{k+1} - X_k) - Y_G^2 S \quad (2.4)$$

$$J_y = \frac{1}{3} \sum_k X_k^3 (Y_{k+1} - Y_k) - X_G^2 S \quad (2.5)$$

$$J_{xy} = \frac{1}{3} \sum_k X_k^2 Y_k (Y_{k+1} - Y_k) - X_G^2 Y_G^2 S \quad (2.6)$$

Then the angle  $\alpha$  of the principal inertial axe can be calculated by the inertial moments  $J_x, J_y$  and  $J_{xy}$  as shown in the equation (2.7):

$$\alpha = \frac{\pi}{2} - \frac{1}{2} \arctan\left(\frac{J_{xy}}{J_y - J_x}\right) \quad (2.7)$$

With the barycentre  $(X_G, Y_G)$  and the angle  $\alpha$ , the principal inertial axe can be fixed. Before calculating the intersections of the internal lip contour and the principal inertial axe, a parabolic curve is used to fit a part of internal lip contour near the intersection. The part of internal lip passing the nearest points (e.g. the nearest three points) to the principal inertial axe is selected to be fitted. Then the intersections of the principal inertial axe and the parabolic curves  $B1$  and  $B2$  were determined. The lip height  $B$  is the distance of the  $B1$  and  $B2$  (see Figure 2.3).

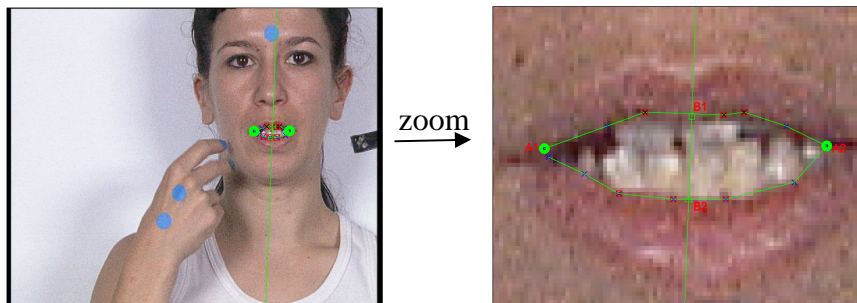


Figure 2.3 The lip parameters extraction based on the points on the internal lip contour.

This method works well in most of time. But when the lip height, i.e. lip parameter B, is very small, there are some problems with this method. For example, in the case of pronouncing vowels [ø, u], the lip height of the internal lip contour is very small and some of the extracted values of the lip parameter B are almost equal to zero. Figure 2.4 shows the errors on a sub corpus.

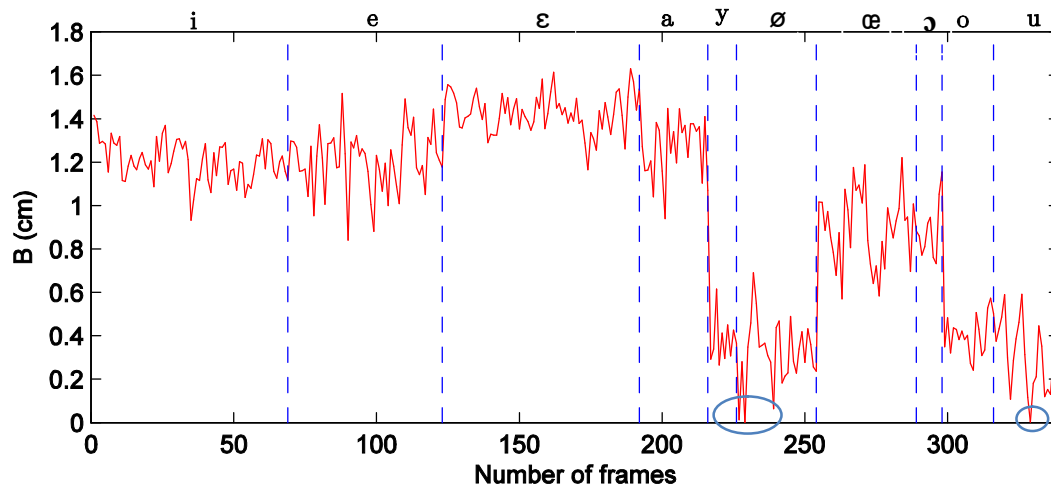


Figure 2.4 The extracted lip parameter B obtained by the method of Lallouache (1991). The red line was the value of the extracted lip parameter B and the blue circles denoted the errors in the case of pronouncing the vowels [ø, u].

The zeros values of lip parameter B were caused by the inappropriate parabolic fitting of the internal lip contour. When the lip aperture was small such as pronouncing the vowels [ø, u], it was difficult to select manually enough points on the internal lip contour for the appropriate parabolic fitting of the lip contour. Thus the wrong intersections of the principal inertial axe and the upper and lower internal lip contour were nearly overlapped and resulted in the distance of two intersections close to zero (see Figure 2.5).

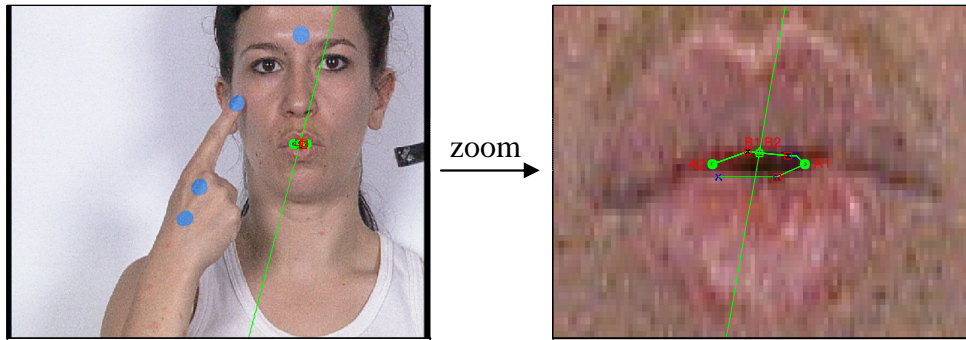


Figure 2.5 The lip parameter  $B$  was miscalculated when the lip aperture was very small such as pronouncing of the vowel  $[\emptyset]$ .

Thus it is better to use the pixels instead of points selected manually on the internal lip contour to calculate the lip parameter  $B$ , since there are enough pixels provided for fitting the lip contour even if the lip aperture was quite small. Following this idea, we firstly used the manually selected points to obtain the internal lip ROI and then used the ellipse to fit the internal lip contour to calculate the lip parameters (Leung et al., 2004). Figure 2.6 shows the fitting procedure based on the internal lip ROI.

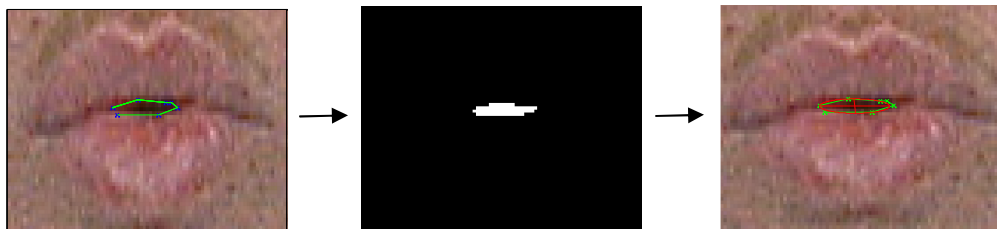


Figure 2.6 The procedure of ellipse fitting of the internal lip contour based on the internal lip ROI extracted by the points selected manually. The white region in the binary image (BW) was extracted by polygon formed by the points selected manually on the internal lip contour and then used as the ROI for ellipse fitting of the internal lip contour.

Then the length of the minor axis (the shorter axis) of the ellipse was used as the lip height, i.e. lip parameter  $B$ . In most of the time, the length of the major axis of the fitting ellipse was almost equal to the lip width  $A$ . But when the lip shape was very flat and long such as pronouncing the vowel  $[i]$ , the major axis of the fitting ellipse would be slightly shorter than the measured one. At this time, the lip width was closer

to the distance of the  $A1$  and  $A2$  (see Figure 2.7). Thus, we just used the fitting ellipse of the internal lip contour to calculate the lip height  $B$ . Figure 2.8 shows the extracted lip parameter  $B$  based on the internal lip ROI.

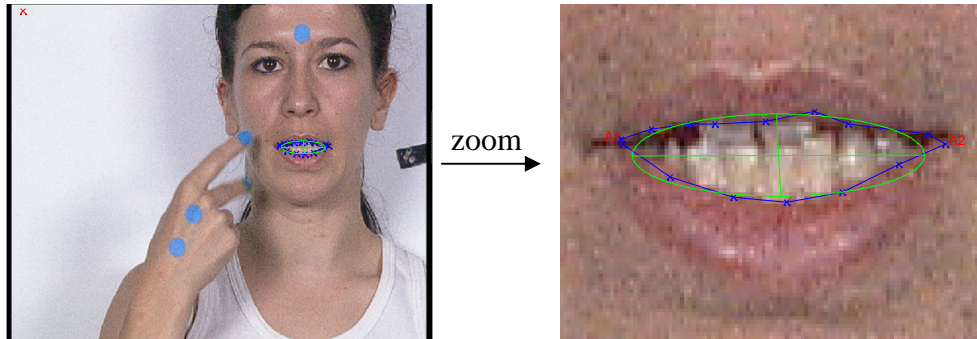


Figure 2.7 The length of the major axis of the fitting ellipse was slightly shorter than the measured lip width, the distance of  $A1$  and  $A2$ , when pronouncing the vowel  $[i]$ .

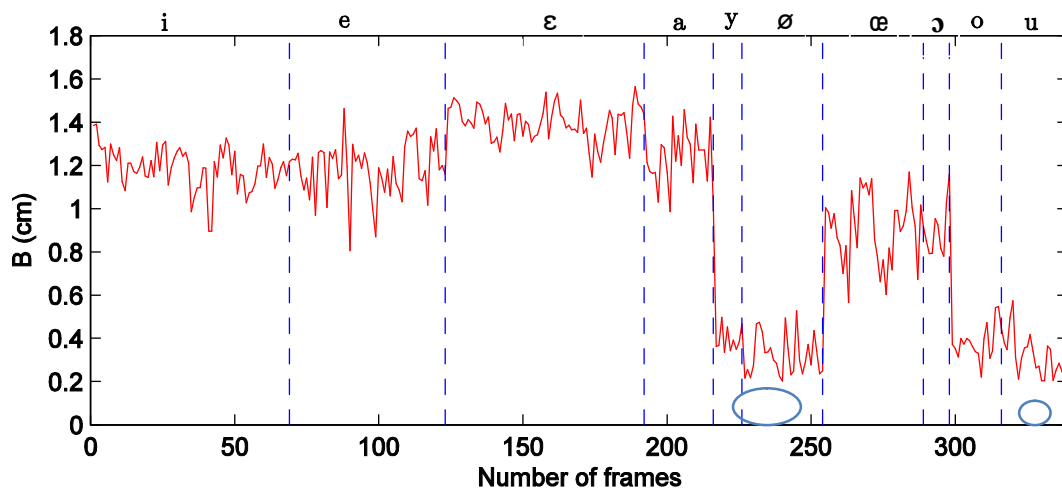
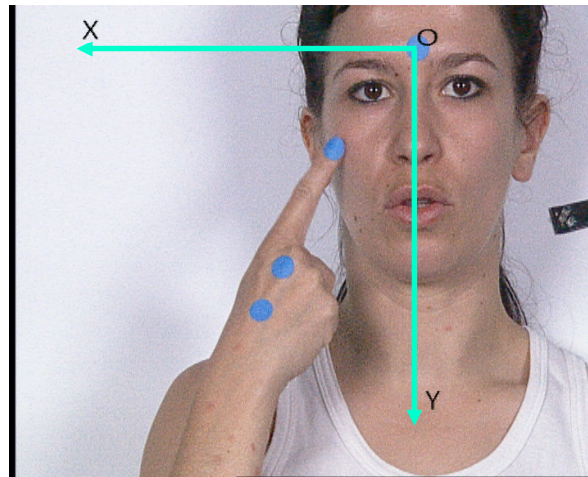


Figure 2.8 The extracted lip parameter  $B$  based on the fitting ellipse of the internal lip contour. The blue circles denote the frames in which the lip height was miscalculated by the method of Lallouache (1991).

### 2.3.2. Extraction of hand visual features

Unlike the lip shape which is normally synchronous with the acoustic speech, the hand position in CS is not always synchronous with the acoustic speech (Beautemps et al., 2012). This is due to the fact that speech is produced directly by the articulators including lips rather than the hand position or configuration used in the CS. Therefore,

we needed to select the instant  $t_{1i}$  at which the hand reaches the target position in synchronous with the acoustic realization at the instant  $t_{0i}$ . The point marked in blue in the middle of the eyebrows was set as original point (see Figure 2.9). The X axe is the horizontal coordinate and Y axe is the vertical coordinate. The hand position is defined as the relative distance of the fingertip marked in blue to the defined original point in our experiment.



*Figure 2.9 The point marked in blue located in the middle of the eyebrows was the original point  $O$ . The  $X$  and  $Y$  axes were the horizontal and vertical coordinates. The relative distance of the fingertip marked in blue relative to the original point  $O$  was defined as the hand position in our experiment.*

Noting that, in some of cases there are more than two fingers in the hand configurations to indicate the consonants in CS. In these cases, we need to choose one finger as guiding finger to determine the hand position. We always select the middle finger as the guiding finger. Otherwise, the index is chosen as the guiding finger (Beautemps et al., 2012) as shown in Figure 2.10.

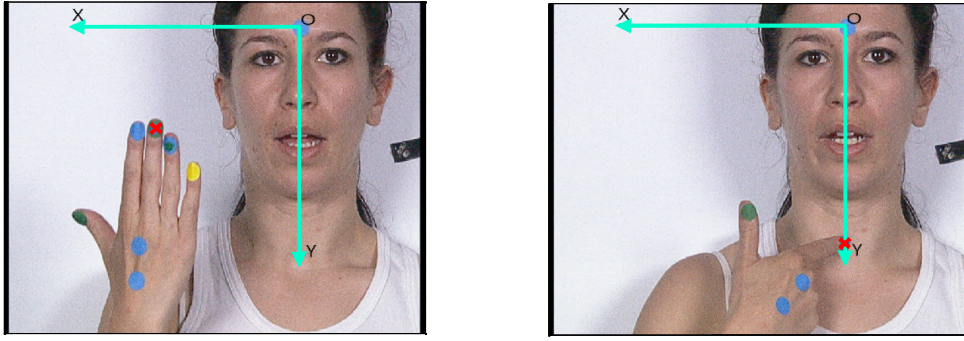


Figure 2.10 The middle finger indicated by the red cross on the left image always chosen as the guiding finger. Otherwise, the index finger was chosen as the guiding finger remarked by the red cross on the right image.

## 2.4. Extraction of spectral parameters

In this work, the audio of the recording was digitalized at 44100 Hz and re-sampled at 16000 Hz. Three kinds of spectral parameters of speech signal were used: formant values, Mel-Frequency Cepstrum Coefficients (MFCCs) and Line Spectral Pairs (LSP). One of our main objectives was to determine which parameters can give best performance when using them to predict the lips and hand parameters. As we know, formant values were the most pertinent to describe vowels and they presented relatively small variance from one speaker to other. However, the small number (2 to 4) of formants values limits the estimation performance. Instead, MFCCs and LSP offer sufficient spectral information for the estimation modelling.

MFCCs are calculated from a Mel-frequency-scaled short-term speech spectrum (in dB) by a DCT on the basis of a 32 ms Hamming window centered on  $t_{0i}$ . In such a way, MFCCs well describe the perceived speech spectrum in ear. The MFCCs are proved more efficient than other spectral parameters for speech recognition and speaker identification systems (Davis et al., 1980).

The calculation of the MFCCs included the following steps.

(1) Firstly, a short-term power spectrum of the windowed speech signal is calculated.

(2) This spectrum is then converted to Mel-frequency scale by using the following relation (Zwicker, 1961; Picone, 1993)

$$Mel(f) = 2595 \log_{10}(1 + f/100) \quad (2.8)$$

The conversion is done by using a set of frequency filters which have the same bandwidth in the Mel scale.

(3) A DCT is applied to the resulted spectrum to obtain MFCCs. Note that only the first 16 MFCCs were retained in this study.

LSP was derived from the Linear Predictive (LP) modelling of short-term speech spectrum (Itakura, 1975) on the basis of a 20 ms Hamming window centered on  $t_{0i}$ . As we know, the short-term spectral envelop can be efficiently obtained by using LP modelling. We used  $A(z)$  to denote the polynomial used in LP model which was determined from short-term correlation coefficients of speech signals. LSP described the roots of the following two polynomials ( $p$  is the order):

$$Q(z) = A(z) + z^{-(p+1)}A(z^{-1}) \quad (2.9)$$

$$R(z) = A(z) - z^{-(p+1)}A(z^{-1}) \quad (2.10)$$

The main property of the LSP was that all roots of  $Q(z)$  and  $R(z)$  are situated in the unit-circle, so that only one angle is sufficient to describe the corresponding root. In addition, if  $A(z)$  has a root which is near the unit-circle (often in the case for the first formants), two roots of  $Q(z)$  and  $R(z)$  would be situated in a small region close to the root of  $A(z)$ . The strong meaning of the LSP make it very robust to any quantification procedure, and very pertinent to describe the short-term speech spectral envelop.

The calculation of the LSP includes the following steps:

- (1) A standard LP analysis procedure by Levinson algorithm would give the  $p$  coefficients of  $A(z)$ .
- (2) Found the roots of  $Q(z)$  and  $R(z)$ .
- (3) Converted the roots' position (described by an angle) into a frequency value.

(4) Sorted the result. Note that only the first 16 LSP were retained in this study.

In practice, the formants were derived from the LSP coefficients. Thus we could use the obtained spectral envelop to check the outliers of the formants and remove the corresponding frame in the corpus. Figure 2.11 shows the spectral envelops of the vowel [ε] of a sub corpus. We could see that the first 4 peaks of the spectral envelop corresponding to the first 4 formants. Most of the frames presented the same waveforms except the ones marked in the blue which were obviously different from the others. These frames with the abnormal waveform of the spectral envelop were considered as the outliers in the corpus.

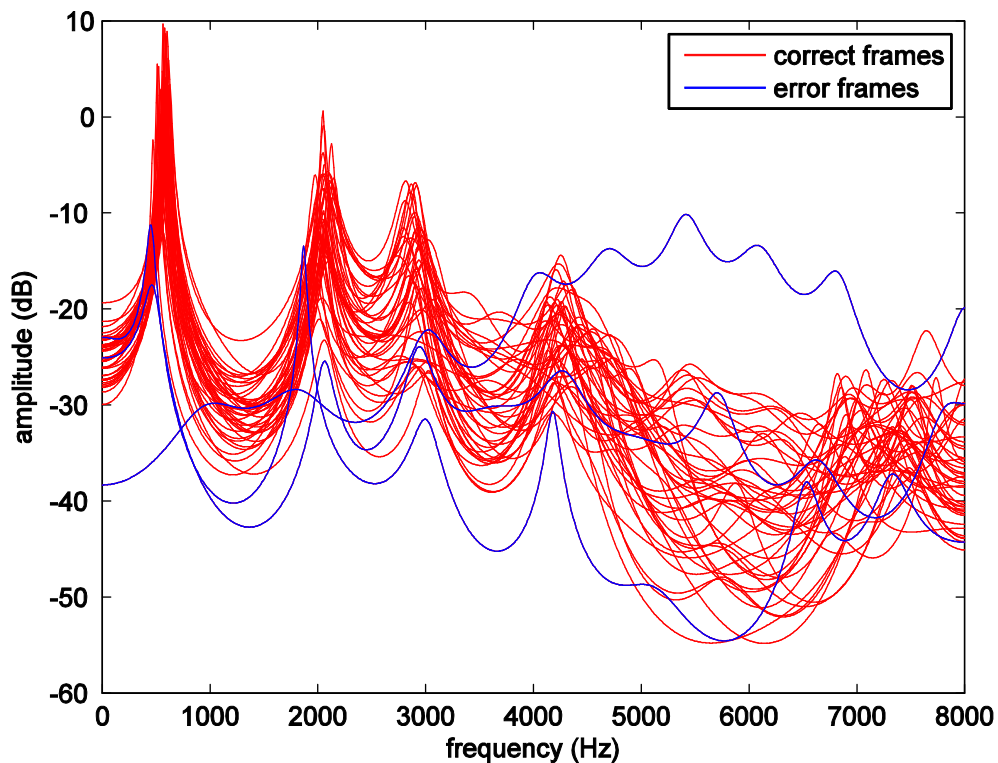


Figure 2.11 The spectral envelop of the vowel [ε] of a sub corpus. The spectral envelops in the red were the normal ones while the spectral envelop in blue were the abnormal ones considered as the outliers of the corpus.



## 2.5. Structure of the database

The work presented in this thesis focused on French vowels. The audio signal was used firstly to locate the  $t_{0i}$  instants corresponding to vowels inside the words. Then the lip images were selected from the video according to the  $t_{0i}$  instants. The lip features, i.e. the lip width (A), the lip aperture (B) and the lip area (S), were extracted respectively from the lip images. Meanwhile the 16 MFCCs and 16 LSP coefficients were calculated by using Hamming window at the  $t_{0i}$  instants. In addition, 4 formants were derived from the spectral envelop (obtained by the LSP coefficients). Since the CS hand position are not always synchronous with speech, the  $t_{1i}$  instants at which hand reaches the target positions were selected. Then the coordinates of the fingertip relative to the landmark between eyebrows were extracted as the hand position. The lip parameters, hand positions and spectral parameters constitute the database made of 1371 occurrences of the 10 French vowels (see Table 2.1 and Figure 2.12).

Table 2.1 List of the ten French vowels with their occurrence

| Vowels      | [i] | [e] | [ɛ] | [a] | [y] | [ø] | [œ] | [ɔ] | [o] | [u] |
|-------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Occurrences | 236 | 255 | 231 | 168 | 37  | 80  | 137 | 83  | 40  | 104 |

In the following work, the evaluation of the mapping approaches will be measured by the 5 folds cross-validation procedure. Thus the 1371 occurrences are divided into 5 partitions. The number of each vowel in each partition is proportional to the total number of vowel in the database. 4 partitions are used for training and the left one for testing in turn until all the partitions are used for both training and testing.

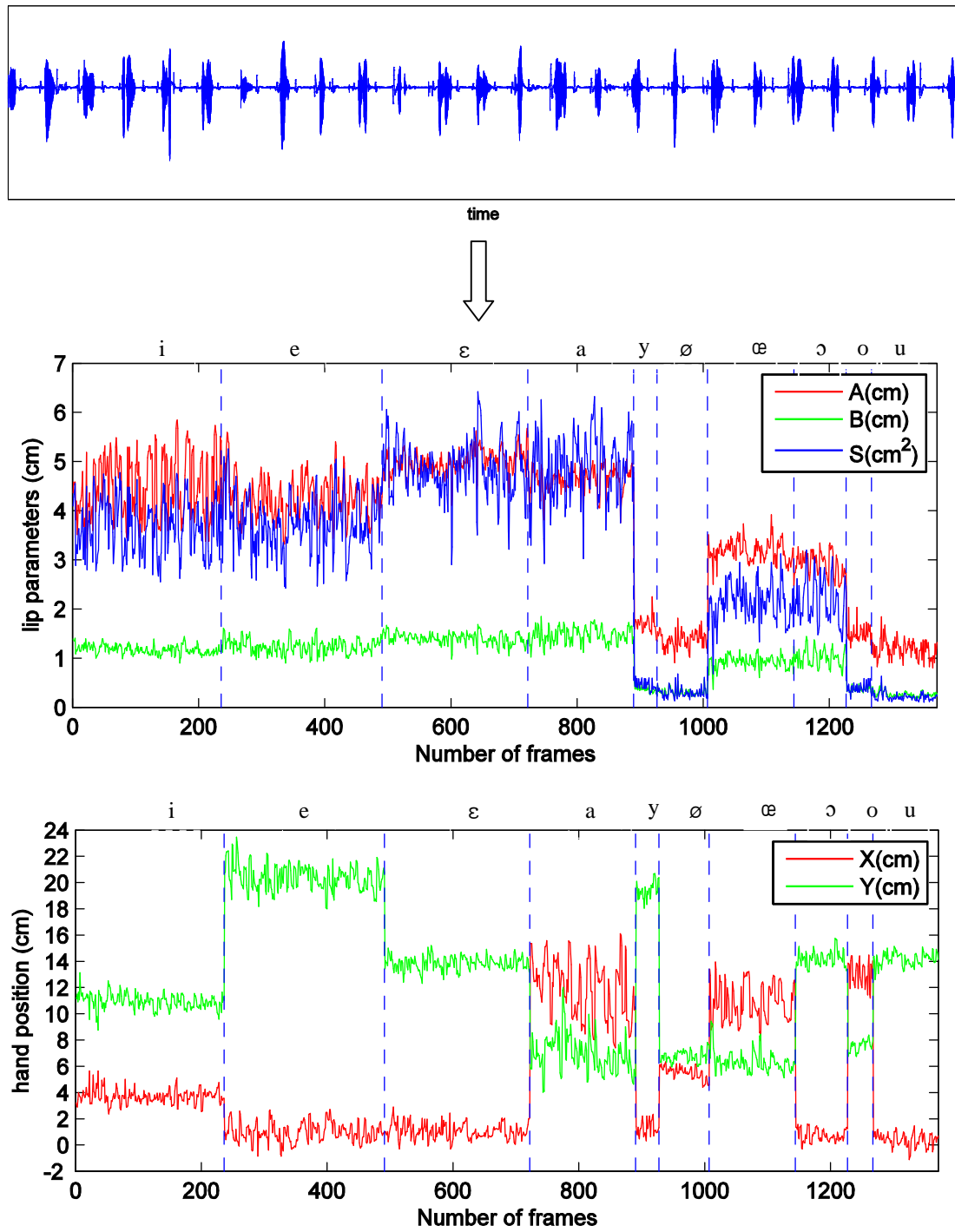


Figure 2.12 The speech signal, the lip parameters A,B,S and X,Y coordinates of hand position.



# Chapter 3. Mapping methods

## 3.1. Introduction

This chapter provides a detailed theoretical description of the approaches used in this work. Firstly, we introduce the general development of the linear mapping methods and the statistical mapping methods. Then the MLR approach based on the PCA processing and the more complicated statistical GMM-based mapping approach are presented in detail. The Expectation–Maximization (EM) training method for GMM is elaborated in the following section. Finally, the Minimum Mean Square Error (MMSE) and Maximum A Posteriori Probability (MAP) used as the regression criteria in the GMM-based mapping approach are presented in the end of the chapter.

### 3.1.1. Linear mapping methods

A straightforward way to evaluate the interrelation of the different sets of data collected is to use linear estimators to measure to what extent one data set can be determined from another (Yehia et al., 1998). Even in the case of the better representation of the relations of features that could be obtained by the elaborate nonlinear statistical methods, the performance of the linear methods could represent a lower bound for evaluating the performance of the nonlinear methods.

The linear regression is a method modelling the relationship between a scalar dependent variable and one or more explanatory variables to fit the parameter of the model, which is called linear regression coefficient. The case of one explanatory variable is called simple regression while more than one explanatory variable is called MLR method. Several procedures have been developed for parameter estimation and inference in linear regression. These methods differ in computational simplicity of algorithms, presence of a closed-form solution, robustness with respect to heavy-tailed distributions, and theoretical assumptions needed to validate desirable statistical properties such as consistency and asymptotic efficiency. The two most common estimation methods for linear regression are the least-square estimation method (Lai et al., 1978) and maximum-likelihood estimation method (Lange et al., 1989). In each method many related technologies were developed.

Some more common technologies related to the least-square estimation method are: (i) *Ordinary Least Squares* (OLS) is the simplest and thus most common estimator. It is conceptually simple and computationally straightforward. The OLS method minimizes the sum of squared residuals, and leads to a closed-form expression for the estimated value of the unknown regression coefficient  $\beta$  (Lai et al., 1978); (ii) *Generalized least squares* (GLS) is an extension of the OLS technology, that allows efficient estimation of  $\beta$  when either heteroscedasticity, or correlations, or both are present among the error terms of the model, as long as the form of heteroscedasticity and correlation is known independently of the data; (iii) *Percentage least squares* focuses on reducing percentage errors, which is useful in the field of forecasting or time series analysis. It is also useful in situations where the dependent variable has a wide range without constant variance, as here the larger residuals at the upper end of the range would dominate if OLS were used. When the percentage or relative error is normally distributed, least squares percentage regression provides maximum likelihood estimates. Percentage regression is linked to a multiplicative error model, whereas OLS is linked to models containing an additive error term (Tofallis, 2009). (iv) *Total least squares* (TLS) is an approach to least squares estimation of the linear regression model that treats the covariates and response variable in a more geometrically symmetric manner than OLS. It is one approach to handling the "errors in variables" problem, and is sometimes used when the covariates are assumed to be error-free (Nievergelt, 1994).

Some of more common technologies related to the maximum-likelihood estimation method are: (i) *Maximum likelihood estimation* can be performed when the distribution of the error terms is known to belong to a certain parametric family  $f_{\theta}$  of probability distributions (Lange et al., 1989). When  $f_{\theta}$  is a normal distribution with mean zero and variance  $\theta$ , the resulting estimate is identical to the OLS estimate. GLS estimates are maximum likelihood estimates when the error  $\epsilon$  follows a multivariate normal distribution with a known covariance matrix; (ii) *Ridge regression* (Draper et al., 1979; Swindel, 1981; Hoerl et al., 1985) and other forms of penalized estimation such as Lasso regression (Tibshirani, 1996), deliberately introduce bias into the estimation of  $\beta$  in order to reduce the variability of the estimate. The resulting estimators generally have lower mean squared error than the OLS estimates,

particularly when multicollinearity is present. They are generally used when the goal is to predict the value of the response variable  $\mathbf{y}$  for values of the predictors  $\mathbf{X}$  that have not yet been observed. These methods are not as commonly used when the goal is inference, since it is difficult to account for the bias; (iii) *Least absolute deviation* (LAD) regression is a robust estimation technique in that it is less sensitive to the presence of outliers than OLS (but is less efficient than OLS when no outliers are present). It is equivalent to maximum likelihood estimation under a Laplace distribution model for error  $\epsilon$  (Narula et al., 1982); (iv) *Adaptive estimation*. If we assume that error terms are independent from the regressors  $\epsilon_i \perp \mathbf{x}_i$ , the optimal estimator is the 2-step *Maximum likelihood estimation*, where the first step is used to non-parametrically estimate the distribution of the error term (Stone, 1975).

Some other methods: (i) *Bayesian linear regression* applies the framework of Bayesian statistics to linear regression. In particular, the regression coefficients  $\beta$  are assumed to be random variables with a specified prior distribution. The Bayesian estimation process produces not a single point estimate for the "best" values of the regression coefficients but an entire posterior distribution, completely describing the uncertainty surrounding the quantity. This can be used to estimate the "best" coefficients using the mean (for example linear MMSE estimation), mode, median and any other function of the posterior distribution (Kay, 1993); (ii) *Principal component regression (PCR)* is used when the number of predictor variables is large, or when strong correlations exist among the predictor variables (Kendall, 1957; Jeffers, 1967). This two-stage procedure first reduces the predictor variables using principal component analysis then uses the reduced variables in an OLS regression fit. While it often works well in practice, there is no general theoretical reason that the most informative linear function of the predictor variables should lie among the components with the big eigenvalues. The components with small eigenvalues can be as important as those with large variance (Jolliffe, 1982).

### 3.1. 2. Statistical mapping methods

The statistical methods for mapping the acoustic speech to the visual features have been used widely in the audio-visual speech recognition, human computer interface and acoustic-to-articulatory inversion etc. The problem of mapping from the audio

space to the visual space can be solved at several different levels according to the speech analysis being used (Huang et al., 1998).

At the first level, frame level, many methods can be used to map one frame of the audio to one frame of visual parameters, which use a large set of audiovisual parameters to train the mapping. The mapping is normally based on the methods such as Vector Quantization (VQ), the Neural Network, and the GMM which is used also in our work.

- (i) VQ: VQ is a classical quantization technique from signal processing which allows the modelling of probability density functions by the distribution of prototype vectors. It works by clustering a large set of points (vectors) into groups closest to them. Each group is represented by its centroid point, as in *k*-means algorithm. Each vector is called code vector or a codeword, and the set of all codewords is called a codebook. VQ is originally used for data compression and now it is also widely used for mapping the speech features to the visual features. Morishima et al. (1991) converted voice signals from a speaker to mouth and jaw motions by VQ to realize an intelligent human-machine interface. The codebook design method is done in two steps. First, the voice training sequences are clustered based on Cepstrum distance. Second, the image centroids are selected to minimize the total Euclidian distance between the centroid and all the image training vectors which belong to the same voice clustering cell. The selection of the image codeword (vector) is on a frame-by-frame basis. The VQ method is simple but it generates undesirable quantization errors. Huang et al. (1998) used VQ to class the 15 dimension audio-visual data into 20 classes and used the centre vector and convariance matrices of each cluster as the initial mean value and covariance of GMM in the audio-to-visual mapping. In the acoustic-to-articulatory inversion, the VQ is widely used: Schroeter et al. (1994) used articulatory codebook approach to build lookup tables consisting of pairs of acoustic and articulatory parameters from parallel recorded articulatory-acoustic data. Hogden et al. (1996) used Electromagnetic Articulography data recorded by one Swedish male

subject to built a codebook of quantized articulatory-acoustic parameter pairs. In their study, the acoustic vectors were categorized into a lookup table with 256 codes by finding the shortest Euclidean distance between the acoustic vectors and each of a small set of numbered reference vectors. A VQ codebook was used to map from acoustic segments to VQ codes, and a lookup table was then used to map from the VQ code to an estimated articulatory configuration. Hogden et al. (1996) reported Root-Mean-Squared errors (RMSE) around 2 mm for coils on the tongue. Being a discrete method, the disadvantage of VQ approach is apparently that it does not give the same level of approximation to the target distribution without significantly increasing the size of the lookup table in comparison with methods employing continuous variables. However, with the low storage and computation costs the VQ method, such as  $k$ -means, is suitable for initialization of some other more sophisticated statistical modelling.

- (ii) **Neural Network:** Neural networks can also be used to convert acoustic parameters to visual parameters. In the training phase, input patterns and output patterns are presented to the network, and an algorithm called back propagation can be used to train the network weights. The design choice lies in selecting a suitable topology for the network. The number of hidden layers and the number of nodes per layer may be experimentally determined. Furthermore, a single network can be trained to reproduce all the visual parameters, or many networks can be trained with each network estimating a single visual parameter (Chen et al., 1997). Morishima et al. (1991) used a neural network based on a three-layer neural network for voice to image conversion. The input layer has 16 units, corresponding to the dimensions for LPC Cepstrum parameters. The hidden layer has 16 units empirically. The mapping is learned by back propagation, with both voices and image parameters which are synchronized to each other. The image sequences are synthesized from the image parameters, which correspond to the output values of the neural network. In case of a neural network, the mouth shape motion can be synthesized very delicately because the output values for network units take continuous values, so



image motion is more natural and closer to the original motion than the synthesized motion obtained by the VQ method. In terms of the acoustic-to-articulatory mapping problem, the neural network has been also extensively studied. Richmond et al. (2003) used the mixture density network (MDN) which has been reported that the multiple representation of articulatory probability density is effective for the inversion mapping. In the most general sense, the MDN can be considered as combining a trainable regression function (typically a non-linear regressor such as an artificial neural network) with a probability density function. A multilayer perceptron (MLP) was used as a trainable non-linear regressor and a GMM. The role of the MLP is to take an input vector in acoustic domain and map to the articulatory domain. Training consists of updating the MLP weights to optimize an error function, defined as the negative log likelihood of the target data. Thus, standard nonlinear optimization algorithms may be used to train the MDN. Since, the MDN gives a model of conditional probability density, it is trivial to augment the target features with derived delta and delta-delta features *i.e.* dynamic features. Once trained, the input sequence of acoustic feature vectors gets an output of a sequence of probability density functions (pdfs) over the static and dynamic articulatory features. More recently, the trajectory mixture density network (TMDN) approach with many more free parameters resulted in a decreased RMSE to 1.40 mm on the same training, validation and testing datasets (Richmond, 2009). Kjellström et al. (2009) implemented an audiovisual-to-articulatory inversion using simple MLR or ANNs. Depending on the type of fusion (early or late) between the audio signal and the video signal (based on independent component images of the mouth region), they obtained RMSE for the tongue shape ranging from 2.5 to 3 mm.

(iii)GMM: The GMM as a stochastic model being a mixture combination of the individual Gaussian could appropriately describe the unknown data distribution. One of the powerful attributes of GMM is its ability to form smooth approximations to arbitrary-shape densities. The GMM not only provides a smooth overall distribution fit, its components also clearly

detail the multi-modal nature of the density (Reynolds et al., 1995). The effectiveness of GMM has been illustrated in many speech research areas, such as in the speech recognition, speaker identification (Reynolds et al., 1995), speech conversion (Kain et al., 1998; Stylianou et al., 1998; Chen et al., 2003; Toda et al., 2007), audiovisual speech (Huang et al., 1998) and acoustic-to-articulator inversion (Toda et al., 2008; Zen et al., 2011). There are several methods for estimating the parameters of GMM (McLachlan et al., 1988). The most used method by far is the EM algorithm. In the application of the speech conversion, Stylianou et al. (1998) used the source data only to estimate the parameters of GMM. But Kain et al. (1998) later used the joint vectors of the source and target data as a training set, which is proved to be more robust especially for a small amount of training data. In this work we used the joint vector as the training set to estimate the parameters of the GMM. Initializing the parameters of model is another critical factor prior to the EM algorithm. Reynolds et al. (1995) indicated that the elaborate initialization schemes are not necessary for training GMM. Usually the  $k$ -means method is used to initialize the GMM. The mapping criteria MMSE or MAP can be used in the GMM-based mapping problems (Huang et al., 1998; Toda et al., 2008; Zen et al., 2011). The MAP mapping criterion needs to be initialized by some other methods, such as MMSE. Toda et al. (2008) proposed dynamic features to the maximum likelihood estimation (MLE)-based mapping method in the framework of GMM, which could alleviate the discontinuities obtained by the MMSE-based mapping method. Zen et al. (2011) proposed a method based on the trajectory GMM for continuous stochastic feature mapping, in which the joint pdf is modeled by a trajectory GMM. The proposed method can resolve the inconsistencies introduced by using dynamic-feature constraints while retaining their benefits. However, it offers entire sequence-level transformation rather than frame-by-frame mapping.

At the second level, phoneme level, the mapping could be found for each phoneme in the speech signal (Huang et al., 1998). The first step of mapping from audio to visual parameters is to segment the speech sequence phonetically. Then we use a lookup-

table to find out the sequence of visual features. The look-up table is predefined for the whole set of phonemes. In this table, each phoneme is associated with one visual feature set.

At the third level, word level, we can explore the context cues in the speech signals. First we use a speech recognizer to segment the speech into words. For each word, we can create a HMM to represent the acoustic state transition in the word. For each state in the HMM model, we can use the methods as in the first level to model the mapping from acoustic to visual feature frame by frame. The vast majority of audio-to-visual mapping systems employ HMM with a continuous observation probability density, modeled as a mixture of Gaussian densities (Rao et al., 1997). Yamamoto et al. (1998) used HMM to map audio feature to lip movements. The HMM is learned from audio training data, and each video training sequence is aligned with the audio sequence using Viterbi optimization. During synthesis, an HMM state sequence is selected for a given novel audio input using the Viterbi algorithm, and the visual output associated with each state in the audio sequence is retrieved. But the performance was limited by the sensitivity to the noise of Viterbi algorithm. Choi et al. (2001) proposed a hidden Markov model inversion (HMMI) method. In HMMI, the visual output is generated directly from the given audio input and the trained HMM by means of an EM iteration, thus avoiding the use of the Viterbi sequence. A different mechanism has been proposed by Brand (1999), in which a minimum-entropy training method is used to learn a concise HMM. As a result, the Viterbi sequence captures a larger proportion of the total probability mass, thus reducing the detrimental effects of noise in the audio input. In the application of acoustic-to-articulatory inversion: Hiroya et al. (2004) developed a method that determines the articulatory movements from speech acoustics using an HMM-based speech production model. After proper labeling of the recorded corpus, each allophone in training corpus is modeled by a context-dependent HMM, and the proper inversion is performed by a state-dependent linear regression between the observed acoustic and the corresponding articulatory parameters. The articulatory parameters of the statistical model are then determined for a given speech spectrum by MAP. In order to assess the importance of phonetics, they tested their method under two experimental conditions, namely with and without phonemic information. In the former, the phone HMMs were assigned according to the correct

phoneme sequence for each test utterance. In the latter, the optimal state sequence was determined among all possible state sequences of the phone HMMs and silence model. Zhang (2009) indicate that the jointly trained acoustic-articulatory models are more accurate (having a lower RMSE) than the separately trained ones, and that Trajectory-HMM training results in greater accuracy compared with conventional Baum-Welch parameter updating. Trajectory-HMM training using the Root Mean Square criteria proves to be better than using the standard Maximum Likelihood criteria. Katsamanis et al. (2009) approximated the audiovisual-to-articulatory mapping by an adaptive piecewise linear model. Model switching was governed by a Markovian discrete process which captures articulatory dynamic information. Each constituent linear mapping is effectively estimated via canonical correlation analysis. For facial analysis, active appearance models demonstrated fully automatic face tracking and visual feature extraction capabilities.

Some other methods for mapping audio features to the visual features are based on machine learning theory such as kernel methods and support vector regression (SVR) methods. Kernel methods solve the problem by mapping the data into a high dimensional intermediate space in which a variety of methods can be used to find relations in the data. Kernel functions enable the kernel methods to operate in the intermediate space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the intermediate space. This operation is often computationally cheaper than the explicit computation of the coordinates (Shawe-Taylor et al., 2004). The best known application of kernel method is the SVR which uses the Support Vector (SV) machines for function estimation/regression (Smola et al., 2004). SVR seeks to find an optimal estimate (in terms of structural risk minimization) for the estimation function. Toutios et al. (2005) employ the SVR on speech inversion and used the radial basis function (RBF) kernel with user-defined parameter as the kernel function. The results are comparable to the ones obtained by Neural Networks to the same task. However, SVR training is a relatively slow process. Training time increases with the cube of the amount of training data.

## 3.2. Multi-linear mapping method based on PCA Regression (PCR)

### 3.2.1. Multi-linear regression method

As mentioned in section 3.1.1, in linear regression, data are modeled using linear predictor functions, and unknown model parameters are estimated from the explanatory variables and dependent value. Normally, linear regression is used to quantify the strength of the relationship between the explanatory variables  $X$  and the dependent values  $\mathbf{y}$ . In our work, linear regression is used to fit a predictive model by an observed data  $X$  and dependent  $\mathbf{y}$  values. After developing such a model, we can use an additional value (test value) of  $X$  to map to the dependent value of  $\mathbf{y}$ , that means the fitted model can be used to make a prediction of the value of  $\mathbf{y}$ .

Given a data set  $\{y_i, x_{i1}, \dots, x_{in}\}$ , where  $i = 1, \dots, m$ , in our work,  $i$  is a variable meaning the  $i$ th frame of the acoustic sequence. A MLR model assumes that the relationship between the dependent variable  $y_i$  and the  $n$ -dimension regressor  $\mathbf{x}_i^T = [x_{i1}, \dots, x_{in}]$  is linear. This relationship is modeled through a disturbance error variable  $\varepsilon_i$  which is an unobserved random variable that adds noise to the linear relationship between the dependent variable and regressor. Thus the model takes the form as,

$$y_i = \beta_1 x_{i1} + \dots + \beta_n x_{in} + \varepsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = (1, 2, \dots, m) \quad (3.1)$$

where  $y_i$  is the dependent variable or the target data to be predicted.  $\mathbf{x}_i^T = [x_{i1}, \dots, x_{in}]$  is the  $n$ -dimension regressor used as predictor variable for forecasting the dependent variable  $y_i$ .  $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_n]^T$  is the  $n$ -dimension regression coefficient which is obtained by regressor  $\mathbf{x}_i^T$  and the dependent  $y_i$  to fit the MLR model.  $\varepsilon_i$  is the error which captures all other factors that influence the dependent variable  $y_i$  other than the regressor  $\mathbf{x}_i^T$ . Finally,  $^T$  denotes the transpose, so that  $\mathbf{x}_i^T \boldsymbol{\beta}$  is the inner product between vectors  $\mathbf{x}_i$  and  $\boldsymbol{\beta}$ .

Further, if we defined,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_m^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}_{m \times n}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{bmatrix} \quad (3.2)$$

Equation (3.1) can be written in the vector form as,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (3.3)$$

In the equation (3.3), the estimation of regression coefficient  $\boldsymbol{\beta}$  is the key problem of the linear regression model. As mentioned in section 3.1.1, owing to its simplicity, and computationally straightforward, the OLS is most common used. Thus we applied OLS to estimate the unknown parameter of the model, i.e. the regression coefficient  $\boldsymbol{\beta}$ , with the assumption that errors are zero-mean, normally distributed and uncorrelated with each other. By minimizing the sum of squared residuals, the OLS leads to a closed-form expression for the estimated value of the regression coefficient  $\boldsymbol{\beta}$  (Kay, 1993).

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (3.4)$$

Note that equations (3.1), (3.3) and (3.4) are in ‘centered’ form, with the assumption that all variable measured about their means, namely mean centering. When we want to reconstruct the measured  $\mathbf{y}$  with the  $\mathbf{X}$  and the estimated  $\hat{\boldsymbol{\beta}}$  we need to add the mean of  $\mathbf{y}$  to the result of equation of (3.1) or (3.3). More generally, the equation (3.5) shows the OLS applying on the data before centered. In the sense of OLS, the residual  $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$  of the estimator is orthogonal to the regressor  $\mathbf{X}$ . This means that the residual is the shortest of all possible vectors  $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ , that is, the variance of the residuals is the minimum. The residuals and predictors are uncorrelated.

$$\mathbf{y} - \bar{\mathbf{y}} = (\mathbf{X} - \bar{\mathbf{X}})\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon} \quad (3.5)$$

### 3.2.2. Multi-linear PCA Regression model

As introduced in the section 3.1.1, while the OLS is the most common and simple regression method, however, when the regressors are close to being correlated, then the variances of some of the estimated regression coefficients can become very large, leading to unstable and potentially misleading estimates of the regression equation

(Jolliffe, 1986). To overcome this problem, instead of regressing dependent variable on the regressor directly, the principal components obtained by the principal component analysis of the regressors are used to estimate regression coefficients. The procedure is called principal component regression (PCR) in statistics. In addition, when the number of the explanatory variables is large, in order to reduce the regressors, only a subset of the principal components is used in the PCR, making a kind of regularized estimation. Often the principal components with the highest variance are selected. However, the low-variance principal components may also be important, in some cases even more important (Jolliffe, 1982). Furthermore, calculation of the least squares estimates via PCR may be numerically more stable than direct calculation (Flury, 1988).

Normally, PCR method can be divided into three steps: (1) The first step is to run a principal components analysis on the table of the explanatory variables.(2) The second step is to run an OSL on the selected components: the factors that are most correlated with the dependent variable will be selected. (3) Finally the parameters of the model are computed for the selected explanatory variables (Jeffers, 1967).

### **3.2.2.1. Principal Components Analysis (PCA)**

Principal component analysis (PCA) is probably the most popular multivariate statistical method and it is used by almost all scientific disciplines, which was invented by Pearson (1901). Its goal is to extract the important information from the data table, to represent it as a set of new orthogonal variables called principal components, and to display the pattern of similarity of the observations and of the variables as points in maps (Abdi et al., 2010). The number of principal components is less than or equivalent to the number of original variables, and the first principal component has the largest possible variance, meanwhile each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

PCA is the simplest of the true eigenvector-based multivariate analyses, which can be done by eigenvalues decomposition of a data covariance matrix or singular value decomposition (SVD) of a data matrix, usually after mean centering the data matrix

for each attribute (Abdi et al., 2010). Often, its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data. If a multivariate dataset is visualized as a set of coordinates in a high-dimensional data space (1 axis per variable), PCA can supply the user with a lower-dimensional picture. This is realized by using only the first few principal components so that the dimensionality of the transformed data is reduced.

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on (Jolliffe, 1986).

Define a data matrix  $\mathbf{X}$ , with mean centering (the sample mean of the distribution has been subtracted from the data set), where each of the  $m$  rows represents an observation of data, and each of the  $n$  columns corresponds a dimension of the data space. The SVD of  $\mathbf{X}$  is:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{A}^T \quad (3.6)$$

where the  $m \times m$  matrix  $\mathbf{U}$  is the matrix of eigenvectors of the covariance matrix  $\mathbf{X}\mathbf{X}^T$  in which each variable is scaled to have its sample variance equivalent to one, the matrix  $\mathbf{\Sigma}$  is an  $m \times n$  rectangular diagonal matrix with nonnegative real numbers on the diagonal, and the  $n \times n$  matrix  $\mathbf{A}$  is the matrix of eigenvectors of  $\mathbf{X}^T\mathbf{X}$ . The PCA transformation that preserves all dimensions is then given by:

$$\begin{aligned} \mathbf{F} &= \mathbf{X}\mathbf{A} \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{A}^T\mathbf{A} \\ &= \mathbf{U}\mathbf{\Sigma} \end{aligned} \quad (3.7)$$

where  $\mathbf{U}$  is not uniquely defined in the usual case when  $n < m - 1$ , but  $\mathbf{F}$  will usually still be uniquely defined. Since  $\mathbf{A}$  is an orthogonal matrix defined by the SVD, each row of  $\mathbf{F}$  is a rotation of the corresponding row of  $\mathbf{X}$ . Therefore, the columns of  $\mathbf{A}$  are a new set of basis vectors for representing of rows of  $\mathbf{X}$ . And the column vectors of  $\mathbf{A}$ , which are the eigenvectors of the covariance matrix  $\mathbf{X}^T\mathbf{X}$ , are the principal components of  $\mathbf{X}$ . Each column of  $\mathbf{F}$  is made up of the "score" with respect to the



corresponding principal component, the first column of  $\mathbf{F}$  is made up of the "score" of the case with respect to the "most principal" component, the next column has the scores with respect to the "second principal" component, and so on.

If we want a reduced-dimensionality representation, we can project  $\mathbf{X}$  down into the reduced space defined by only the first  $p$  singular vectors, that is the first  $p$  columns of  $\mathbf{A}$  denoted by  $\mathbf{A}_p$ :

$$\begin{aligned}
 \mathbf{F}_p &= \mathbf{X}\mathbf{A}_p \\
 &= \mathbf{X}\mathbf{A}\mathbf{I}_{n \times p} \\
 &= \mathbf{U}\mathbf{\Sigma}\mathbf{A}^T\mathbf{A}\mathbf{I}_{n \times p} \\
 &= \mathbf{U}\mathbf{\Sigma}_p
 \end{aligned} \tag{3.8}$$

where  $\mathbf{I}_{n \times p}$  is the  $n \times p$  rectangular identity matrix.

Given a set of points in Euclidean space, the first principal component corresponds to a line that passes through the multidimensional mean and minimizes the sum of squares of the distances of the points from the line. The second principal component corresponds to the same concept after all correlation with the first principal component has been subtracted from the points (Jolliffe, 1986). The singular values in  $\mathbf{\Sigma}$  are the square roots of the eigenvalues of the matrix  $\mathbf{X}^T\mathbf{X}$ . Each eigenvalue is proportional to the portion of the "variance" which is the sum of the squared distances of the points from their multidimensional mean. The sum of all the eigenvalues is equivalent to the variance. PCA essentially rotates the set of points around their mean in order to align with the principal components. This moves as much of the variance as possible using an orthogonal transformation into the first few dimensions. The values in the remaining dimensions, therefore, tend to be small and may be dropped with minimal loss of information. PCA is often used in this manner for dimensionality reduction. PCA has the distinction of being the optimal orthogonal transformation for keeping the subspace that has largest "variance".

### 3.2.2.2. Modelling multiple PCR

The idea of using principal components rather than the original predictor variables is not new (Kendall, 1957), and it has a number of advantages. Given the  $\mathbf{A}$  in equation

(3. 7) is orthogonal and its column vectors are the eigenvectors of  $\mathbf{X}^T\mathbf{X}$ , thus  $\mathbf{A}\mathbf{A}^T = \mathbf{I}$  and  $\mathbf{X}\boldsymbol{\beta}$  can be rewritten as  $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\mathbf{A}\mathbf{A}^T\boldsymbol{\beta}$ . The equation (3. 3) can be substituted in these results and equation (3. 7):

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \mathbf{X}\mathbf{A}\mathbf{A}^T\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \mathbf{F}\mathbf{A}^T\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \mathbf{F}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\end{aligned}\tag{3. 9}$$

We can see in equation (3. 9), the predictor variable (regressor)  $\mathbf{X}$  is replaced by the scores of the principal components  $\mathbf{F}$  in the linear regression. Thus the multiple PCR can be defined as equation (3. 9) and the PCR coefficient is:

$$\boldsymbol{\gamma} = \mathbf{A}^T\boldsymbol{\beta}\tag{3. 10}$$

which could be estimated by:

$$\hat{\boldsymbol{\gamma}} = \mathbf{A}^T\hat{\boldsymbol{\beta}}\tag{3. 11}$$

or

$$\hat{\boldsymbol{\gamma}} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{y}\tag{3. 12}$$

Note that equations (3. 9) are in ‘centered’ form, with the assumption that all variable measured about their means. The  $\mathbf{F}$  is centered by the centered  $\mathbf{X}$  in advance for PCA. When we want to reconstruct the measured  $\mathbf{y}$  with the  $\mathbf{F}$  and estimated  $\hat{\boldsymbol{\gamma}}$ , we need to add the mean of  $\mathbf{y}$  to the result of equation of (3. 9).

When using principal components instead of the original explanatory variables in linear regression, the contribution of each principal component to the equation can be more easily interpreted than the contributions of the original variables. Because of uncorrelatedness, the contribution of a principal component is unaffected by other principal components which are also included in the regression. Whereas, for the original variables both contributions and regression coefficients can be changed dramatically when another variable is added to, or deleted from, the equation. It is more important that when multicollinearities are present by deleting a subset of the

principal components, especially those with small variances, much more stable estimates of the regression coefficient can be obtained.

### 3.2.2.3. Regressors selection in PCR

As in equation (3. 8), the first few principal components corresponding to the first highest variances are often selected to make up a subspace to realize the dimensionality reduction. While in the PCR, the principal components selected in this way are not truly the most appropriate regressors for regression because PCA constructs components to explain variation in the explanatory variable  $\mathbf{X}$  but not in the dependent variable  $\mathbf{y}$ . Factors that are most correlated with the dependent variable will be selected, but not necessarily the ones with the highest variance (Jolliffe, 1982). Relations between the dependent variable and all of the components should be examined since it is always possible that one of the components with small variance maybe related to the dependent variable. Thus the first few principal components which have the largest correlation coefficients with the dependent variable are selected as the regressors but not the ones with largest variance. The PCR on the reduction subspace of selected  $p$  principal components is defined as:

$$\mathbf{y} = \mathbf{F}_p \boldsymbol{\gamma}_p + \boldsymbol{\varepsilon}_p \quad (3. 13)$$

where  $\boldsymbol{\gamma}_p$  being a subset of  $\boldsymbol{\gamma}$  has  $p$  elements,  $\mathbf{F}_p$  is an  $(m \times p)$  matrix whose columns are the scores of the first selected  $p$  principal components, and  $\boldsymbol{\varepsilon}_p$  is the appropriate error term. As same as equation (3. 9), this equation is in the centred form so we need to add the mean value of  $\mathbf{y}$  to the result of (3. 13) when we reconstruct the measured variable  $\mathbf{y}$ .

## 3.3. Gaussian Mixture Model (GMM) mapping method

### 3.3.1. Gaussian Mixture Model (GMM)

The GMM-based mapping method whose efficiency has been illustrated by the plenty of applications in the domains of speech recognition, speech synthesis and acoustic-to-articulatory inversion has both the classification and regression properties which inspire us to employ it in our work. Meanwhile GMM is the most studied case of the finite mixture model. In statistics, a mixture model is a probabilistic model for

representing the presence of subpopulations within an overall population, without requiring that an observed data-set should identify the sub-population to which an individual observation belongs. Finite mixture models are particularly suitable for modelling distributions where the measurements arise from separate groups, but individual membership is unknown. A finite mixture model is a distribution with probability density function of the form (Jain et al., 2000; Webb, 2011):

$$p(x) = \sum_{m=1}^M \alpha_m p(x; \theta_m) \quad (3.14)$$

where

- $M$  is the number of the components;
- $\alpha_m \geq 0$  are the mixture component probability (also referred to as mixing proportion), which satisfy  $\sum_{m=1}^M \alpha_m = 1$ ;
- $p(x; \theta_m)$ ,  $m = 1, 2, \dots, M$  are the component density functions (each of which depends on a parameter vector  $\theta_m$ ).

Although the component densities can be built from many different types of components, the majority of the literature focuses on Gaussian mixtures. In the GMM, each component is a multivariate normal distribution, i.e.  $p(x; \theta_m)$  is the probability density for the multivariate normal distribution with mean vector  $\mu_m$  and covariance matrix  $\Sigma_m$ , so that  $\theta_m = \{\mu_m, \Sigma_m\}$ . The GMM therefore has probability density function:

$$p(\mathbf{x}) = \sum_{m=1}^M \alpha_m N(\mathbf{x}; \mu_m, \Sigma_m) \quad (3.15)$$

The principal motivation for using the Gaussian mixture densities as a representation of the unknown data distribution is that a GMM with a sufficient number of components may be a suitable approximation to the data distribution, even if the data are not known to be built up from a series of normal components. The individual component densities of the GMM may model some underlying set of classes and the

linear combination of Gaussian basis function is capable of representing a large class of sample distribution. The GMM, inheriting the attributes of unimodal Gaussian model and VQ model not only providing a smooth overall distribution fit but also clearly detail the multi-model nature of the density by the discrete set of positions (representing by the mean vector) and the elliptic shape (representing by the covariance matrix), acts as a hybrid between these two models by using a discrete set of Gaussian functions (Reynolds et al., 1995).

### 3.3.2. GMM parameters estimation

If the number of the components  $M$  is pre-specified, the parameters of GMM to be estimated are the mixture component probabilities  $[\alpha_1, \alpha_2, \dots, \alpha_M]$  and the parameters of basis Gaussian function,  $\theta_m = [\mu_m, \Sigma_m], m = 1, 2, \dots, M$ . The most prevalent method for optimizing mixture-density parameter given a set of  $n$  independent observations  $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$  (the training data) is an iterative algorithm known as the EM algorithm which was proposed initially by Dempster et al. (1977). The common application of EM in the pattern recognition community is simplifying the likelihood function by assuming the existence of the latent variables for additional but missing (or hidden) parameters, when optimizing directly the maximum-likelihood function of the original (incomplete) data is intractable.

#### (1) Initialization

Since the initial values such as the mean vectors and covariance matrices are needed for estimating the parameters of GMM with EM algorithm, a clustering algorithm such as  $k$ -means clustering is used to provide reasonable initial values. The  $k$ -means algorithm is one of the unsupervised classification algorithms using the method of iterative squared-error component clustering. The  $k$ -means algorithm attempts to obtain the partition which minimizes the within-cluster scatter based on the squared-error criterion. To guarantee that an optimum solution has been obtained, one has to examine all possible partitions of the  $n$   $d$ -dimensional elements into  $k$  clusters (for a given  $k$ ). The  $k$ -means algorithm can terminate naturally, when no further changes are made to the cluster assignments, or can be set to stop when a maximum number of iterations have taken place (reducing computational expense). However the  $k$ -means

algorithm often leads to a local optimum clustering solution, the algorithm is often used as a preprocessing of a subsequent processing but not recommended as an independent clustering algorithm. The procedure of the  $k$ -means algorithm is as follow (Webb, 2011):

1. Initialization: Pick  $M$  measurements from  $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$  randomly without replacement, and use these as initial values for the mean vectors  $[\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_M]$  of components.
2. Iterative step:
  - Find the closest centre to each training data vector. For  $i = 1, 2, \dots, n$ , determine  $\psi_i \in [1, 2, \dots, M]$ ,  $\boldsymbol{\mu}_{\psi_i}$  is the closest centre to the training data  $\mathbf{x}_i$ :

$$|\mathbf{x}_i - \boldsymbol{\mu}_{\psi_i}|^2 = \min |\mathbf{x}_i - \boldsymbol{\mu}_j|^2, \quad j \in [1, 2, \dots, M] \quad (3.16)$$

- Compute new cluster centers as the centroids of the clusters. If any value  $\psi_i, i = 1, 2, \dots, n$ , has changed from the previous iteration, recalculate the component means as follows:

$$\boldsymbol{\mu}_j = \frac{1}{\sum_{i=1}^n I(\psi_i = j)} \sum_{i=1}^n I(\psi_i = j) \mathbf{x}_i, \quad j \in [1, 2, \dots, M] \quad (3.17)$$

where  $I(\psi_i = j)$  is the indicator function, taking value 1 if  $\psi_i = j$ , and 0 otherwise,  $I(\psi_i = j) \mathbf{x}_i$  indicates the ones belonging to the component  $j$  obtained from the previous iteration. If none of the  $\psi_i$  have changed from the previous iteration then move forward to step 4.

3. If the maximum number of iterations has been exceeded then move to step 4, otherwise repeat step 2.
4. Initialize the mixture component covariance matrices as follows:

$$\Sigma_j = \frac{1}{\sum_{i=1}^n I(\psi_i = j)} \sum_{i=1}^n I(\psi_i = j) (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T, \quad j \in [1, 2, \dots, M] \quad (3.18)$$

The calculated mean vectors  $\boldsymbol{\mu}_j, j \in [1, \dots, M]$ , covariance matrices  $\Sigma_j, j \in [1, \dots, M]$  by  $k$ -means and the given number of components are used as the initialization values in the EM algorithm for estimation the parameters of the GMM.

## (2) EM algorithm

The mixture-density parameter estimation problem is probably one of the most widely used applications of the EM algorithm in the computational pattern recognition community. The probabilistic model of the mixture-density is showed in the equation (3.14), the maximum-likelihood estimate function for the parameters of the mixture-density distribution is:

$$L(\boldsymbol{\Theta}|\mathbf{X}) = P(\mathbf{X}|\boldsymbol{\Theta}) = \prod_{i=1}^N p(x_i) = \prod_{i=1}^N \left\{ \sum_{j=1}^M \alpha_j p_j(x_i; \theta_j) \right\} \quad (3.19)$$

where the parameters are  $\boldsymbol{\Theta} = (\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M)$  such that  $\sum_{j=1}^M \alpha_j = 1$  and each  $p_j$  is the density function parameterized by  $\theta_j$ . The log-likelihood expression of the equation (3.19) is given by:

$$\log(L(\boldsymbol{\Theta}|\mathbf{X})) = \sum_{i=1}^N \log \left\{ \sum_{j=1}^M \alpha_j p_j(x_i; \theta_j) \right\} \quad (3.20)$$

In the equation (3.20) we can see that the log expression of the maximum-likelihood estimation function of mixture-density distribution is difficult to optimize because it contains the log of the sum. The EM algorithm is introduced to simplify the likelihood function by considering  $\mathbf{X}$  as incomplete data. The ‘missing’ or latent data items  $\mathbf{Y} = [y_1, y_2, \dots, y_N]$  whose values inform us which component density “generated” each data item is also introduced to simplify the log sum function.  $y_i = k$  ( $y_i \in [1, 2, \dots, M]$ ) if the  $i^{th}$  sample was generated by the  $k^{th}$  mixture component (Bilmes, 1998). If the values of  $\mathbf{Y}$  are known, the log likelihood function of the complete data  $[\mathbf{X}, \mathbf{Y}]$  is:

$$\log(L(\Theta|\mathbf{X}, \mathbf{Y})) = \sum_{i=1}^N \log(p(x_i|y_i)p(y_i)) = \sum_{i=1}^N \log(p_{y_i}(x_i|\theta_{y_i})\alpha_{y_i}) \quad (3.21)$$

In the above equation, we can see that the log sum function in the equation (3.20) was simplified by the conditional probability given the latent variable  $y_i$  which indicates the  $i^{th}$  component generated the sample  $x_i$ . But in fact we do not know the value of  $y_i$ . However, if we assume variable  $\mathbf{Y}$  is a random vector governed by an underlying distribution, we can continue to estimate the parameters. Thereby we can think the equation (3.21), namely the complete-likelihood function, is the function of the random variable  $\mathbf{Y}$  where  $\mathbf{X}$  and  $\Theta$  are constant. The first step of EM algorithm is finding the expected value of the complete-data log-likelihood with respect to random variable  $\mathbf{Y}$  given the observed data  $\mathbf{X}$  and the current parameter estimates  $\Theta^{old}$ . We define (Bilmes, 1998; Bishop, 2006):

$$Q(\Theta, \Theta^{old}) = E[\log(L(\Theta|\mathbf{X}, \mathbf{Y}))|\mathbf{X}, \Theta^{old}] \quad (3.22)$$

In the equation (3.22),  $\Theta$  are the new parameters we used to increase the  $Q$  with the current parameter estimates  $\Theta^{old}$ . Given the expected value of the random variable  $\mathbf{Y}$  and the definition of the expectation of a discrete random variable, the equation (3.22) can be written as:

$$Q(\Theta, \Theta^{old}) = \sum_{\mathbf{Y} \in \Upsilon} [\log(P(\mathbf{X}, \mathbf{Y}|\Theta)P(\mathbf{Y}|\mathbf{X}, \Theta^{old}))] \quad (3.23)$$

where  $P(\mathbf{Y}|\mathbf{X}, \Theta^{old})$  is the marginal distribution of the latent variable  $\mathbf{Y}$  dependent of the observed data  $\mathbf{X}$  and the current parameters  $\Theta^{old}$ ; the  $\Upsilon$  is the range of  $\mathbf{Y}$  values. Introducing the equation (3.21) to the equation (3.23), we can write the equation (3.23) as:

$$Q(\Theta, \Theta^{old}) = \sum_{\mathbf{Y} \in \Upsilon} \left[ \sum_{i=1}^N \log(p_{y_i}(x_i|\theta_{y_i})\alpha_{y_i}) \prod_{t=1}^N p(y_t|x_t, \Theta^{old}) \right] \quad (3.24)$$

In the work of Bilmes (1998), it is also proved that the equation (3.24) can be simplified as:



$$Q(\Theta, \Theta^{old}) = \sum_{m=1}^M \sum_{i=1}^N \log(\alpha_m) p(m|x_i, \Theta^{old}) + \sum_{m=1}^M \sum_{i=1}^N \log(p_m(x_i|\theta_m)) p(m|x_i, \Theta^{old}) \quad (3.25)$$

where  $\alpha_m$  is the prior probability of each mixture component,  $\sum_m \alpha_m = 1$ ; the  $p_m(x_i|\theta_m)$  is the distribution of  $x_i$  given the parameters  $\theta_m$ ; the  $p(m|x_i, \Theta^{old})$  is the a posteriori probability of  $m$  given the current parameters indicating the probability of  $x_i$  generated by the component  $m$ . We can compute the posteriori probability by the Bayesian theorem:

$$p(m|x_i, \Theta^{old}) = \frac{\alpha_m^{old} p_m(x_i|\theta_m^{old})}{\sum_{k=1}^M \alpha_k^{old} p_k(x_i|\theta_k^{old})} \quad (3.26)$$

The second step of the EM algorithm is to maximize the expectation  $Q(\Theta, \Theta^{old})$  by maximizing the terms including  $\alpha_m$  and the terms including  $\theta_m$  independently in equation (3.25). To estimate the parameter of  $\alpha_m$ , by introducing the Lagrange multiplier  $\lambda$  with the constraint that  $\sum_m \alpha_m = 1$  and setting the derivatives of the Lagrange function with respect  $\alpha_m$  to zero, we obtained the new parameter of  $\alpha_m^{new}$  dependent on the current parameters  $\Theta^{old}$  and  $\alpha_m^{old}$ .

$$\alpha_m^{new} = \frac{1}{N} \sum_{i=1}^N p(m|x_i, \Theta^{old}) \quad (3.27)$$

The expression of the parameter of  $\theta_m$  is dependent on the component distribution, for example in our case, the component distribution is the  $d$ -dimensional Gaussian distribution with the mean vector  $\mu_m$  and covariance matrix  $\Sigma_m$ , that is  $\theta_m = (\mu_m, \Sigma_m)$ :

$$p_m(x_i|\theta_m) = N(x_i|\mu_m, \Sigma_m) = \frac{1}{(2\pi)^{d/2}|\Sigma_m|^{1/2}} e^{-\frac{1}{2}(x-\mu_m)^T \Sigma_m^{-1}(x-\mu_m)} \quad (3.28)$$

Introducing the equation (3.28) to the (3.25) and then taking the partial derivative of the parameter  $\mu_m$  and  $\Sigma_m$  separately (A. Bilmes, 1998), we obtain:

$$\mu_m^{new} = \frac{\sum_{i=1}^N x_i p(m|x_i, \Theta^{old})}{\sum_{i=1}^N p(m|x_i, \Theta^{old})} \quad (3.29)$$

$$\Sigma_m^{new} = \frac{\sum_{i=1}^N p(m|x_i, \Theta^{old})(x_i - \mu_m^{new})(x_i - \mu_m^{new})^T}{\sum_{i=1}^N p(m|x_i, \Theta^{old})} \quad (3.30)$$

After the expectation step and the maximization step, we obtain the new parameters which maximizing the maximum-likelihood function  $Q(\Theta, \Theta^{old})$ . However, the equations (3.27), (3.29) and (3.30) do not provide closed-form solutions for estimating parameters of the mixture model. Thus an iterative scheme for finding the maximum likelihood is taken as a solution which is also proved to be convergent. We first choose some initial values (e.g. initializing by the  $k$ -mean procedure) for the parameters of the mixture-density distribution, for example mean vector, covariance matrices and mixing coefficients in the context of the GMM. Then we alternate between the following two updates which are the expectation (E) step and the maximization (M) step. In the expectation step, or E step, we use the current values for the parameters to evaluate the a posteriori probabilities, given by (3.26). We then use these probabilities in the maximization step, or M step, to re-estimate the mixing coefficients, mean vectors and covariance matrices using the results (3.27), (3.29) and (3.30). In practice, the iterative procedure will stop when the increment of the maximum-likelihood function  $\Delta Q(\Theta, \Theta^{old})$  falls below some threshold. Given a GMM, the EM algorithm is summarized as below:

1. Initialize the mean vectors  $\mu_m$ , covariance matrices  $\Sigma_m$  and mixing coefficients  $\alpha_m$ , and evaluate the initial value of the log likelihood  $Q$ .
2. Expectation (E) step. Evaluate the a posteriori probability of  $m$  given the current parameters denoted as  $p(m|x_i, \Theta^{old})$ , i.e. so-called responsibility in

some literature (Bishop, 2006), using the current parameter values  $\alpha_m^{old}$ ,  $\mu_m^{old}$  and  $\Sigma_m^{old}$ :

$$\gamma_{im} = p(m|x_i, \Theta^{old}) = \frac{\alpha_m^{old} N(x_i|\mu_m^{old}, \Sigma_m^{old})}{\sum_{k=1}^M \alpha_k^{old} N(x_i|\mu_k^{old}, \Sigma_k^{old})} \quad (3.31)$$

3. Maximization (M) step. Re-estimate the parameters using the current responsibilities, i.e. the a posteriori probabilities:

$$\alpha_m^{new} = \frac{N_m}{N} \quad (3.32)$$

$$\mu_m^{new} = \frac{\sum_{i=1}^N \gamma_{im} x_i}{N_m} \quad (3.33)$$

$$\Sigma_m^{new} = \frac{\sum_{i=1}^N \gamma_{im} (x_i - \mu_m^{new})(x_i - \mu_m^{new})^T}{N_m} \quad (3.34)$$

where

$$N_m = \sum_{i=1}^N \gamma_{im} \quad (3.35)$$

4. Evaluated the log likelihood  $Q$  shown in (3. 25). If the convergence condition

$$\Delta Q = Q^{k+1} - Q^k < Threshold \quad (3.36)$$

is not satisfied return to step 2, where  $k$  is the number of iterative times.

Since the E step computing the a posteriori probability requires the covariance matrices to be inverted, the problem would occur when one or more component covariance matrices are singular or near-singular. This problem is more likely caused by the number of the components is large relative to the number of the training data or the mixture components are not well separated. A more principled solution to this problem than the addition of a multiple of the identity matrix to near singular covariance matrices is to impose constraints on the covariance matrices structure (Webb, 2011). The most often three constrains are:

- Diagonal covariance matrices : using the diagonal covariance matrices  $Diag(\sigma_{m,1}^2, \sigma_{m,2}^2, \dots, \sigma_{m,p}^2)$  instead of the original covariance matrices  $\Sigma_m$  and the update equation for the diagonal elements is:

$$\sigma_{m,l}^{2(new)} = \frac{\sum_{i=1}^N \gamma_{im} (x_{i,l} - \mu_{m,l}^{new})^2}{N_m}, \quad l \in [1, \dots, p] \quad (3.37)$$

where  $p$  is the dimensionality of the data and  $\sigma_{m,l}^{2(new)}$  is  $l$ th diagonal element of the updated diagonal covariance matrices,  $x_{i,l}$  and  $\mu_{m,l}$  are the  $l$ th dimension of the  $x_i$  and  $\mu_m$  respectively.

- Spherical covariance matrices : using the spherical covariance, i.e. a singer variance  $\sigma_m^2$  multiplying the  $p \times p$  identity matrix  $I_{p,p}$  denoted as  $\sigma_m^2 I_{p,p}$ , instead of the original covariance matrices  $\Sigma_m$ . The updated equation for the singer variance  $\sigma_m^2$  is:

$$\sigma_m^{2(new)} = \frac{\sum_{i=1}^N \gamma_{im} (x_i - \mu_m^{new})^T (x_i - \mu_m^{new})}{p N_m} \quad (3.38)$$

- Common covariance matrix across mixture components: using a common covariance matrix across all mixture components instead of the original covariance matrices  $\Sigma_m$ . The updated equation for the common covariance matrix is:

$$\Sigma_m^{new} = \frac{\sum_{m=1}^M \sum_{i=1}^N \gamma_{im} (x_i - \mu_m^{new}) (x_i - \mu_m^{new})^T}{N} \quad (3.39)$$

In our work, the common covariance matrix is used as a constraint to avoid the singular covariance matrices.

The above discussion is given on a fix number of the components. In fact the maximum-likelihood criterion cannot be used to estimate the number of mixture components because the maximized likelihood is a nondecreasing function of the number of components, thereby making it useless as a model selection criterion (Jain et al., 2000). Choosing the number of components in a mixture model depends on many factors including the actual distribution of the data being modeled, the shape of

clusters, and the amount of available training data. We can train a number of different mixture model, each using a different number of mixture components, and apply each to a separate test set. The model order can then be selected as that giving rise to the best performance on this test set.

### **3.3.3. GMM-based mapping method**

The GMM-based continuous stochastic mapping methods have been prevalent for years in various speech applications, such as the speech conversion, articulatory-to-acoustic mapping, acoustic-to-articulatory inversion, speaking aids (Kain et al., 2004; Nakamura et al., 2006), noise compensation (Afify et al., 2007; Cui et al., 2008) and so on. Summarizing the mapping methods used in literature, the GMM-based mapping methods use the Bayesian criteria such as MMSE and MAP to estimate the target vectors in the context of Gaussian mixture distribution. The principal procedure of the GMM-based mapping method is:

- (1) The joint probability density function of source and target vector sequence is modeled by a GMM which is trained by EM. It is proved that using the joint source and target vectors rather than the source vectors only to train the GMM is more robust for small amounts since the joint density should lead to a more judicious clustering for the regression problem (Kain et al., 1998).
- (2) The conditional probability density function of a target vector sequence given a source vector sequence is estimated by the joint probability density function.
- (3) The mapped target vector sequence is determined with the Bayesian approaches to minimize its mean-square error or maximize its a posteriori probability based on the conditional probability density function.

The mapping procedure is illustrated in Figure 3.1:

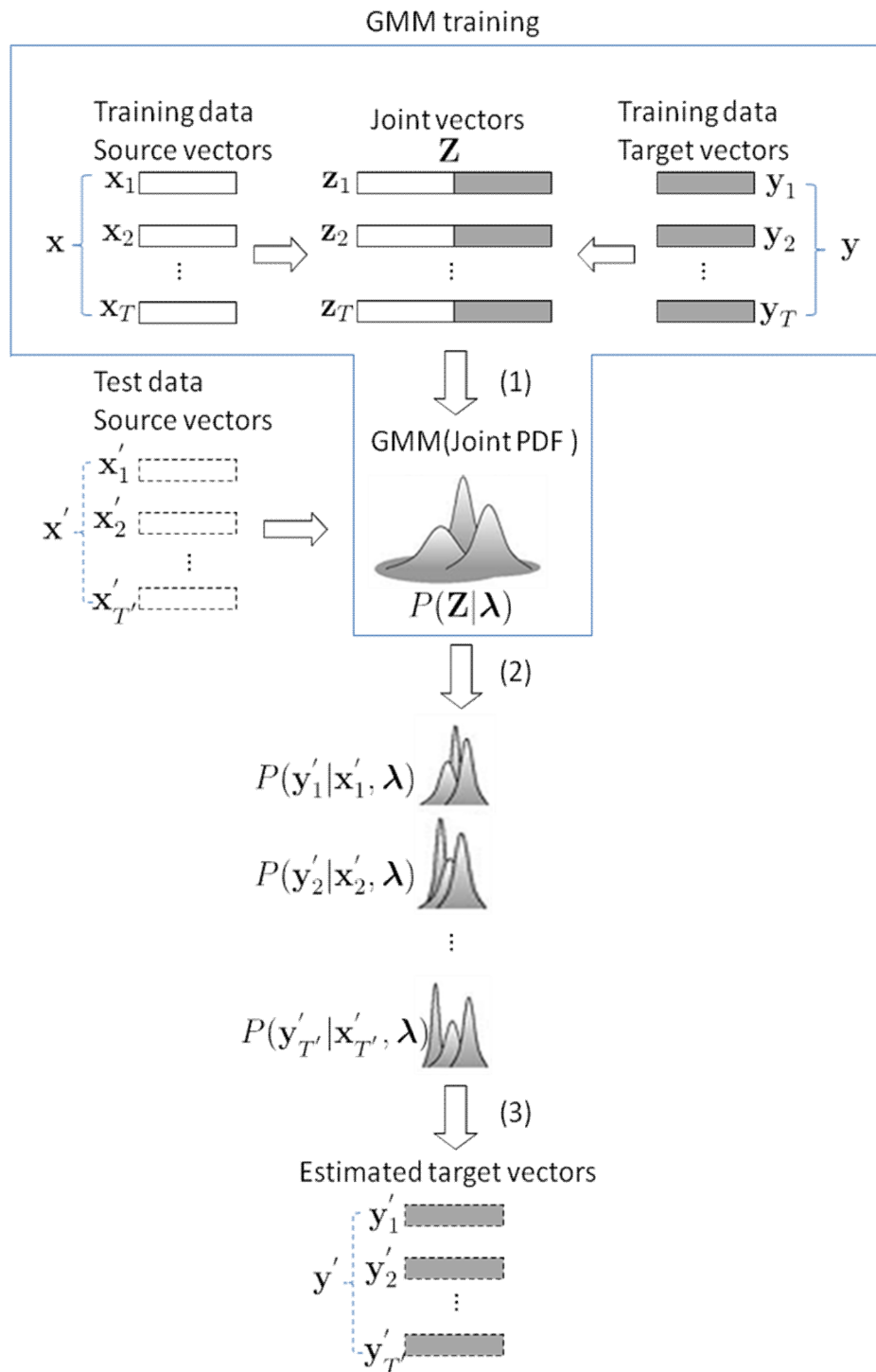


Figure 3.1: The procedure of the GMM-based mapping method.

### 3.3.3.1. GMM-based mapping approach under the criterion of MMSE

Given the  $p$ -dimension source vector  $\mathbf{x}_t = [x_t(1), x_t(2), \dots, x_t(p)]^T$  and a  $p$ -dimension target vector  $\mathbf{y}_t = [y_t(1), y_t(2), \dots, y_t(p)]^T$  ( $t$  denotes the frame number in the temporal sequence), the joint vector is  $\mathbf{z}_t = [\mathbf{x}_t^T, \mathbf{y}_t^T]^T$ . The joint probability density of the source and target vector is modeled by GMM as follows:

$$p(\mathbf{z}_t | \boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^M \alpha_m \mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}) \quad (3.40)$$

where

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \quad (3.41)$$

$\boldsymbol{\mu}_m^{(x)}$  and  $\boldsymbol{\mu}_m^{(y)}$  are the mean vector of the  $m$ th mixture component for the source and that for the target, respectively. The matrices  $\boldsymbol{\Sigma}_m^{(xx)}$  and  $\boldsymbol{\Sigma}_m^{(yy)}$  are the covariance matrix of the  $m$ th mixture component for the source and that for target, respectively. The matrices  $\boldsymbol{\Sigma}_m^{(xy)}$  and  $\boldsymbol{\Sigma}_m^{(yx)}$  are the cross-covariance matrices of the  $m$ th mixture component for the source and target. The GMM is parameterized by a set of triple parameters: the mean vector  $\boldsymbol{\mu}_m$ , covariance matrix  $\boldsymbol{\Sigma}_m$  and mixture weight  $\alpha_m$ .

$$\boldsymbol{\lambda}^{(z)} = \{ \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}, \alpha_m \} \quad m = [1, 2, \dots, M] \quad (3.42)$$

After training the parameters of the GMM using EM algorithm, the target data  $\mathbf{y}_t$  could be estimated by the given source data  $\mathbf{x}_t$ . In the sense of MMSE, the estimated target vector  $\hat{\mathbf{y}}_t$  is the conditional expectation of the  $\mathbf{y}_t$  given the observed source data  $\mathbf{x}_t$ , that is  $E[\mathbf{y}_t | \mathbf{x}_t]$  (Kay, 1993; Toda et al., 2008):

$$\begin{aligned}
\hat{\mathbf{y}}_t &= E[\mathbf{y}_t | \mathbf{x}_t] = \int p(\mathbf{y}_t | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)}) \mathbf{y}_t d\mathbf{y}_t \\
&= \int \sum_{m=1}^M p(\mathbf{y}_t | \mathbf{x}_t, m, \boldsymbol{\lambda}^{(z)}) p(m | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)}) \mathbf{y}_t d\mathbf{y}_t \\
&= \sum_{m=1}^M p(m | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)}) E_{m,t}^{(y)}
\end{aligned} \tag{3.43}$$

where

$$p(m | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)}) = \frac{\alpha_m N(\mathbf{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{n=1}^M \alpha_n N(\mathbf{x}_t; \boldsymbol{\mu}_n^{(x)}, \boldsymbol{\Sigma}_n^{(xx)})} \tag{3.44}$$

$$E_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} (\mathbf{x}_t - \boldsymbol{\mu}_m^{(x)}) \tag{3.45}$$

$$\mathbf{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \boldsymbol{\Sigma}_m^{(xy)} \tag{3.46}$$

The  $p(m | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)})$  is a posteriori probability density indicating the probability of the  $m$ th component generates the  $\mathbf{x}_t$ . Indeed  $E_{m,t}^{(y)}$  is the estimated value of the target  $\mathbf{y}_t$  in the sense of least-square error contributed by the  $m$ th component. The final estimated vector is the sum of the weighted conditional mean vectors. Note that the fact that the equation (3.45) is same to the equation (3.5) indicates that the MLR model described in the section 3.2 is the special case of GMM-based mapping approach when the number of component is equal to one. Thereby the MLR could be regarded as a rough regression model based on the unimodal Gaussian. This point explains theoretically that the GMM-based mapping method as a local regression method should be superior to the global linear regression method.

### 3.3.3.2. GMM-based mapping approach under the criterion of MAP

In Bayesian statistics, a MAP estimate is a mode of the posterior distribution. The MAP is closely related to the maximum likelihood estimation (MLE), which could be seen as a regularization of MLE. In the MAP estimation approach we choose  $\hat{\mathbf{y}}_t$  to maximize the posterior probability density function (pdf) as below (Kay, 1993):



$$\hat{\mathbf{y}}_t = \arg \max_{\mathbf{y}_t} p(\mathbf{y}_t | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)}) \quad (3.47)$$

The MAP can be computed in many ways: computing analytically when the posterior can be given in closed form; computing by an EM algorithm, this does not require derivatives of the posterior density or by a Monte Carlo method using simulated annealing. In this work the MAP estimator is in the context of mixture Gaussian density distribution, thus EM algorithm is adopted for computing  $\mathbf{y}_t$ . A auxiliary function which is indeed the expectation of the latent variable is defined as follow (Afify et al., 2007; Toda et al., 2008):

$$\begin{aligned} Q(\mathbf{y}_t, \hat{\mathbf{y}}_t) &= \sum_m^M p(m | \mathbf{x}_t, \mathbf{y}_t, \boldsymbol{\lambda}^{(z)}) \log p(\hat{\mathbf{y}}_t, m | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)}) \\ &= \sum_{m=1}^M p(m | \mathbf{x}_t, \mathbf{y}_t, \boldsymbol{\lambda}^{(z)}) \log p(\hat{\mathbf{y}}_t | \mathbf{x}_t, m, \boldsymbol{\lambda}^{(z)}) p(m | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)}) \\ &= \sum_{m=1}^M p(m | \mathbf{x}_t, \mathbf{y}_t, \boldsymbol{\lambda}^{(z)}) [\log p(\hat{\mathbf{y}}_t | \mathbf{x}_t, m, \boldsymbol{\lambda}^{(z)}) + \log p(m | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)})] \\ &= \sum_{m=1}^M p(m | \mathbf{x}_t, \mathbf{y}_t, \boldsymbol{\lambda}^{(z)}) \log p(\hat{\mathbf{y}}_t | \mathbf{x}_t, m, \boldsymbol{\lambda}^{(z)}) + \sum_{m=1}^M p(m | \mathbf{x}_t, \mathbf{y}_t, \boldsymbol{\lambda}^{(z)}) \log p(m | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)}) \end{aligned} \quad (3.48)$$

The estimated value  $\hat{\mathbf{y}}_t$  is the one which mixmizes the function  $Q(\mathbf{y}_t, \hat{\mathbf{y}}_t)$

$$\hat{\mathbf{y}}_t = \arg \max_{\mathbf{y}_t} Q(\mathbf{y}_t, \hat{\mathbf{y}}_t) \quad (3.49)$$

Thereby we can see that in the equation (3.48), the second part of the equation which is independent of the  $\hat{\mathbf{y}}_t$  could be ignored in the estimation of the  $\hat{\mathbf{y}}_t$ . The  $Q(\mathbf{y}_t, \hat{\mathbf{y}}_t)$  can be define as:

$$Q(\mathbf{y}_t, \hat{\mathbf{y}}_t) = \sum_{m=1}^M p(m | \mathbf{x}_t, \mathbf{y}_t, \boldsymbol{\lambda}^{(z)}) \log p(\hat{\mathbf{y}}_t | \mathbf{x}_t, m, \boldsymbol{\lambda}^{(z)}) \quad (3.50)$$

where the  $\hat{\mathbf{y}}_t$  is governed as the unimodal Gaussian distribution given the component  $m$  and source data  $\mathbf{x}_t$  with the parameters  $\boldsymbol{\lambda}^{(z)}$ :

$$p(\hat{\mathbf{y}}_t | \mathbf{x}_t, m, \boldsymbol{\lambda}^{(z)}) = \frac{1}{e^{-\frac{1}{2}(\hat{\mathbf{y}}_t - \mu_{\hat{\mathbf{y}}|\mathbf{x},m,\boldsymbol{\lambda}^{(z)}})^T \Sigma_{\hat{\mathbf{y}}|\mathbf{x},m,\boldsymbol{\lambda}^{(z)}}^{-1} (\hat{\mathbf{y}}_t - \mu_{\hat{\mathbf{y}}|\mathbf{x},m,\boldsymbol{\lambda}^{(z)}})}} (2\pi)^{d/2} |\Sigma_{\hat{\mathbf{y}}|\mathbf{x},m,\boldsymbol{\lambda}^{(z)}}|^{1/2}} \quad (3.51)$$

where  $\mu_{\hat{\mathbf{y}}|\mathbf{x},m,\boldsymbol{\lambda}^{(z)}}$  being the conditional expectation of  $\hat{\mathbf{y}}_t$  given the component  $m$  and source data  $\mathbf{x}_t$ , is equal to the  $E_{m,t}^{(y)}$  showed in equation (3.45);  $\Sigma_{\hat{\mathbf{y}}|\mathbf{x},m,\boldsymbol{\lambda}^{(z)}}$  being the conditional covariance, is equal to the  $\mathbf{D}_m^{(y)}$  showed in equation (3.46). Introducing  $E_{m,t}^{(y)}$ ,  $\mathbf{D}_m^{(y)}$  and equation (3.51) into equation (3.50):

$$\begin{aligned} Q(\mathbf{y}_t, \hat{\mathbf{y}}_t) &= -\frac{1}{2} \sum_{m=1}^M p(m|\mathbf{x}_t, \mathbf{y}_t, \boldsymbol{\lambda}^{(z)}) \log |\mathbf{D}_m^{(y)}| \\ &\quad - \frac{1}{2} \sum_{m=1}^M p(m|\mathbf{x}_t, \mathbf{y}_t, \boldsymbol{\lambda}^{(z)}) (\hat{\mathbf{y}}_t - E_{m,t}^{(y)})^T \mathbf{D}_m^{(y)-1} (\hat{\mathbf{y}}_t - E_{m,t}^{(y)}) \end{aligned} \quad (3.52)$$

Noting that both  $E_{m,t}^{(y)}$  and  $\mathbf{D}_m^{(y)}$  which are fixed by the GMM training processing are independent of  $\hat{\mathbf{y}}_t$ . Thus the first part of the auxiliary function  $Q(\mathbf{y}_t, \hat{\mathbf{y}}_t)$  shown in equation (3.52) could be ignored and finally the function  $Q(\mathbf{y}_t, \hat{\mathbf{y}}_t)$  is simplified as:

$$Q(\mathbf{y}_t, \hat{\mathbf{y}}_t) = -\frac{1}{2} \overline{\hat{\mathbf{y}}_t^T \mathbf{D}_t^{(y)-1} \hat{\mathbf{y}}_t} + \overline{\hat{\mathbf{y}}_t^T \mathbf{D}_t^{(y)-1} \mathbf{E}_t^{(y)}} + \bar{K}_t \quad (3.53)$$

where

$$\overline{\mathbf{D}_t^{(y)-1}} = \sum_{m=1}^M \gamma_{m,t}^{(z)} \mathbf{D}_m^{(y)-1} \quad (3.54)$$

$$\overline{\mathbf{D}_t^{(y)-1} \mathbf{E}_t^{(y)}} = \sum_{m=1}^M \gamma_{m,t}^{(z)} \mathbf{D}_m^{(y)-1} E_{m,t}^{(y)} \quad (3.55)$$

$$\gamma_{m,t}^{(z)} = p(m|\mathbf{x}_t, \mathbf{y}_t, \boldsymbol{\lambda}^{(z)}) \quad (3.56)$$

$\bar{K}_t$  is a item independent of  $\hat{\mathbf{y}}_t$ . By differentiating equation (3.53) with respect to  $\hat{\mathbf{y}}_t$ , setting the derivative to zero, and solving for  $\hat{\mathbf{y}}_t$ , we arrive at the estimate of target  $\hat{\mathbf{y}}_t$  given by:

$$\hat{\mathbf{y}}_t = \overline{(\mathbf{D}_t^{(y)})^{-1}}^{-1} \overline{\mathbf{D}_t^{(y)} \mathbf{E}_t^{(y)}} \quad (3.57)$$

However the equation of the estimation of  $\hat{\mathbf{y}}_t$  shown in the (3.57) is not a close form due to the calculation of the a posteriori probability  $\gamma_{m,t}^{(z)}$  including the old  $\mathbf{y}_t$ . Thus the EM algorithm is used to estimate  $\hat{\mathbf{y}}_t$ . The E(expectation) step is to compute the a posteriori probability  $\gamma_{m,t}^{(z)}$  given the current parameter  $\mathbf{y}_t$  which is the estimated  $\hat{\mathbf{y}}_t$  in the last step, and the successive M(maximization) step is to estimate the target  $\hat{\mathbf{y}}_t$  given the calculated  $\gamma_{m,t}^{(z)}$ . When the result of auxiliary function  $Q(\mathbf{y}_t, \hat{\mathbf{y}}_t)$  satisfy the convergence condition given the predefined threshold:

$$\Delta Q = Q^{k+1} - Q^k < \text{Threshold} \quad (3.58)$$

the iterative EM procedure stop and we obtain the final estimated target vector  $\hat{\mathbf{y}}_t$ . Note that  $E_{m,t}^{(y)}$  and  $\mathbf{D}_m^{(y)}$  are considered as constant items in the EM procedure. In addition, the initialization of EM algorithm for estimating parameters  $\hat{\mathbf{y}}_t$  is achieved by the MMSE mapping method.

The EM algorithm for estimating target  $\hat{\mathbf{y}}_t$  is summarized as follow:

1. Initializing the parameter  $\hat{\mathbf{y}}_t$  by the MMSE mapping method;
2. Expectation (E) step: Evaluate the a posteriori probability of  $m$  given the current parameters  $\mathbf{y}_t$  (noting that  $\mathbf{x}_t$  and  $\boldsymbol{\lambda}^{(z)}$  are independent with  $\hat{\mathbf{y}}_t$ ):

$$\gamma_{m,t}^{(z)} = p(m|\mathbf{x}_t, \mathbf{y}_t, \boldsymbol{\lambda}^{(z)}) \quad (3.59)$$

3. Maximization (M) step: Re-estimate the parameters given the current the a posteriori probabilities ( $E_{m,t}^{(y)}$  and  $\mathbf{D}_m^{(y)}$  determined in GMM training are constant items):

$$\hat{\mathbf{y}}_t = \overline{(\mathbf{D}_t^{(y)})^{-1}}^{-1} \overline{\mathbf{D}_t^{(y)} \mathbf{E}_t^{(y)}} = \left( \sum_{m=1}^M \gamma_{m,t}^{(z)} \mathbf{D}_m^{(y)} \right)^{-1} \sum_{m=1}^M \gamma_{m,t}^{(z)} \mathbf{D}_m^{(y)} \mathbf{E}_{m,t}^{(y)} \quad (3.60)$$

4. Evaluate the log likelihood  $Q(\mathbf{y}_t, \hat{\mathbf{y}}_t)$  by using the equation (3.53). If the convergence condition:

$$\Delta Q = Q^{k+1} - Q^k < \text{Threshold} \quad (3.61)$$

is not satisfied return to step 2, otherwise use the equation (3.57) obtain the final estimated target parameters  $\hat{\mathbf{y}}_t$ .

It is shown in equation (3.57) that the MAP mapping method uses not only the mean vectors  $E_{m,t}^{(y)}$  but also the covariance matrices  $\mathbf{D}_m^{(y)}$ , the conditional probability density for the estimating the target vector. That is to say the MAP estimator includes more information than the MMSE-based estimator which only uses the mean vectors. These covariance matrices in the MAP estimator are used as weights in the weighted sum of the conditional mean vectors. They are regarded as a confidence measure of the conditional mean vectors from individual mixture components (Toda et al., 2008).



# **Chapter 4. Estimation of lip features from natural image of mouth region-of-interest**

## **4.1. Introduction**

Human speech perception is bimodal in nature: Humans combine audio and visual information in deciding what has been spoken, especially in noisy environments (Potamianos et al., 2012). Visual information from lip image is widely used to help speech recognition leading to audiovisual speech recognition system. The problem of obtaining the visual features from lip image has two issues: the first issue is the lips or mouth tracking; the second issue is visual speech representation in terms of a small number of informative features (Potamianos et al., 2012). After the mouth or part of lower face region is located, the algorithms for the lip contour tracking can be implemented (Adjoudani et al., 1996; Li et al., 2004; Garcia et al., 2007). One of the popular methods for lip or mouth tracking is snakes, which is an elastic curve represented by a set of control points. The control point coordinates are iteratively updated, by converging towards the local minimum of an energy function, defined on basis of curve smoothness constraints and a matching criterion to the desired features of the image (Kass et al., 1988; Chiou et al., 1997). Another widely used method for lip tracking is by means of lip templates (Yuille et al., 1992; Chandramohan et al., 1996; Aleksic et al., 2002) which constitute parameterized curves that are fitted to the desired shape by minimizing an energy function, defined similarly to snakes. In addition, active shape and appearance models construct a lip shape or ROI appearance statistical model for lip tracking (Cootes et al., 1998). This assumes that, given small perturbations from the actual fit of the model to a target image, a linear relationship exists between the difference in the model projection and image and the required updates to the model parameters (Potamianos et al., 2012). The lip tracking is aiming to extract the visual features of the lip for the automatic speech reading in the audiovisual recognition. Various sets of visual features for automatic speech reading

have been proposed in the recent two decades. They are grouped generally in three categories (Potamianos et al., 2012): (a) appearance (i.e. video pixel) based features; (b) shape (i.e. lip contour) based features; and (c) features that are a combination of both appearance and shape. Typically in the appearance based features, the feature vector is obtained by the appropriate transformations for compressing the dimension of the image of the speaker's mouth ROI, such as DCT (Potamianos et al., 1998; Neti et al., 2000; Nefian et al., 2002; Barker et al., 2009) or PCA (Dupont et al., 2000; Hazen et al., 2004), applying to the pixels of the image of the speaker's mouth ROI. In contrast to appearance based features, the shape based features assumes that most speech reading information is contained in the shape (contours) of the speaker's lips or more generally in the face contour (Matthews et al., 2001). Geometric features and shape-model based features are the two types of the lip shape-based features. In terms of the geometric type ones, the feature of lip contour are considered as the features, such as the contour height, width, perimeter as well as the area contained within the contour (Adjoudani et al., 1996; Zhang et al., 2000; Heckmann et al., 2001; Heracleous et al., 2009; Ming et al., 2010). In terms of the shape model based features, the parameters of the parametric models are used as visual features, such as using a number of snake radial vectors as visual features in the snake based lip contour estimation model (Chiou et al., 1997) or using the lip template parameters (Chandramohan et al., 1996; Su et al., 1996). Various combinations of the appearance-based and shape-based features are used in the automatic speech reading system ranging from simple concatenation to joint modelling, for example, combining the geometric lip features with the PCA projection of a subset of pixels contained in the corresponding mouth ROI and so on (Dupont et al., 2000; Chan, 2001; Aleksic et al., 2005; Papandreou et al., 2009). In our work, we present a method that estimates directly speech lip geometric features by concentrated information obtained by the DCT coefficients of the natural image of mouth ROI. Fontecave Jallon et al. (2009) applied a similar procedure to estimate the geometrical vocal tract contours based on the key images of X-ray data. In the following, we first present the experimental setup and the lip material and then we present the image processing. The MLR estimation model and the GMM-based estimation model are evaluated respectively

for estimating lip geometric features. Finally, the summary is drawn at the end of this chapter.

## **4.2. Experimental set-up and lip material**

The database used in this experiment is composed by the images of a continuous speech rather than merely vowels. The database is different from the one mentioned in the chapter 2 and it is divided into two partitions for training and test the model respectively. The lip geometric features, the lip width (A, Aext), the lip height (B, Bext) and the lip area (S, Sext) for both inner and outer contours, are extracted from the image of the mouth ROI. The method for extracting the lip geometric features is same as the one described in chapter 2. The whole process led to a database made of a set of 1439 (A, B, S, Aext, Bext, Sext) sextuplets. The first 768 frames are for training the model and the left 671 frames are for evaluation the model.

## **4.3. Image processing and DCT**

This section presents the different steps needed to build the model that allows the prediction of the six lip geometric features by an appropriate transformation of the image of the mouth ROI. The transformation was based on the 2-D DCT of the mouth ROI.

### **4.3.1. Detection of the mouth region-of-interest (ROI)**

In this work, since the estimation model of the lip geometric features is more interested than the lip tracking approach, the lip detection is realized simply by the chroma-key method. The mouth ROI has been defined as a 101x101 pixels square containing only the mouth. With the color segmentation method, the blue landmark locating in the center of the speaker's eyebrows is easily detected. Then the mouth ROI with a given distance from the landmark can be located.(see Figure 4.1).





Figure 4.1: Image of the speaker (left) and the associated mouth ROI (right). The landmark in the center of the speaker's eyebrows is used to locate the mouth ROI.

### 4.3.2. The 2D discrete cosine transform (DCT)

The mouth ROI obtained from the lip detection step was converted into a  $101 \times 101$  grayscale intensity image matrix. The DCT was chosen as the image transform instead of other transform methods because of excellent energy compaction for highly correlated images and widely used in the image compression methods. Also this method allows staying in the real numbers domain. The 2-D DCT was used in this work.

$$B_{rs} = \alpha_r \alpha_s \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} A_{jk} \cos \frac{\pi(2j+1)r}{2J} \cos \frac{\pi(2k+1)s}{2K}, \quad (4.1)$$

$$0 \leq r \leq J-1, 0 \leq s \leq K-1$$

where

$$\alpha_r = \begin{cases} \frac{1}{\sqrt{J}}, & r = 0 \\ \sqrt{2/J}, & 1 \leq r \leq J-1 \end{cases}, \quad \alpha_s = \begin{cases} \frac{1}{\sqrt{K}}, & s = 0 \\ \sqrt{2/K}, & 1 \leq s \leq K-1 \end{cases} \quad (4.2)$$

The matrix  $A_{jk}(J=101, K=101)$  is the set of pixels in grayscale contained in the mouth ROI and  $B_{rs}$  is the resulting transformed coefficients matrix.

### 4.3.3. Using a mask

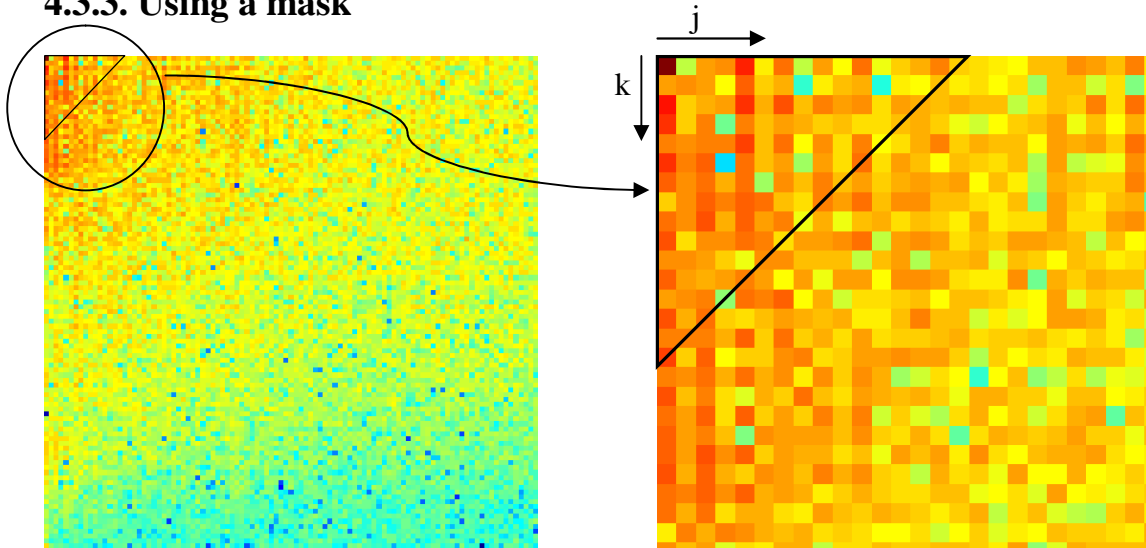


Figure 4.2: DCT coefficients matrix of the natural image of the mouth ROI (on the left) and the partial enlargement of the triangle mask area in the DCT matrix (on the right). Each pixel in the image is corresponding to a coefficient obtained by the DCT of the natural image of mouth ROI. The more darker (red) color of the pixel, the larger the coefficient. The top-left pixels indicate the largest coefficients in the DCT matrix. The triangle mask selected the most significant coefficients located at the top left of the DCT matrix.

At this stage, we need to recall that the aim of the work is to predict the high level geometric features by a limited set of coefficients that concentrate the main lip information. The DCT for its part packs energy in the low frequency regions, so some of the high frequency content can be ignored without significant quality degradation. Therefore we used a mask to select the most significant coefficients located in the top left of the DCT matrix in order to reduce the dimensionality (see Figure 4.2). We tested a triangle and a square mask both to compare the results.

Preliminary trials showed that the best results were obtained with masks containing about 100 DCT coefficients, i.e. only 1% of the total number of DCT coefficients (101x101). The optimal size and the shape of the mask will be discussed in the last section, with respect to the precision of the prediction.

## 4.4. Modelling

### 4.4.1. The MLR estimation modelling

This subsection establishes the MLR relation between the set of DCT coefficients included inside a mask and the six geometric lip features. For this, we used a subset of 768 elements of the database. The theoretical description of the MLR model is elaborated in the chapter 3. Here we present the application of the approach in the context of geometric lip features estimation by selected DCT vectors.

Given the matrix  $D = [D_1, D_2, \dots, D_M]$ ,  $M$  is the size of the mask. Each line contains the  $M$  DCT coefficients of the mask region for a frame instant. The number of lines is 768. Considering that the matrix  $D$  contains large image information and the generality of the linear fitting, we could use the linear combinations of vectors  $D_i$  to estimate the geometric lip features, by example for lip height  $B$ :

$$\hat{B} = f(D_1, D_2, \dots, D_p) \quad 1 \leq p \leq M \quad (4.3)$$

But in order to have a set of predictors with a number largely less than  $M$ , it is important to analyze the order of importance of each predictor. For this aim, we applied PCA on the DCT coefficients. The first principal component accounts for the maximum variance of the DCT coefficients, and each succeeding component accounts for the maximum of the residual variance (RVAR). The set of  $F$  factors of the PCA are the projections of the DCT coefficients on the principal axes. A subset of these factors has been used to predict  $B$  efficiently:

$$\hat{B} = f(F_1, F_2, \dots, F_p) \quad 1 \leq p \leq M \quad (4.4)$$

Thus the geometric lip features can be estimated the geometric lip features can be estimated step by step since the orthogonality property of the  $F$  factors. Thus for  $B$ , we obtained (see Figure 4.3):

$$\widehat{B} = k_1 F_1 + \overline{B} \quad (4.5)$$

where  $k_1$  is the linear fitting coefficient in the least square sense and  $\overline{B}$  is the mean value of  $B$ . We defined  $r_1 = B - \widehat{B}_1$  as the residual error of the first estimation, and we used  $F_2$  to estimate  $r_1$ , (Figure 4.4):

$$\widehat{r}_1 = k_2 F_2 \quad (4.6)$$

The estimation continues following the same procedure until the order  $p$  of prediction is attained. Finally, the estimated  $B$  values can be expressed as following:

$$\widehat{B} = k_1 F_1 + k_2 F_2 + \dots + k_p F_p + \overline{B} \quad (4.7)$$

$$\widehat{\mathbf{k}} = (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{B} \quad (4.8)$$

where  $\widehat{\mathbf{k}} = [k_1, k_2, \dots, k_p]$ ,  $\mathbf{F} = [F_1, F_2, \dots, F_p]$ ,  $\mathbf{B} = B - \overline{B}$  is the centered form of  $B$ .

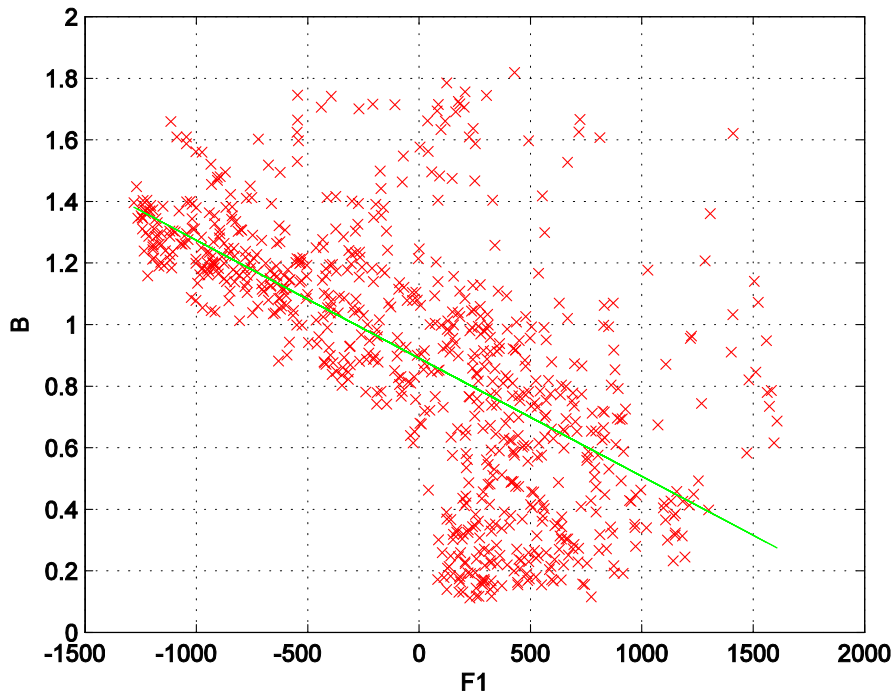


Figure 4.3:  $B$  values (red crosses) and its estimation (straight line) in function of  $F_1$  (with the use of a right-angled triangle mask of side 15).

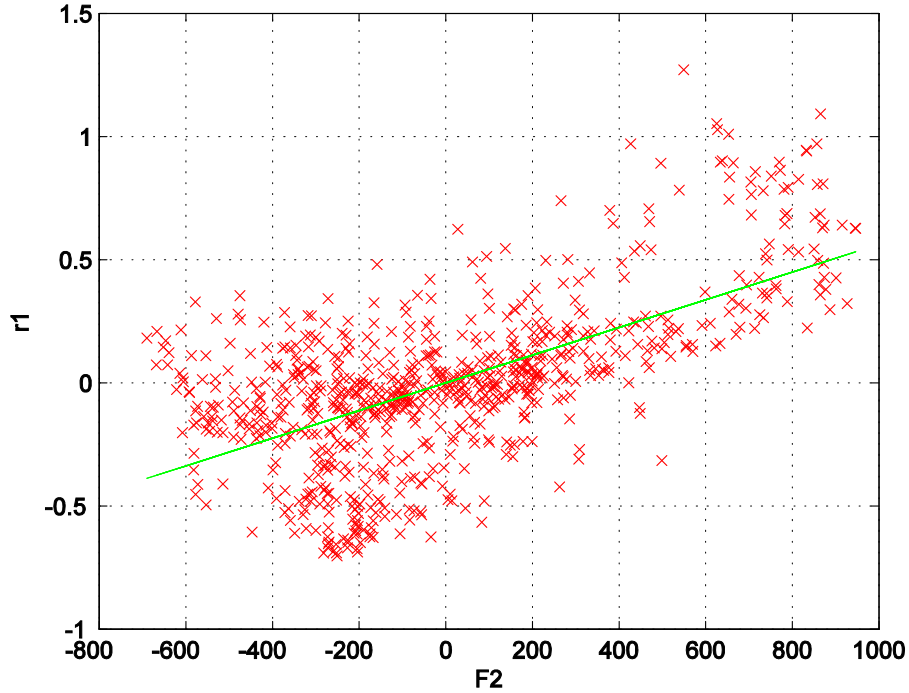


Figure 4.4:  $r_1$  values (crosses) and its estimation (straight line) in function of  $F_2$  (with the right-angled triangle mask of side 15).

#### 4.4.2. The GMM-based estimation modelling

As mentioned in the chapter 3, the GMM-based estimation model has two criteria on the regression or estimation phase: the criterion of MMSE or the criterion of MAP. We used the GMM-based estimation model under the criterion of MMSE as an estimator which is more correlated to the MLR model, so that we can compare the two models more intuitively. To recall the modelling process, we briefly present the formulas in the context of the experimental data used in this chapter. Given the  $p$ -dimension factor  $\mathbf{F}_t = [F_{1t}, F_{2t}, \dots, F_{pt}]^T$  as source data being used also in the MLR modelling and the geometric lip features vector  $\mathbf{y}_t = [A_t, B_t, S_t, Aext_t, Bext_t, Sext_t]^T$  as target data at frame  $t$ , the joint vector is  $\mathbf{z}_t = [\mathbf{F}_t^T, \mathbf{y}_t^T]^T$ . The joint probability density of the joint vector is modeled by GMM as follows:

$$P(\mathbf{z}_t | \lambda^{(z)}) = \sum_{m=1}^M \alpha_m N(\mathbf{z}_t; \mu_m^{(z)}, \Sigma_m^{(z)}) \quad (4.9)$$

where

$$\mu_m^{(z)} = \begin{bmatrix} \mu_m^{(F)} \\ \mu_m^{(y)} \end{bmatrix}, \quad \Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(FF)} & \Sigma_m^{(Fy)} \\ \Sigma_m^{(yF)} & \Sigma_m^{(yy)} \end{bmatrix} \quad (4.10)$$

$\mu_m^{(F)}$  and  $\mu_m^{(y)}$  are the mean vector of the  $m$ th mixture component for the source and target data respectively. The matrices  $\Sigma_m^{(FF)}$  and  $\Sigma_m^{(yy)}$  are the covariance matrix of the  $m$ th mixture component for the source and target data respectively. The matrices  $\Sigma_m^{(Fy)}$ ,  $\Sigma_m^{(yF)}$  are the cross-covariance matrices of the  $m$ th mixture component for the source and target. Every covariance matrix is full here. The GMM is parameterized by a set of triple parameters: the mean vector  $\mu_m$ , covariance matrix  $\Sigma_m$  and mixture weight  $\alpha_m$ .

$$\lambda^{(z)} = \{\mu_m, \Sigma_m, \alpha_m\} \quad m = [1, 2, \dots, M] \quad (4.11)$$

After training the parameters of the GMM with EM algorithm initialized by  $k$ -means method, the target data  $\mathbf{y}_t$  could be estimated by the given source data  $\mathbf{F}_t$  with the conditional expectation  $E[\mathbf{y}_t | \mathbf{F}_t]$  in the sense of the MMSE as follows:

$$\begin{aligned} \mathbf{y}_t = E[\mathbf{y}_t | \mathbf{F}_t] &= \int P(\mathbf{y}_t | \mathbf{F}_t, \lambda^{(z)}) \mathbf{y}_t d\mathbf{y}_t \\ &= \int \sum_{m=1}^M P(\mathbf{y}_t | \mathbf{F}_t, m, \lambda^{(z)}) P(m | \mathbf{F}_t, \lambda^{(z)}) \mathbf{y}_t d\mathbf{y}_t \\ &= \sum_{m=1}^M P(m | \mathbf{F}_t, \lambda^{(z)}) E_{m,t}^{(y)} \end{aligned} \quad (4.12)$$

where

$$P(m|F_t, \lambda^{(z)}) = \frac{\alpha_m N(F_t; \mu_m^{(F)}, \Sigma_m^{(FF)})}{\sum_{n=1}^M \alpha_n N(F_t; \mu_n^{(F)}, \Sigma_n^{(FF)})} \quad (4. 13)$$

$$E_{m,t}^{(y)} = \mu_m^{(y)} + \sum_m^{(yF)} \sum_m^{(FF)^{-1}} (F_t - \mu_m^{(F)}) \quad (4. 14)$$

The  $P(m|F_t, \lambda^{(z)})$  is a posteriori probability density of  $m$  indicating the probability of  $\mathbf{x}_t$  being generated by the  $m$ th Gaussian component. Indeed  $E_{m,t}^{(y)}$  is estimated value of the target  $\mathbf{y}_t$  in the sense of the least-square error contributed by the  $m$ th component. The final estimated target vector is the weighted sum of the conditional mean vectors. The equation (4. 12) also indicates that the MLR is the special case when the number of Gaussian is equal to one, thus the MLR could be considered as a rough regression model based on the unimodal Gaussian. Instead, the GMM which uses numbers of Gaussians to describe the distribution of the data has a precise local regression.

## 4.5. Evaluation

To measure the efficiency and accuracy of the regression by the MLR model and GMM-based regression model, we use the residual variance (RVAR) and RMSE to evaluate the models.

The efficiency of the estimation model can be measured by the RVAR as follows:

$$RVAR = Var_i / Var_0 \quad (4. 15)$$

where  $Var_i$  is the variance of the residual by using  $F_1$  to  $F_i$  predictors and  $Var_0$  is the variance of the lip feature to be predicted (see Figure 4.5).

The accuracy of the estimation model can be measured by the RMSE as follows:

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^T (\mathbf{y}_t - \hat{\mathbf{y}}_t)^2} \quad (4. 16)$$

where  $t$  denotes the frame number and  $T$  is the total number of frames,  $\mathbf{y}_t$  is the real value of the target and  $\hat{\mathbf{y}}_t$  is the estimated value of the target.

### 4.5.1. Evaluation of MLR model

The MLR model has been established in the previous section. But the choice of the system parameters, such as the shape and size of the mask, affects the performance of the model. The optimization consists to minimize RVAR, the shape and the size of the mask and the prediction order. If we fix the explained variance to 95%, in other words the RVAR to 5%, we find the optimal mask for the modelling. According to the results of the experiment (Table 4.1), we found that the triangle mask of side 8 is too small to contain enough information to explain the variance, the same remark as for the square mask of side 7. And we observe that the triangle mask of side 15 and the square mask of side 11 have almost the same good performance.

*Table 4.1: Performance of different masks. The line corresponding to each mask means the number of the predictors that explain 95% of the variance. If using all of the predictors still cannot explain 95% of the variance, the asymptotic value of explained variance is shown in the table.*

RMSE = Root Mean Square error (cm), T = Triangle Mask, S= Square Mask

|              | A    | B    | S    | Aext  | Bext | Sext |
|--------------|------|------|------|-------|------|------|
| T of side 15 | 43   | 17   | 13   | 93%   | 82   | 53   |
| RMSE         | 0.24 | 0.1  | 0.32 | 0.10  | 0.08 | 0.39 |
| T of side 8  | 90%  | 94%  | 13   | 91%   | 91%  | 93%  |
| RMSE         | 0.41 | 0.17 | 0.31 | 0.13  | 0.12 | 0.49 |
| S of side 7  | 92%  | 21   | 12   | 91%   | 92%  | 94%  |
| RMSE         | 0.38 | 0.1  | 0.36 | 0.128 | 0.11 | 0.45 |
| S of side 11 | 63   | 19   | 11   | 93%   | 95%  | 34   |
| RMSE         | 0.29 | 0.1  | 0.35 | 0.12  | 0.09 | 0.41 |



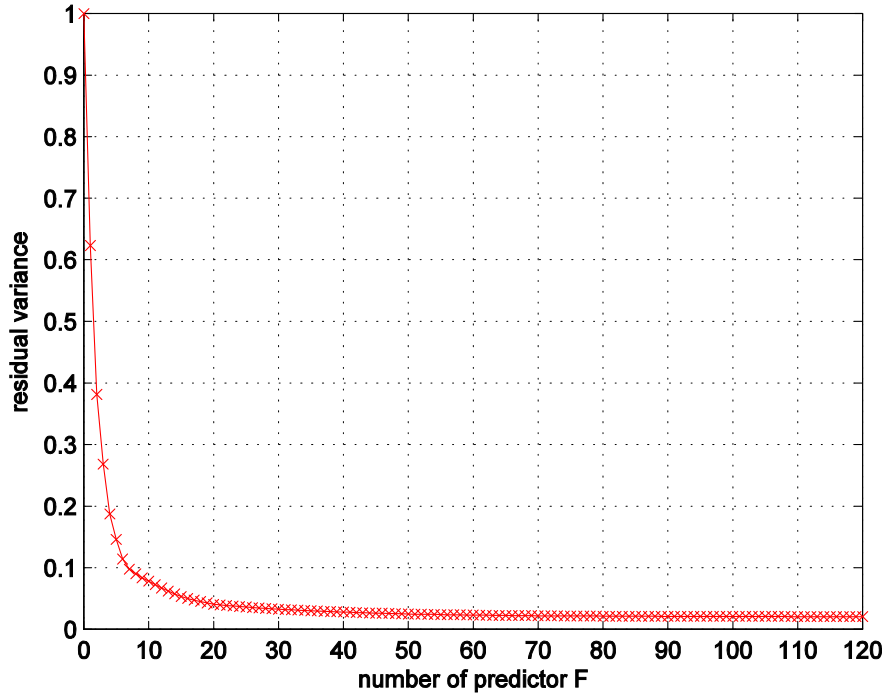


Figure 4.5: The RVAR of the training data for lip parameter  $B$  in function of the number of the ordered predictors  $F$  in the case of a triangle mask with side length 15.  $Var_0 = 0.17 \text{ cm}^2$ .

We can observe from the table that for  $B$ , only 17 or 21 predictors are needed to explain 95% of the variance for both masks. It means that less than 20% of the predictors could efficiently estimate the feature.

In order to validate our model, we used it to predict the test data made of the remaining 671 elements of the database that have not been included in the modelling process. We show in Figure 4.6 the RVAR in function of the prediction order for the test data for  $B$ . We can see a rapid decrement of the RVAR to attain a minimum value of 7 % similar to that obtained for the modelling data (see Figure 4.5). The estimated values for  $B$  of the test data obtained by the first 20 predictors is shown in Figure 4.7. We note that after the order of 20, no significant improvement can be obtained. Further, we extended the test to the case of images containing fingers near the mouth, as it is the case in CS (Cornett, 1967). The result of this preliminary test shows that the effect of the fingers in the mouth ROI is not significant (see also Figure 4.8), the RMSE being similar for  $B$  (0.12 cm with fingers vs. 0.10 cm without).

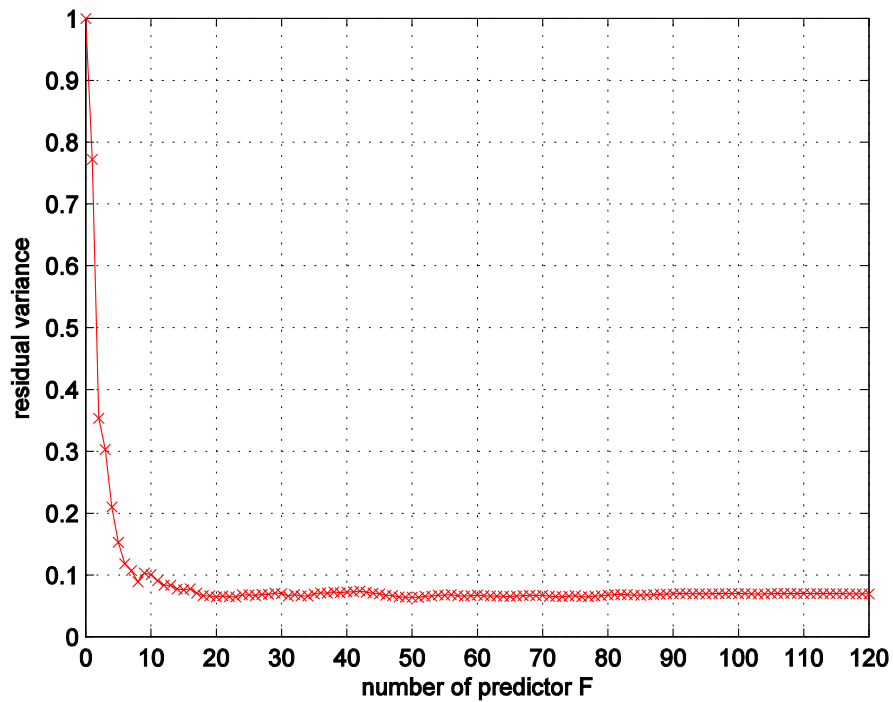


Figure 4.6: The RVAR of test data for lip parameter  $B$  in function of the number of the ordered predictors  $F$  in the case of a triangle mask with side length 15.  $\text{Var}_0 = 0.16 \text{ cm}^2$ .

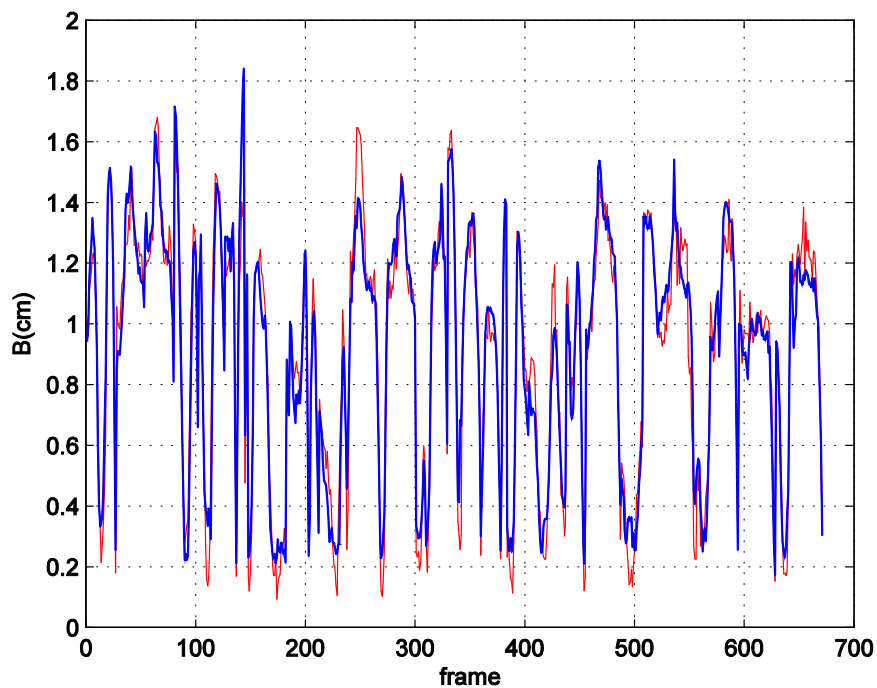


Figure 4.7: Estimation curve of test data for lip parameter  $B$  given by the first 20 ordered predictors, the red line is the test data, the blue line are the estimated values.

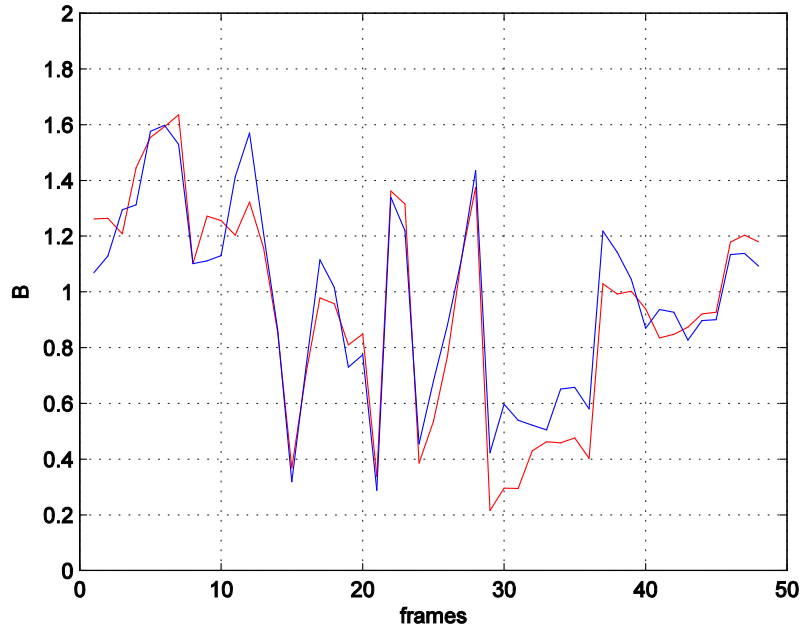


Figure 4.8: Estimation curve of test data for lip parameter  $B$  in the case of images containing fingers near the mouth given by the first 20 ordered predictors, the red line is the test data, the blue line are the estimated values.

#### 4.5.2. Evaluation of GMM-based estimation model

The GMM-based estimation model was established in the previous section. In this experiment, we used a triangle mask with side length 15 to extract the top left DCT coefficients in the DCT matrix, thus  $\mathbf{F}_t = [F_t(1), F_t(2), \dots, F_t(120)]$ . In order to compare with the performance of the MLR model, we used the same database for the training and test of the GMM-based estimation model. Figure 4.9 shows the RVAR for the training data obtained by the GMM-based estimation model with 3 components for parameter  $B$ . We can see that the RVAR of GMM-based estimation model attains a value less than 5 % quickly at the 7th predictor in comparison with the 17th predictor in the MLR model. This indicates that the GMM-based estimation model is more efficient than the MLR model. When the number of predictors is higher than 20, the RVARs of the two models decrease to their asymptotic values which are almost the same.

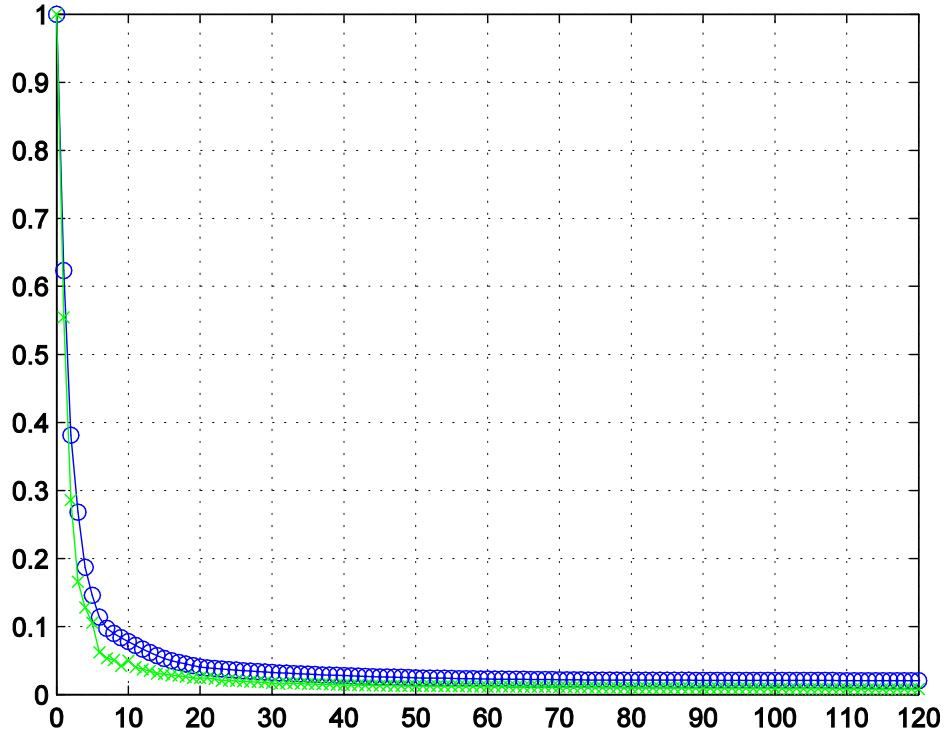


Figure 4.9: The RVARs of training data for lip parameter  $B$  in function of the number of the ordered predictors  $F$ . The green line corresponds to the GMM-based estimation model with 3 components and the blue dotted line is corresponding to the MLR model.  $Var_0 = 0.17cm^2$ .

For test data, the evaluation results are shown in Figure 4.10. The RVARs of the first 20 predictors also indicate that the performance of the GMM-based estimation model is slightly better than the MLR model. The RVAR of GMM-based estimation model decreases to the minimum value of 6% at 10th predictor in comparison with the 10% of the MLR model. The RMSEs of the two models shown in Figure 4.11 are in coherence with the RVARs shown in Figure 4.10. We can compare the two models more intuitively with the estimated values of parameter  $B$  shown in the Figure 4.12. We can observe that the GMM-based estimation model is slightly superior to MLR model especially when the value of target data is low. Besides the  $B$  parameters, Figure 4.13 , Figure 4.14 and Figure 4.15 show respectively the RVARs and RMSEs of the lip parameters  $A, S, A_{ext}, B_{ext}, S_{ext}$ .

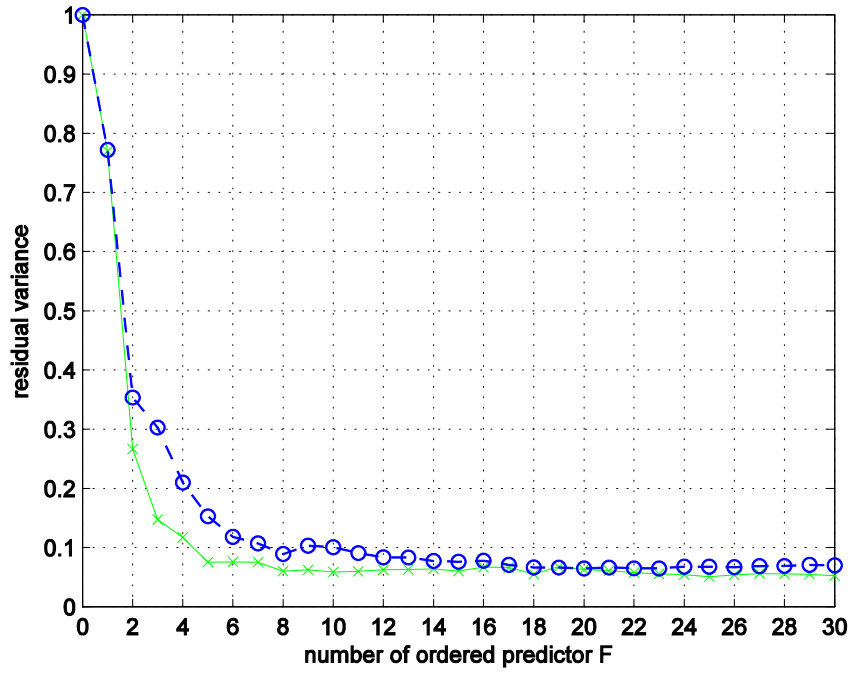


Figure 4.10: The RVARs of test data for lip parameter  $B$  in function of the number of the ordered predictors  $F$ . The green line is corresponding to the GMM-based estimation model and the blue dotted line is corresponding to the MLR model.  $\text{Var}_0 = 0.16 \text{ cm}^2$ .

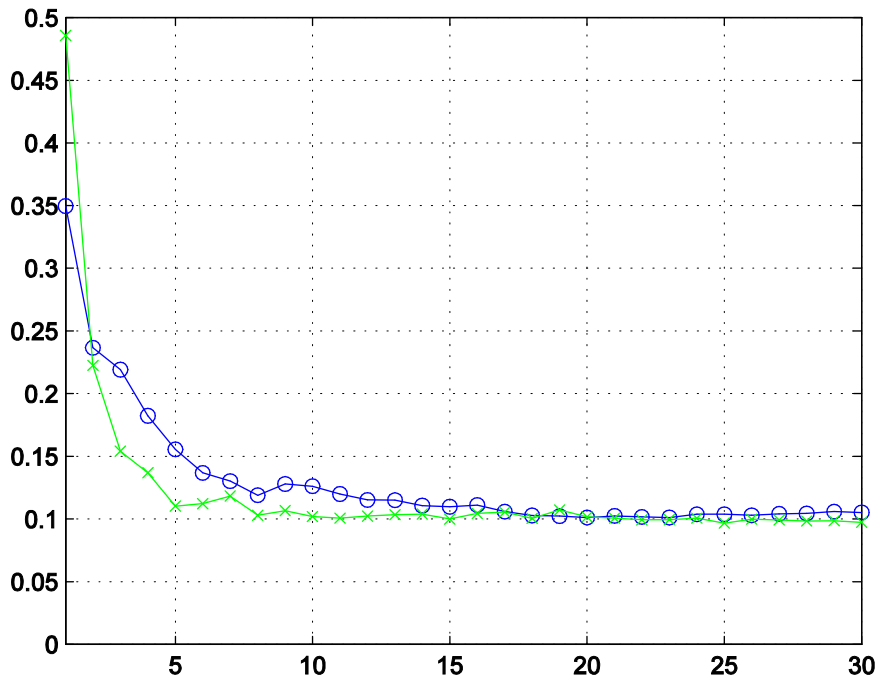


Figure 4.11: The RMSEs (cm) of test data for lip parameter  $B$  in function of the number of the ordered predictors  $F$ . The green line is corresponding to the GMM-based estimation model and the blue dotted line is corresponding to the MLR model.

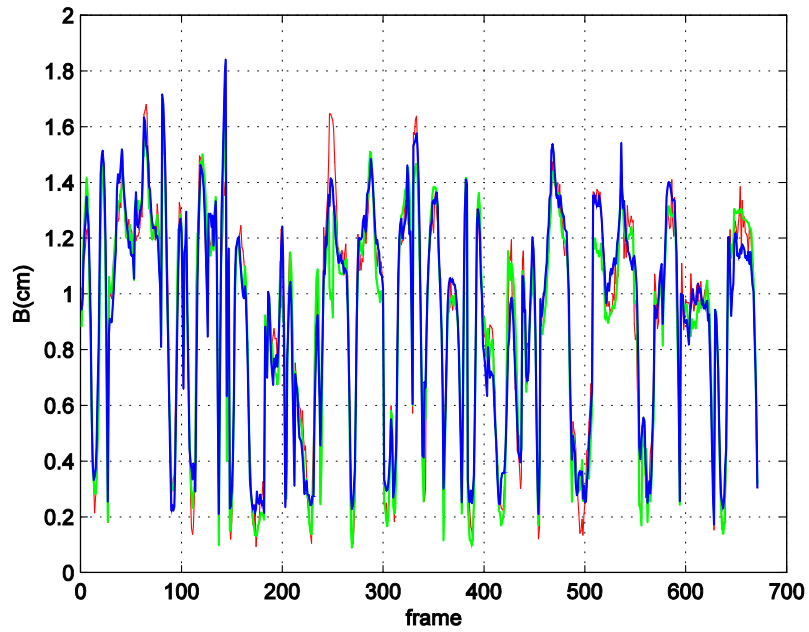


Figure 4.12: Estimated value of test data for lip parameter  $B$  given by the first 20 ordered predictor. The red line denotes the test data, the green line denotes the estimated value obtained by the GMM-based estimation model and the one obtained by the MLR model.

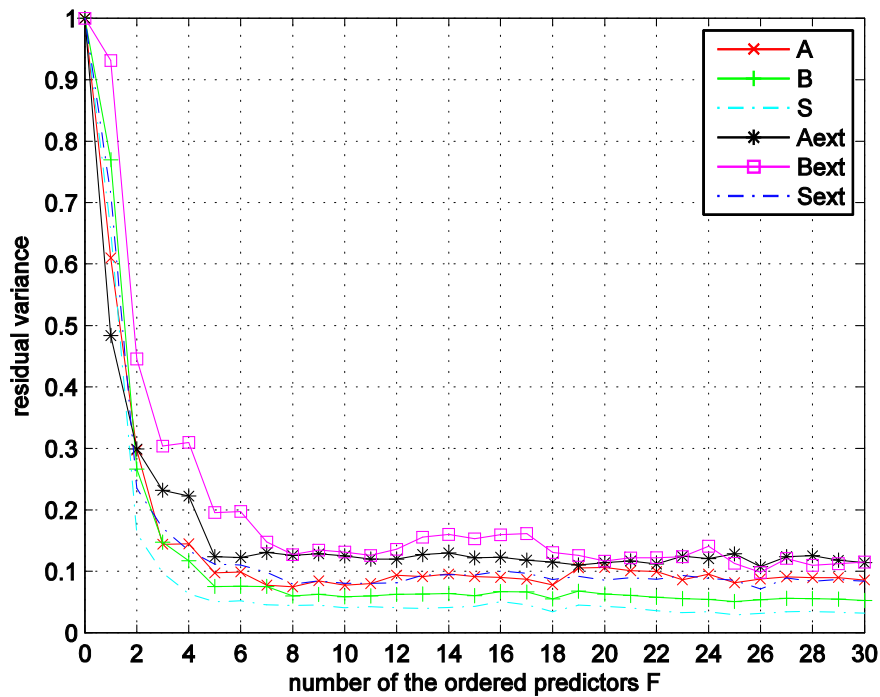


Figure 4.13: The RVARs of the test data for lip features ( $A$ ,  $B$ ,  $S$ ,  $A_{ext}$ ,  $B_{ext}$ ,  $S_{ext}$ ) in function of the number of the ordered predictors  $F$ .

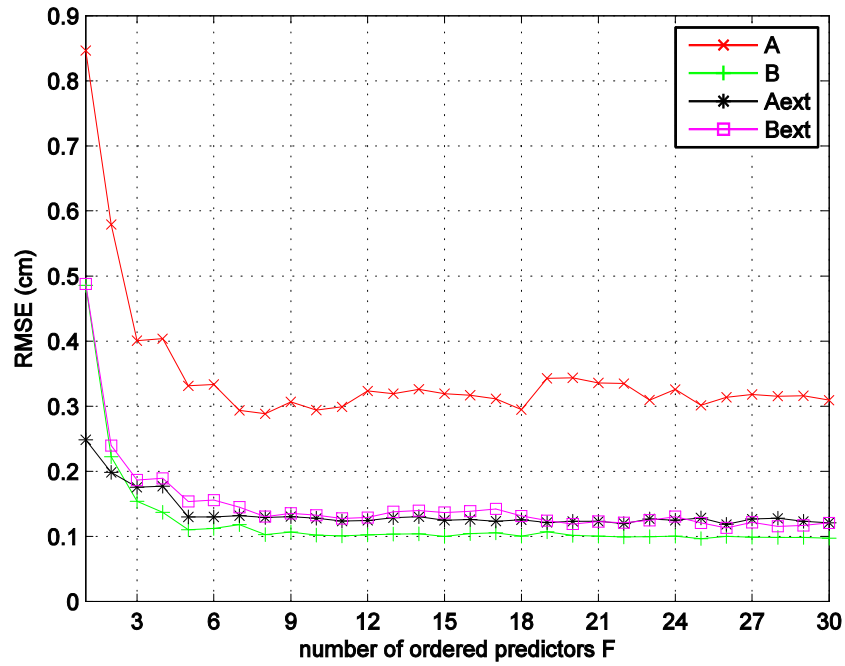


Figure 4.14: The RMSEs(cm) of the test data for lip features (A, B, Aext, Bext) in function of the number of the ordered predictors F.

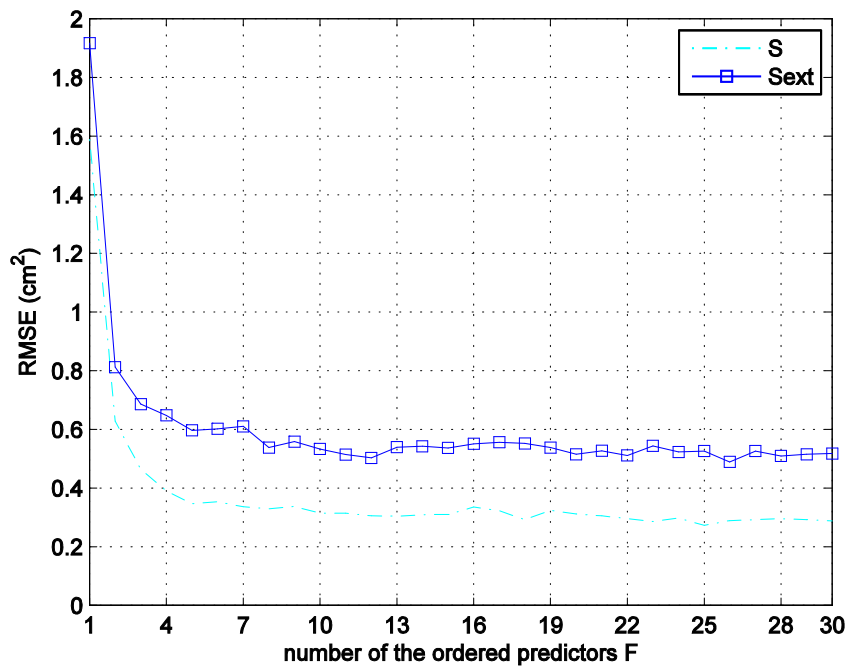


Figure 4.15: The RMSEs(cm<sup>2</sup>) of the test data for lip features (S, Sext) in function of the number of the ordered predictors F.

Besides training the GMM with 3 components (Gaussians), we also tried to train GMM composed by 4 and 5 mixture Gaussians separately. When we use more than 5 Gaussians, the training process is affected by the singularity problem. Although we used the variance limit to moderate the problem (Reynolds et al., 1995), the effect is not very obvious. In Figure 4.16 we can see that when we increase the number of the components from 3 to 5, the RMSE of the training set denoted by the red circle decrease, while the RMSE of test set are not in coherence with the one of training set. For the test set, the RMSE of the GMM with 3 mixture Gaussians is the lowest ones in comparison with the RMSE of the GMM with 4 or 5 Gaussians. That is to say, the generalization ability of GMM with 3 mixture Gaussians is best.

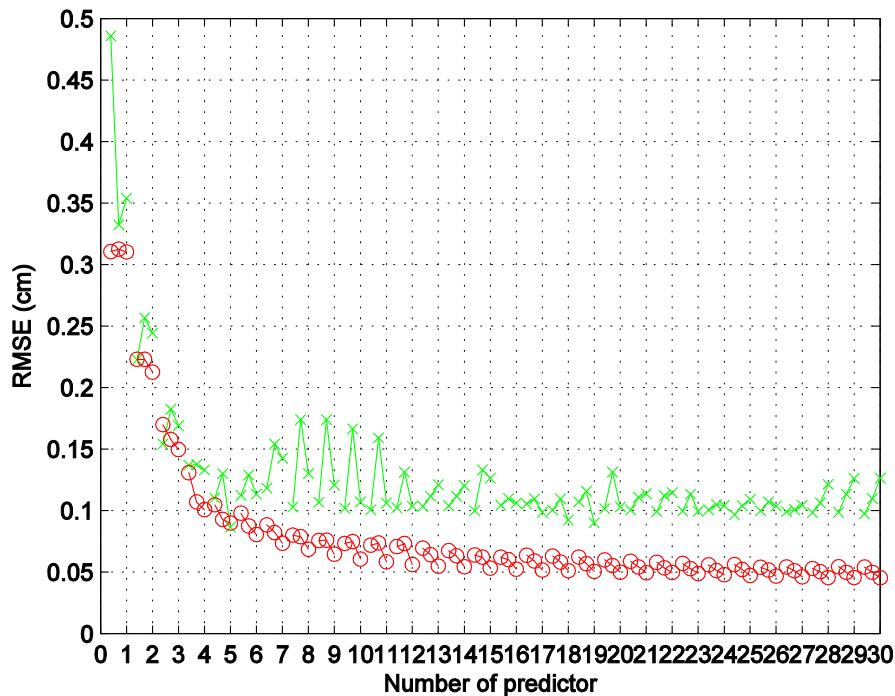


Figure 4.16: The RMSE of lip parameter  $B$  in function of number of the ordered predictors. The red ones are the RMSE of training data and the green ones are the RMSE of test data. From left to right within one set, the number of components is 3, 4 and 5.

As an elaborate model, the training processing of GMM is more complicated than the MLR model. One of the training problems is the over-fitting. We observed this problem in the test processing when the dimension of GMM is too high. See in Figure 4.17, when the dimension of GMM is more than 65, the RMSEs of the lip features ( $A$ ,  $B$ ,  $A_{ext}$ ,  $B_{ext}$ ) increase drastically. This is due to the model trying to fit the noise



in the training process when the dimension is very high, so that the outliers are produced when applying the trained model on the test data. However, the dimension of the model used in practice in our work (less than 10 normally) is far away from the one producing the over-fitting problem.

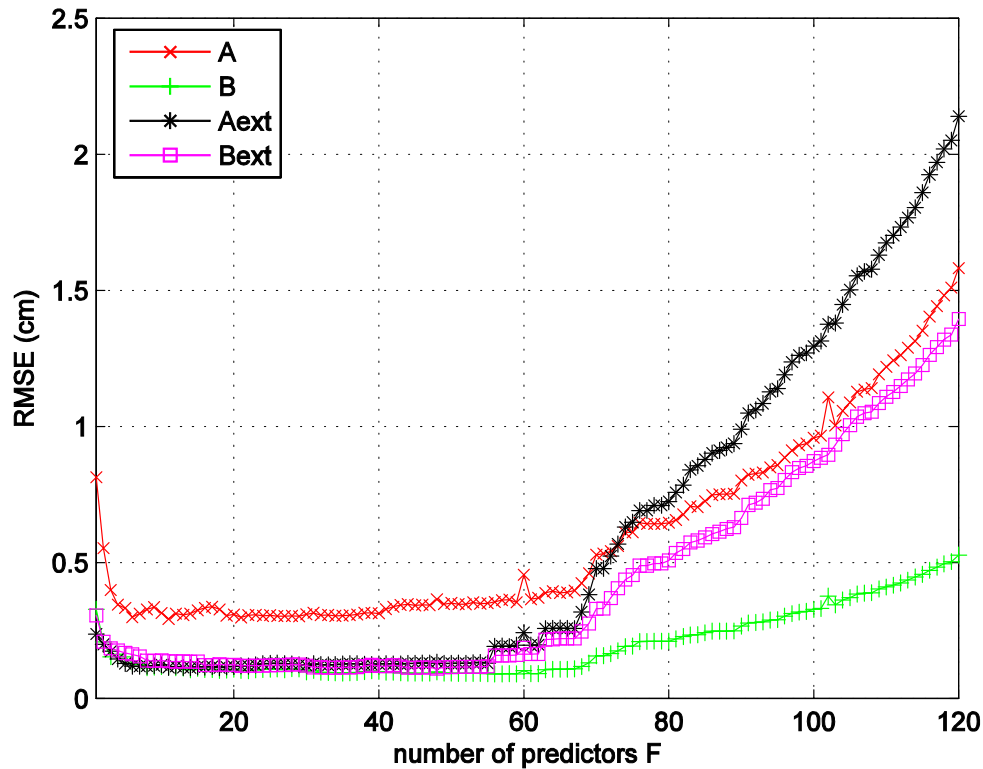


Figure 4.17: Over-fitting problem of GMM with 3 components. The RMSEs of the lip features (A, B, Aext, Bext) of test data are divergence drastically when the dimension of GMM is more than 65.

## 4.6. Summary

This chapter presents the two models, MLR model and GMM-based estimation model for the prediction of geometric lip features by concentrated information obtained by DCT coefficients of the natural image of mouth ROI. The PCA is also applied on DCT aiming to reduce the order of the predictors. With the MLR model, the order of predictors could be reduced to 17 or 21 accounted for just 2% of the 120 predictors derived from the DCT coefficients respectively explaining of 95% and 93% of total variance of the modelling and test data, as it has been observed in the case of lip

parameter  $B$  (see Figure 4.5 and Figure 4.6). But the MLR model still has some limitations when the target data is close to zero as shown in Figure 4.11. The GMM-based estimation model overcomes this problem in some extent that the estimated values are apparently closer to the target data as shown in the Figure 4.11. Further, the GMM-based estimation model is also more efficient than the MLR model (see in Figure 4.10), which only uses 8 predictors to converge in contrast to the 17 predictors of the MLR model. It is also proved that the GMM with 3 Gaussians has the best generalization ability in comparison with the GMM with 4 or 5 Gaussians. In addition, the high dimensional GMM-based estimation model is also affected by the over-fitting problem. However, the dimension of the GMM used in practice in this work (less than 10 normally) is far away from the one producing the over-fitting problem.



# Chapter 5. Speech to Cued Speech mapping: linear approach

## 5.1. Introduction

As introduced in the chapter 1, the performance of perception is best when lip-reading is associated with coding manual of CS. Thus the acoustic-to-CS mapping problem can be divided in two parts: acoustic-to-lip shape mapping and acoustic-to-hand parameter mapping (in our work the hand parameter is related to hand position representing the vowels in CS). In terms of the acoustic-to-lip shape mapping, the problem can be linked to the subject acoustic-to-articulatory (A-to-A) mapping since the lip shape is considered as a part of the articulatory parameters. But in terms of the acoustic-to-hand position mapping, it is a new problem. As the linear approach is the most researched approach and we want to know the limit in terms of performance of the approach, we first use MLR approach which was described in chapter 3 to estimate the lip shape or hand position. Since the hand position, which is used to disambiguate similar lip shapes, has no relation to the acoustic feature in essence, the linear approach does not quite suit for estimating the hand position. In order to establish a “linear correlation” between the hand position and acoustic spectrum, the intermediate space in which the hand position is relocated in order to simulate a similar topology structure of the acoustic space composed by formant1-formant2 (i.e. formant triangle) is introduced. Then the classification method is introduced for remapping the estimation results from the intermediate space to the original space. In the following sections of this chapter, we will first briefly present the MLR approach in our applications in section 5.2 and then evaluate the approach for estimating the lip parameters and hand position based on the database introduced in the chapter 2 in section 5.3 and 5.4 separately. In the following section 5.5, we extend the MLR approach to the intermediate space. Finally, we draw a summary in section 5.6.

## 5.2. Modelling

The objective of this work is to map the acoustic feature represented by the spectral parameters of speech signals to the lip shape and the hand position. As a start of the work, we briefly present the MLR approach combining with our applications.

### 5.2.1. Definition of the predictors

The theoretical description of the MLR approach was introduced in detail in the chapter 3. Here, we apply the approach in our experiment and define the parameters of the model. As mentioned in chapter 2, we used several kinds of spectral parameters of speech signal, 4 dimension formant coefficients, 16 dimension MFCCs, 16 dimension LSP coefficients and the concatenation of the MFCCs and LSP coefficients, to represent the acoustic feature respectively, meanwhile the lip width ( $A$ ), lip height ( $B$ ) and lip area ( $S$ ) of the inner lip contour are used to represent the lip shape. Thus the spectral parameters are the explanatory variables and the lip parameters or the coordinates ( $x, y$ ) of the hand position are the dependent variable or target to be predicted in the MLR model. In order to overcome the problem of collinearity of the explanatory variable, instead of using the spectral parameters to predict the target directly, the scores of principal components obtained by the PCA of the spectral parameters are used to estimate the model parameters and predict the target. Given the matrix  $\mathbf{F} = [F_1, F_2, \dots, F_p]$ ,  $F_p$  is the vector of scores of principal components,  $p$  is the number of selected principal components. Instead of the spectral parameters, the vectors  $[F_1, F_2, \dots, F_p]$  are chosen as the predictors to estimate the model parameters and predict the target.

### 5.2.2. The selection of the predictors

The PCA can compress the information into the first few principal components corresponding to the first highest variance to realize the dimensionality reduction. Thus when we use the scores of principal components  $F_i$  as the predictors in the MLR model, we can use the most informative predictors to estimate the target. As in the context of applying the linear model, we use the linear correlation between the target and the predictors as the criterion to select the predictors (Jolliffe, 1982). In practice,

the predictors are ordered by the square linear correlation coefficients  $\rho^2$ , then the first  $p$  predictors with the  $p$  largest  $\rho^2$  was selected.

### 5.2.3. The prediction

Here, the objective is to predict the target by the set of ordered predictors by MLR approach. Because the predictors are independent to each other, we can use an iterate method to estimate the regression coefficients of the model and then predict the target. First, the target (after being centered) is submitted to a linear regression with the first predictor  $F_1$ . The linear coefficient  $k_1$  is obtained as to minimize the residual error between the real values of the target and the predicted ones. The residual error is then submitted to a linear regression with the second predictor  $F_2$ . This procedure is thus successively applied to all the predictors. Finally, the target to be predicted, for example lip parameter  $B$  can be expressed as follows, where  $p$  is the number of predictors and  $\bar{B}$  is the mean value of the target  $B$ :

$$\hat{B} = k_1 F_1 + k_2 F_2 + \dots + k_p F_p + \bar{B} \quad (5.1)$$

$$\hat{B} = f(\mathbf{k}, \mathbf{F}) \quad (5.2)$$

$$\hat{\mathbf{k}} = (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{B} \quad (5.3)$$

where  $\mathbf{k} = [k_1, k_2, \dots, k_p]$  is the regression coefficients of the MLR model,  $\mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_p]$  are predictors derived from the PCA of the spectral parameters,  $\mathbf{B} = B - \bar{B}$  is the centered form of the lip parameter  $B$ .

## 5.3. Evaluation of lip feature prediction

We present the evaluation the prediction of lip feature in this section. The prediction of each parameter is independent to each other. As mentioned in chapter 2, cross-validation, sometimes called rotation estimation, is used as the evaluation method in our experiment. This method could estimate how accurately a predictive model will perform in practice especially where further samples are costly or difficult to collect (Kohavi, 1995). We apply the 5-fold cross-validation in our work in practice. The original sample is randomly partitioned into 5 subsamples. Of the 5 subsamples, a

single subsample is retained as the validation data for testing the model, and the remaining 4 subsamples are used as training data. The cross-validation process is then repeated 5 times (the *folds*), with each of the 5 subsamples used exactly once as the validation data. The 5 results from the folds then were averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once.

As same as the chapter 4, we used the residual variance (RVAR) and root-mean-square error (RMSE) as the evolution measures to evaluate the prediction efficiency and accuracy of the model. The RVAR is defined as follows:

$$RVAR = \frac{Var(\mathbf{y} - \hat{\mathbf{y}})}{Var(\mathbf{y})} \quad (5.4)$$

where  $Var(\mathbf{y} - \hat{\mathbf{y}})$  is the variance of the residual and  $Var(\mathbf{y})$  is the variance of the target data  $\mathbf{y}$ . The RMSE is defined as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (\mathbf{y}_t - \hat{\mathbf{y}}_t)^2} \quad (5.5)$$

where  $t$  denote the frame number and  $N$  is the total number of the frames,  $\mathbf{y}_t$  is the real value of the target and  $\hat{\mathbf{y}}_t$  is the estimated value of the target.

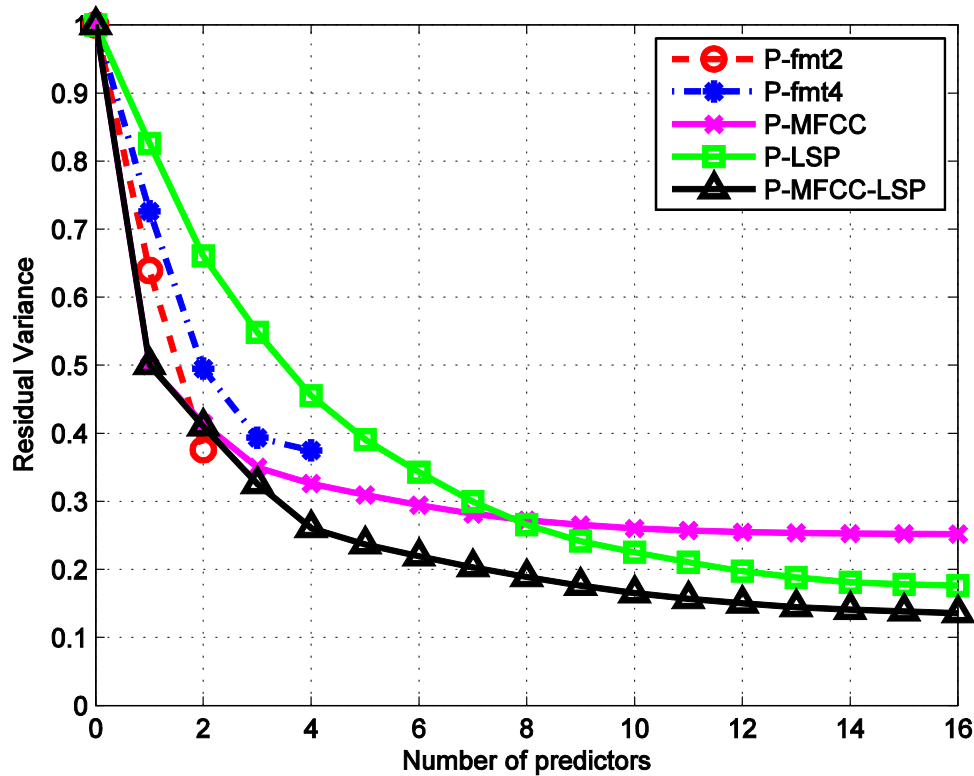


Figure 5.1: The average RVAR by 5-fold cross-validation of the training data for the lip parameter  $B$  in function of the number of predictors obtained by the different audio speech spectral parameters: 2 or 4 formant (fmt), MFCCs, LSP and mixture of MFCCs-LSP.

Figure 5.1 plots the average RVAR of the training data by 5-fold cross-validation for estimating the lip parameter  $B$  as a function of the number of selected predictors. The RVAR is calculated as the complement of the explained variance expressed in percentage. From the figure, it can be first observed that the RVAR decreases as the number of used predictors increases. One can then notice that the RVAR remains high with the use of formants (37% of the initial variance by using 4 predictors from formants). This is probably due to a lack of dimensions. Indeed the 16-dimension MFCCs and LSP coefficients significantly improve the performances of the prediction (a RVAR of 25% is obtained by using all 16 predictors from MFCCs meanwhile 18% for LSP). The MFCCs allow a quicker decrease while the LSP coefficients attain a lower RVAR. The prediction based on the mixture of the MFCCs and LSP has the advantage of the quick decrease property of the MFCCs and the low residual of the LSP (a RVAR of 14% by using first 16 predictors from the mixture of MFCCs and LSP). This mixture of MFCCs and LSP is thus considered as the best parameters for



this prediction. In terms of precision, using the first 16 predictors from the mixture of MFCCs and LSP the RMSE of  $B$  decreased to 1.55 mm which is relative low in comparison of the mean value of  $B$  (10.75 mm). The predictions of other lip parameters (A, S) have also been analyzed. Similar results have been obtained (see Table 5.1, Table 5.2).

Table 5.1: The RVARs of the training data for the lip parameters with the predictors obtained by 2 Formant (Formant1, Formant2), 4 Formant (Formant1-Formant4), MFCCs, LSP and mixture of MFCCs and LSP respectively.

| RVAR                       | Number of predictors | A   | B   | S   |
|----------------------------|----------------------|-----|-----|-----|
| Formant(Formant1,Formant2) | 2                    | 42% | 37% | 43% |
| Formant(Formant1-Formant4) | 4                    | 42% | 36% | 42% |
| MFCCs                      | 16                   | 27% | 26% | 28% |
| LSP                        | 16                   | 18% | 18% | 18% |
| MFCCs+LSP                  | 16                   | 16% | 14% | 15% |

Table 5.2: The RMSEs of the training data for the lip parameters with the predictors obtained by 2 Formant (Formant1, Formant2), 4 Formant (Formant1-Formant4), MFCCs, LSP and mixture of MFCCs and LSP respectively.

| RMSE(mm)                   | Number of predictors | A    | B    | S( $mm^2$ ) |
|----------------------------|----------------------|------|------|-------------|
| Formant(Formant1,Formant2) | 2                    | 8,70 | 2,53 | 11,17       |
| Formant(Formant1-Formant4) | 4                    | 8,69 | 2,52 | 11,12       |
| MFCCs                      | 16                   | 6,99 | 2,12 | 9,03        |
| LSP                        | 16                   | 5,74 | 1,76 | 7,31        |
| MFCCs+LSP                  | 16                   | 5,31 | 1,55 | 6,67        |

Figure 5.2 plots the average RVAR of test data by the 5-fold cross-validation. In the Figure 5.2, we can see that the test result is in coherence with the training result. The only difference between the training and test result is that the curves in test result (Figure 5.2) are not so smooth as in the training result (Figure 5.1). This is due to the selection order of the predictors learned by the training process probably being slightly different from the order obtained by the real test data. Table 5.3 and Table 5.4 demonstrate the RVARs and the RMSEs of the test data.

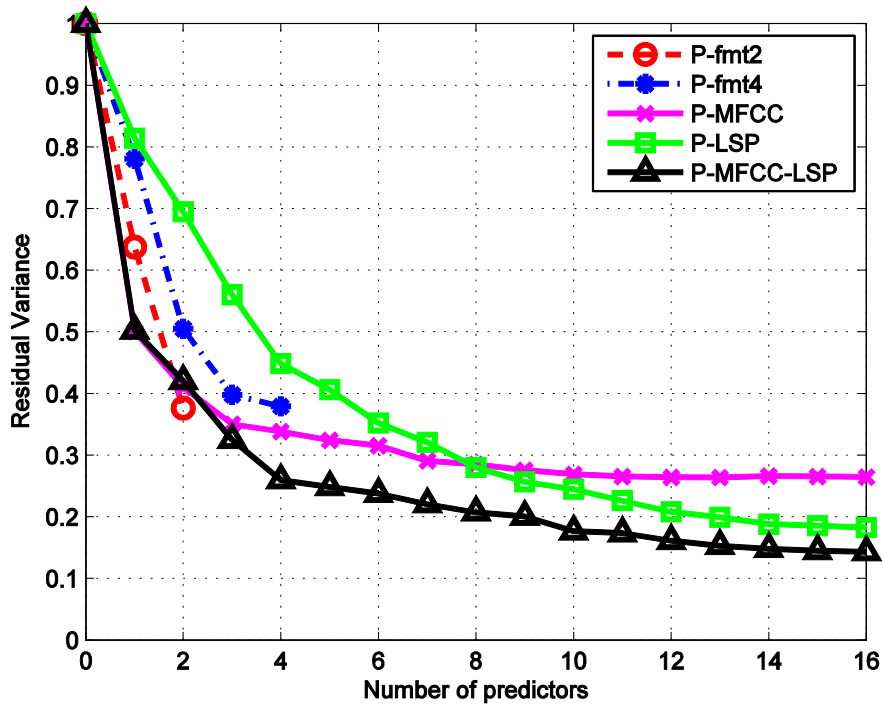


Figure 5.2 : Average RVAR by 5-fold cross-validation of the test data for the lip parameter  $B$  in function of the number of predictors obtained by the different speech spectral parameters: 2 or 4 formant ( $fmt$ ), MFCCs, LSP and MFCCs-LSP Mixture.

Table 5.3: The RVARs of the test data for the lip parameters with the predictors obtained by 2 Formant (Formant1, Formant2), 4 Formant (Formant1-Formant4), MFCCs, LSP and mixture of MFCCs and LSP respectively.

| RVAR                       | Number of predictors | A   | B   | S   |
|----------------------------|----------------------|-----|-----|-----|
| Formant(Formant1,Formant2) | 2                    | 52% | 41% | 48% |
| Formant(Formant1-Formant4) | 4                    | 52% | 42% | 49% |
| MFCCs                      | 16                   | 30% | 24% | 26% |
| LSP                        | 16                   | 20% | 17% | 19% |
| MFCCs+LSP                  | 16                   | 17% | 12% | 14% |

Table 5.4: The RMSEs of the test data for the lip parameters with the predictors obtained by 2 Formant (Formant1, Formant2), 4 Formant (Formant1-Formant4), MFCCs, LSP and mixture of MFCCs and LSP respectively.

| RMSE(mm)                   | Number of predictors | A    | B    | S(mm <sup>2</sup> ) |
|----------------------------|----------------------|------|------|---------------------|
| Formant(Formant1,Formant2) | 2                    | 9,96 | 2,66 | 11,95               |
| Formant(Formant1-Formant4) | 4                    | 9,99 | 2,67 | 12,07               |
| MFCCs                      | 16                   | 7,62 | 2,06 | 8,81                |
| LSP                        | 16                   | 6,28 | 1,75 | 7,47                |
| MFCCs+LSP                  | 16                   | 5,76 | 1,49 | 6,50                |

## 5.4. Evaluation of hand position prediction

The  $(x, y)$  coordinates of the hand position are considered as the hand features of CS. They are predicted by the audio speech spectral parameters. The final RVAR of the estimated value on training data is 39% for  $x$  and 29% for  $y$  with the best 16 predictors derived from the mixture of MFCCs and LSP (see Figure 5.3, Figure 5.4). And the RVAR of test data is also close to the one of training data, which is 43% for  $x$  and 31% for  $y$  predicted by the mixture of MFCCs and LSP. In terms of the RMSE, the training results by the 16 best predictors from the mixture of MFCCs and LSP is 2.85 cm for  $x$  and 2.67 cm for  $y$ . Figure 5.5 shows the estimated value of the hand coordinates  $x$  and the hand coordinates  $y$  in the training procedure by the 16 predictors derived from the mixture of MFCCs and LSP.

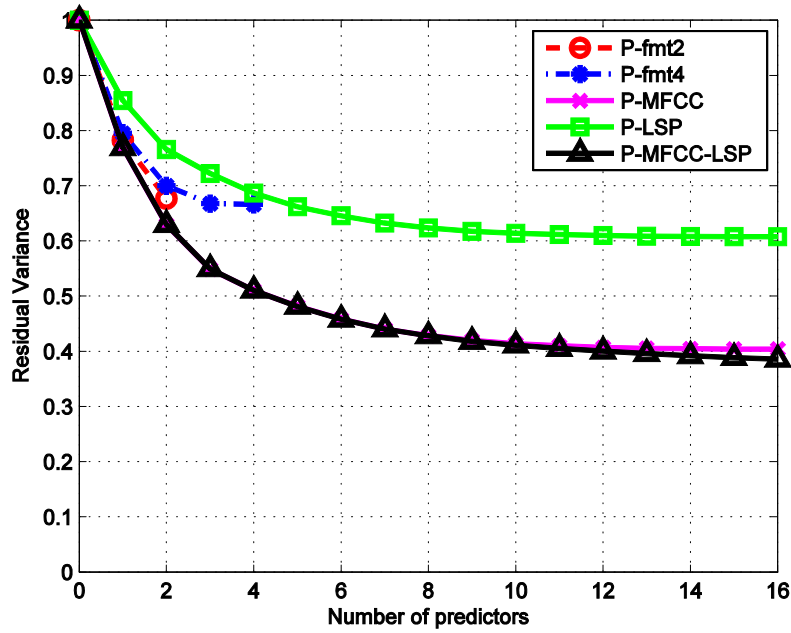


Figure 5.3: Average RVAR by 5-fold cross-validation of the training data for the hand coordinate  $x$  in function of the number of predictors obtained by the different audio speech spectral parameters: 2 or 4 formant (*fmt*), MFCCs, LSP and mixture of MFCCs-LSP.

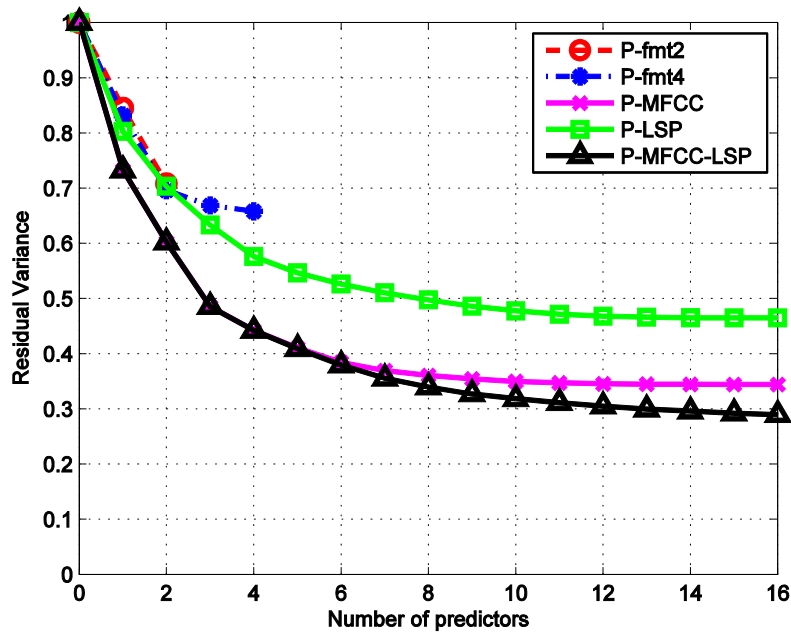


Figure 5.4: Average RVAR by 5-fold cross-validation of the training data for the hand coordinate  $y$  in function of the number of predictors obtained by the different audio speech spectral parameters: 2 or 4 formant (*fmt*), MFCCs, LSP and mixture of MFCCs-LSP.

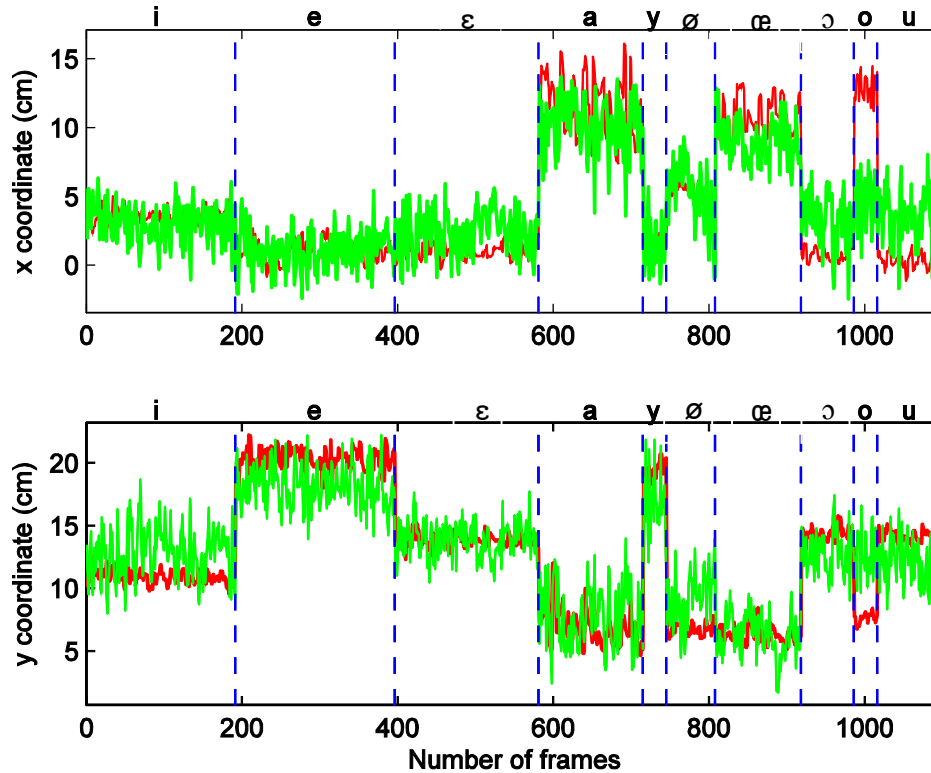


Figure 5.5: The estimated value of hand  $x$ ,  $y$  coordinates by the first 16 predictors derived from the mixture spectral parameters MFCCs and LSP on training data. The upper one is corresponding to the  $x$  coordinate and the lower one is corresponding to the  $y$  coordinate. The red line denotes the measured value and the solid green line denotes the estimated value. The vertical blue dash lines separate the hand positions belonging to different vowels.

In the Figure 5.5, we can see that the estimated values obviously do not follow the measure data well, especially in terms of the vowels [œ, ɔ, o, u] for  $x$  coordinate and vowels [e, ø, o] for  $y$  coordinate. For a more intuitive evaluation, we can plot the estimation hand position in the  $x$ - $y$  space in reference to the origin located in the centre of eyebrows (see Figure 5.6). Due to the poor estimation of the hand coordinates  $(x, y)$ , the distribution (in the green ellipses) of estimated values is obviously deviated from the original measured data (in the red ellipses). In comparison with the estimation of the lip parameters, the poor performance of the estimation of hand coordinates can be explained by the low level of the correlation values between the CS hand features and the audio spectral parameters. The maximum correlation coefficient (absolute value) between the  $x, y$  coordinates and

the first selected predictor from the mixture of MFCCs and LSP is 0.47 and 0.52 respectively in contrast to 0.70 between the lip parameters  $B$  and the predictor.

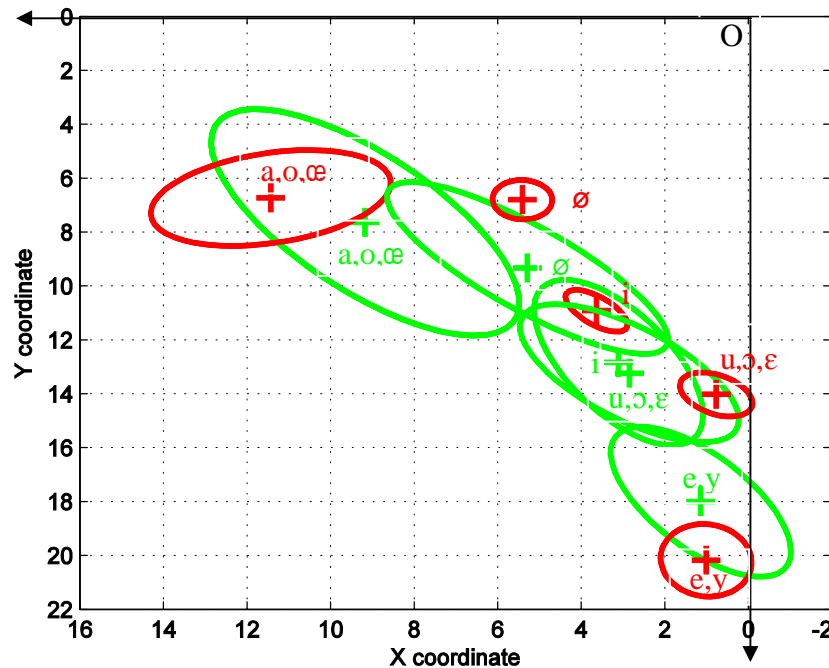


Figure 5.6: Estimation of hand position shown in the  $x$ - $y$  space, using 1.5 standard deviation ellipses to denote the distribution of estimated values (in green) and measured data (in red) for each CS hand position, the '+' are the centers of groups of the coordinates.

## 5.5. Prediction of the hand position in the intermediate space

### 5.5.1. Intermediate space

The lack of inherent correlation between the speech spectral parameters and the hand position affects the performance of the MLR for estimating the hand coordinates. Given the strong linear correlation between the formant values and the spectral parameters MFCCs or LSP (the two maximum correlation coefficients between Formant1, Formant2 and LSP are 0.97, 0.87 respectively, and the ones between Formant1, Formant2 and MFCCs are 0.82, 0.72 respectively), we introduce an intermediate space in which the original hand positions are translated to simulate the distribution of the formant values in the formant space (see Figure 5.7). The MLR approach is applied to estimate the new hand position in the intermediate space.

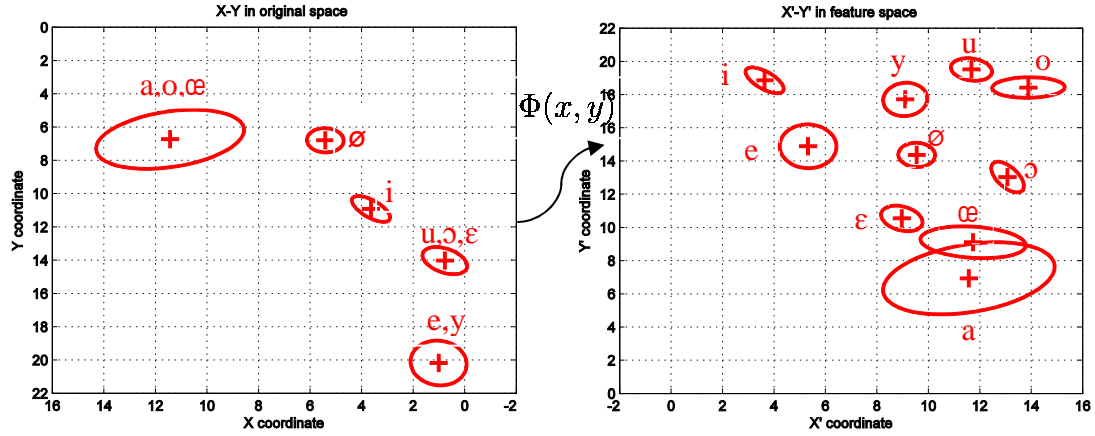


Figure 5.7: The left figure plots the hand position  $(x, y)$  (cm) in the original space and the right one shows the new coordinates  $(x', y')$  (cm) in the intermediate space in which the original hand positions are translated to simulate the distribution of the formant values in the formant space. The red ellipses ( $\text{std}=1.5$ ) denote the distribution of the coordinates, the '+' denotes the center of each group.

The translation function  $\Phi(x, y)$  is:

$$(x', y') = \Phi(x, y) = (x + \Delta x, y + \Delta y) = \mathbf{A} + \mathbf{v} \quad (5.6)$$

where  $(x', y')$  is the new coordinate in the intermediate space,  $\mathbf{A} = [x, y]^T$  is the original coordinates vector,  $\mathbf{v} = (\Delta x, \Delta y)^T$  is the shift vector that is varied depending on the different vowels (see Table 5.5). The shift vector  $\mathbf{v}$  is initialized manually according to the formant values (Formant1, Formant2) distribution in formant space and then it is optimized by the gradient decent method aiming to minimize the RVAR of the estimated value of hand coordinates in the intermediate space by the MLR approach.

Table 5.5: Shift vector  $\mathbf{v} = (\Delta x, \Delta y)^T$  (cm) which is varied depending on the different vowels.

| vowels     | a    | i    | u     | ø    | o     | ɔ     | y     | e     | ε     | œ    |
|------------|------|------|-------|------|-------|-------|-------|-------|-------|------|
| $\Delta x$ | 0,00 | 0,00 | 11,33 | 4,14 | 0,91  | 12,35 | 7,76  | 4,36  | 7,99  | 0,91 |
| $\Delta y$ | 0,00 | 7,94 | 5,29  | 7,57 | 10,80 | -1,16 | -1,75 | -5,40 | -3,34 | 2,86 |

After mapping the original hand coordinate to the intermediate space, we obtain the new hand coordinate and we could make a linear regression between the new hand coordinates in the intermediate space and the predictors derived from the spectral parameters.

### 5.5.2. Prediction procedure in intermediate space

In the intermediate space, we can use the predictors derived from PCA of the spectral parameters to predict the new hand coordinates by the MLR approach as mentioned in the previous section. After obtaining the estimation results in the intermediate space, the results needed to be remapped into the original space. However, due to the shift value in the translation function varying for different vowels, we need to determine the shift value of a new estimated hand coordinate in the intermediate space for remapping it to the original space. We turn to the classification method to achieve the purpose, which will classify the estimated coordinate into a group corresponding to a specific vowel so that we can obtain the associated shift value (see Figure 5.8).

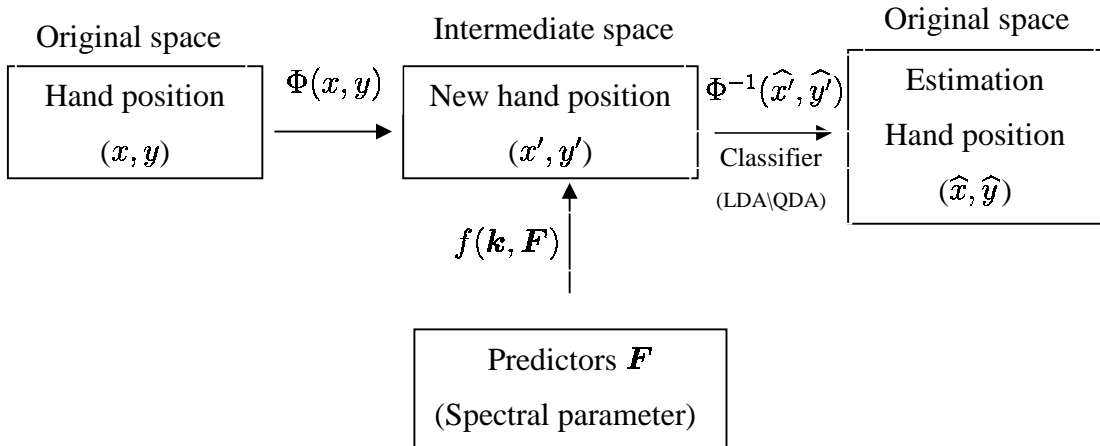


Figure 5.8: The estimation procedure in the intermediate space. The estimated values of the hand coordinates in the intermediate space are finally remapped to the original space.

In the Figure 5.8, the  $\Phi(x, y)$  is the translation function from the original space to the intermediate space,  $\Phi^{-1}(\hat{x}, \hat{y})$  is the inverse of the translation function, the  $f(\mathbf{k}, \mathbf{F})$  is the MLR function defined as equation (5. 2) used to predict the hand coordinates in the intermediate space. The classifier is used for classifying the estimated hand coordinates in the intermediate space to determine the shift values in the function



$\Phi^{-1}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$  for remapping the estimated coordinates to the original space. Thus the performance of the estimation is not only decided by the performance of the MLR approach but also by the classification accuracy of the classifier. In the sense of the minimum-error-rate classification, the classifier can be achieved easily and naturally by using the a posteriori probability of Bayes classifier as the discriminant function (Duda et al., 2000). The following equation is one of the discriminant functions in the sense of the minimum-error-rate classification derived from the a posteriori probability of Bayes.

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i) \quad (5.7)$$

where  $\mathbf{x}$  is a vector to be classified,  $g_i(\mathbf{x})$  is the discriminant function of the classifier,  $\omega_i$  denotes the class  $\omega_i$ ,  $p(\mathbf{x}|\omega_i)$  is the probability densities of class  $\omega_i$ ,  $P(\omega_i)$  is prior probability for the class  $\omega_i$ . The classifier is said to assign a feature vector  $\mathbf{x}$  to the class  $\omega_i$  if

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \text{for all } j \neq i \quad (5.8)$$

If the densities  $p(x|\omega_i)$  are multivariate and normal, i.e., if  $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ , the discriminant function  $g_i(\mathbf{x})$  is

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \quad (5.9)$$

Since the  $(d/2) \ln 2\pi$  term in the equation above is independent of  $i$ , it can be ignored by the superfluous additive constant. The resulting discriminant function is

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \quad (5.10)$$

In the resulting discriminant function (equation (5. 10)), two categories of classification can be derived from the two different cases of the covariance  $\boldsymbol{\Sigma}_i$ : linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) (Hastie et al., 2005).

(1) Linear discriminant analysis (LDA)

The first case is a simplification of the covariance matrices  $\Sigma_i$  in the way that covariance for all of the classes is identical, that is  $\Sigma_i = \Sigma$ . Thus the term  $\Sigma_i$  in equation (5. 10) is independent of  $i$ , it can be ignored as superfluous additive constant. This simplification leads equation (5. 10) to the new discriminant functions:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + \ln P(\omega_i) \quad (5. 11)$$

Expansion of the quadratic form  $(\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$  results in a sum involving a quadratic term  $\mathbf{x}^t \Sigma^{-1} \mathbf{x}$  which here is independent of  $i$ . After removing this term from equation (5. 11), the resulting discriminant function is a linear discriminant function:

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0} \quad (5. 12)$$

where

$$\mathbf{w}_i = \Sigma^{-1} \boldsymbol{\mu}_i \quad (5. 13)$$

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^t \Sigma^{-1} \boldsymbol{\mu}_i + \ln P(\omega_i) \quad (5. 14)$$

We can infer from the equation (5. 12) that the decision surface between any two of the classes is the plane so called ‘hyperplane’ which is orthogonal to the direction of  $\mathbf{w}_i$  and intersects the line connecting the means  $\boldsymbol{\mu}_i, \boldsymbol{\mu}_j$  of the two classes (Duda, 2000). The decision rule to define the specific class  $i$  to which  $\mathbf{x}$  will be assigned is the maximum value of the discriminant function  $g_i(\mathbf{x})$  shown in equation (5. 15).

In practice, we need the training data to estimate the parameters of the Gaussian distributions(Hastie et al., 2005):

- $\hat{\pi}_k = P(\omega_k) = N_k/N$ , where  $N_k$  is the number of the class- $k$  observation;
- $\hat{\boldsymbol{\mu}}_k = \Sigma_{g_i=k} \mathbf{x}_i / N_k$ , where  $g_i$  means the class of the observation  $\mathbf{x}_i$ ,  $g_i = k$  means  $\mathbf{x}_i$  is belonging to class- $k$ ;

- $\widehat{\Sigma} = \sum_{k=1}^K (N_k - 1) \widehat{\Sigma}_k / (N - K) = \sum_{k=1}^K \sum_{g_i=k} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^T / (N - K)$ , where  $\widehat{\Sigma}$  is the pooled estimated covariance,  $\widehat{\Sigma}_k$  is covariance matrix of class- $k$  (Thomaz et al., 2000).

(2) Quadratic discriminant analysis (QDA)

In the general multivariate normal case, the covariance matrices are different for each class. The covariance matrices  $\boldsymbol{\Sigma}_i$  for all of the classes are arbitrary. So the discriminant function is same to the equation (5. 10), the discriminant function is quadratic:

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0} \quad (5. 15)$$

where

$$\mathbf{W}_i = -\frac{1}{2} \boldsymbol{\Sigma}_i^{-1} \quad (5. 16)$$

$$\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i \quad (5. 17)$$

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_0) \quad (5. 18)$$

We can see that the discriminant function (equation (5. 15)) is not linear but quadratic, so that the decision surfaces will be the ‘hyperquadratics’ as the general forms of the hyperplanes, hyperspheres, hyperellipsoids and etc. The decision rule to define the specific class  $i$  to which  $\mathbf{x}$  will be assigned is the maximum value of the discriminant function  $g_i(\mathbf{x})$  shown in equation (5. 15).

As it uses the arbitrary covariance matrices for each class instead of the common covariance matrix, the QDA is the preferred approach and is a convenient substitute of the LDA method. But the differences are generally small. However, since the decision boundaries are the function of the parameters of the densities, the computation of the QDA will dramatically increase when the dimension of the model is high.

### 5.5.3. Evaluation of the hand position prediction in the intermediate space

As described in the Figure 5.8, the prediction procedure begins with mapping the hand coordinate to the intermediate space from the original space (see Figure 5.9), and then uses the predictors from the PCA of the spectral parameters to predict the new hand coordinates in the intermediate space by the MLR approach.

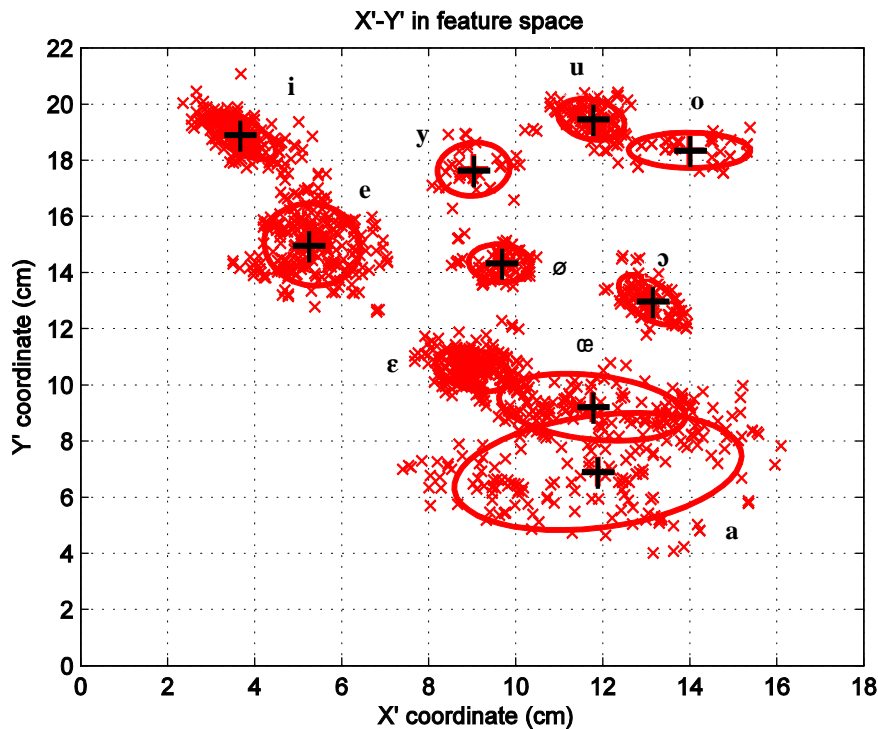


Figure 5.9: The new hand coordinates (red crosses) in the intermediate space. The red ellipses ( $std=1.5$ ) denote the distributions of the hand coordinates corresponding to each vowel and the black '+' denotes the center of each group corresponding to each vowel.

The 5-fold cross-validation is still used for the evaluation in the intermediate space. The RVAR and RMSE are used as two criteria of the evaluation. Figure 5.10 and Figure 5.11 show the average RVARs of the estimated value of the hand coordinates X and Y on the training data.

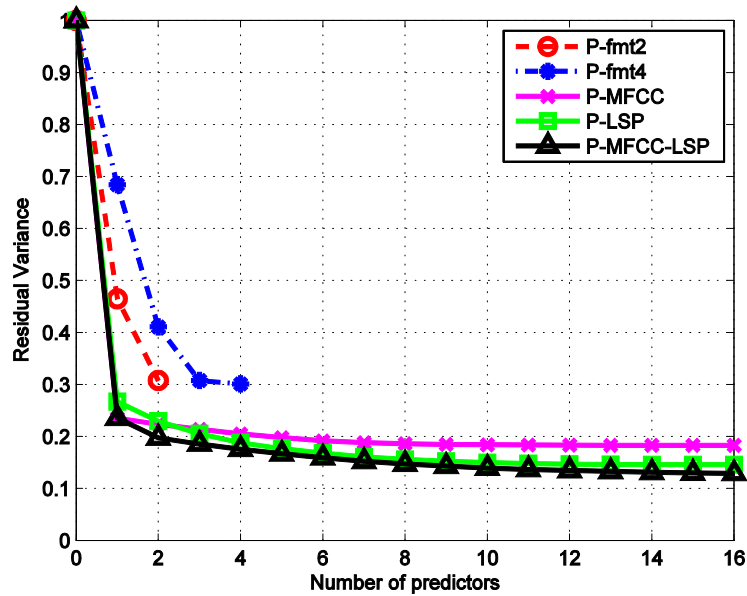


Figure 5.10: Average RVAR in the intermediate space of the training data for the hand coordinate  $X$  in function of the number of predictors obtained by the different audio speech spectral parameters: 2 or 4 formant (fmt), MFCCs, LSP and mixture of MFCCs-LSP.

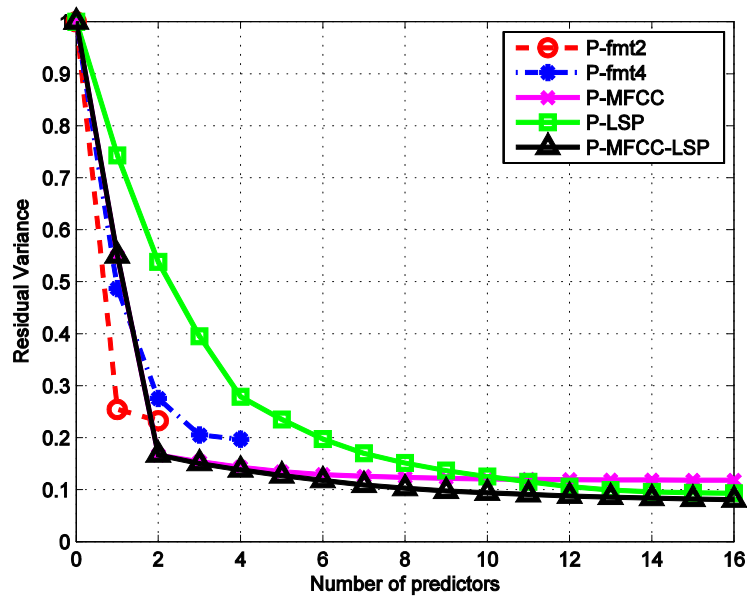


Figure 5.11: Average RVAR in the intermediate space of the training data for the hand coordinate  $Y$  in function of the number of predictors obtained by the different audio speech spectral parameters: 2 or 4 formant (fmt), MFCCs, LSP and mixture of MFCCs-LSP.

In comparison with the Figure 5.3, Figure 5.4, we can see that the average RVARs obtained in the intermediate space shown in the Figure 5.10, Figure 5.11 decrease

significantly for all 5 types of predictors. For example, in terms of the prediction of the X coordinate, the average RVAR obtained by the best 16 predictors from the mixture of the MFCCs-LSP decreased from 43% in Figure 5.3 to 13% in Figure 5.10 and the corresponding RMSE also decreased from 2.83 cm to 1.27 cm. Meanwhile for the Y coordinate, the average RVAR decreased from 29% in Figure 5.4 to 7% in Figure 5.11 and the corresponding RMSE also decreased from 2.68 cm to 1.22 cm. As in the original space, the predictors from the mixture of MFCCs-LSP inherit the advantages of the MFCCs and LSP i.e. decreasing quicker and lower. Thus the predictors derived from the mixture of MFCCs-LSP are chosen as the best predictors to estimate the hand coordinates in the intermediate space (see Figure 5.12).

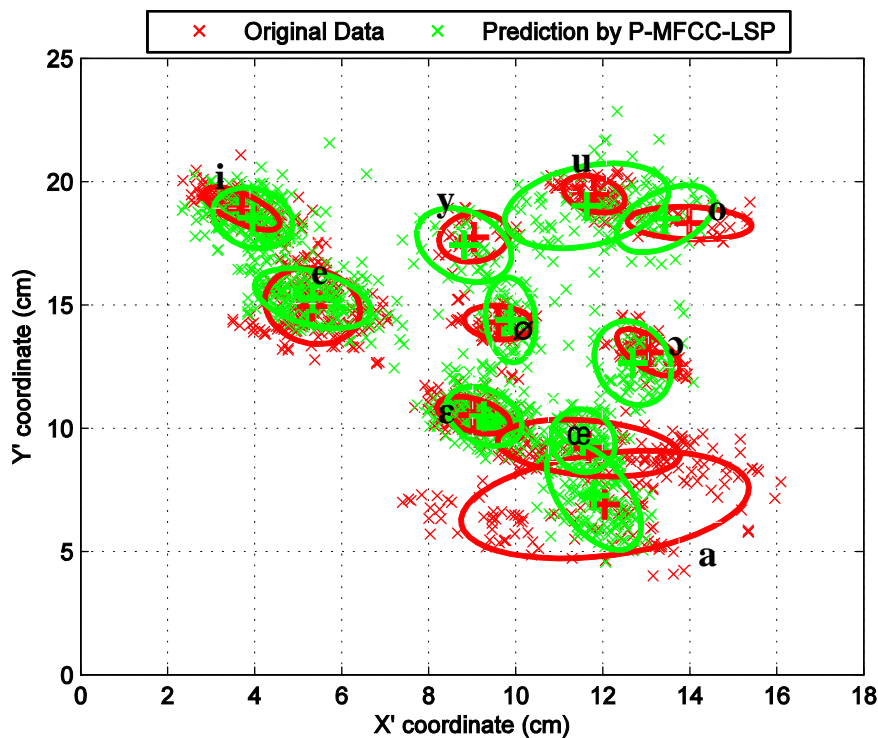


Figure 5.12: The hand coordinates (red crosses) and the estimated values (green crosses) in the intermediate space on the training data. The coordinates were estimated by 16 predictors from the mixture of MFCCs and LSP. The red ellipses ( $std=1.5$ ) denote the distributions of data and the '+' denotes the center of each ellipse.

Figure 5.12 shows that the target data and estimated values are close in comparison to the ones shown in the Figure 5.6. It indicates that the MLR approach is effective for estimating the hand position in the intermediate space. Figure 5.13 shows the

examples of the hand coordinates X and Y estimated by the 16 predictors of the mixture of MFCCs and LSP on the training data.

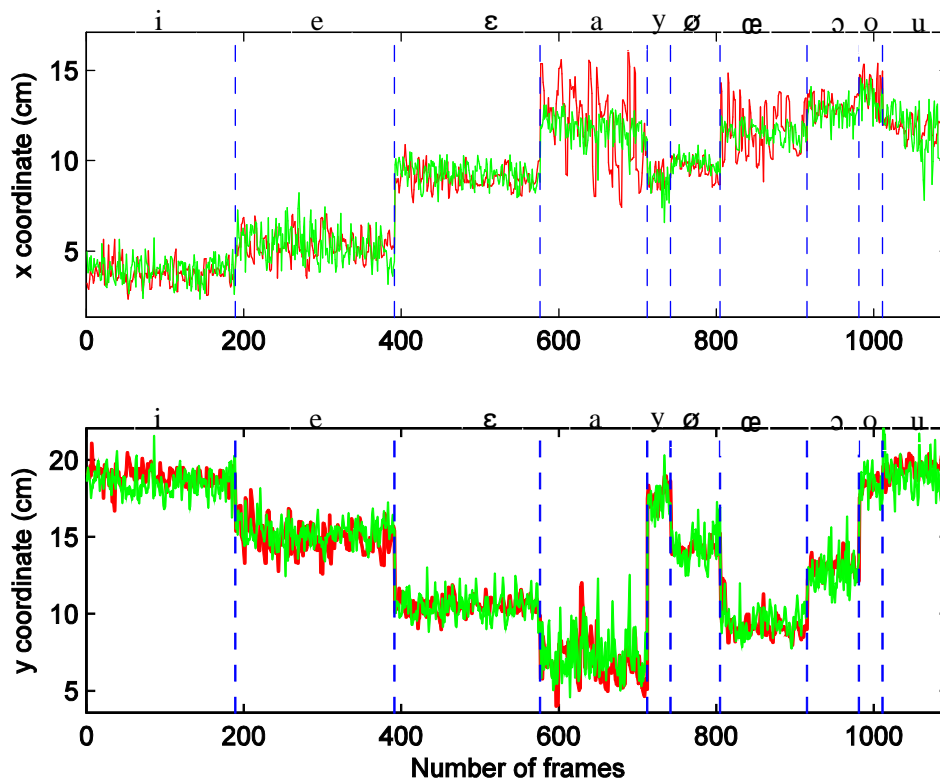


Figure 5.13: The estimated value of hand  $x$ ,  $y$  coordinates by the first 16 predictors derived from the mixture spectral parameters MFCCs and LSP on training data in the intermediate spaces. The upper one is corresponding to the  $x$  coordinate and the lower one is corresponding to the  $y$  coordinate. The red solid line denotes the measured value and the green solid line denotes the estimated value. The vertical blue dash lines separate the groups of coordinates belonging to the different vowels.

The training data used in the Figure 5.13 is same to in the Figure 5.5 which shows the estimated value of hand coordinates in the original space. We can see that the estimation performance in Figure 5.13 is improved significantly in comparison with the ones in the original space especially for the vowels [ɔ,u,o]. The estimated value shown in Figure 5.5 obviously deviated from the reference data in contrast to the one in intermediate space shown in Figure 5.13. The RMSE for the X and Y in Figure 5.5 are 2.84 cm and 2.66 cm while the ones shown in Figure 5.13 are 1.26 cm and 1.18

cm. The average RVARs and RMSEs of the X and Y coordinates estimated by the different predictors in intermediate space on training data and test data are shown in the Table 5.6 and Table 5.7 respectively. The significant improvement also proves that the similar topology structure between the acoustic space and hand position space is the key to the estimation performance of the MLR approach.

Table 5.6: The average RVARs and RMSEs of the X and Y coordinates estimated by the different predictors in intermediate space on training data.

| RVAR                       | Number of predictors | X   | Y   |
|----------------------------|----------------------|-----|-----|
| Formant(Formant1,Formant2) | 2                    | 31% | 23% |
| Formant(Formant1-Formant4) | 4                    | 30% | 20% |
| MFCCs                      | 16                   | 18% | 12% |
| LSP                        | 16                   | 15% | 9%  |
| MFCCs+LSP                  | 16                   | 13% | 7%  |

| RMSE(cm)                   | Number of predictors | X    | Y    |
|----------------------------|----------------------|------|------|
| Formant(Formant1,Formant2) | 2                    | 1,95 | 2,08 |
| Formant(Formant1-Formant4) | 4                    | 1,93 | 1,91 |
| MFCCs                      | 16                   | 1,51 | 1,48 |
| LSP                        | 16                   | 1,35 | 1,31 |
| MFCCs+LSP                  | 16                   | 1,27 | 1,22 |

Table 5.7: The average RVARs and RMSEs of the X and Y coordinates estimated by the different predictors in intermediate space on test data.

| RVAR                       | Number of predictors | X   | Y   |
|----------------------------|----------------------|-----|-----|
| Formant(Formant1,Formant2) | 2                    | 29% | 22% |
| Formant(Formant1-Formant4) | 4                    | 29% | 19% |
| MFCCs                      | 16                   | 20% | 12% |
| LSP                        | 16                   | 15% | 10% |
| MFCCs+LSP                  | 16                   | 14% | 8%  |



| RMSE(cm)                   | Number of predictors | X    | Y    |
|----------------------------|----------------------|------|------|
| Formant(Formant1,Formant2) | 2                    | 1,86 | 2,04 |
| Formant(Formant1-Formant4) | 4                    | 1,85 | 1,88 |
| MFCCs                      | 16                   | 1,56 | 1,51 |
| LSP                        | 16                   | 1,38 | 1,37 |
| MFCCs+LSP                  | 16                   | 1,35 | 1,26 |

#### 5.5.4. Remapping hand position to the original space

In the above section we can see that the performance of MLR for estimating the hand position in the intermediate space improves significantly in comparison with the one in the original space. But keep in mind that the estimated coordinates in the intermediate space are not the ones what we want in the end. The estimated values of the hand coordinates in intermediate space are not the real positions in the CS. Therefore we need to remap the estimated coordinates from the intermediate space to the original space where the hand position can be used to indicate vowels in CS.

As mentioned in the section 5.5.2, the shift value in the inverse of the translation function  $\Phi^{-1}(\hat{x}, \hat{y})$  varies for different vowels. Hence we need to determine the shift value of a new estimated hand position in the intermediate space for remapping it to the original space. For that, we turn to the Bayes classifier LDA and QDA. We use the estimated coordinates  $(\hat{x}, \hat{y})$  in the intermediate space to train 10 different classes corresponding to the 10 different vowels. Then we can obtain the shift value of a new estimated hand position by classifying it into a class. The LDA and QDA have been elaborated in the section 5.5.2. Here we evaluate the two classification methods separately and then present the remapping results in the original space. The 5-fold cross-validation is also used in this evaluation. In addition, the estimated coordinates are obtained by the predictors derived from the mixture of MFCCs and LSP which are the best predictors.

##### (1) Evaluation of LDA

Before evaluation the classification method LDA, we need to train the parameters of the model: prior probability  $\hat{\pi}$ , mean vector  $\hat{\mu}$  and covariance  $\hat{\Sigma}$  first.

As mentioned in section 5.5.2:

- $\hat{\pi}_k = P(\omega_k) = N_k/N,$

where  $\hat{\pi}_k$  is prior probability of class- $k$ ,  $N_k$  is the number of the observations of class- $k$ ,  $N$  is the total number of frames of the training data;

- $\hat{\boldsymbol{\mu}}_k = \sum_{g_i=k} \mathbf{x}_i / N_k,$

where  $\hat{\boldsymbol{\mu}}_k$  is prior probability of class- $k$ ,  $\mathbf{x}_i = (X, Y)_i$  is the coordinate vector  $\mathbf{x}$  at  $i$ th frame and belonging to class- $k$ ,  $N_k$  is the number of the observations of class- $k$ ;

- $\hat{\boldsymbol{\Sigma}} = \sum_{k=1}^K (N_k - 1) \hat{\boldsymbol{\Sigma}}_k / (N - K) = \sum_{k=1}^K \sum_{g_i=k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T / (N - K),$

where  $\hat{\boldsymbol{\Sigma}}$  is the identical pooled estimated covariance,  $\hat{\boldsymbol{\Sigma}}_k$  is covariance matrix of class- $k$ ,  $K$  is the number of the categories of classes, in our case  $K = 10$ .

Figure 5.14 shows the classification of the training data by LDA. In the Figure 5.14, the symbol ‘square’ shows the misclassification of LDA. We can see that the squares are mainly located at the boundaries of the adjacent groups, for example, between the vowel [i] and [e] or [u] and [o], and so on. In these areas, it is difficult to classify the estimated hand positions due to the fact that the individuals of different groups are mixed up. Moreover, we do not have any information from other dimensions rather than the two-dimension coordinates of the hand position for helping to classify. In addition, due to the limited estimation performance of the MLR, some estimated values of hand positions of vowel [u], which are misclassified as the ones of vowel [y], are extended to the inside of the group of vowel [y]. Table 5.8 shows the average score of the LDA classification of the training data and test data respectively with the 5-fold cross validation:

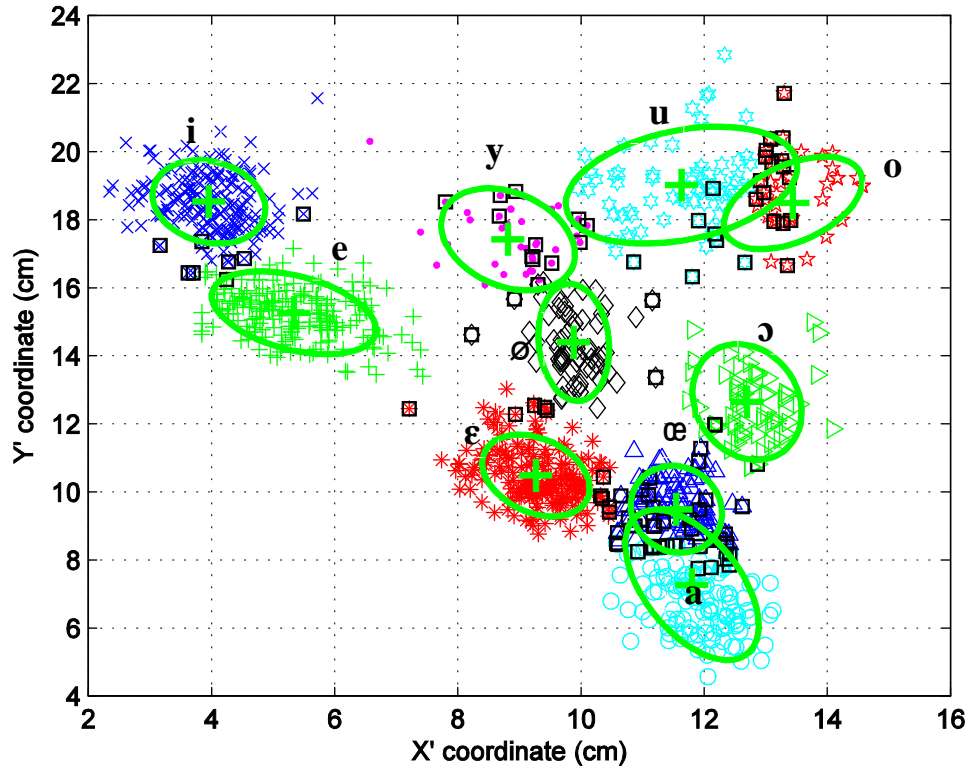


Figure 5.14: The classification results of estimated hand coordinates by LDA in intermediate space on training data. The symbol 'square' shows the misclassification. The coordinates were estimated by 16 predictors from the mixture of MFCCs and LSP. The green ellipses (std=1.5) denote the distributions of the estimated hand coordinates. Each class of estimated hand coordinates is labeled by different symbols.

Table 5.8: The average score of the LDA classification on the training and test data with the 5-fold cross-validation.

|          | Score1 | Score2 | Score3 | Score4 | Score5 | score |
|----------|--------|--------|--------|--------|--------|-------|
| Training | 0.90   | 0.90   | 0.90   | 0.91   | 0.93   | 0.91  |
| Test     | 0,88   | 0,88   | 0,89   | 0,87   | 0,84   | 0.87  |

## (2) Evaluation of QDA

As same as the LDA, we need to train firstly the prior probability  $\hat{\pi}$  and mean vector  $\hat{\boldsymbol{\mu}}$  in QDA. While for the covariance matrix, we trained it as follows:

- $\hat{\Sigma}_k = \sum_{g_i=k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T / (N_k - 1),$

where  $\hat{\Sigma}_k$  is covariance matrix of class- $k$ ,  $K$  is the number of the categories of classes, in our case  $K = 10$ ,  $\mathbf{x}_i = (X, Y)_i$  is the coordinate vector  $\mathbf{x}$  at  $i$ th frame and belonging to class- $k$ .

Since using the arbitrary covariance matrix  $\hat{\Sigma}_k$  instead of the common covariance matrix  $\hat{\Sigma}$  in the discriminant function, the discriminant function of QDA should be more precise for classifying the individuals than the LDA. Figure 5.15 shows the classification result by QDA on training data. The training data used in Figure 5.15 is same as the one used in the Figure 5.14, so we can compare the classification performance of the LDA and QDA by the two figures.

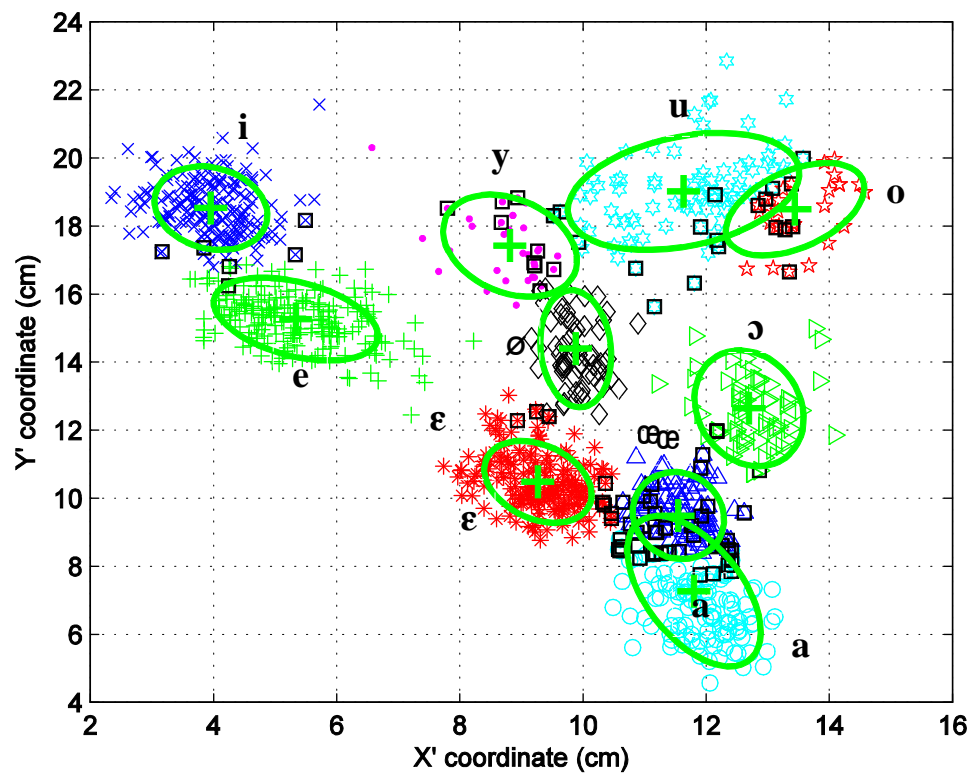


Figure 5.15: The classification results of estimated hand coordinates by QDA in intermediate space on training data. The symbol 'square' shows the misclassification. The coordinates were estimated by 16 predictors from the mixture of MFCCs and LSP. The green ellipses ( $std=1.5$ ) denote the distributions of the estimated hand coordinates. Each class of estimated hand coordinates is labeled by different symbols.

In Figure 5.15 we can see that the misclassification is less especially at the boundaries between vowel [i] and [e] or [u] and [o]. This indicates that the QDA performs better than LDA especially at boundaries of the adjacent groups. The misclassifications in the group of vowel [y] still exist due to the limited estimation performance of the MLR which results in the estimated values of the coordinates of vowel [u] extending to the area of group of vowel [y].

*Table 5.9: The average score of the QDA classification on the training and test data with the 5-fold cross-validation.*

|          | Score1 | Score2 | Score3 | Score4 | Score5 | score |
|----------|--------|--------|--------|--------|--------|-------|
| Training | 0,93   | 0,93   | 0,90   | 0,92   | 0,93   | 0,92  |
| Test     | 0,89   | 0,89   | 0,93   | 0,91   | 0,85   | 0,90  |

Table 5.9 shows that the average classification score of QDA is better than the LDA, thus we chose the QDA as the classifier. Then we can obtain the shift value of an estimated hand position and then use the inverse translation function to remap the estimated hand position from the intermediate space to the original space.

$$(x, y) = \Phi^{-1}(x', y') = (x' - \Delta x, y' - \Delta y) = \mathbf{A}' - \mathbf{v} \quad (5.19)$$

where  $(x, y)$  are the coordinates in the original space,  $\mathbf{A}' = [x', y']^T$  is the coordinates vector estimated in intermediate space,  $\mathbf{v} = (\Delta x, \Delta y)^T$  is the shift vector that varies for the different vowels defined in Table 5.5. Figure 5.16 shows the remapping results in the original space by using the QDA classifier.

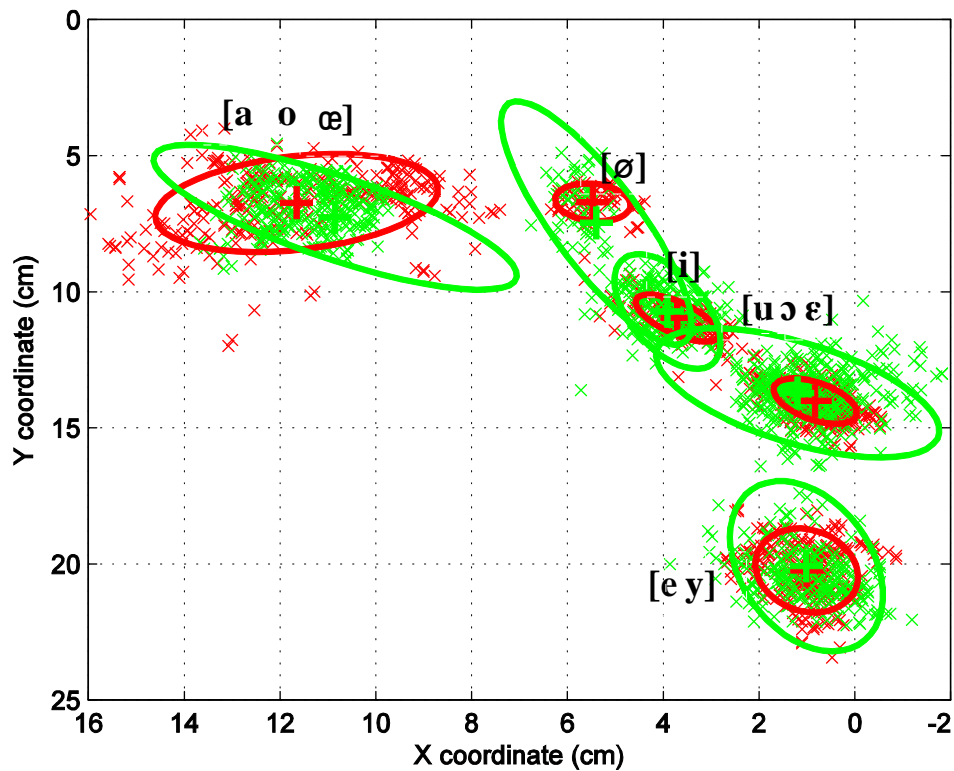


Figure 5.16: Remapping estimated hand position from the intermediate space to the original space. Using 1.5 standard deviation ellipses to denote the distribution of estimated values (in green) and measured data (in red) for each CS hand location. The '+' is the center of each group of the coordinates.

The remapping results shown in the Figure 5.16 are much better than the ones in Figure 5.6. The estimated coordinates are closer to the reference ones. However, some estimated values of the vowel [ø] and vowel [u ɔ ε] are still far from the reference ones. Generally speaking, the remapping estimated coordinates are located close to the reference coordinates in the original space. More precisely, we can see the Figure 5.17 which shows the remapping coordinates in function of the frames.

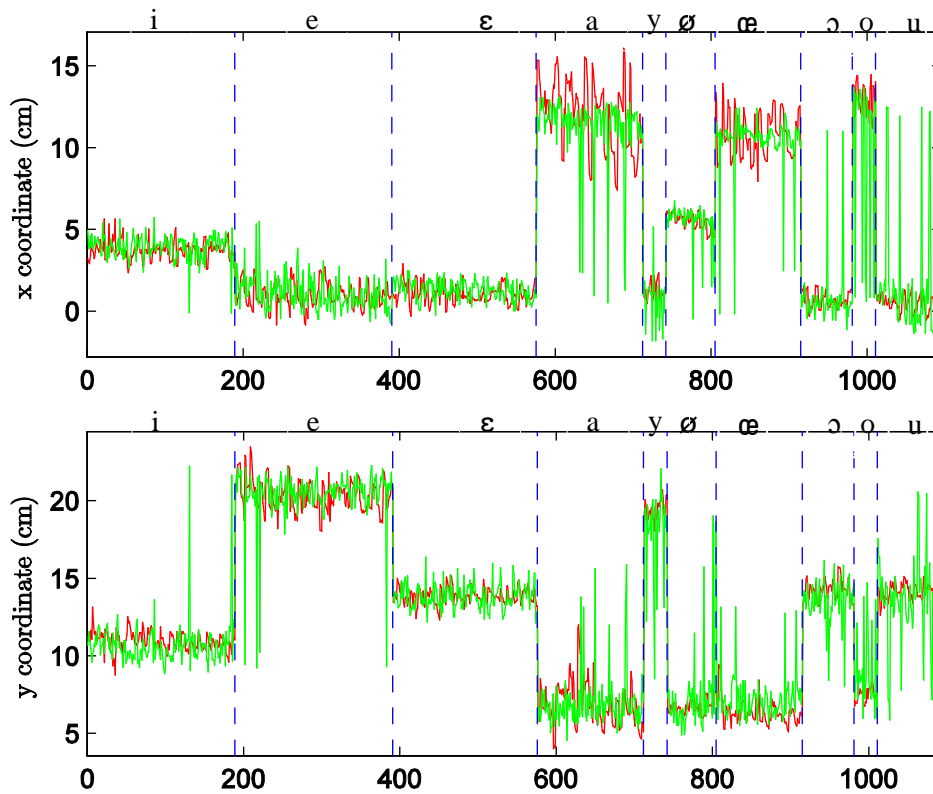


Figure 5.17: The remapping results of estimated hand  $x$ ,  $y$  coordinates on training data in original space. The upper one is corresponding to the  $x$  coordinate and the lower one is corresponding to the  $y$  coordinate. The red line denotes the measured value and the solid green line denotes the estimated value. The vertical blue dash lines separate the groups of coordinates belonging to the different vowels.

The Figure 5.17 shows the deviations caused by the misclassifications at the boundaries of the adjacent groups in the intermediate space (see Figure 5.15). The remapping process would probably enlarge the errors of the misclassified elements. For example, the misclassified individuals of vowels [i] and [e] are adjacent in the intermediate space but they are far away from each other in the original space (See Figure 5.18, Figure 5.19).

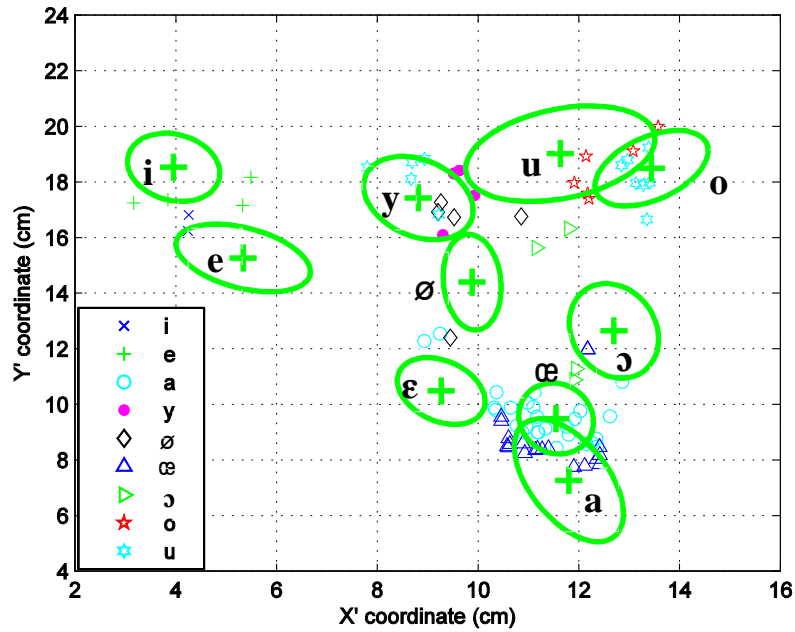


Figure 5.18: The misclassification individuals (obtained by QDA) in intermediate space. Individuals belonging to the different vowels are labeled with the different symbols. The green ellipse ( $std=1.5$ ) denote the distribution of estimated hand positions.

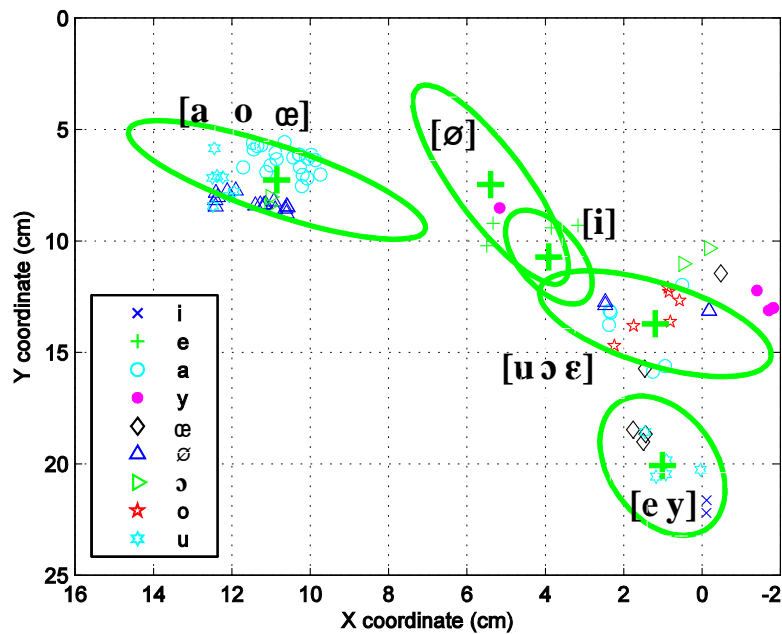


Figure 5.19: The misclassification individuals (obtained by QDA) in original space. Individuals belonging to the different vowels are labeled with the different symbols. The green ellipse ( $std=1.5$ ) denote the distribution of estimated hand positions.



Table 5.10 shows the average RVAR and the RMSE of the remapping coordinates on the training and test data of the 5-fold cross validation. Note that, the error of the final estimation results includes two parts: the first part is the error of the MLR in the intermediate space, and the second part is the error introduced by the classification method.

*Table 5.10: The average RMSE (cm) and RVAR of the remapping coordinates on the training and test data of the 5-fold cross-validation.*

| X coordinate | $\overline{\text{RMSE}}$ | $\overline{\text{RVAR}}$ |
|--------------|--------------------------|--------------------------|
| Training     | 2,30                     | 26%                      |
| Test         | 2,50                     | 31%                      |

| Y coordinate | $\overline{\text{RMSE}}$ | $\overline{\text{RVAR}}$ |
|--------------|--------------------------|--------------------------|
| Training     | 2,20                     | 20%                      |
| Test         | 2,37                     | 22%                      |

The average RMSE and RVAR (RVAR) increase in comparison with the ones obtained in the intermediate space by the MLR approach with the same predictors derived from the mixture MFCCs and LSP (see Table 5.6). This is due to the error introduced by the misclassification. Figure 5.20 shows the remapping results of estimated x, y coordinates on test data in the original space. Except the errors shown in the figure, the prediction value is quite close to the reference data. The errors caused by the misclassification in the intermediate fail to achieve the good performance in the original space.

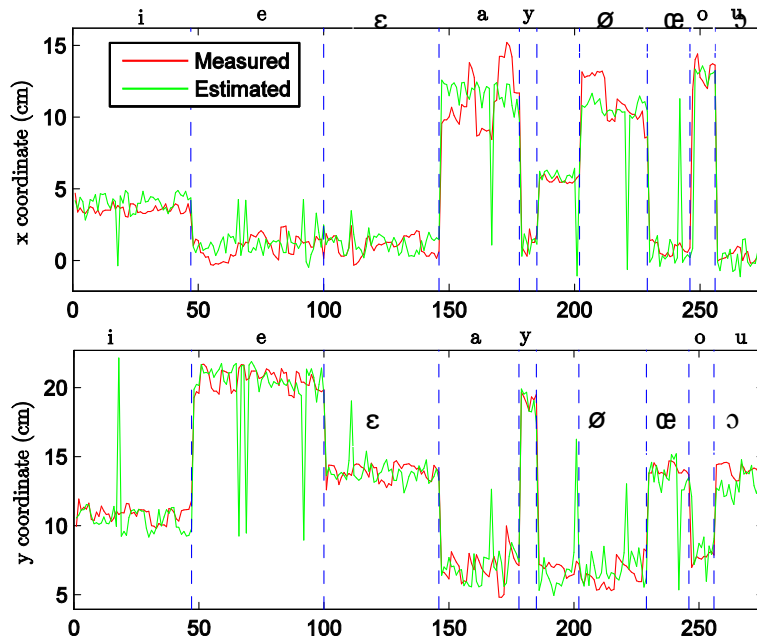


Figure 5.20: The remapping results of estimated  $x$ ,  $y$  coordinates on test data in original space. The upper one is corresponding to the  $x$  coordinate and the lower one is corresponding to the  $y$  coordinate. The red line denotes the measured value and the solid green line denotes the estimated value. The vertical blue dash lines separate the groups of coordinates belonging to the different vowels.

Although the misclassifications which increase the RMSE and RVAR, the indirect prediction with the intermediate space still improves the estimation performance in comparison with the direct prediction in the original space (see Table 5.11, Table 5.12).

Table 5.11: The average  $\overline{\text{RMSE}}_0$  (cm) of the direct MLR approach. The average  $\overline{\text{RMSE}}_1$  (cm) of MLR approach obtained in the intermediate space. The average  $\overline{\text{RMSE}}_2$  (cm) of the indirect MLR approach.

| X coordinate | $\overline{\text{RMSE}}_0$ | $\overline{\text{RMSE}}_1$ | $\overline{\text{RMSE}}_2$ |
|--------------|----------------------------|----------------------------|----------------------------|
| Training     | 2,83                       | 1,27                       | 2.30                       |
| Test         | 3.00                       | 1,35                       | 2.50                       |

| Y coordinate | $\overline{\text{RMSE}}_0$ | $\overline{\text{RMSE}}_1$ | $\overline{\text{RMSE}}_2$ |
|--------------|----------------------------|----------------------------|----------------------------|
| Training     | 2.68                       | 1,22                       | 2.20                       |
| Test         | 2.78                       | 1,26                       | 2.37                       |

Table 5.12: The average  $\overline{\text{RVAR}}_0$  of the direct MLR approach. The average  $\overline{\text{RVAR}}_1$  of MLR approach obtained in the intermediate space. The average  $\overline{\text{RVAR}}_2$  of the indirect MLR approach.

| X coordinate | $\overline{\text{RVAR}}_0$ | $\overline{\text{RVAR}}_1$ | $\overline{\text{RVAR}}_2$ |
|--------------|----------------------------|----------------------------|----------------------------|
| Training     | 39%                        | 13%                        | 26%                        |
| Test         | 43%                        | 14%                        | 31%                        |

| Y coordinate | $\overline{\text{RVAR}}_0$ | $\overline{\text{RVAR}}_1$ | $\overline{\text{RVAR}}_2$ |
|--------------|----------------------------|----------------------------|----------------------------|
| Training     | 29%                        | 7%                         | 20%                        |
| Test         | 31%                        | 8%                         | 22%                        |

## 5.6. Summary

In this chapter, we applied the MLR approach for predicting the lip parameter and hand position of CS by the acoustic spectral parameters. The best predictors are a set of 16 parameters predictors derived from PCA of the LSP and MFCCs coefficients. In the case of lip parameter estimation, the strong linear correlation between the lip movements and the acoustic spectrum gives good estimation results. However in the case of the hand position, the direct MLR approach shows a poor performance. This is due to the hand positions which are used to disambiguate similarity of the lip shapes in CS actually having no relation to the speech signal. For the acoustically close vowels (for example, vowel [i] and [e]), their corresponding position in X-Y plan probably is very far; on the contrary, two far vowels in the acoustic space may share the same hand position, such as vowel [a] and [o]. It indicates that the hand positions space has a different topology structure from the acoustic space. This is the reason why the direct MLR approach obtains a poor performance for estimating the hand position. In order to establish a similar topology structure of the acoustic space, the intermediate space is introduced where the hand position is relocated corresponding to the distribution of vowels in the F1-F2 formant space. The estimation performance of the hand position in the intermediate space improves significantly. However, it is necessary to remap the estimated hand positions in the intermediate space to the original space. The classification methods are introduced to classify the hand position in intermediate space to find the shift value in the translation function by which the

hand coordinates in the intermediate space can be remapped into the original space. The Bayes LDA and QDA classifiers were both evaluated, by using the arbitrary covariance matrix instead of the identical covariance matrix, the performance of QDA is slightly better than the LDA. However the misclassification in the intermediate space probably results in a large deviation in the original space during the remapping processing. Therefore the errors of the classification method have failed to achieve the good estimation performance in the original space. Thus we have to turn to another approach that both releases the linearity constraint and overcomes the problem of binary decision. In addition, we did not verify the residual of MLR are normally distributed. In a linear model, if the residuals of the MLR are normally distributed, the estimators are also the maximum likelihood estimators. However, if the residuals are not normally distributed, a central limit theorem often nonetheless implies that the parameter estimates will be approximately normally distributed so long as the sample is reasonably large. For this reason, given the important property that the residual mean is independent of the independent variables, the distribution of the residual term is not an important issue in regression analysis. Specifically, it is not typically important whether the residual term follows a normal distribution (Rao et al., 1999).



# Chapter 6. Speech to Cued Speech mapping: GMM approach

## 6.1. Introduction

In chapter 5, we applied the MLR approach to estimate the lip parameters and hand positions of CS by the acoustic spectral parameters. In terms of the lip parameters (lip width, lip height and lip area), the performance of the direct MLR approach with the predictors derived from the PCA of the speech parameters, i.e. formant, MFCCs, LSP or the mixture of the MFCCs and LSP, is quite acceptable. Indeed the RVAR could decrease to 12% for lip parameter  $B$  in the case of the best results. However, due to the lack of relationship between the hand positions and the acoustic speech, the performance of the direct MLR approach for mapping the acoustic speech to the hand positions is poor. The hand positions defined in CS to differentiate the vowels are different from the lips since the lips are a part of the vocal tract articulator for producing acoustic speech. In order to establish a similar topology between the hand space and the acoustic speech space, the intermediate space in which the hand positions are relocated to simulate the distribution of vowels in the 2 first formants space is introduced. It shows that the RVAR of the hand positions estimated by the MLR approach decreases significantly. However, the misclassifications introduced by the classification method used in intermediate space fail to achieve the good performance in the original space. In order to release the linear constraint of the MLR method, we introduce the GMM-based mapping approach which is particularly suitable for modelling the distribution where the measurements arise from separate groups. Moreover this method includes the both the regression and the classification properties. The theory of the GMM-based mapping method is described in chapter 3.

This chapter is organized as follows. In Section 6.2 we present three different training methods for estimating the parameters of GMM. In Section 6.3 and 6.4 the evaluation results of the lip parameters and hand positions estimated under the criteria of MMSE and MAP for the three different GMMs are presented. In Section 6.5 we have a discussion of the different estimation approaches used in this work. Section 6.6

discusses briefly of the RVARs produced by the different mapping approaches. The summary of this chapter is presented in Section 6.7.

## 6.2. GMM-based mapping approach

Since the lip parameters and the hand positions have different relationship with the acoustic speech, we discuss the GMM-based mapping approach in terms of lip parameters and hand positions separately.

### 6.2.1. GMM-based mapping approach for estimating lip parameters

In the chapter 5, we used the MLR approach to estimate the lip parameters. In the MLR approach, we implicitly assume that the probability distribution of lip parameters are the same on the whole set of data. This hypothesis is slightly arbitrary and in fact it is probably the reason of the estimation limit of MLR approach. Thus if we want to improve further the performance of the estimation, we need to describe the probability density of the data more precisely and estimate the targets by a local regression method. The finite mixture density model is used to present the probability of the overall population and the GMM is the most studied case of the finite mixture model. The effectiveness of GMM has been illustrated in many applications, such as speech recognition, speech synthesis and so on. There are many methods to estimate the parameters of GMM. In this work, we present mainly three methods in the framework of supervised, unsupervised and semi-supervised procedure separately in the view of the machine learning theory. Then the regression processing is implemented based on the different trained GMM.

#### 6.2.1.1. Supervised training method

The supervised training/learning is the machine learning task of inferring a function from labeled training data. The training data  $\mathbf{z}$  consists of a set of training examples which normally are pairs consisting of input object vectors  $\mathbf{x}$  and desired output values  $\mathbf{y}$ , namely  $\mathbf{z} = [\mathbf{x}, \mathbf{y}]$ . In our case, the input vector is the predictor derived from PCA of the speech parameter :  $\mathbf{x}_t = [F_{t,1}, F_{t,2}, \dots, F_{t,p}]$  ( $t$  refers to the frame number in the corpus,  $p$  is the number of selected principal components) and the target  $\mathbf{y}$  is the lip feature such as the lip width  $A$ , height  $B$  and area  $S$ ,  $\mathbf{y}_t = [A_t, B_t, S_t]$ , ( $t$  refers

to the frame number in the corpus). By the individual phonetic label, we can gather the training data into 10 different groups corresponding to the 10 different vowels. Or we can gather the training data into 3 groups corresponding to the 3 lip visemes which represent three different mouth shapes. The 3 lip visemes are corresponding to 3 groups of vowels [i,e,ɛ,a],[y,ø,o,u] and [œ,ɔ](see in the Figure 6.1, Figure 6.2).

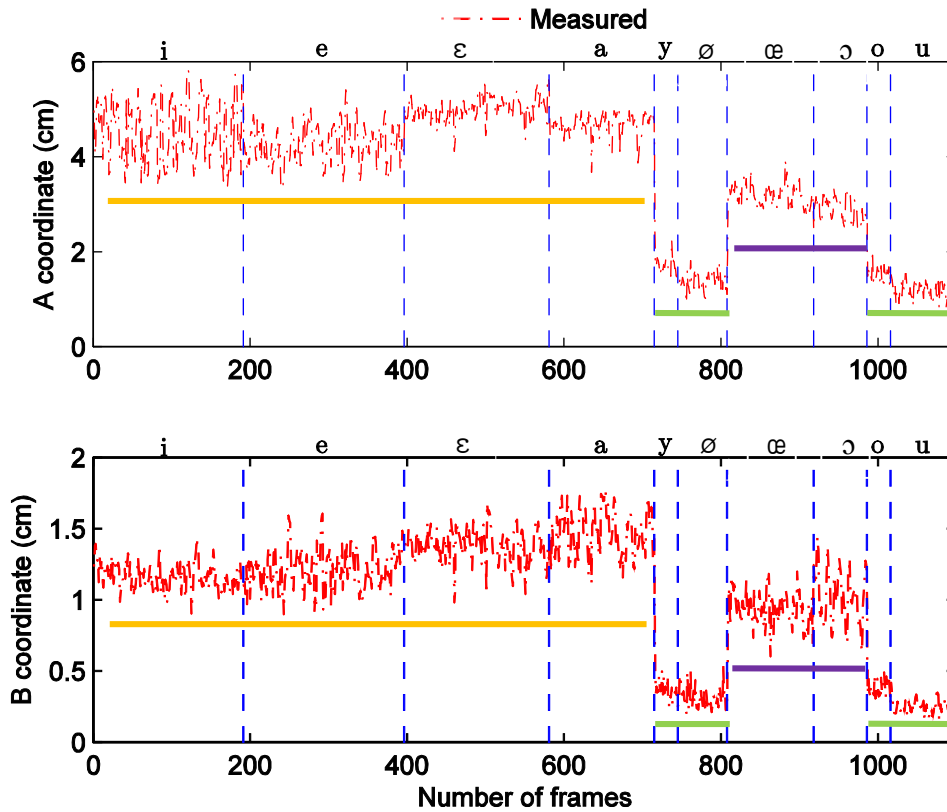


Figure 6.1: The 3 different lip visemes are corresponding to the three groups of vowels highlighted by the three underlines with different colors. The measured values of lip width (A) and lip height (B) are denoted by the red dotted line. The vertical blue dash lines separate the groups of coordinates belonging to the different vowels.



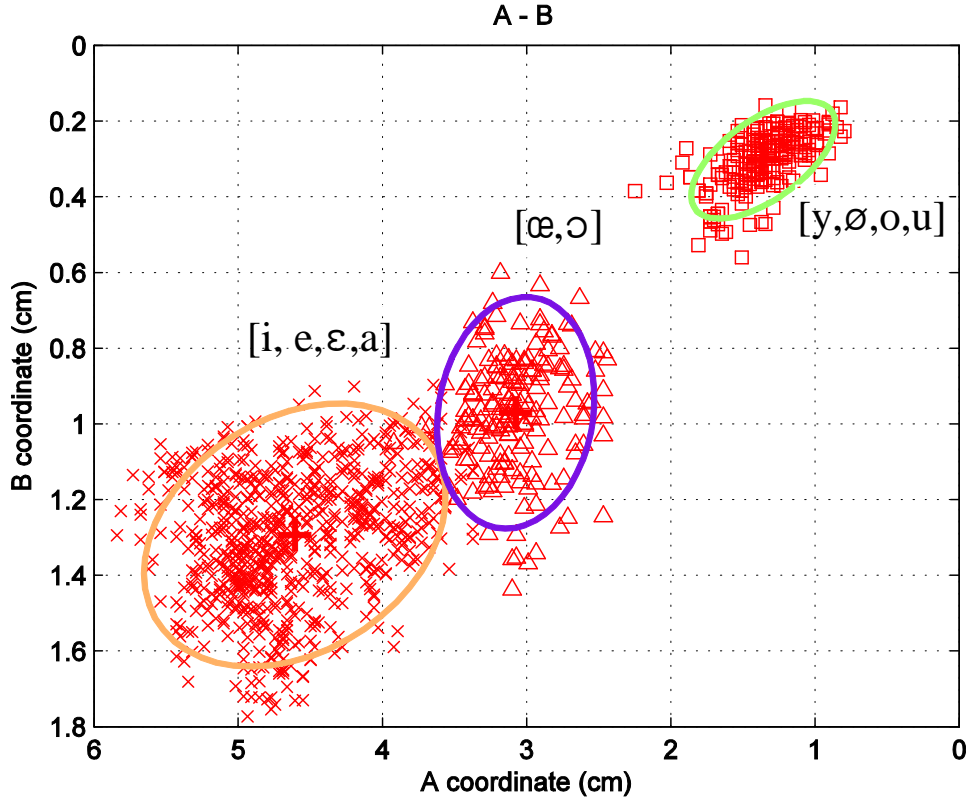


Figure 6.2: The three groups of the lip parameters in the A-B distribution plan are corresponding to the three lip visemes. The points in the plan are denoted by the red crosses, triangles and squares respectively corresponding to the different groups. The ellipses (std=2) in three different colors denote the distributions of the three groups of lip parameters and the red '+' denote the centers of the ellipses.

Either gathering the training set into 10 groups corresponding to the 10 vowels or gathering the training set into three different groups corresponding to the lips visemes, each group is corresponding to a Gaussian (i.e. component) in GMM. With the phonetic labels of the individuals, the gathering processing could be realized quickly. When the gathering procedure is finished, the training of the GMM in the sense of supervised method is completed. That is to say, the parameters of GMM which are the mean vectors  $\boldsymbol{\mu}_m^{(x)}$  and  $\boldsymbol{\mu}_m^{(y)}$ , covariance matrices  $\boldsymbol{\Sigma}_m^{(xx)}$ ,  $\boldsymbol{\Sigma}_m^{(yy)}$ ,  $\boldsymbol{\Sigma}_m^{(xy)}$ ,  $\boldsymbol{\Sigma}_m^{(yx)}$  and the mixture weights  $\alpha_m$  of each Gaussian in GMM are estimated:

$$\boldsymbol{\lambda}^{(z)} = \{ \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}, \alpha_m \} \quad m = [1, 2, \dots, M] \quad (6.1)$$

where

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \quad (6.2)$$

$M = 10$  if using the first supervised GMM training method corresponding to the 10 different vowels or  $M = 3$  if using the second supervised GMM training method corresponding to the 3 different lips visemes. Then the lip parameters could be estimated under the criterion of MMSE or MAP described in the chapter 3 based on the estimated GMM. To recall the regression equations, we present briefly the formulas as follows:

Under the criteria of MMSE:

$$\begin{aligned} \hat{\mathbf{y}}_t &= E[\mathbf{y}_t | \mathbf{x}_t] = \int p(\mathbf{y}_t | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)}) \mathbf{y}_t d\mathbf{y}_t \\ &= \int \sum_{m=1}^M p(\mathbf{y}_t | \mathbf{x}_t, m, \boldsymbol{\lambda}^{(z)}) p(m | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)}) \mathbf{y}_t d\mathbf{y}_t \\ &= \sum_{m=1}^M p(m | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)}) E_{m,t}^{(y)} \end{aligned} \quad (6.3)$$

where

$$p(m | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)}) = \frac{\alpha_m N(\mathbf{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{n=1}^M \alpha_n N(\mathbf{x}_t; \boldsymbol{\mu}_n^{(x)}, \boldsymbol{\Sigma}_n^{(xx)})} \quad (6.4)$$

$$E_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} (\mathbf{x}_t - \boldsymbol{\mu}_m^{(x)}) \quad (6.5)$$

The  $p(m | \mathbf{x}_t, \boldsymbol{\lambda}^{(z)})$  is the a posteriori probability of the  $m$ th Gaussian (i.e. component) indicating the probability of  $\mathbf{x}_t$  being generated by the  $m$ th Gaussian, which is the weight of the regression result produced by the  $m$ th Gaussian.  $E_{m,t}^{(y)}$  is the conditional expectation of the  $\hat{\mathbf{y}}_t$  given the component  $m$  and source data  $\mathbf{x}_t$ .  $M$  equals to 10 or 3 depending on the different supervised training methods.

Under the criteria of MAP:

$$\hat{\mathbf{y}}_t = \overline{(\mathbf{D}_t^{(y)})^{-1}}^{-1} \overline{\mathbf{D}_t^{(y)} \mathbf{E}_t^{(y)}} = \left( \sum_{m=1}^M \gamma_{m,t}^{(z)} \mathbf{D}_m^{(y)} \right)^{-1} \sum_{m=1}^M \gamma_{m,t}^{(z)} \mathbf{D}_m^{(y)} \mathbf{E}_{m,t}^{(y)} \quad (6.6)$$

where

$$\gamma_{m,t}^{(z)} = p(m | \mathbf{x}_t, \mathbf{y}_t, \boldsymbol{\lambda}^{(z)}) \quad (6.7)$$

$$\mathbf{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \boldsymbol{\Sigma}_m^{(xy)} \quad (6.8)$$

Noting that, the solution for estimating  $\hat{\mathbf{y}}_t$  shown in equation (6.6) is not a closed form which includes the current value of  $\mathbf{y}_t$  in right part of the equation for computing the a posteriori probability  $\gamma_{m,t}^{(z)}$ . Thus an EM iterative procedure is applied for estimating the  $\hat{\mathbf{y}}_t$ . The iterative procedure will not stop until the convergence condition is satisfied, namely the increment of the auxiliary function  $\Delta Q(\mathbf{y}_t, \hat{\mathbf{y}}_t)$  falls below a threshold.

$$Q(\mathbf{y}_t, \hat{\mathbf{y}}_t) = -\frac{1}{2} \hat{\mathbf{y}}_t^T \left( \sum_{m=1}^M \gamma_{m,t}^{(z)} \mathbf{D}_m^{(y)} \right) \hat{\mathbf{y}}_t + \hat{\mathbf{y}}_t^T \left( \sum_{m=1}^M \gamma_{m,t}^{(z)} \mathbf{D}_m^{(y)} \mathbf{E}_{m,t}^{(y)} \right) + \bar{K}_t \quad (6.9)$$

where  $\bar{K}_t$  is a term independent of  $\hat{\mathbf{y}}_t$  considered as a constant during computing the  $Q(\mathbf{y}_t, \hat{\mathbf{y}}_t)$ .  $M$  equals to 10 or 3 depending on the different supervised training methods. We can see that the weight of the  $m$ th Gaussian is  $\gamma_{m,t}^{(z)} \mathbf{D}_m^{(y)}$  in the MAP criterion. We can see that the Gaussians defined explicitly have clear sense in the supervised training method. Thus it is reasonable that the supervised training method has a relative more stable performance in comparison with the unsupervised training method. This will be proved in the evaluating results.

### 6.2.1.2. Unsupervised training method

The most used unsupervised training method is EM as mentioned in the chapter 3. The EM needs, for example, the  $k$ -mean algorithm to initialize the parameters of the GMM, i.e. mean vector, covariance matrices and mixing coefficients. Then the EM algorithm alternates between the following two steps: the expectation (E) step and the maximization (M) step. In the expectation step, the current values of the parameters (mean vector, covariance matrices and mixing coefficients) are used to evaluate the a

posteriori probabilities, given by (3. 26). Then the a posteriori probabilities are used in the maximization step to re-estimate the parameters of the GMM, i.e. mixing coefficients, mean vectors and covariance matrices. Normally, the iterative procedure will not stop until the convergence condition is satisfied, i.e. the increment of the auxiliary function  $\Delta Q(\Theta, \Theta^{old})$  falls below some threshold or the iteration times reach a maximum value. The unsupervised training method is a procedure without the label information. Therefore the unsupervised training method is better than the supervised training method to find the components of GMM underlying the data by using the automatic iteration procedure. Meanwhile the automatic clustering will be affected by the outliers in the data. In addition, the common covariance matrix is used in this work to avoid the matrices to be singular when the number of components is too large in comparison with the dimension or the size of the training data. As the supervised methods, when the training processing of GMM finishes the MMSE and MAP are applied to estimate the lip parameters.

### 6.2.1.3. Semi-supervised training method

As mentioned above, the supervised training method has the advantage of being more stable benefiting from the explicit classification. Meanwhile it also limits the ability to find the underlying components hidden in the data. The unsupervised training method without any label information of the training data could cluster automatically meanwhile relying too much on the training data gives rise to the unstable performance. In order to include advantages of the two training methods and ameliorate the disadvantages, the semi-supervised training method is proposed. Both of the supervised and unsupervised training methods are used in the semi-supervised training method. In the framework of semi-supervised training method, firstly, the training data are split into 10 or 3 big groups by the phonetic label as the supervised training method; secondly, inside of each group, the parameters of Gaussians are trained by the unsupervised method such as the EM method. Finally, the whole training set is divided into many Gaussians (see Figure 6.3). As same as the other two training methods, we can use MMSE and MAP to estimate the lip parameters when the Gaussians of the GMM are fixed.

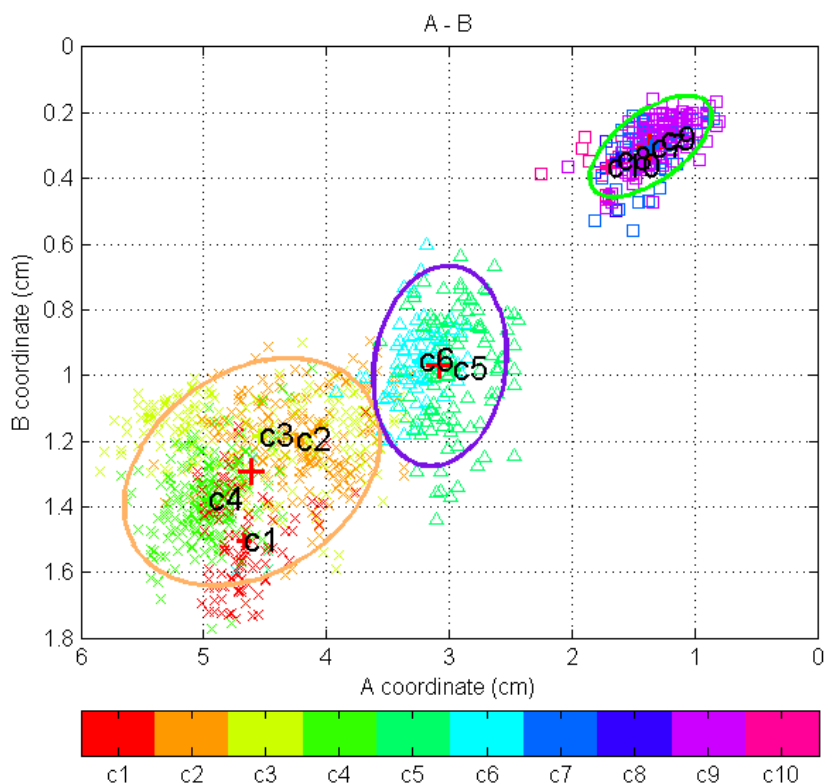


Figure 6.3: The semi-supervised method for training GMM. The three groups trained by the supervised method are denoted by the three ellipses ( $std=2$ ) with different colors. The Gaussians inside each group are trained by the unsupervised method denoted by the different colors. The color bar in the bottom of the figure indicates the colors corresponding to the Gaussians. The centers of the Gaussians are indicated by the labels 'c1', 'c2', ..., 'c10'.

The conventional semi-supervised method in machine learning is to combine a small amount of labeled data with a large amount of unlabeled data to improve the learning accuracy of the unlabeled data (Zhu et al., 2003). However in this work, we take advantage of the limited label information of the data to improve the training accuracy. In the conventional semi-training method, the incomplete information is the large amount of unlabeled data, while in this work the incomplete information is the incomplete labeled information of the data, that is to say the label of the data just indicates the group to which the individual roughly belongs rather than the real Gaussian inside each group. The common point between the conventional semi-supervised method and our approach is that the training method uses the limited label information to train the unlabeled data.

## 6.2.2. GMM-based mapping approach for estimating hand positions

The procedure for estimating hand positions by GMM-based mapping approach is similar to the case of lip parameters estimation. But due to the different relationship with the acoustic speech, the GMM training processing for hand positions is different from the lip parameters case. As what are used for estimating the lip feature, three different training methods, i.e. supervised, unsupervised and semi-supervised methods are used for training GMM for estimating hand positions. Then the same criteria MMSE and MAP are used as the regression approaches.

### 6.2.2.1. Supervised training method

As the supervised training method used for estimating lip parameters, the training set which includes the source data and target ones, can be divided into groups by using the individual phonetic label. We can gather the training set according to the feature of the source data, namely gathering the training set into 10 groups corresponding to the 10 different vowels. Or taking advantage of the feature of the target data, which are the five hand positions defined in the CS to differentiate the vowels with similar lip shapes, to gather the training set into 5 groups corresponding to the five positions (see Figure 6.4, Figure 6.5).

### 6.2.2.2. Unsupervised training method

The unsupervised training method for estimating hand positions is similar to the case of lip parameters estimation. The only difference is that using the hand position coordinates as the target  $\mathbf{y}_t$  instead of the lip parameters in the estimation procedure.

### 6.2.2.3. Semi-supervised training method

The semi-supervised training method for estimating hand positions is exactly identical to the case of lip parameters estimation: use the supervised training method to classify preliminary the training data into groups and then use the unsupervised training method to estimate the parameters of Gaussians inside each group. When the Gaussians are fixed, the MMSE and MAP regression approaches are used to map the hand positions from acoustic spectral parameters.

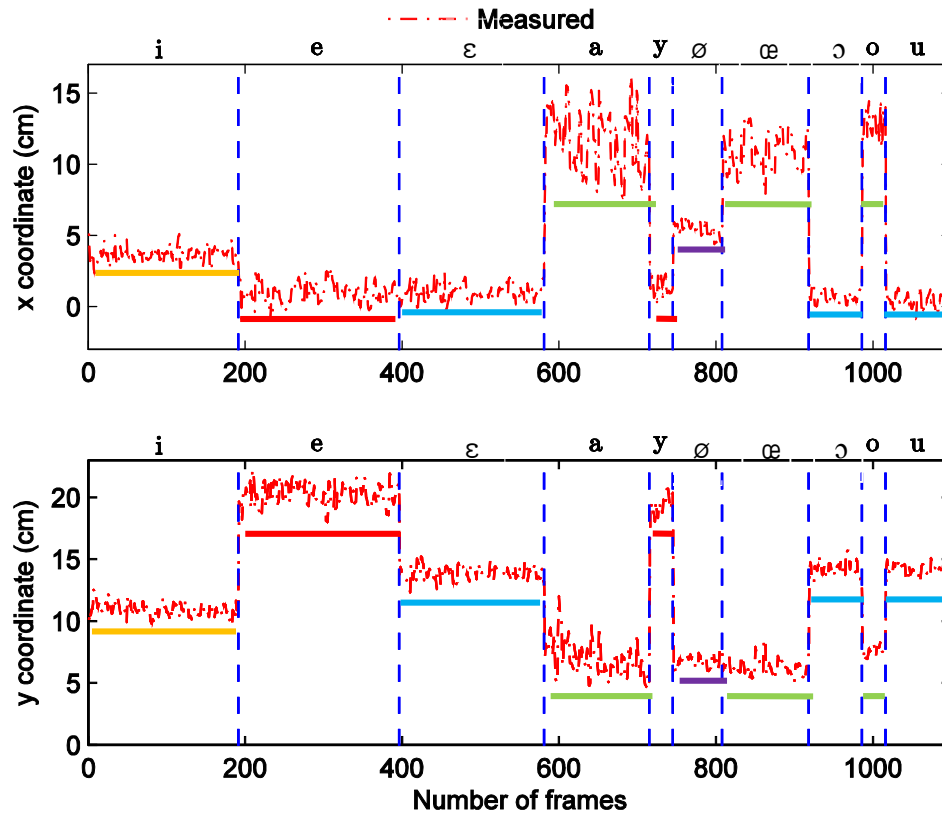


Figure 6.4: The target data, i.e. the hand coordinates, gather in five groups highlighted by the five underlines with different colors by the CS rule. The vertical blue dash lines separate the groups of coordinates belonging to the different vowels.

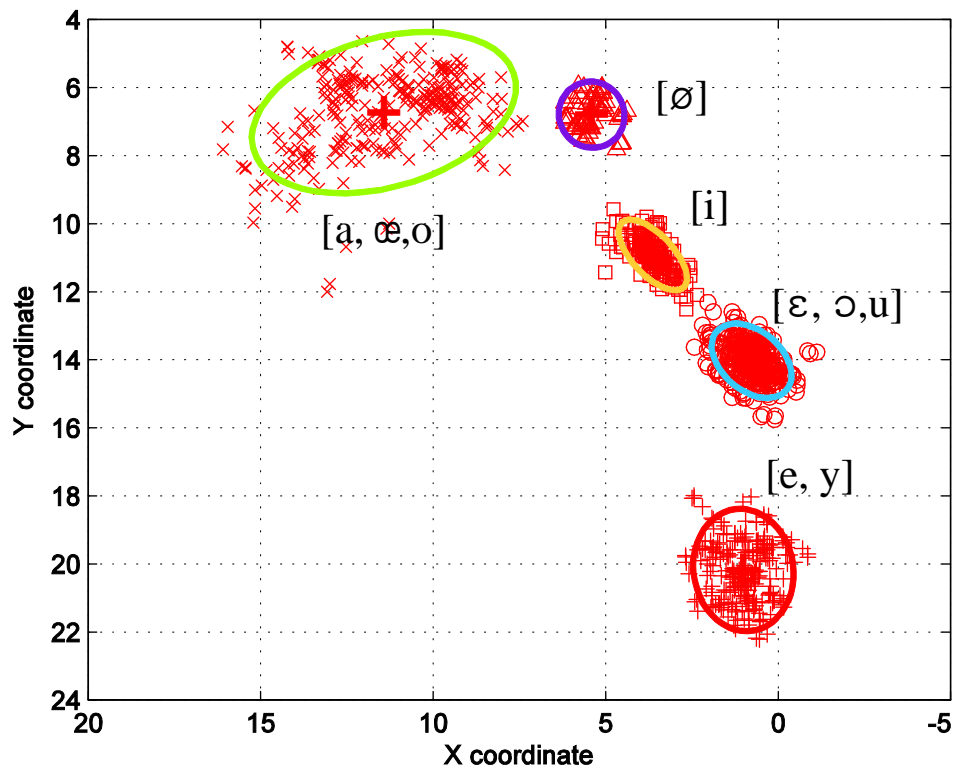


Figure 6.5: The five groups of the hand position shown in X-Y plan. The points in the plan are denoted in the red crosses, triangles, squares, circle, plus sign corresponding to the different five groups. The ellipses (std=2) in five different colors denote the distributions of the five groups of hand position and the red '+' denote the centers of the ellipses.

### 6.2.3. The selection of the predictors

As what has been done in chapter 5, we use the PCA scores of the speech spectral parameters as the predictors to estimate the targets. But unlike the MLR approach applied in chapter 5, the regression in GMM-based mapping approach is implemented on each Gaussian of GMM rather than on the global data. Thus we cannot use the correlation coefficient squared  $\rho^2$  between the predictors and the global targets as the criterion to order the predictors. Instead, we use the RVAR of the estimated value as the criterion to order the predictors, that is to say the predictor which can obtain the minimum RVAR of the estimated value is selected firstly. Note that, using the RVAR as the criterion is same to the criterion used in chapter 5 when the GMM has only one Gaussian. Since in the case of unimodal Gaussian, the minimum of the RVAR corresponds to the maximum of the correlation coefficient squared  $\rho^2$  as shown in the



equation  $RVAR = Var(\mathbf{y} - \hat{\mathbf{y}})/Var(\mathbf{y}) = 1 - \rho^2$ . Therefore we can consider the RVAR as an extension of the criterion used in the MLR.

### 6.3. Evaluation of GMM-based mapping approach for estimating lip parameters

The database used for evaluating the GMM-based mapping approach is the one used in the chapter 5 for evaluating the MLR approach. The 5 fold cross validation is also used in the evaluation of the GMM-based mapping approach. We also use RVAR and the RMSE as the evaluation criteria:

$$RVAR = \frac{Var(\mathbf{y} - \hat{\mathbf{y}})}{Var(\mathbf{y})} \quad (6.10)$$

where  $Var(\mathbf{y} - \hat{\mathbf{y}})$  is the variance of the residual and  $Var(\mathbf{y})$  is the variance of the target data  $\mathbf{y}$ .

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (\mathbf{y}_t - \hat{\mathbf{y}}_t)^2} \quad (6.11)$$

where  $N$  is the total number of the frames in the corpus,  $\mathbf{y}_t$  is the target to be estimated and  $\hat{\mathbf{y}}_t$  is the estimated value of the target. Given that the predictors derived from the PCA of the mixture of the MFCCs and LPC having the best performance for estimating lip parameters and hand positions in chapter 5, the ones are used as predictors to estimate the lip parameters and hand positions in this chapter.

#### 6.3.1. Evaluation of GMM-based mapping approach with supervised trained GMM for estimating lip parameters

As mentioned in the section 6.2.1, there are two supervised training methods: one is based on the 10 different vowels; the other one is based on the 3 lips visemes. The two training methods are evaluated separately and each supervised training processing is followed by the MMSE and MAP regression approaches.

### 6.3.1.1. Evaluation of GMM-based mapping approach with supervised trained GMM based 10 vowels

We can see that in the Figure 6.6 the RVARs decrease with increasing the number of the predictors and the RVARs decrease to about 5% for all of lip parameters  $A$ ,  $B$  and  $S$ . The GMM-based mapping approach has both the classification and regression properties. As increasing the number of the predictors, namely the dimension of the GMM, the GMM-based mapping approach improves the classification rate of the individuals. That is to say, the a posteriori probability of  $m$ th component which most probably ‘generated’ the individual is improved with increasing the dimension of the GMM. Figure 6.7 also shows the decrease of the RMSE by increasing the number of the predictors.

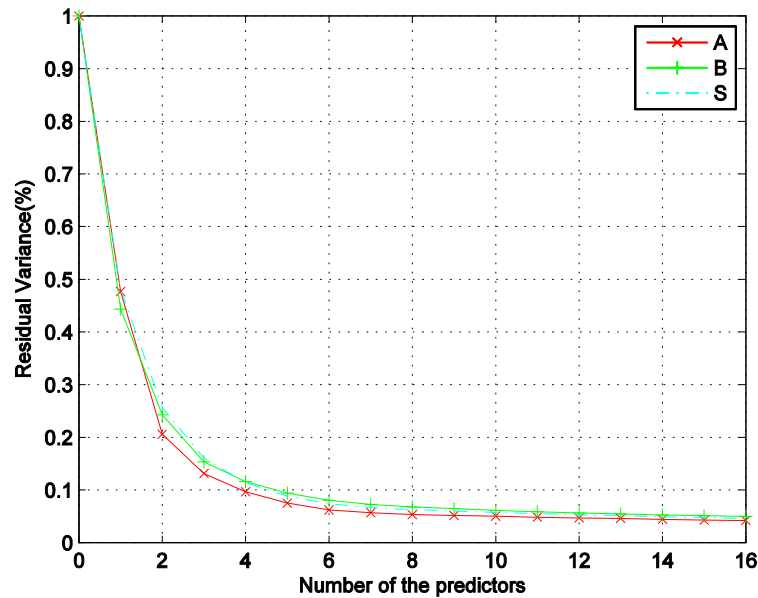


Figure 6.6: Average RVARs of the training data based on the supervised trained GMM with 10 Gaussians for estimating the lip parameters  $A, B$  and  $S$  in function of the number of predictors. MMSE is used as the regression criterion.

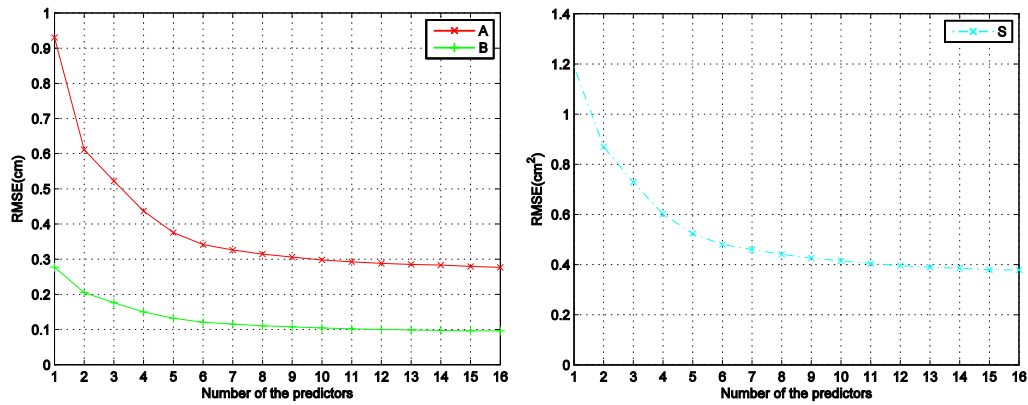


Figure 6.7: Average RMSEs of the training data based on the supervised trained GMM with 10 Gaussians for estimating the lip parameters A, B and S in function of the number of predictors. MMSE is used as the regression criterion.

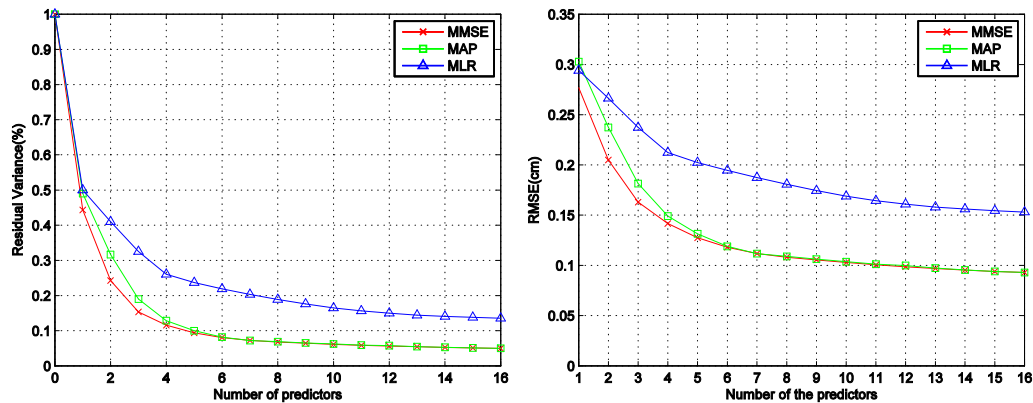


Figure 6.8: Average RVARs and RMSEs of the training data for estimating the lip parameter B in function of the number of predictors based on the MLR approach and GMM-based mapping approach with 10 supervised trained Gaussians corresponding to the 10 vowels in the regression criteria of MMSE and MAP separately.

We can see that in the Figure 6.8, both the RVAR and RMSE obtained by the GMM-based mapping approach for estimating the lip parameter B are lower than the ones obtained by the MLR approach. The RVAR of GMM-based mapping approach is almost lower 10 percents than the MLR approach by using 16 predictors. Figure 6.8 also shows that the criteria of MMSE and MAP are almost the same for estimating the lip parameter B when the number of the predictors is more than 6. That is to say when the dimension of the GMM is enough for classifying the individual to the proper Gaussian, the regression approaches of MMSE and MAP are almost the same. But

when the dimension of the GMM is low, the MMSE performs slightly better than MAP. Figure 6.9 shows the average RVAR of the test data in the regression criterion of MMSE. Figure 6.10 compares the average RVAR and RMSE of the test data obtained respectively by the GMM-based mapping approach in the different regression criteria and the MLR approach.

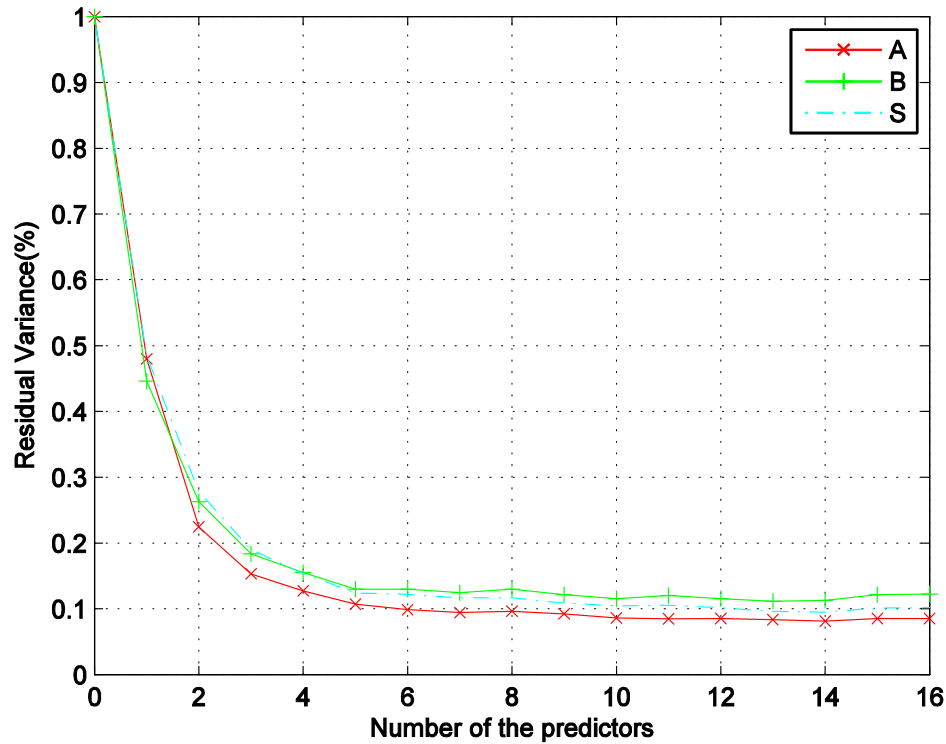


Figure 6.9: Average RVARs of the test data based on the supervised trained GMM with 10 Gaussians for estimating the lip parameters A, B and S in function of the number of predictors. MMSE is used as the regression criterion.

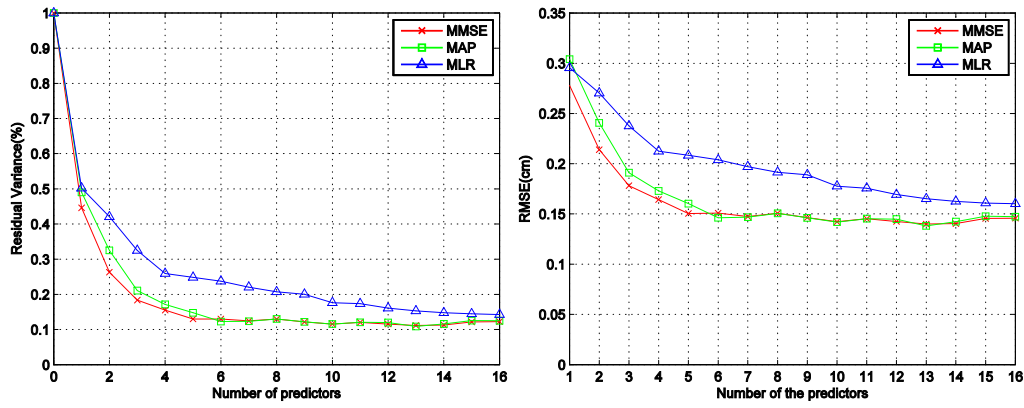


Figure 6.10: Average RVARs and RMSEs of the test data for estimating the lip parameter  $B$  in function of the number of predictors based on the MLR approach and GMM-based mapping approach with 10 supervised trained Gaussians corresponding to the 10 vowels in the regression criteria of MMSE and MAP separately.

Figure 6.10 shows that the RVAR or RMSE obtained by the GMM-based mapping approach in the test processing are higher than the ones obtained in the training processing. The generalization ability of the supervised trained GMM with 10 components is limited by the misclassifications in the test processing, which gives rise to the increment of the RVARs (see the Figure 6.11).

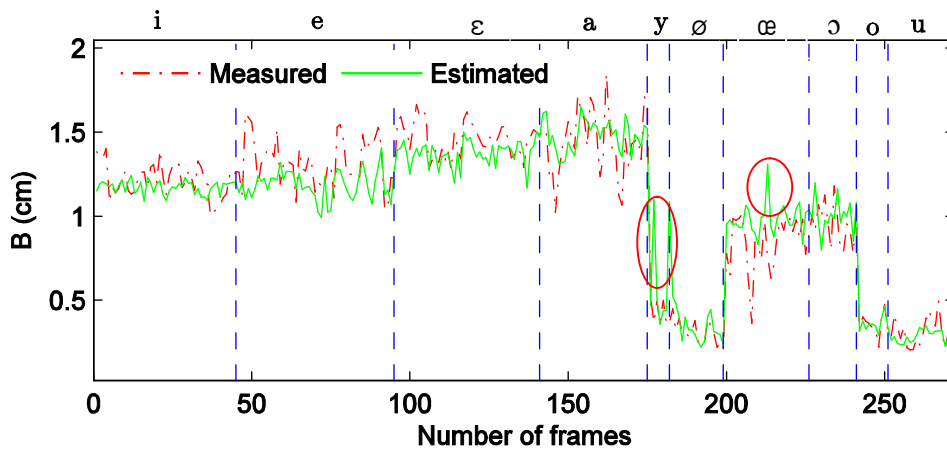


Figure 6.11: The estimated value of the lip parameter  $B$  obtained by the GMM-based mapping approach with 10 supervised trained Gaussians. MMSE is used as the regression criterion. The red dotted line denotes the measured value of test data and the solid green line denotes the estimated value. The red circles indicate the misclassified individuals.

### 6.3.1.2. Evaluation of GMM-based mapping approach with supervised trained GMM based on 3 lip visemes

In the previous section, we can see that the GMM-based mapping approach with 10 supervised trained Gaussians can obtain a good performance in the training process. However, the misclassifications of the test data degrade the generalization ability of the model. Aiming to balance the estimation performance and the generalization ability of the model, the GMM-based mapping approach with 3 Gaussians which are trained according to the 3 lips visemes (see Figure 6.1 and Figure 6.2) is introduced.

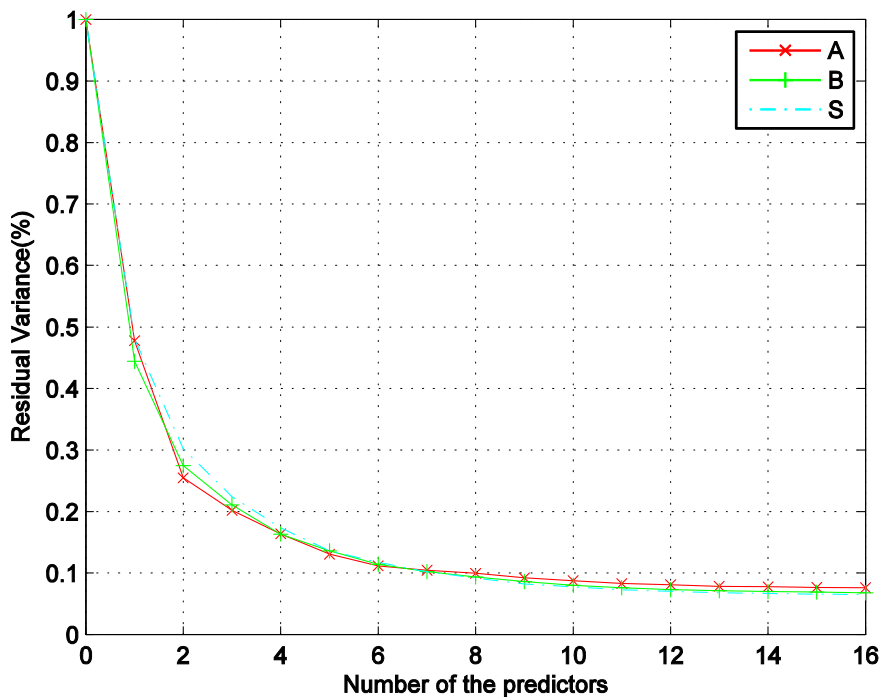


Figure 6.12: Average RVARs of the training data based on the supervised trained GMM with 3 Gaussians for estimating the lip parameters A, B and S in function of the number of predictors. MMSE is used as the regression criterion.

Figure 6.12 and Figure 6.13 show that the RVARs and RMSEs of the supervised trained GMM with 3 Gaussians are slightly higher (increased about 2% for RVAR and 0.02 cm for RMSE) than the ones obtained by the GMM with 10 Gaussians in the training process.

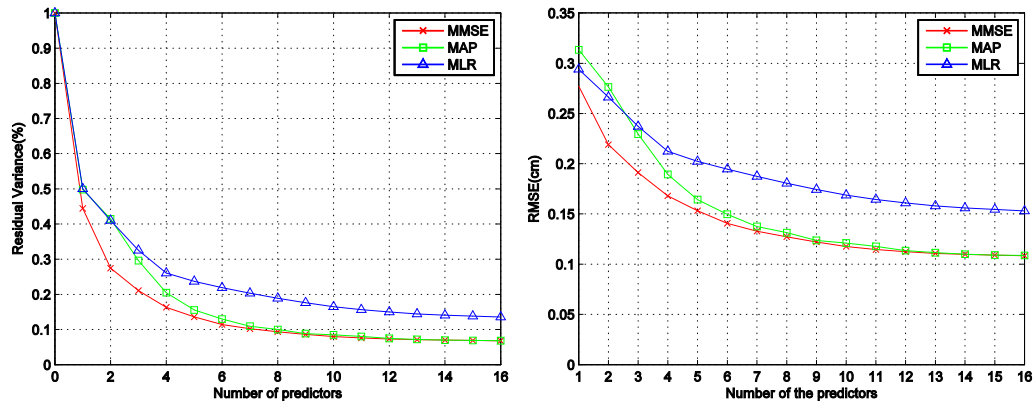


Figure 6.13: Average RVARs and RMSEs of the training data for estimating the lip parameter  $B$  in function of the number of predictors based on the MLR approach and GMM-based mapping approach with 3 supervised trained Gaussians corresponding to the 3 lip visemes in the regression criteria of MMSE and MAP separately.

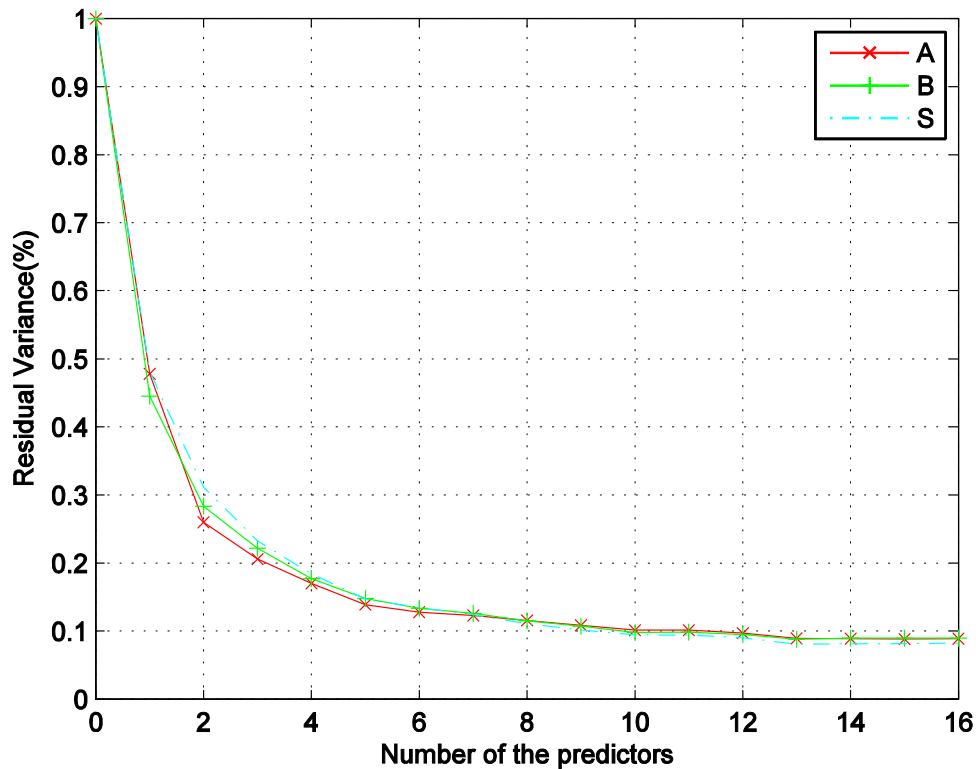


Figure 6.14: Average RVARs of the test data based on the supervised trained GMM with 3 Gaussians for estimating the lip parameters  $A$ ,  $B$  and  $S$  in function of the number of predictors. MMSE is used as the regression criterion.

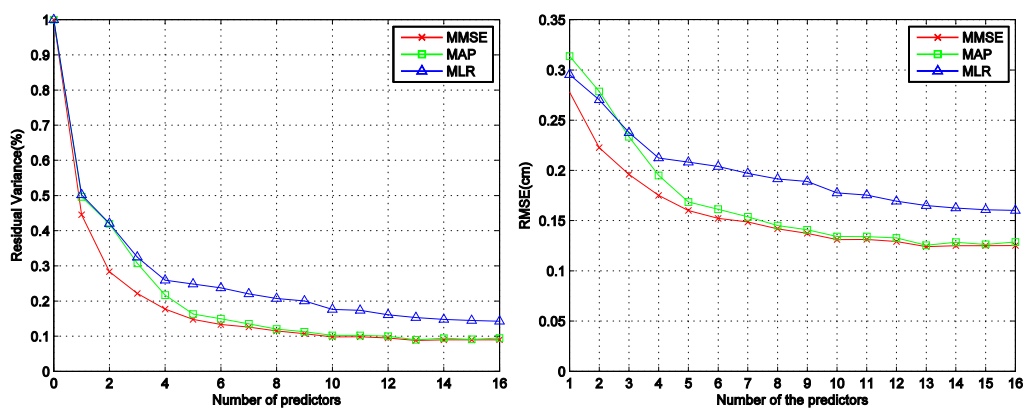


Figure 6.15: Average RVARs and RMSEs of the test data for estimating the lip parameter  $B$  in function of the number of predictors based on the MLR approach and GMM-based mapping approach with 3 supervised trained Gaussians corresponding to the 3 lip visemes in the regression criteria of MMSE and MAP separately.

In Figure 6.14 and Figure 6.15, we can see that the test results of RVAR and RMSE of lip parameter  $B$  based on the GMM with 3 Gaussians are apparently better than the results obtained by the GMM with 10 Gaussians. This is due to the misclassification obtained by the GMM composed by 3 components is less (see Figure 6.16). That is to say, although the estimating performance of the GMM with 3 supervised trained Gaussians is slightly inferior to the GMM with 10 Gaussians in the training process, the generalization ability of the GMM with 3 Gaussians is superior to the GMM with 10 Gaussians. Therefore the GMM with 3 Gaussians is more robust.

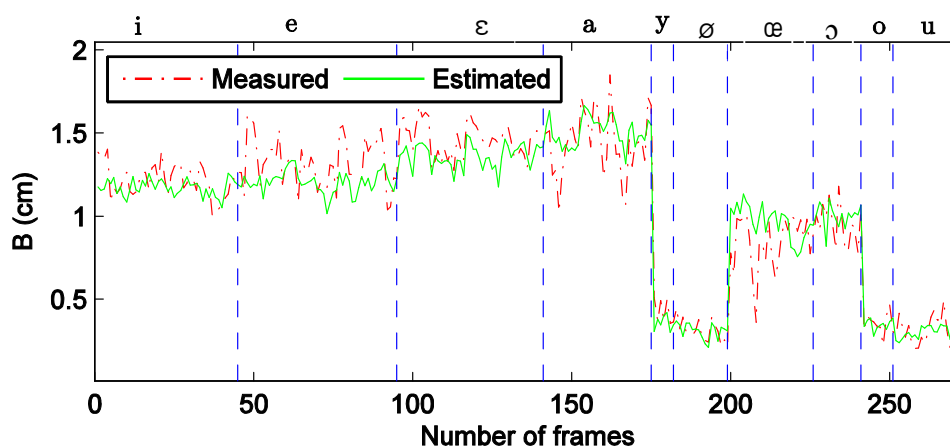


Figure 6.16: The estimated value of the lip parameter  $B$  obtained by GMM-based mapping approach with 3 supervised trained Gaussians in the regression criterion of MMSE. The red dotted line denotes the measured value and the solid line denotes the estimated value.



### 6.3.2. Evaluation of GMM-based mapping approach with unsupervised trained GMM for estimating lip parameters

The main character of the unsupervised training method is that the Gaussians of GMM are trained automatically by EM algorithm given the initialization parameters obtained by the  $k$ -means method. We can see in Figure 6.17 and Figure 6.18 that the RVARs of the estimated lip parameters ( $A$ ,  $B$  and  $S$ ) almost no longer decrease when the number of the Gaussians reached to 36 and finally the RVARs of the estimated lip parameters are 4% for  $A$ , 6% for  $B$  and 5% for  $S$  respectively in the training process.

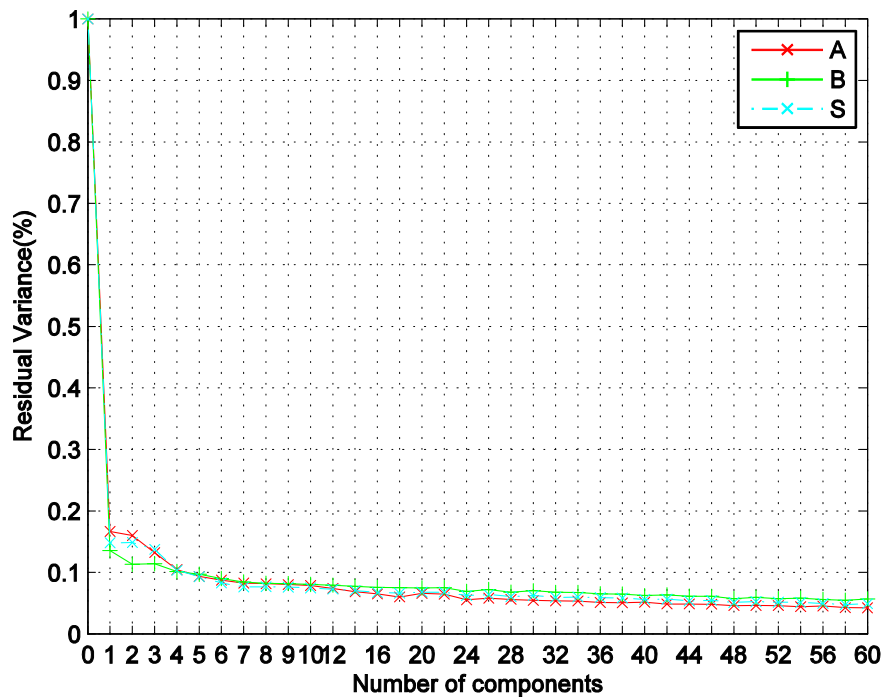


Figure 6.17: Average RVARs of the training data based on the unsupervised trained GMM for estimating the lip parameters in function of the number of the Gaussians/components. MMSE is used as the regression criterion.

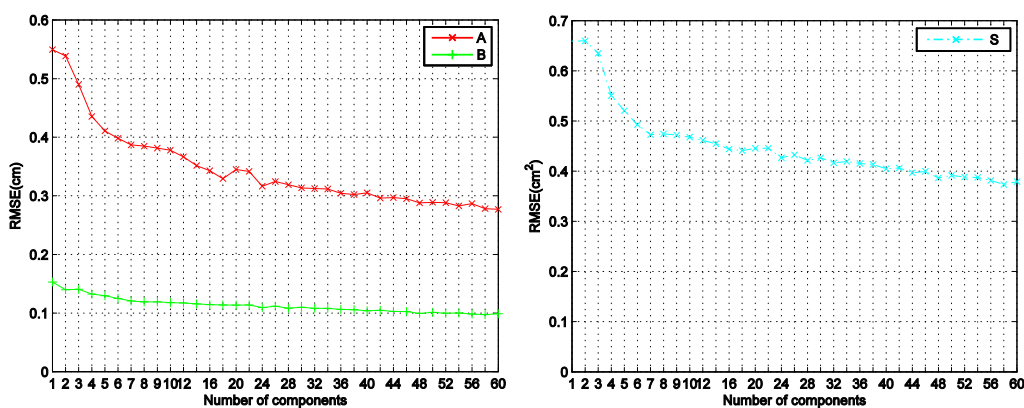


Figure 6.18: Average RMSEs of the training data based on the unsupervised trained GMM for estimating the lip parameters in function of the number of the Gaussians/components. MMSE is used as the regression criterion.

Figure 6.19 shows that the regression criteria of MMSE and MAP based on the unsupervised GMM almost perform in the same way in the training process. In addition, we can see that the MLR approach is exactly equal to the GMM-based mapping approach when the GMM has only one Gaussian. That is to say the MLR approach is indeed a special case of the GMM-based mapping approach when the number of the Gaussians equals to one (noting that the MMSE and MAP mapping criteria are identical when GMM has only one Gaussian). Although the unsupervised training method can improve the estimation performance by increasing the number of the Gaussians, the estimation efficiency of the unsupervised training method is inferior to the supervised method both in the training and test processing. We can see in the Figure 6.19 that the RVAR and RMSE obtained by the supervised trained GMM with 3 Gaussians are comparable to the results obtained by the unsupervised model with 36 Gaussians for estimating lip parameter  $B$ . It indicates that the supervised training method benefiting from a priori knowledge can train the Gaussians more effectively than the unsupervised training method. However the unsupervised training method can train the Gaussians automatically without a priori knowledge such as the individual phonetic label.

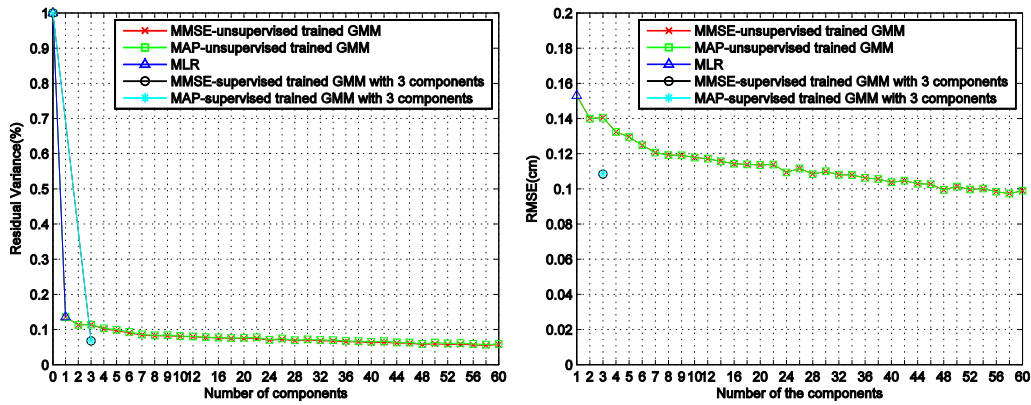


Figure 6.19: Average RVARs and RMSEs of the training data for estimating the lip parameter  $B$  in function of the number of Gaussians based on the MLR approach, GMM-based mapping approach with 3 supervised Gaussians and unsupervised Gaussians in the regression criteria MMSE and MAP respectively.

Figure 6.20 shows the corresponding RVARs and RMSEs of the test data. It shows that the RVARs and RMSEs of test data are slightly higher than the ones obtained from the training data. In the test processing, the RVARs or RMSEs do not always decrease as increasing the number of the Gaussians. For example, when the number of the Gaussians reached to 40, the RVARs and RMSEs begin to increase. This is due to the over-fitting problem in the training process. Generally, the RVAR of estimated lip parameter  $B$  retains at about 9% (7% for  $A$  and 8% for  $S$ ).

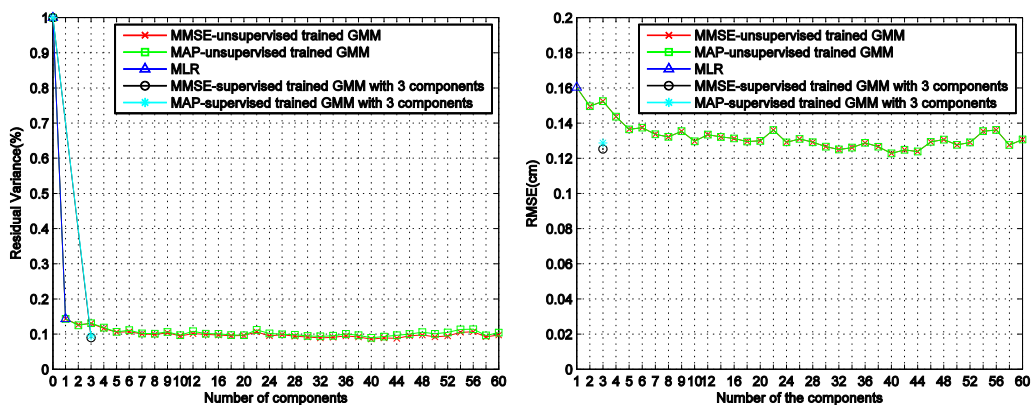


Figure 6.20: Average RVARs and RMSEs of the test data for estimating the lip parameter  $B$  in function of the number of Gaussians based on the MLR approach, GMM-based mapping approach with 3 supervised Gaussians and unsupervised Gaussians in the regression criteria MMSE and MAP respectively.

### 6.3.3. Evaluation of GMM-based mapping approach with semi-supervised trained GMM for estimating lip parameters

The semi-supervised training method aims to include the advantages of unsupervised and supervised training method to improve the estimation performance. Note that the Gaussians inside the same group share a common covariance matrix obtained by the pool estimation method. Figure 6.21 and Figure 6.22 show the average RVARs of the training data and the test data respectively for estimating the lip parameters  $A$ ,  $B$  and  $S$  based on the semi-supervised trained GMM. Table 6.1 shows the different configuration of the number of Gaussians in the 3 groups trained by supervised training method. The configuration of the number of Gaussians in each group is according to the size of each group.

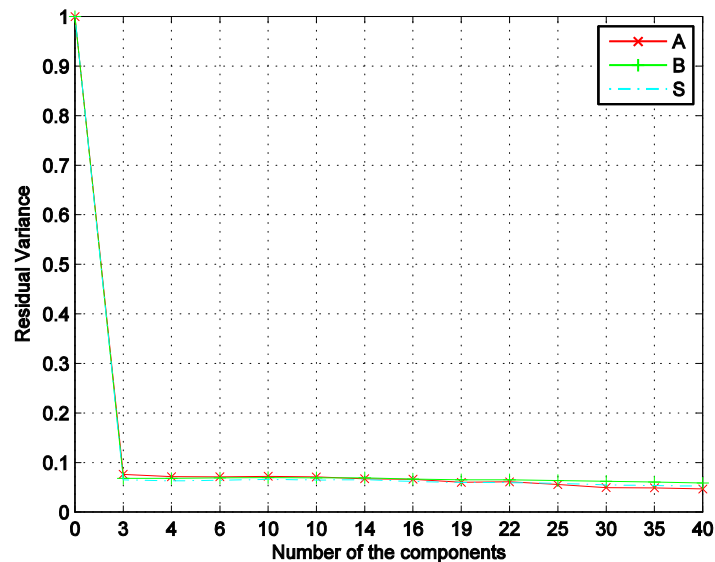


Figure 6.21: Average RVARs of the training data based on the semi-supervised trained GMM for estimating the lip parameters in function of the number of the Gaussians. MMSE is used as the regression criterion.

Table 6.1: The number of the components within each group in the training of the semi-supervised GMM..

| Group<br>[a,ɛ,e,i] | Group<br>[œ,ɔ] | Group<br>[y,o,ø,u] | total<br>components |
|--------------------|----------------|--------------------|---------------------|
| 1                  | 1              | 1                  | 3                   |
| 2                  | 1              | 1                  | 4                   |
| 4                  | 1              | 1                  | 6                   |
| 4                  | 2              | 4                  | 10                  |
| 6                  | 2              | 2                  | 10                  |
| 8                  | 3              | 3                  | 14                  |
| 10                 | 3              | 3                  | 16                  |
| 12                 | 3              | 4                  | 19                  |
| 14                 | 4              | 4                  | 22                  |
| 16                 | 4              | 5                  | 25                  |
| 20                 | 5              | 5                  | 30                  |
| 24                 | 5              | 6                  | 35                  |
| 28                 | 5              | 7                  | 40                  |

We can see in Figure 6.21 and Figure 6.22 that the RVARs decrease only slightly as increasing the number of Gaussians. The decrease is slow due to the common covariance matrix rather than the arbitrary covariance matrices used in the GMM-based mapping approach. The common covariance matrix could keep the main feature of the group to make the approach stable. However, it also weakens the individual fitting ability of each Gaussian inside each group. Therefore the RVARs do not decrease apparently as increasing the number of Gaussians inside each group. If we apply the arbitrary covariance matrix instead of common covariance matrix to estimate the lip parameters, the estimation performance of the test data would be deteriorated by the over-fitting problem (see Figure 6.23).

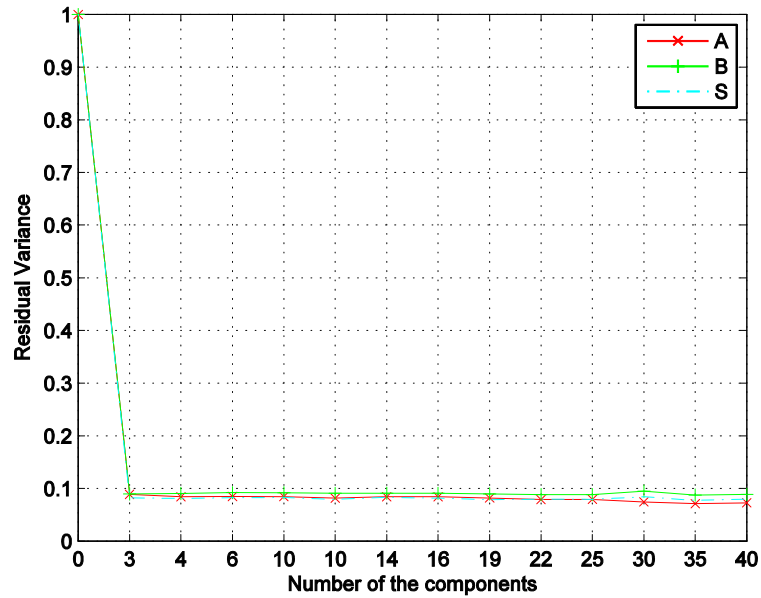


Figure 6.22: Average RVARs of the test data based on the semi-supervised trained GMM for estimating the lip parameters in function of the number of the Gaussians. MMSE is used as the regression criterion.

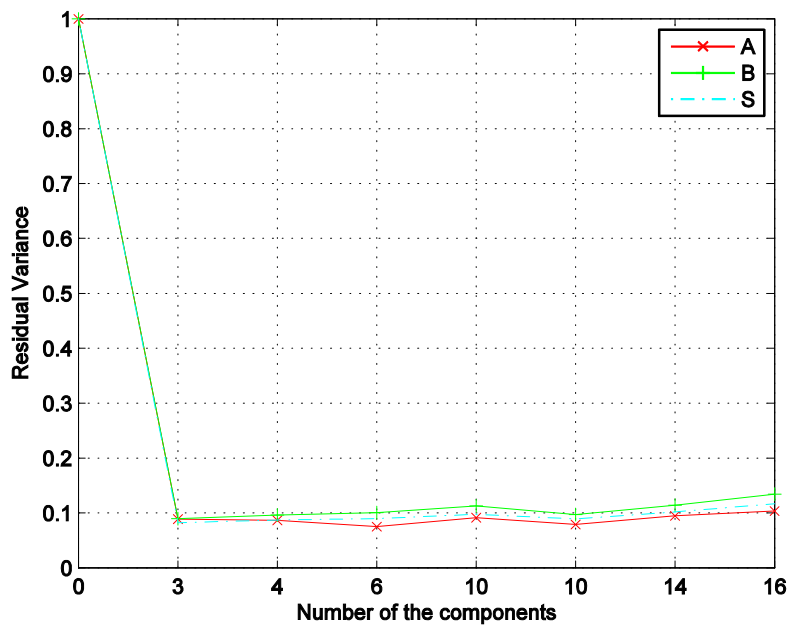


Figure 6.23: Average RVARs of the test data based on the semi-supervised trained GMM for estimating the lip parameters in function of the number of the Gaussians. The arbitrary covariance matrices instead of the common matrix are used in the MMSE regression method.

Table 6.2 and Table 6.3 list the best RVARs and RMSEs of the test data obtained by the GMM-based mapping approach with different GMMs and MLR approach for

estimating the lip parameters. Since the MMSE and MAP perform almost in the same way for estimating the lip parameters, the tables only list the results obtained by the MMSE regression approach. In the tables, we can see that the results obtained by the supervised trained GMM with 10 Gaussians/components are obviously inferior to the other three approaches. This is due to the over-fitting problem. The estimation results obtained by the unsupervised trained GMM and the semi-supervised trained GMM are very close and are slightly better than the supervised trained GMM with 3 Gaussians. We can see that the supervised trained GMM with 3 Gaussians is a robust and effective approach by using much less Gaussians to obtain comparable results. But the GMM-based mapping approach with either of the GMM performs better than the MLR.

Table 6.2: The best RVARs obtained by the GMM-based mapping approach with different GMMs and MLR approach.

| (%) | MLR | Supervised GMM (3 components) | Supervised GMM (10 components) | Unsupervised GMM | Semi-supervised GMM (35 comp.) |
|-----|-----|-------------------------------|--------------------------------|------------------|--------------------------------|
| A   | 17% | 9%                            | 9%                             | 7% (38 comp.)    | 7%                             |
| B   | 12% | 9%                            | 12%                            | 9% (40 comp.)    | 9%                             |
| S   | 14% | 8%                            | 10%                            | 8% (40 comp.)    | 8%                             |

Table 6.3: The best RMSEs obtained by the GMM-based mapping approach with different GMMs and MLR approach.

| (cm)      | MLR  | Supervised GMM (3 components) | Supervised GMM (10 components) | Unsupervised GMM | Semi-supervised GMM(35 comp.) |
|-----------|------|-------------------------------|--------------------------------|------------------|-------------------------------|
| A         | 0.56 | 0.40                          | 0.39                           | 0.35 (38 comp.)  | 0.36                          |
| B         | 0.15 | 0.13                          | 0.15                           | 0.12 (40 comp.)  | 0.12                          |
| $S(cm^2)$ | 0.55 | 0.49                          | 0.55                           | 0.48 (40 comp.)  | 0.48                          |

\* 'comp.' is the abbreviation of the 'component'.

## 6.4. Evaluation of GMM-based mapping approach for estimating hand positions

### 6.4.1. Evaluation of GMM-based mapping approach with supervised trained GMM for estimating hand positions

As mentioned before, there are two supervised training methods in terms of the estimation the hand positions: one is based on the 10 vowels; the other is based on the five hand positions defined in the CS. The supervised training method is described in detail in the section 6.2.2. The two training methods are evaluated separately and each supervised training processing is followed by the MMSE and MAP regression approaches.

#### 6.4.1.1. Evaluation of GMM-based mapping approach with supervised trained GMM based on 10 vowels

Figure 6.24 and Figure 6.25 show respectively the RVARs and RMSEs of the training and test data for estimating the hand position based on the supervised trained GMM with the 10 Gaussians. We can see that the RVARs and RMSEs decrease with increasing the dimension of the GMM, i.e. the number of the predictors. The test results increase slightly in comparison with the training results (the increment is about 11% for x and 6% for y in terms of RVAR, 0.72 cm for x and 0.87 cm for y in terms of RMSE).

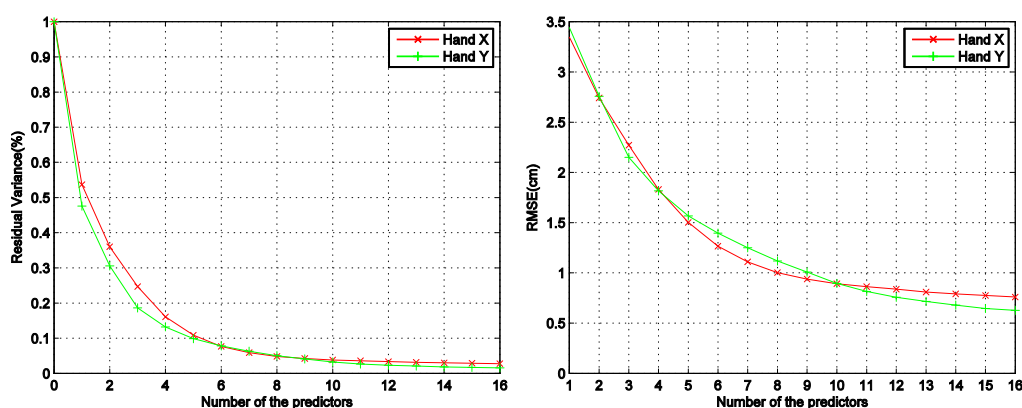


Figure 6.24: Average RVARs (on left) and RMSEs (on right) of the training data based on the supervised trained GMM with 10 Gaussians for estimating the hand position in function of the number of predictors. MMSE is used as the regression criterion.



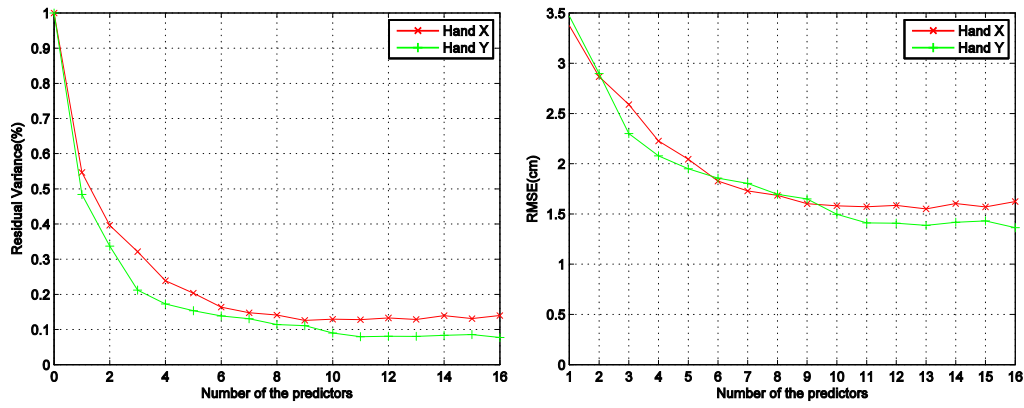


Figure 6.25: Average RVARs (on left) and RMSEs (on right) of the test data based on the supervised trained GMM with 10 Gaussians for estimating the hand position in function of the number of predictors. MMSE is used as the regression criterion.

Figure 6.26 and Figure 6.27 compare RVAR of the GMM-based mapping approach with the direct and indirect MLR approaches in the training and test processing respectively. We can see that either in the training or test processing the RVAR obtained by GMM with 10 supervised trained Gaussians is significant lower than the ones obtained by the direct or indirect MLR approach. The evaluation results indicate that the GMM-based mapping approach performs much better than the MLR approach for estimating the hand position.

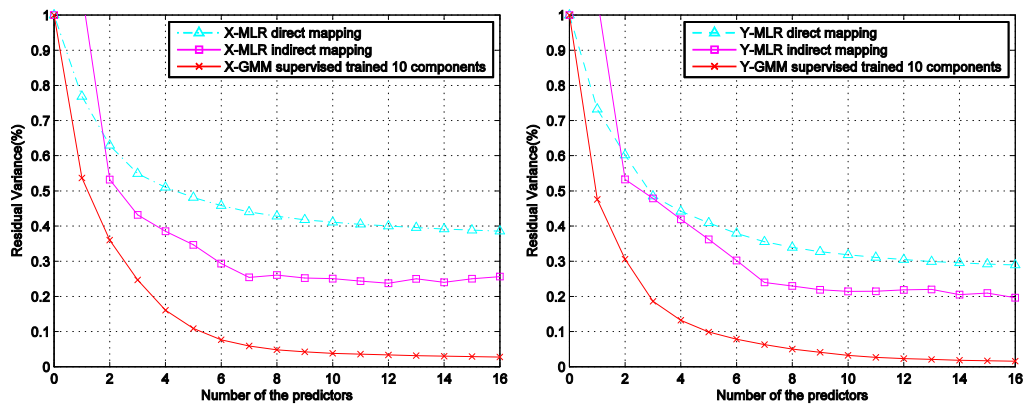


Figure 6.26: Average RVARs of the training data for estimating X (on left) and Y (on right) coordinates in function of the number of predictors based on the direct and indirect MLR approaches and the GMM-based mapping approach with 10 supervised trained Gaussians in the regression criterion of MMSE.

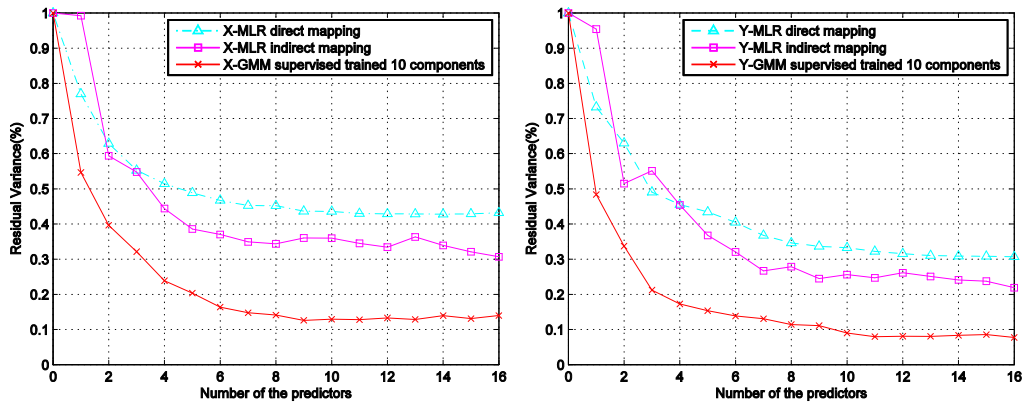


Figure 6.27: Average RVARs of the test data for estimating  $X$  (on left) and  $Y$  (on right) coordinates in function of the number of predictors based on the direct and indirect MLR approaches and the GMM-based mapping approach with 10 supervised trained Gaussians in the regression criterion of MMSE.

#### 6.4.1.2. Evaluation of GMM-based mapping approach with supervised trained Gaussians based on 5 hand positions in CS

Figure 6.28 and Figure 6.29 show the estimation results based on the supervised trained GMM with 5 Gaussians corresponding to the 5 hand positions defined in the CS. We can see that the RVARs and RMSEs decrease as increasing the dimension of the GMM, namely increasing the number of the predictors.

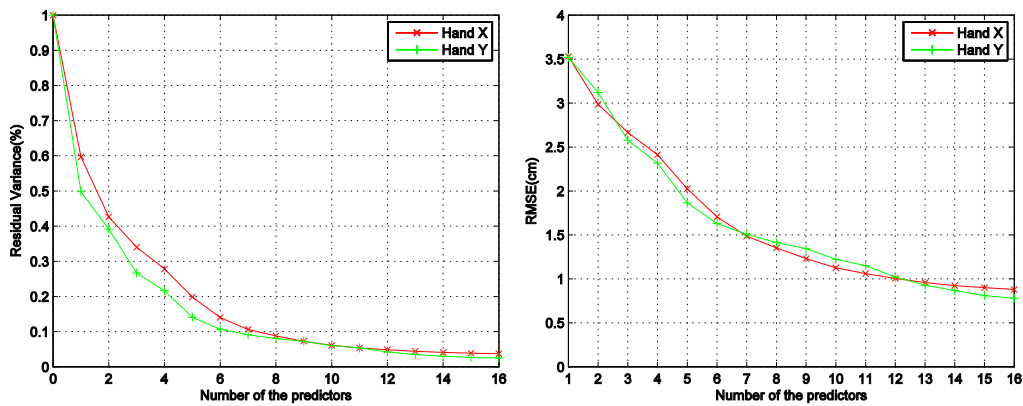


Figure 6.28: Average RVARs (on left) and RMSEs (on right) of the training data based on the supervised trained GMM with 5 Gaussians for estimating the hand position in function of the number of predictors. MMSE is used as the regression criterion.

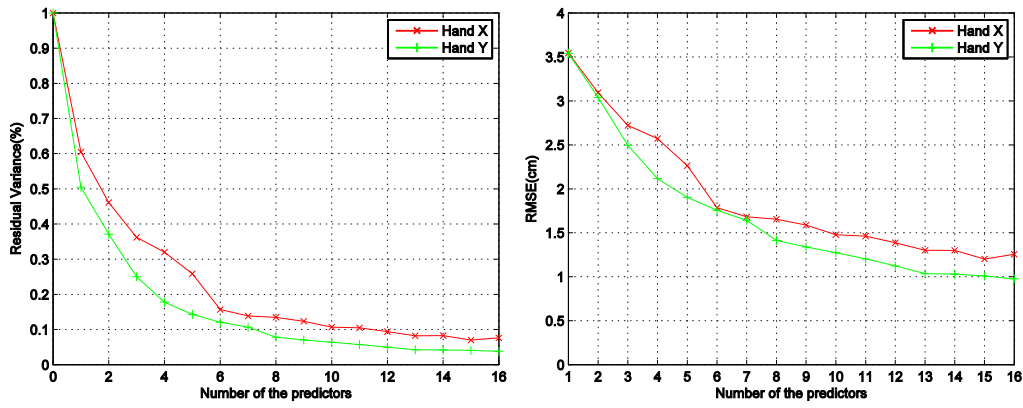


Figure 6.29: Average RVARs (on left) and RMSEs (on right) of the test data based on the supervised trained GMM with 5 Gaussians for estimating the hand position in function of the number of predictors. MMSE is used as the regression criterion.

Figure 6.29 shows that the RVARs and RMSEs of the test data are only slightly higher in comparison with the training results, which indicate that the GMM-based mapping approach based on the supervised trained 5 Gaussians is robust.

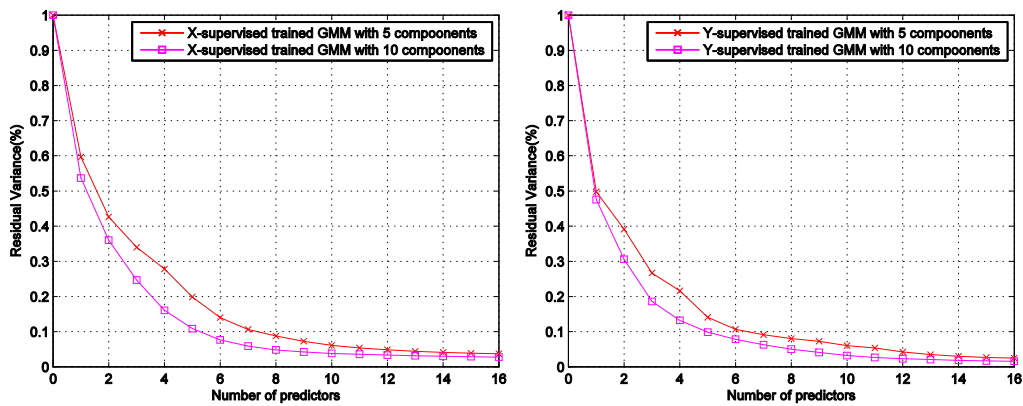


Figure 6.30: Average RVARs of the training data for estimating X (on left) and Y (on right) coordinates in function of the number of predictors based on supervised trained GMM with 5 Gaussians and supervised trained GMM with 10 Gaussians.

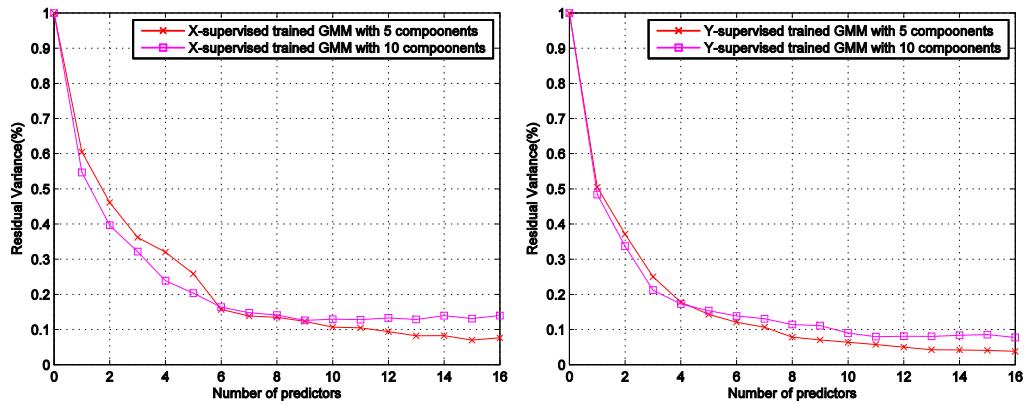


Figure 6.31: Average RVARs of the training data for estimating  $X$  (on left) and  $Y$  (on right) coordinates in function of the number of predictors based on supervised trained GMM with 5 Gaussians and supervised trained GMM with 10 Gaussians.

Figure 6.30 compares the RVARs of the supervised trained GMM having 5 Gaussians with the one having 10 Gaussians for estimating the hand position on the training data. It shows that when the dimension of the Gaussians is low, the mapping performance of the GMM with 10 Gaussians is better than the one with 5 Gaussians. This is because more Gaussians can help to classify the individuals properly when the dimension of GMM is low. However, when the dimension of GMM is enough for classifying properly (for example, 12 dimension shown in the Figure 6.30), the estimation performances based on the two GMMs are almost the same. While in the test processing shown in the Figure 6.31, the RVARs obtained by the supervised trained GMM with 5 Gaussians is better than the results obtained by the supervised trained GMM with 10 Gaussians. This is due to the over-fitting problem, which produces more misclassifications in the test processing degrading the estimation performance of the model (see Figure 6.32). Thus the supervised trained GMM with 5 Gaussians is more robust than the supervised trained GMM with 10 Gaussians for estimating the hand position.

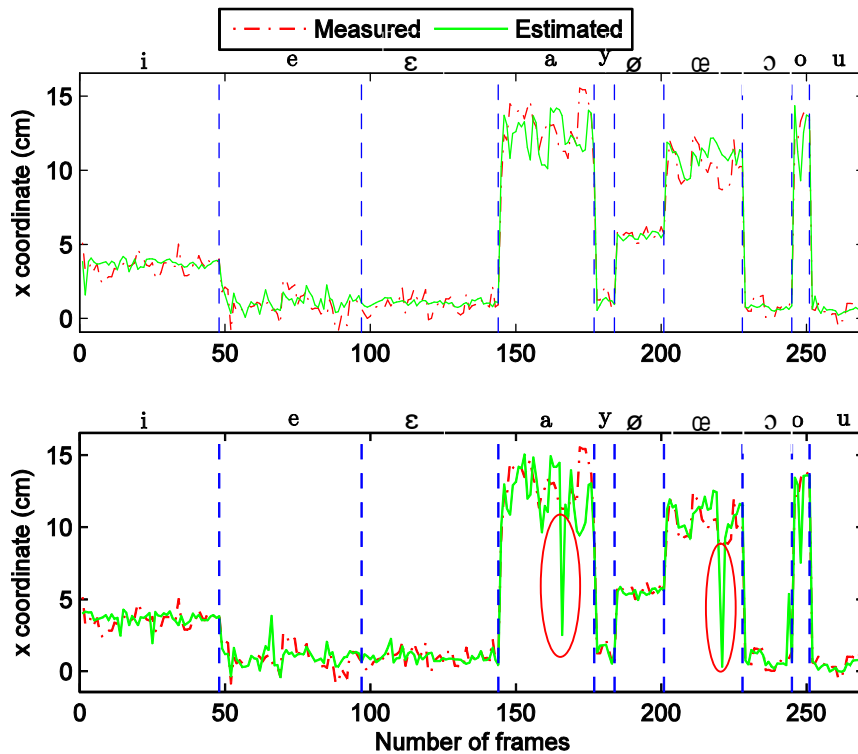


Figure 6.32: The estimated value of X coordinates of hand positions obtained by the supervised trained GMM with 5 Gaussians (the upper plan) and 10 Gaussians (the lower plan). The red dotted line denotes the measured value and the green solid line denotes the estimated value. The red circles denote the misclassified individuals of the supervised trained GMM with 10 Gaussians.

Figure 6.33 and Figure 6.34 compare the regression criterion MMSE with MAP based on the supervised trained GMM with 5 Gaussians. We can see that both in the training and test processing, the performance of MAP regression criterion is almost the same to MMSE when the number of predictors is higher than 6. The results obtained by the MMSE regression criterion are used as the initial parameters for the MAP regression criterion, which accelerate the classification processing. Meanwhile, the initial parameters can also aggravate the misclassification which results in the larger deviation (see Figure 6.35). In the Figure 6.35, we can see that the minor errors (marked by the red circles in the figure) obtained by the MMSE could be eliminated in the MAP. Meanwhile, some deviations (marked by the blue circles in the figure) may be amplified.

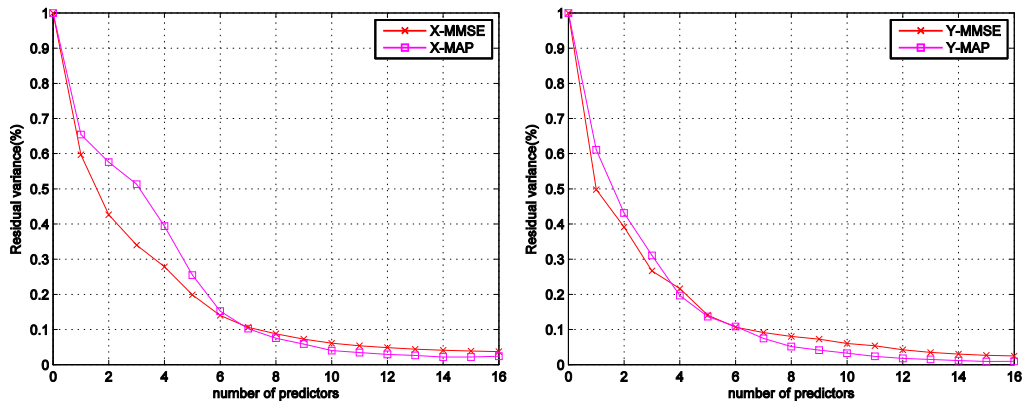


Figure 6.33: Average RVARs of the training data for estimating  $X$  (on left) and  $Y$  (on right) coordinates in function of the number of predictors based on supervised trained GMM with 5 Gaussians in the regression criteria MMSE and MAP respectively.

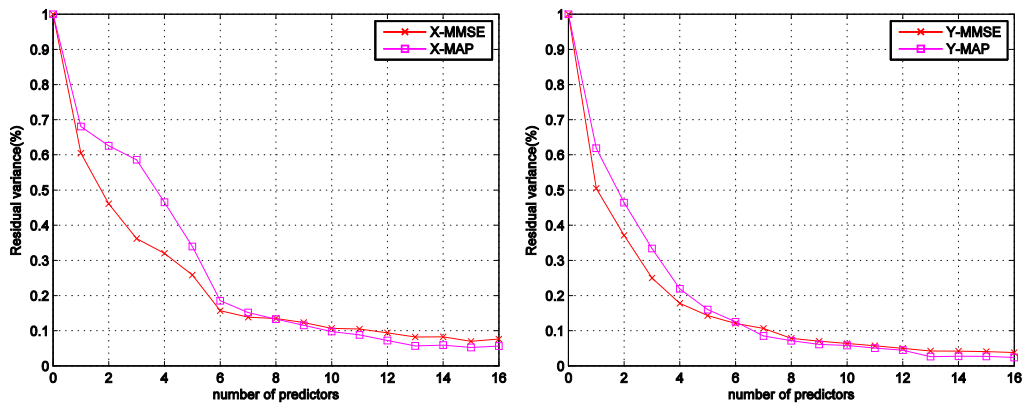


Figure 6.34: Average RVARs of the test data for estimating  $X$  (on left) and  $Y$  (on right) coordinates in function of the number of predictors based on supervised trained GMM with 5 Gaussians in the regression criteria MMSE and MAP respectively.

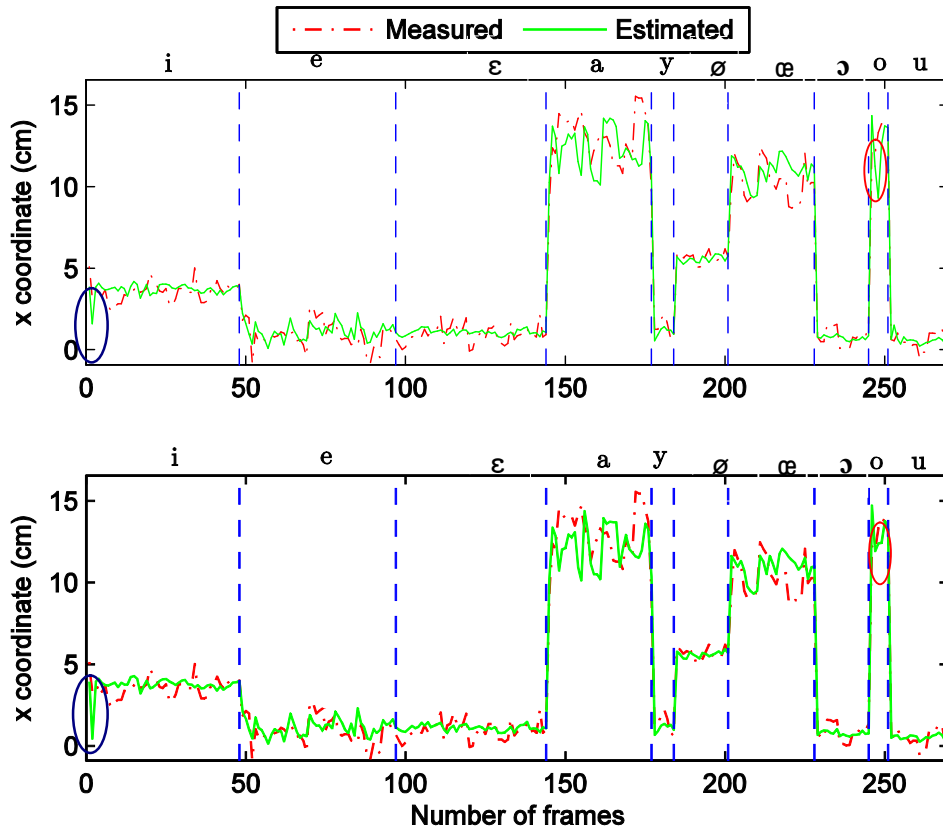


Figure 6.35: The estimated value of  $X$  coordinates of hand position obtained by the supervised trained GMM with 5 Gaussians under the criteria of MMSE (the upper plan) and MAP (the lower plan). The red dotted line denotes the measured value and the green solid line denotes the estimated value. The red circles denote the misclassification which is eliminated in the MAP method while the blue circles denote the misclassification which is aggravated in the MAP method.

#### 6.4.2. Evaluation of GMM-based mapping approach with unsupervised trained GMM for estimating hand positions

In the previous section, we discussed about the effect of the dimension of the GMM for estimating the hand position based on the supervised trained GMM. In this section, we discuss about the effect of the number of the Gaussians of GMM for estimating the hand position based on the unsupervised trained GMM. We fix the dimension of the GMM to 18 (16-dimension predictors and 2-dimension coordinates). The Gaussians of the GMM are trained automatically by the unsupervised training method-EM. The common covariance matrix is used to avoid the singular matrix.

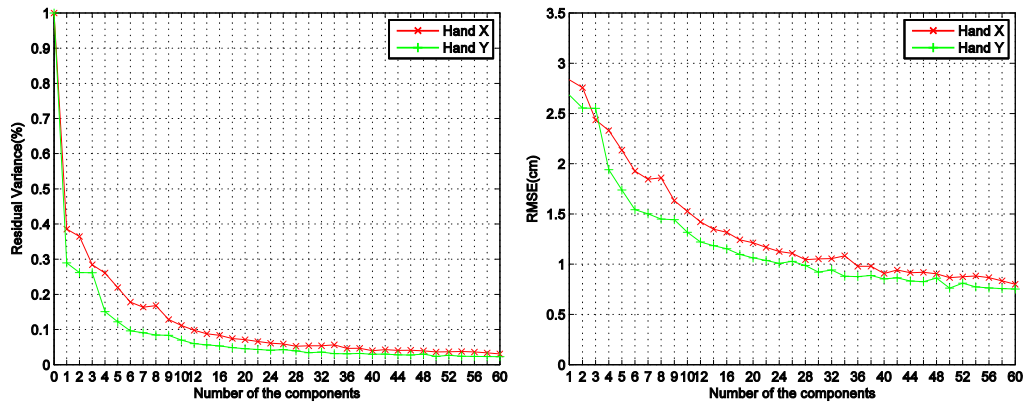


Figure 6.36: Average RVARs (on left) and RMSEs (on right) of the training data based on unsupervised trained GMM for estimating the hand position in function of the number of Gaussians. MMSE is used as the regression criterion.

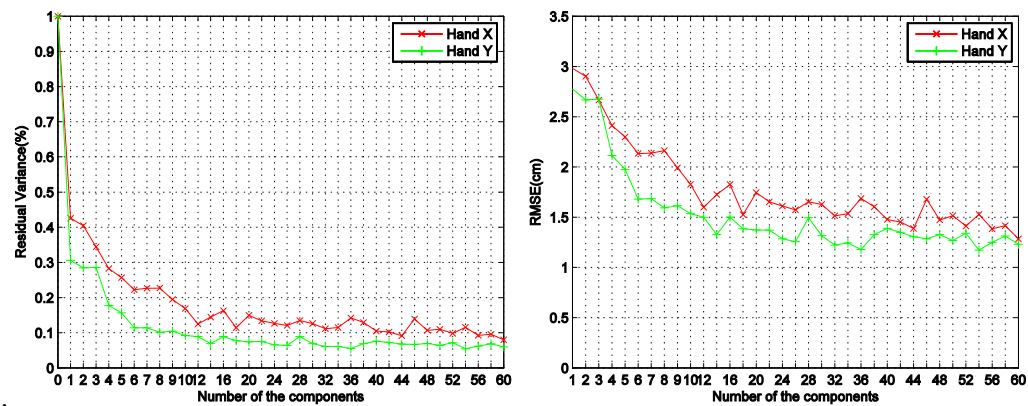


Figure 6.37: Average RVARs (on left) and RMSEs (on right) of the test data based on unsupervised trained GMM for estimating the hand position in function of the number of Gaussians. MMSE is used as the regression criterion.

Figure 6.36 and Figure 6.37 show the RVARs and RMSEs of the training and test data separately. We can see that the RVARs decrease as the number of Gaussians increases until 40 for coordinate X and 28 for coordinate Y. As in the case of the lip parameters estimation, the GMM-based mapping approach with one Gaussian is identical to the MLR approach for estimating the hand position. We can see that the RVARs and RMSEs obviously decrease both in training and test processing with increasing the number of the Gaussians. This proves that GMM can describe the joint probability distribution of the data more precisely with more Gaussians.



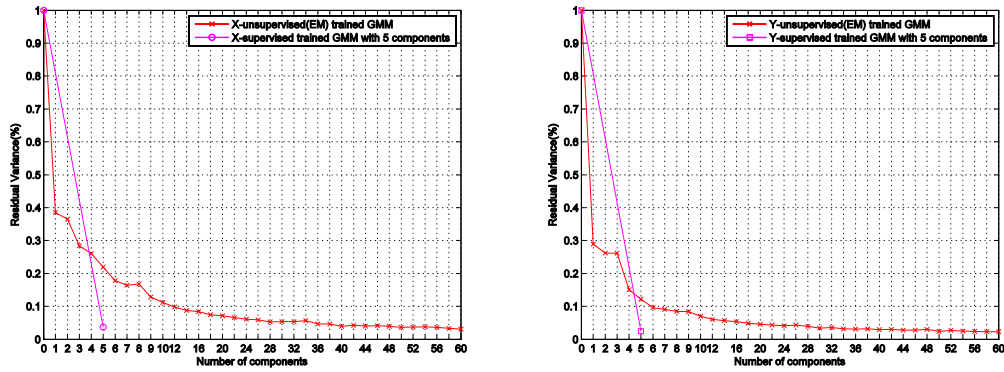


Figure 6.38: Average RVARs of the training data for estimating X (on left) and Y (on right) coordinates in function of the number of Gaussians based on supervised trained GMM with 5 Gaussians and unsupervised trained GMM..

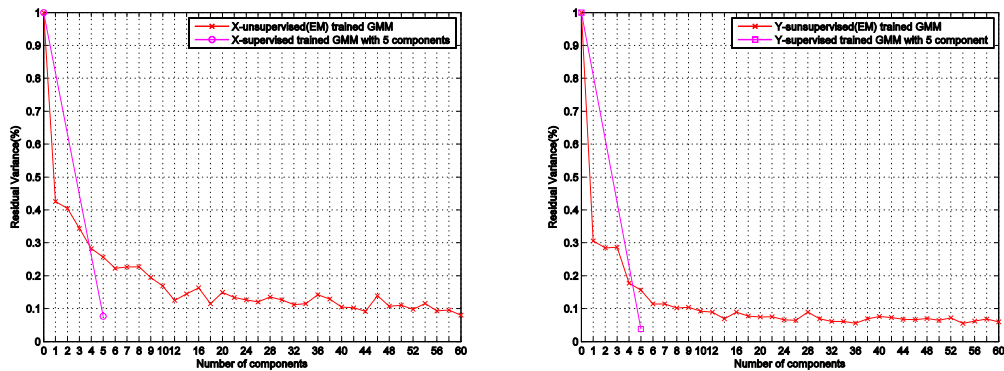


Figure 6.39: Average RVARs of the test data for estimating X (on left) and Y (on right) coordinates in function of the number of Gaussians based on supervised trained GMM with 5 Gaussians and unsupervised trained GMM..

Figure 6.38 and Figure 6.39 compare the RVAR of the unsupervised trained GMM with the supervised trained GMM with 5 Gaussians for estimating the hand position both on the training and test data. We can see that the estimation based on the supervised trained GMM with 5 Gaussians shows very high efficiency in comparison with the unsupervised trained GMM. The RVARs obtained by the supervised trained GMM with only 5 Gaussians are comparable to the best values of the unsupervised trained GMM with more than 54 Gaussians both for X and Y coordinates in the training processing, as it does in the test processing. It proves that a priori knowledge come from the CS rules can lead to a very effective and robust clustering for

estimating the hand position. Figure 6.40 and Figure 6.41 explain the reason why the efficiencies of the two GMMs are different. In the Figure 6.40 we can see that the 5 Gaussians trained by the supervised training method are clustered correctly without the effect of the outliers. However, if we use unsupervised training method to train 5 Gaussians, the result shown in Figure 6.41 are not as good as the supervised training method (for example, the cluster 4 in the Figure 6.41). This is due to the fact that the unsupervised training method depends completely on the distribution of the data, which is probably affected by the outliers. However, the inappropriate clustering can be adjusted gradually by increasing the number of the Gaussians so that the mapping performance can be improved gradually. Therefore the unsupervised training method can explore the underlying Gaussians of the data but at the cost of the low convergence efficiency.

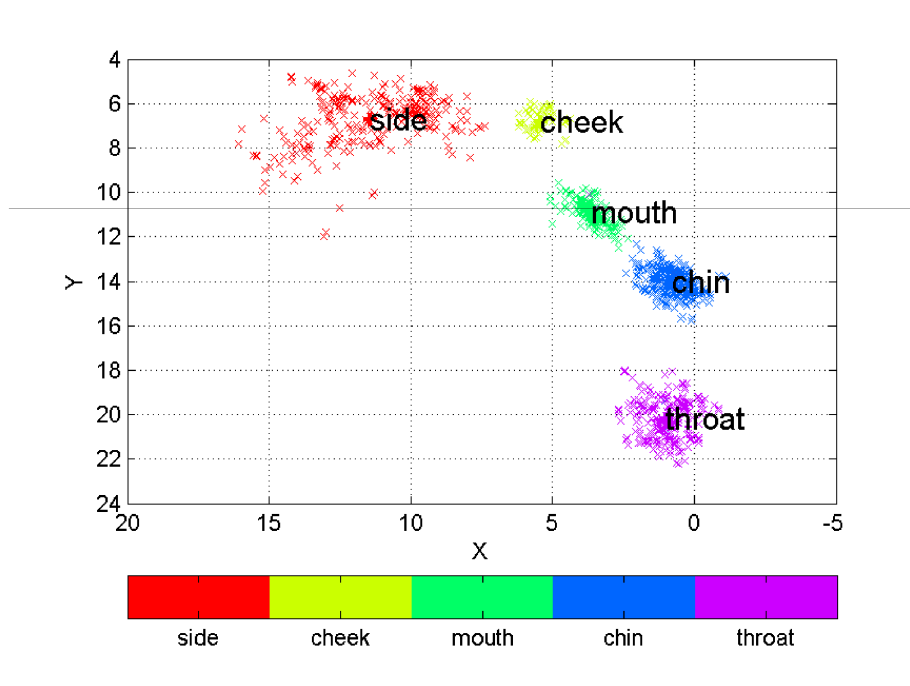


Figure 6.40: The supervised trained GMM with 5 Gaussians projects on the  $(X, Y)$  coordinates plan. The five different colors indicated the different Gaussians corresponding to the five hand positions defined in the CS: side, cheek, mouth, chin and throat.

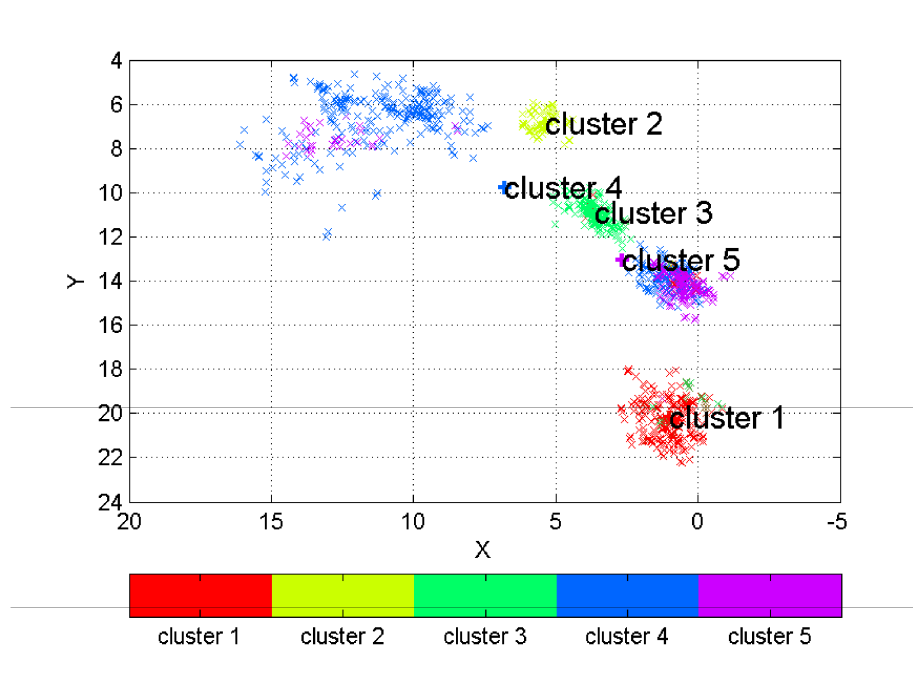


Figure 6.41: The unsupervised trained GMM with 5 Gaussians projects on the  $(X, Y)$  coordinates plan. The five different colors indicate the five different Gaussians trained automatically by EM algorithm.

### 6.4.3. Evaluation of GMM-based mapping approach with semi-supervised trained GMM for estimating hand positions

The semi-supervised training method aims to include the advantages of the supervised training method and unsupervised training method. Thus we combine the two methods into the semi-supervised training method. Firstly we gather the data into five groups according to the five hand positions of CS. Then we use the unsupervised training method to train the Gaussians inside each group. We fix the dimension of the GMM to 18 (16 predictors and 2 dimension coordinates) and evaluate the estimation performance by changing the number of the Gaussians inside each group. Table 6.4 lists the configuration of the number of the Gaussians belonging to the groups in the semi-supervised training process.

Table 6.4: The configuration of the number of the Gaussians belonging to the groups in the semi-supervised training process.

| Group<br>[a,o,œ] | Group<br>[ø] | Group<br>[i] | Group<br>[ε,u,ɔ] | Group<br>[y,e] | total<br>components |
|------------------|--------------|--------------|------------------|----------------|---------------------|
| 1                | 1            | 1            | 1                | 1              | 5                   |
| 2                | 2            | 2            | 2                | 2              | 10                  |
| 3                | 1            | 1            | 3                | 2              | 10                  |
| 4                | 1            | 3            | 6                | 5              | 19                  |
| 5                | 2            | 5            | 7                | 5              | 24                  |
| 6                | 3            | 6            | 8                | 6              | 29                  |
| 7                | 3            | 7            | 8                | 7              | 32                  |
| 8                | 3            | 8            | 9                | 8              | 36                  |

We can see that in the training processing as shown in the Figure 6.42, the supervised trained GMM with 5 Gaussians (corresponding to the first item in the figures) performs so efficiently that in fact there is no much space for the improvement. The results of the test processing (Figure 6.43) don't show any improvement while increasing the number of the Gaussians. Generally speaking, the clustering inside the groups does not change the estimation performance apparently in comparison with the supervised trained GMM with 5 Gaussians. This is due to using a common covariance matrix of the Gaussians for estimating the hand position. If we use the arbitrary matrices instead of the common covariance matrix for estimating the hand position, it will produce the singular matrix and over-fitting problem as the case in the lip parameter estimation.

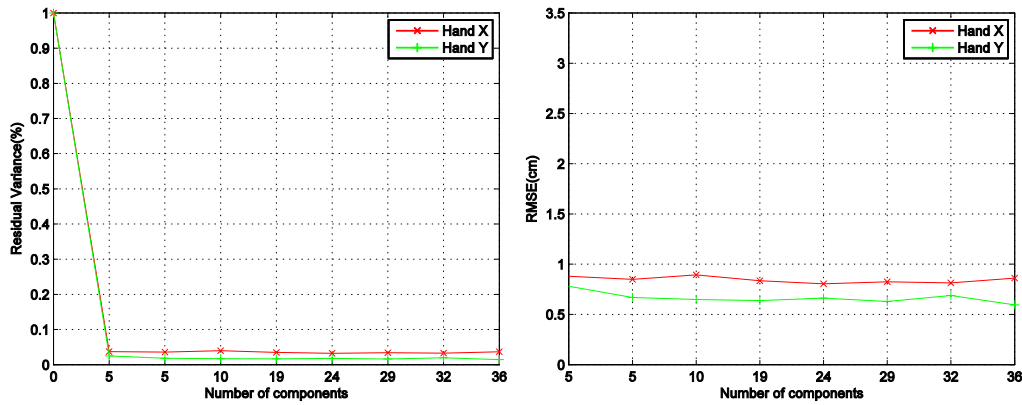


Figure 6.42: Average RVARs (on left) and RMSEs (on right) of the training data based on the semi-supervised trained GMM for estimating the hand position in function of the number of Gaussians. MMSE is used as the regression criterion.

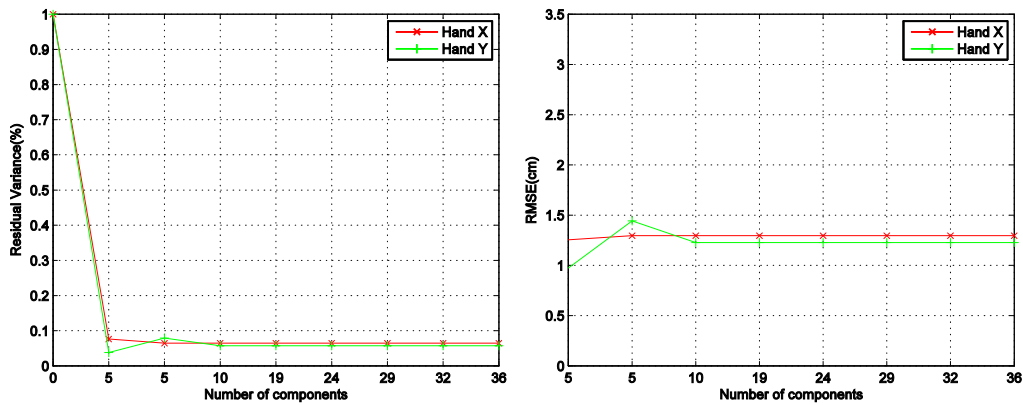


Figure 6.43: Average RVARs (on left) and RMSEs (on right) of the test data based on the semi-supervised trained GMM for estimating the hand position in function of the number of Gaussians. MMSE is used as the regression criterion.

#### 6.4.4. Discussion about the GMM-based classification method and the GMM-based mapping approach for estimating hand position

The definition of the five hand positions in CS inspires us using the GMM-based classification method to estimate the hand positions. In the GMM-based classification method, each Gaussian is corresponding to a class. The five Gaussians as defined in the supervised training method are used as five classes for classifying the source data  $\mathbf{x}$ , i.e. the predictor derived from PCA of the spectral parameter. Then the a posteriori probability is used as the discriminant function. The classifier decision is obtained as the class/Gaussian with the maximum a posteriori probability given the observation.

Figure 6.44 shows the comparison of the GMM-based classification method and the GMM-based mapping approach with 5 supervised Gaussians in the regression criteria of MMSE and MAP in the training stage. Figure 6.45 shows the same comparison in the test stage.

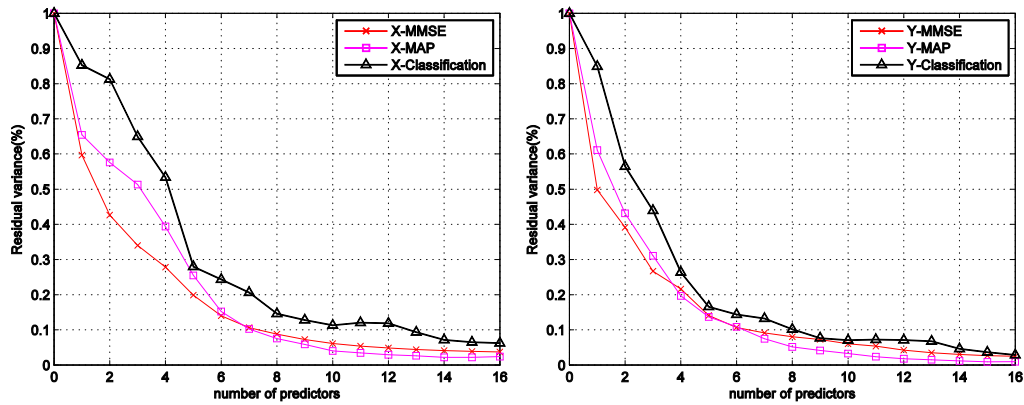


Figure 6.44: Average RVARs of X (on left) and Y (on right) of the training data in function of the number of predictors based on GMM-based classification method and the GMM-based mapping approach with 5 supervised Gaussians in the regression criteria of MMSE and MAP respectively.

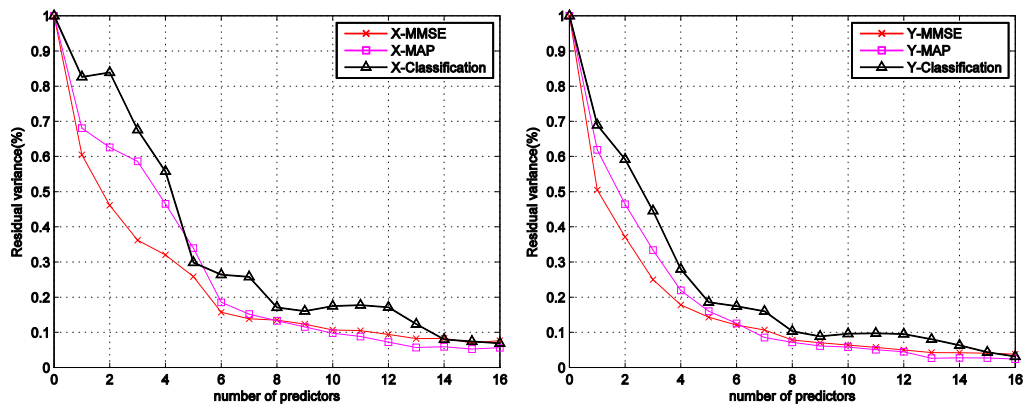


Figure 6.45: Average RVARs of X (on left) and Y (on right) of the test data in function of the number of predictors based on GMM-based classification method and the GMM-based mapping approach with 5 supervised Gaussians in the regression criteria of MMSE and MAP respectively.

The comparison results shown in the above figures suggest that when the dimension of the GMM is low (the number of the predictors less than 14 for X and 15 for Y in terms of the test processing) the performance of the GMM-based mapping approach is

superior to the GMM-based classification method. This is because the GMM-based classification method produces many misclassifications when the dimension of GMM is low. When the dimension of the GMM is high, the performances of the two methods are similar.

Although the performances of GMM-based mapping approach and the classification method are very close, the GMM-based mapping approach gives different result with the classification method in terms of the perception of the CS. Because several deviated individuals obtained by the GMM-based mapping approach (See Figures 6.46, Figure 6.47) are still near the target hand position. However, the misclassified individuals obtained by the classification method definitely locate in the wrong hand position (see Figure 6.47, Figure 6.48). That is to say, people can probably understand the errors obtained by the GMM-based mapping approach. While the errors caused by the classification method are difficult to be understood.

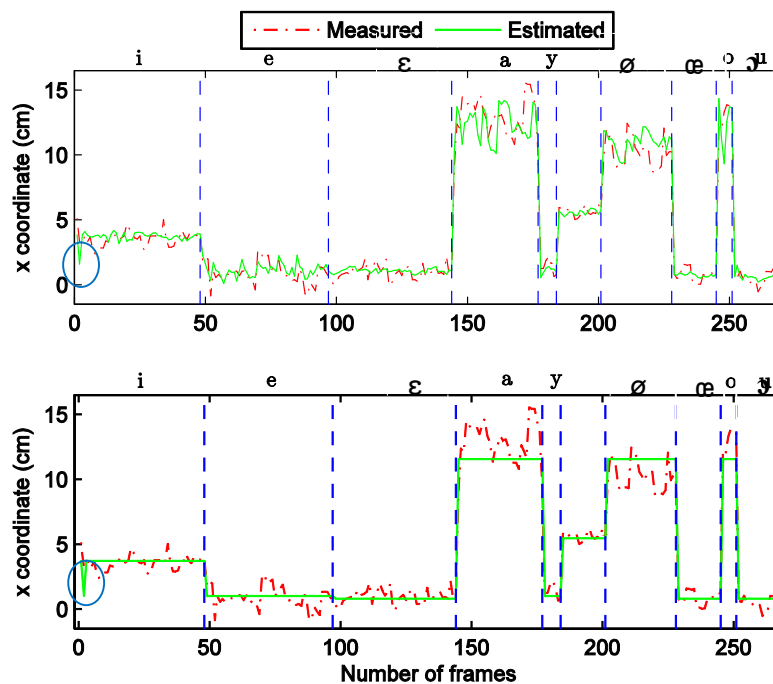


Figure 6.46: The estimated value of X coordinates of hand positions based on the GMM-based mapping approach (the upper plan) and the GMM-based classification method (the lower plan). The red dotted line denotes the measured value and the green solid line denotes the estimated value. The blue circles indicate the errors obtained by GMM-based mapping approach and classification method respectively.

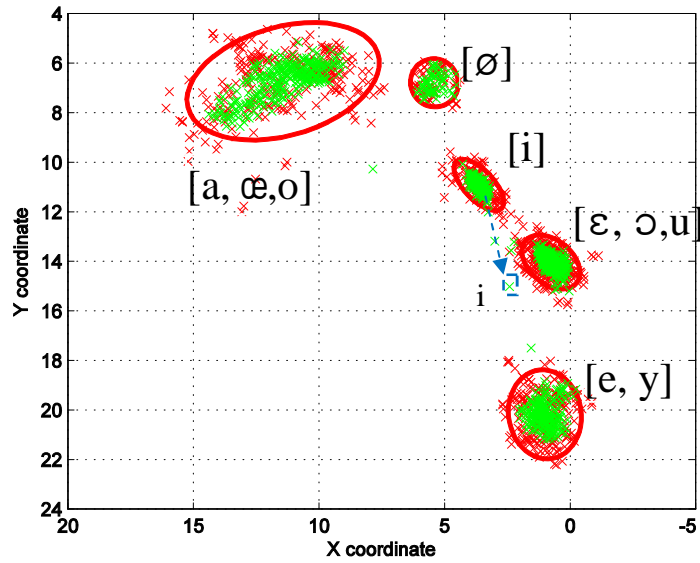


Figure 6.47: The estimated hand positions obtained by the GMM-based mapping approach. The reference positions are denoted in the red crosses and the estimated values are denoted in the green crosses. The blue arrow and the blue square with dotted line denote the deviation of a estimated hand position corresponding to the vowel [i].

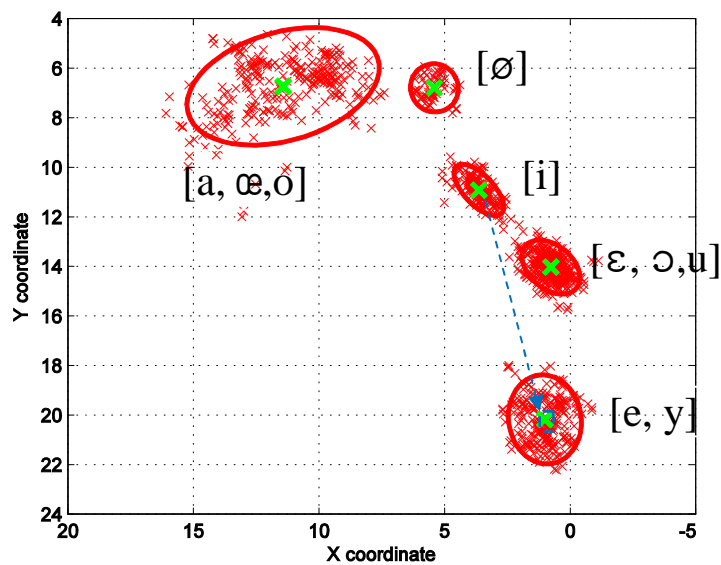


Figure 6.48: The estimated hand positions obtained by the GMM-based classification method. The reference positions are denoted in the red crosses and the estimated values are denoted in the green crosses. The blue arrow and the blue square with dotted line denote that a hand position corresponding to the vowel [i] is misclassified as a hand position corresponding to the vowel [e].



Table 6.5 and Table 6.6 list the best evaluation results (RVARs and RMSEs) obtained by the direct and indirect MLR approach and GMM-based mapping approach with different GMMs in the regression criterion MMSE..

Table 6.5: The best RVARs obtained by the GMM-based mapping approach with different GMMs and MLR approach.

| (%) | MLR direct | MLR indirect | Supervised GMM (5 components) | Supervised GMM (10 components) | Unsupervised GMM | Semi-supervised GMM |
|-----|------------|--------------|-------------------------------|--------------------------------|------------------|---------------------|
| X   | 42.57%     | 30.67%       | 7.65%                         | 13.97%                         | 7.98% (60 comp.) | 5.07% (19 comp.)    |
| Y   | 30.59%     | 21.88%       | 3.78%                         | 7.73%                          | 5.48% (54 comp.) | 3.78% (5 comp.)     |

\* 'comp.' is the abbreviation of the 'component.'

Table 6.6: The best RMSEs obtained by the GMM-based mapping approach with different GMMs and MLR approach.

| (cm) | MLR direct | MLR indirect | Supervised GMM (5 components) | Supervised GMM (10 components) | Unsupervised GMM | Semi-supervised GMM |
|------|------------|--------------|-------------------------------|--------------------------------|------------------|---------------------|
| X    | 2.98       | 2.50         | 1.25                          | 1.62                           | 1.28(60 comp.)   | 1.17 (19 comp.)     |
| Y    | 2.77       | 2.37         | 0.98                          | 1.36                           | 1.17(54 comp.)   | 0.98 (5 comp.)      |

## 6.5. Discussion of the approaches used for estimating the lip parameters and hand position

After our hard working on mapping the acoustic spectrum to the lip parameters and the hand positions, we can see that it is much more difficult to estimate the hand position than lip parameters from the spectral parameters. Several different approaches, from the direct MLR approach to the sophisticate GMM-based mapping approach, have been tested to solve this problem. Actually the source of the difficulty is that there is no relationship between the hand position and the spectral parameters unlike the case of the lips as a vocal articulatory with corresponding acoustic consequences. More specifically, there are two key points of the meaning of “no relationship”: (1) there is no structural topological relationship between the acoustic space and the hand position space. That is to say, two closed vowels in the acoustic space may be very far in the hand position space, such as the vowel [e] and [i]. On

the contrary, two far vowels in the acoustic space may be corresponding to the same hand position, such as vowel [a] and [o]. It indicates that the two spaces have totally different topology structure since the hand position is determined by the rules of CS but not the acoustic parameters. This is the real reason why we obtain a large RVAR when we use the direct MLR approach to map the spectral parameters to the hand position. And this is also the motivation to introduce the intermediate space where the hand positions are relocated aiming to simulate a similar topology structure of the acoustic space. It proves that the RVARs of the estimated values of hand positions decrease significantly in the intermediate space benefiting from the similar topology structure. But the classification problem introduced in the intermediate space fails to maintain the good performance in the original space; (2) there is no relationship of the variance within group between the acoustic space and the hand position space. That is to say, the tiny variation of the sound will not consequently change the hand position of the speaker even if in proportion. Indeed the hand position around the center within group is random from person to person. Thus it is impossible to establish a global linear relationship by the MLR approach or even a local linear relationship by the GMM to predict the movement of the hand around the center within group by the acoustic spectrum.

In order to have a further understanding and compare the different approaches used in our work, we study a continuous transition achieved by a continuous linear interpolation in the acoustic spectral parameter (i.e. the mixture of MFCCs and LSP) between the vowels [a] and [i]. We project the 16-dimension acoustic spectral parameters onto the first two PCA components to view the continuous linear interpolation in the acoustic space (see Figure 6.49). In fact there are many ways to go to vowel [i] from vowel [a] in the acoustic space when we chose different start and end, but here we only present one of them as an example to show the corresponding transition obtained by the different mapping approaches.

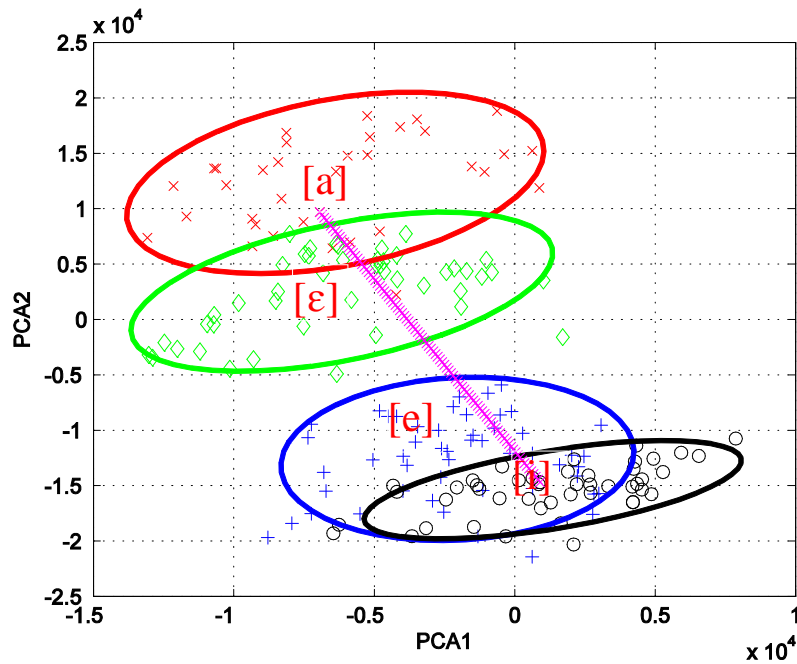


Figure 6.49 The linear interpolation in the acoustic space between vowels [a] and [i].

The corresponding transitions of the estimated lip parameter and hand positions are shown in Figure 6.50 and Figure 6.51. For the MLR approach, the figures present a reasonable linear relationship between the linear interpolation spectral parameters and the estimated hand position or lip parameter. For the GMM-based mapping method (in the MMSE regression criterion, with 5 Gaussians corresponding to the five hand position in CS in the case of hand position estimation and 10 Gaussians corresponding to the ten vowels in the case of lip parameter estimation), the four stable phases in the transition are corresponding to the passing vowels ([a],[ε],[e],[i]) during the spectral parameters changing linearly from vowel [a] to [i] in the acoustic space. From the four stable phases, the GMM-based mapping approach shows classification-partitioning property which helps the approach to significantly decrease the RVAR in comparison with the MLR approach in the case of hand position estimation. However, unlike the GMM-based classification method which cannot project the variance of the source data at all (since the GMM-based classification method always projects the source data onto the mean values of target groups as shown in the Figure 6.50 and Figure 6.51, there is no variance of the estimated values within group), the GMM-based mapping approach can still reflect the linear relationship locally between the source and target data as shown in the Figure 6.51. The effective local linear

regression of GMM-based mapping method enables the approach to improve the performance in comparison with the MLR approach in the case of lip parameters estimation. However in the case of the hand position estimation the local regression is meaningless due to the lack of relationship between the hand position and the acoustic spectral parameters (see Figure 6.50). In addition, we can see that the GMM-based mapping approach produces a smooth changing between the stable phases by weighting the contributions of Gaussians. This is different from the classification method. Therefore the GMM-based mapping approach can perform well thanks to the classification-partitioning property when the source and target data has “no relationship” such as the case of the hand position estimation; and also it can improve the performance by the local regression property when the source and target data has strong correlation such as the case of the lip parameter estimation.

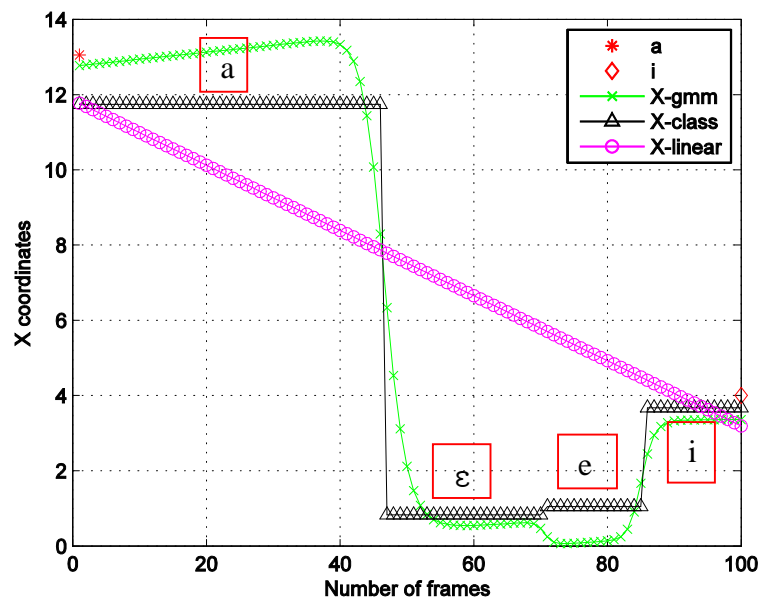


Figure 6.50 The continuous transition of X coordinates of hand position achieved by linear interpolation between the vowels [a] and [i].

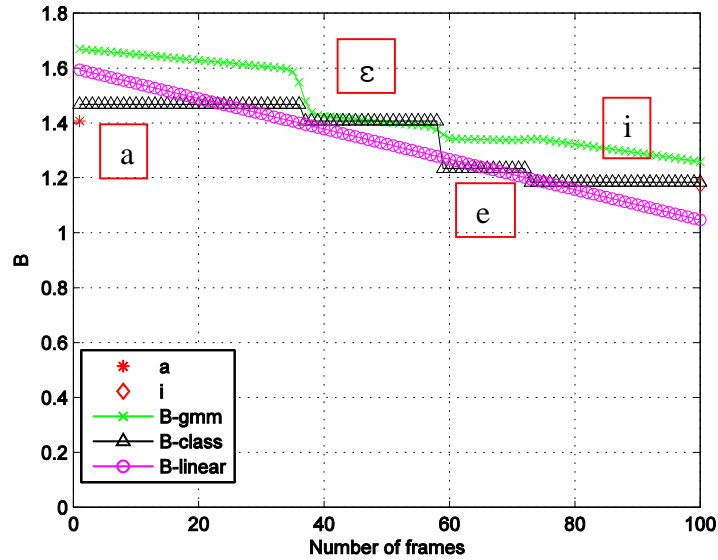


Figure 6.51 The continuous transition of lip parameter  $B$  achieved by linear interpolation between the vowels  $[a]$  and  $[i]$ .

## 6.6. Discussion of the residual variance obtained by the different methods

With both the properties of the classifier and the regression estimator, the GMM-based mapping method decreases the RVAR of the estimated value in comparison with the MLR approach. When the GMM-based mapping method more inclines to the classification method, the RVAR may be no longer appropriate for evaluating the approach since the errors probably cannot be understood at all even with a low RVAR. The results of speech perception tasks with CS users may be an alternative evaluation criterion. But note that in the GMM-based mapping method, the contributions of all the Gaussians are always considered and weighted to produce the final results. This is essentially different from the classification method.

## 6.7. Summary

In this chapter we applied GMM-based mapping approach for estimating the lip parameters and hand position. GMM is trained on the supervised, unsupervised trained (EM algorithm) and semi-supervised training method in the view of the machine learning theory. The mapping approach based on the supervised trained

GMM shows a high efficiency and a good robustness benefiting from the a priori knowledge in comparison with the unsupervised trained GMM which may be affected more by the outliers due to an excessive dependency on the data itself. But the supervised training method also limited by requiring a priori knowledge such as the label information of each individual. While the unsupervised training method without requiring a priori knowledge, can explore the latent Gaussians underlying the training data automatically. But the disadvantages of this training method are the low efficiency (such as using much more Gaussians to reach the same results as the supervised training method) and the relative complicated training processing. The semi-supervised training method which is proposed to include advantages of the efficiency and robustness of the supervised training method and the flexibility of the unsupervised training method does not show apparent improvement in comparison with the other two methods. This is due to using the common covariance matrix which weakens the individual fitting ability of each Gaussian instead of arbitrary matrix in the estimation processing. The regression criteria of MMSE and MAP do not show great difference in this work.

In our work, we also indicate that the MLR approach which conducts the linear regression in the sense of least square error on the overall data set is indeed a special case of GMM-based mapping approach with the unimodal Gaussian. Thus the GMM-based mapping approach can perform better than MLR with increasing the number of the Gaussians.

Finally, a continuous transition achieved by the linear interpolation between the vowels [a] and [i] in the acoustic space is introduced to compare the different mapping approaches used in this work. The MLR approach produces a continuous linear result with a larger dispersion especially in the case of the hand position estimation. This is because there is no relationship between the acoustic space and the hand position space. The classification method eliminates the variance of the source data within group and removes the continuous property of the source data. The GMM-based mapping approach presents classification-partition and regression properties at same time. The GMM-based mapping approach produces a smooth changing between the stable phases by weighting the contributions of Gaussians and the local regression

enables to locally project the variance of the source data. Due to the hand movement being no relationship to the spectral parameters within group, the local regression is meaningless of the GMM-based mapping method for estimating the hand position estimation. However, in the case of lip parameters estimation, the GMM-based mapping method shows effective local regression which enables the approach to improve the performance in comparison with the MLR approach. Therefore the GMM-based mapping approach can perform well thanks to the classification-partitioning property when the source and target data has “no relationship” such as the case of the hand position estimation; and also it can improve the performance by the local regression property when the source and target data has strong correlation such as the case of the lip parameter estimation.

## Chapter 7. Conclusion and perspectives

In this work we focus on the problem of mapping the acoustic spectral speech parameters to the hand position in CS and the accompanying lip shapes (lip width, lip height and lip area). This study is situated in the general framework of speech to CS conversion.

In terms of the acoustic-to-lip shapes mapping, the problem can be considered partly as a problem of acoustic-to-articulatory (A-to-A) mapping. But in terms of the acoustic-to-hand position, it is a new problem since there is no relationship between the hand position and the acoustic spectral parameters. We first applied the MLR approach in order to know the limit in terms of performance and also to understand why the linear mapping cannot offer a sufficient estimation precision. In order to use the speech spectral parameters to estimate the lip parameters and hand position, we first applied the PCA on the speech spectral parameters in order to remove the correlation between the spectral parameters. The best performance can be obtained by using the 16 predictors derived from PCA of the mixture spectral parameters of MFCCs and LSP in comparison with other spectral parameters.

By using the MLR approach, the lip parameters can be estimated with a favorable precision of 12% in terms of RVAR of the lip height parameter. But for the hand position estimation, the RVAR reaches only 39% in the best case.

The reasons for this poor performance are mainly: 1). there is no relationship between the hand position and the acoustic parameters of the speech signal; and 2). the hand positions for each vowel in CS are specially made to remove any ambiguity with the lip-reading. A direct consequence is: for two “acoustically near” vowels (for example, vowels [i] and [e]), their corresponding position in the hand space are very far! However, two far vowels in the acoustic space may share the same hand position, such as vowels [a] and [o]. After several simulation studies, it has been shown that the hand space and the acoustic spectrum space have totally different topology structure which cause the poor performance of the MLR approach for estimating the hand position.



In order to resolve this problem, we have introduced an intermediate space for estimating hand position, in which the hand position of each vowel is relocated following the rule: their relative location should be in coherence with their “acoustical position”. Consequently, the redistributed hand position has a similar topological structure with the acoustic spectrum space, i.e. the “acoustic triangle” composed by F1-F2 in formant space.

It proves that the performance of the MLR approach for estimating the hand position has improved in the intermediate space benefiting from the similar topology structure. However the classification problem in the intermediate space has failed to achieve the good performance in the original space. The Bayes LDA and QDA classifiers are both evaluated. By using the arbitrary covariance matrix instead of the common covariance matrix, the performance of QDA (with the classification score 0.90) is slightly superior to the LDA (with the classification score 0.87). Finally the RVAR of estimated hand position in original space has decreased to 32% for X coordinate and 21% for Y coordinate respectively.

All of these results show that even if a certain level of linearity is conserved, it is at the price of high variance of the linear estimator. When using the intermediate space, the problem of high RVAR is solved. But the necessity of using classifier for remapping causes some big classification errors. In order to continue to release the linearity constraints we introduced the mapping method based on GMM which is particularly suitable for modelling the distribution where the measurements arise from separate groups and the regression processing is implemented on the group of GMM instead of the overall population. This is exactly our case.

Due to its statistical nature and the number of parameters, the GMM-based mapping approach requires more complex training before the phase of regression than for the MLR case. In this work, the GMM was trained by the three methods in the view of the machine learning theory. The supervised training method is grouping the data according to the visemes of the target data (i.e. 3 visemes for lip parameters and 5 visemes for hand positions) or the 10 vowels. The supervised training method is efficient, simple and robust. But it is also constrained by acquiring a priori knowledge (i.e. label information of elements) before training. The unsupervised training method

EM overcomes the problem in supervised training method clustering the data automatically to form the Gaussians of GMM. However the convergence rate of the EM method is slow since the clustering is entirely dependent on the data itself. The outliers in the data would probably affect the process of clustering. The semi-supervised training method is proposed to include advantages of the supervised training method and the unsupervised training method. The semi-supervised does not show the obvious improvement in comparison with the other two training method. This is due to using the common covariance matrix which weakens the individual fitting ability of each Gaussian instead of arbitrary matrix in the estimation processing. When the GMM training processing finished, the regression method in the sense of MMSE or in the sense of MAP are applied based on the trained GMM.

In our work, it also indicates that the MLR mapping approach actually is the special case of GMM-based mapping approach with unimodal Gaussian. This point is shown in the evaluation results and proved in the regression equations in this work. The regression of GMM-based mapping approach is conducted locally on the Gaussians instead of the overall dataset. Thus the GMM-based mapping approach can improve the performance by increasing the number of the Gaussians in comparison with the MLR approach. In terms of the regression criteria, the MMSE and MAP did not show great differences in any of the mapping approach for estimating lip parameters and hand position.

Another problem we researched in this work is estimating the lip geometric features, i.e. the lip width, lip height and lip area, from the lip appearance-based features. The lip appearance-based feature vectors are obtained by the PCA of the DCT coefficients of the natural images of the speaker's mouth ROI. With the MLR model, we can use only 17 predictors accounted for just 2% of the 120 predictors to explain 95% of the total variance of the lip parameters B. This is due to the great concentration abilities of DCT and PCA. But the MLR approach also has limit for estimating the data close to zero. With the GMM-based mapping approach the estimation performance has been improved in comparison with the MLR approach in the case of target data being low. The estimation efficiency of the GMM-based mapping approach is also higher

than the MLR approach. However the GMM-based mapping approach also introduces the over-fitting problem when the dimension of GMM is too high.

In the future, the GMM-based mapping approach should be extended to the continuous situation. In our work, we already have a preliminary discussion on the performance of different mapping approaches in the continuous situation realized by the linear interpolation in the acoustic spectrum space. From the transitions achieved by the linear interpolation, we can see clearly the different properties of the MLR approach, GMM-based mapping approach and the GMM-based classification approach. The MLR always produces a linear continuous results but the performance depends on the linear correlation between the source and target. The GMM-based classification method obtains a step response locating the data in the center of each group of GMM. And the classification method cannot project the variance of the source data within group. The GMM-based mapping approach having both the properties of local regression and classification-partitioning can perform well in different cases. When there is strong linear correlation between the source and target data such as the case of the lip parameter estimation, the GMM-based mapping approach can improve the performance by the local regression property; when there is “no relationship” between the source and target data such as the case of the hand position estimation, it can work well thanks to the classification-partitioning property. Meanwhile, the transition between the phases of GMM-based mapping method is continuous due to weighting the contributions of all the Gaussians unlike the binary classification method. Thus it is reasonable to extend the GMM-based mapping approach to the continuous estimation problem which includes the vowels and consonants. Besides, we can also explore new approaches for estimating the CS components in the continuous context, for example the approaches proposed recently by (Zen et al., 2011; Zen et al., 2012; Shannon et al., 2013) based on the trajectory GMMs or trajectory HMMs in the domain of acoustic-to-articulatory inversion, speech synthesis, which consider the trajectory of the neighbor frames rather than the single frame in the estimating processing.

# Bibliography

- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433-459.
- Aboutabit, N. (2007). *Reconnaissance de la Langue Française Parlée Complétée (LPC): décodage phonétique des gestes main-lèvres*. (Ph.D thesis), Institut National Polytechnique de Grenoble-INPG.
- Adjoudani, A., & Benoit, C. (1996). On the integration of auditory and visual parameters in an HMM-based ASR. In D. Stork & M. Hennecke (Eds.), *Speechreading by humans and machines: NATO ASI series, series F: Computer and systems science*. Berlin: Springer, Vol. 150, pp. 461-472.
- Afify, M., Cui, X., & Gao, Y. (2007). Stereo-based stochastic mapping for robust speech recognition. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol.4, pp.IV-377-IV-380.
- Alegria, J., Charlier, B. L., & Mattys, S. (1999). The role of lip-reading and cued speech in the processing of phonological information in French-educated deaf children. *European Journal of Cognitive Psychology*, 11(4), 451-472.
- Aleksic, P., Potamianos, G., & Katsaggelos, A. (2005). Exploiting visual information in automatic speech processing. In A. C. Bovik (Ed.), *Handbook of Image and Video Processing*: Academic Press, pp. 1263-1289.
- Aleksic, P., Williams, J., Wu, Z., & Katsaggelos, A. (2002). Audio-visual speech recognition using MPEG-4 compliant visual features. *EURASIP Journal on Applied Signal Processing*, 2002(1), 1213-1227.
- Attina, V. (2005). *La Langue française Parlée Complétée: production et perception*. (Ph.D thesis), Institut National Polytechnique de Grenoble-INPG.
- Attina, V., Beautemps, D., & Cathiard, M.-A. (2002). Coordination of hand and orofacial movements for CV sequences in French Cued Speech. in *Proc. Seventh International Conference on Spoken Language Processing*, Denver, Colorado, USA, pp.1945-1948.
- Attina, V., Beautemps, D., Cathiard, M.-A., & Odisio, M. (2004). A pilot study of temporal organization in Cued Speech production of French syllables: rules for a Cued Speech synthesizer. *Speech Communication*, 44(1), 197-214.
- Barker, J., & Shao, X. (2009). Energetic and informational masking effects in an audiovisual speech recognition system. *IEEE Transactions on Speech and Audio Processing*, 17(3), 446-458.
- Beautemps, D., Cathiard, M.-A., Attina, V., & Savariaux, C. (2012). Temporal organization of Cued Speech production. In G. Bailly, P. Perrier & E. Vatikiotis-Bateson (Eds.), *Audiovisual speech processing*: Cambridge University, pp. 104-121.
- Beautemps, D., Girin, L., Aboutabit, N., Bailly, G., Besacier, L., Breton, G., Burger, T., Caplier, A., Cathiard, M.-A., Chêne, D., Clarke, J., Elisei, F., Govokhina, O., Jutten, C., Le, V.-B., Marthouret, M., Mancini, S., Mathieu, Y., Perret, P., Rivet, B., Sacher, P., Savariaux, C., Schmerber, S., Ségnat, J.-F., Tribout, M., & Vidal, S. (2007). Telma: Telephony for the hearing-impaired people. From models to user tests. in *Proc. ASSISTH'2007*, pp.201-208.

- Benoit, C., Lallouache, T., Mohamedi, T., Tseva, A., & Abry, C. (1991). Nineteen ( $\pm$ Two) French Visemes for Visual Speech Synthesis. *in Proc. The ESCA Workshop on Speech Synthesis*, Autrans, France, pp.253-256.
- Bernstein, L. E., Tucker, P. E., & Demorest, M. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, 62(2), 233-252.
- Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models *Technical Report ICSI-TR-97-021*. University of Berkeley.
- Bishop, C. M. (2006). *Pattern recognition and machine learning* (Vol. 1): springer New York.
- Brand, M. (1999). Voice puppetry. *in Proc. The 26th annual conference on Computer graphics and interactive techniques*, pp.21-28.
- Chan, M. T. (2001). HMM-based audio-visual speech recognition integrating geometric and appearance-based visual features. *in Proc. IEEE Fourth Workshop on Multimedia Signal Processing*, pp.9-14.
- Chandramohan, D., & Silsbee, P. L. (1996). A multiple deformable template approach for visual speech recognition. *in Proc. Fourth International Conference on Spoken Language*, Philadelphia, PA, Vol.1, pp.50-53.
- Charlier, B., Hage, C., Alegría, J., & Périer, O. (1990). Evaluation d'une pratique prolongée du LPC sur la compréhension de la parole par l'enfant atteint de déficience auditive. *Glossa*, 22, 28-39.
- Chen, T., & Rao, R. R. (1997). Audio-visual integration in multimodal communication. *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol.1, pp.179-182.
- Chen, Y., Chu, M., Chang, E., Liu, J., & Liu, R. (2003). Voice conversion with smoothed GMM and MAP adaptation. *in Proc. Eurospeech-2003*, Geneva, Switzerland, pp.2413-2416.
- Chiou, G. I., & Hwang, J.-N. (1997). Lipreading from color video. *IEEE Transactions on Image Processing*, 6(8), 1192-1195.
- Choi, K., Luo, Y., & Hwang, J.-N. (2001). Hidden Markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system. *Journal of VLSI signal processing systems for signal, image and video technology*, 29(1-2), 51-61.
- Clarke, B. R., & Ling, D. (1976). The Effects of Using Cued Speech: A Follow-Up Study. *Volta Review*, 78(1), 23-34.
- Cootes, T. F., Edwards, G. J., & Taylor, C. J. (1998). Active appearance models. *in Proc. European Conference on Computer Vision*, Freiburg, Germany, pp.484-498.
- Cornett, R. O. (1967). Cued speech. *American Annals of the Deaf*, 112, 3-13.
- Cornett, R. O. (1988). Cued speech, manual complement to lipreading, for visual reception of spoken language. Principles, practice and prospects for automation. *Acta Oto-Rhino-Laryngologica Belgica*, 42(3), 375.
- Cotton, J. C. (1935). Normal" visual hearing.". *Science*, 82(2138), 592-593.
- Cui, X., Afify, M., & Gao, Y. (2008). MMSE-based stereo feature stochastic mapping for noise robust speech recognition. *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.4077-4080.

- Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4), 357-366.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1-38.
- Denes, P. B. (1963). On the statistics of spoken English. *The Journal of the Acoustical Society of America*, 35, 892.
- Destombes, F. (1982). Aides manuelles a la lecture labiale et perspectives d'aides automatiques (pp. 35 -36). Le Projet VIDVOX: Centre Scientifique IBM-France.
- Dodd, B. (1977). The role of vision in the perception of speech. *Perception*, 6(1), 31-40.
- Draper, N. R., & Van Nostrand, R. C. (1979). Ridge regression and James-Stein estimation: review and comments. *Technometrics*, 21(4), 451-466.
- Duchnowski, P., Lum, D. S., Krause, J. C., Sexton, M. G., Bratakos, M. S., & Braida, L. D. (2000). Development of speechreading supplements based on automatic speech recognition. *IEEE Transactions on Biomedical Engineering*, 47(4), 487-496.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2000). Pattern classification and scene analysis. Part 1: John Wiley, 2 ed.
- Dupont, S., & Luetin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3), 141-151.
- Flury, B. (1988). *Common principal components & related multivariate models*. New York: John Wiley.
- Fontecave Jallon, J., & Berthommier, F. (2009). A semi-automatic method for extracting vocal tract movements from X-ray films. *Speech Communication*, 51(2), 97-115.
- Garcia, C., Cootes, T., & Ostermann, J. (2007). Facial Image Processing. *EURASIP Journal on Image and Vision Processing*, 127, 1-2.
- Gibbon, D., Mertins, I., & Moore, R. K. (2000). Audio-visual and multimodal speech-based systems. In: *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*: Springer, pp. 102-203.
- Gibert, G. (2006). *Conception et évaluation d'un système de synthèse 3D de Langue française Parlée Complétée (LPC) à partir du texte*. (Ph.D thesis), Institut National Polytechnique de Grenoble-INPG.
- Gibert, G., Bailly, G., Beautemps, D., Elisei, F., & Brun, R. (2005). Analysis and synthesis of the three-dimensional movements of the head, face, and hand of a speaker using cued speech. *The Journal of the Acoustical Society of America*, 118, 1144.
- Gray, M. S., Movellan, J. R., & Sejnowski, T. S. (1997). Dynamic features for visual speechreading: A systematic comparison. in *Proc. 3rd Joint Symposium on Neural Computation.*, La Jolla, CA, Vol.6, pp.222-230
- Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2), 83-85.

- Hazen, T. J., Saenko, K., La, C.-H., & Glass, J. R. (2004). A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. *in Proc. The 6th international conference on Multimodal interfaces*, pp.235-242.
- Heckmann, M., Berthommier, F., & Kroschel, K. (2001). A hybrid ANN/HMM audio-visual speech recognition system. *in Proc. AVSP 2001-International Conference on Auditory-Visual Speech Processing*, pp.189-194.
- Hennecke, M. E., Stork, D. G., & Venkatesh Prasad, K. (1996). Visionary speech: Looking ahead to practical speechreading systems. In D. Stork & M. Hennecke (Eds.), *Speechreading by humans and machines: NATO ASI series, series F: Computer and systems science*. Berlin: Springer, Vol. 150, pp. 331-350.
- Heracleous, P., Aboutabit, N., & Beutemps, D. (2009). Lip shape and hand position fusion for automatic vowel recognition in cued speech for french. *IEEE Signal Processing Letters*, 16(5), 339-342.
- Hiroya, S., & Honda, M. (2004). Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Transactions on Speech and Audio Processing*, 12(2), 175-185.
- Hoerl, A. E., Kennard, R. W., & Hoerl, R. W. (1985). Practical use of ridge regression: a challenge met. *Applied Statistics*, 34, 114-120.
- Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P., & Saltzman, E. (1996). Accurate recovery of articulator positions from acoustics: New conclusions based on human data. *The Journal of the Acoustical Society of America*, 100, 1819.
- Huang, F. J., & Chen, T. (1998). Real-time lip-synch face animation driven by human voice. *in Proc. IEEE Second Workshop on Multimedia Signal Processing*, Los Angeles,CA, pp.352-357.
- Itakura, F. (1975). Line spectrum representation of linear predictor coefficients of speech signals. *The Journal of the Acoustical Society of America*, 57, S35.
- Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 4-37.
- Jeffers, J. (1967). Two case studies in the application of principal component analysis. *Applied Statistics*, 3, 225-236.
- Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*, 31, 300-303.
- Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.
- Kain, A., & Macon, M. W. (1998). Spectral voice conversion for text-to-speech synthesis. *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle,WA, Vol.1, pp.285-288.
- Kain, A., Niu, X., Hosom, J.-P., Miao, Q., & Santen, J. P. v. (2004). Formant re-synthesis of dysarthric speech. *in Proc. Fifth ISCA Workshop on Speech Synthesis*, Pittsburgh, USA, pp.25-30.
- Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active contour models. *International journal of computer vision*, 1(4), 321-331.
- Katsamanis, A., Papandreou, G., & Maragos, P. (2009). Face active appearance modeling and speech acoustic information to recover articulation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3), 411-422.

- Kay, S. M. (1993). *Fundamentals of Statistical signal processing: Estimation Theory* (E. Cliffs Ed. Vol. 1). NJ: Prentice Hall PTR.
- Kendall, M. G. (1957). *A course in multivariate analysis*. New York: Hafner Publishing Co.
- Kjellström, H., & Engwall, O. (2009). Audiovisual-to-articulatory inversion. *Speech Communication, 51*(3), 195-209.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. in *Proc. International joint Conference on artificial intelligence*, Vol.14, pp.1137-1145.
- Lai, T. L., Robbins, H., & Wei, C. Z. (1978). Strong consistency of least squares estimates in multiple regression. in *Proc. The National Academy of Sciences, USA*, Vol.75(7), pp.3034-3036.
- Lallouache, M. T. (1991). *Un poste" Visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres*. (PhD thesis), Institut National Polytechnique, Grenoble, France.
- Lange, K. L., Little, R. J., & Taylor, J. M. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association, 84*(408), 881-896.
- Leung, S.-H., Wang, S.-L., & Lau, W.-H. (2004). Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. *IEEE Transactions on Image Processing, 13*(1), 51-62.
- Leybaert, J. (2000). Phonology acquired through the eyes and spelling in deaf children. *Journal of experimental child psychology, 75*(4), 291-318.
- Leybaert, J., & Charlier, B. (1996). Visual speech in the head: the effect of cued-speech on rhyming, remembering, and spelling. *Journal of Deaf Studies and Deaf Education, 1*(4), 234-248.
- Li, S. Z., & Zhang, Z. (2004). Floatboost learning and statistical face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(9), 1112-1123.
- Liew, A. W. C., Leung, S. H., & Lau, W. H. (2002). Lip contour extraction from color images using a deformable model. *Pattern Recognition, 35*(12), 2949-2962.
- Ling, D., & Clarke, B. R. (1975). Cued Speech: An Evaluative Study. *American Annals of the Deaf, 120*(5), 480-488.
- Luettin, J., Thacker, N. A., & Beet, S. W. (1996). Visual speech recognition using active shape models and hidden Markov models. in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol.2, pp.817-820.
- Matthews, I. (1998). *Features for audio-visual speech recognition*. (Ph.D thesis), University of East Anglia.
- Matthews, I., Bangham, J. A., & Cox, S. (1996). Audiovisual speech recognition using multiscale nonlinear image decomposition. in *Proc. Fourth International Conference on Spoken Language*, Philadelphia, PA, Vol.1, pp.38-41.
- Matthews, I., Potamianos, G., Neti, C., & Luettin, J. (2001). A comparison of model and transform-based visual features for audio-visual LVCSR. in *Proc. International Conference on Multimedia and Expo*, Tokyo, Japan, pp.22-25.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748.
- McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker.



- Ming, Z., Beautemps, D., Feng, G., & Schmerber, S. A. (2010). Estimation of Speech Lip Features from Discrete Cosinus Transform. *in Proc. Interspeech 2010*, Makuhari, Japan, pp.1612-1615.
- Montgomery, A. A., & Jackson, P. L. (1983). Physical characteristics of the lips underlying vowel lipreading performance. *The Journal of the Acoustical Society of America*, 73, 2134.
- Morishima, S., & Harashima, H. (1991). A media conversion from speech to facial image for intelligent man-machine interface. *IEEE Journal on Selected Areas in Communications*, 9(4), 594-600.
- Nakamura, K., Toda, T., Saruwatari, H., & Shikano, K. (2006). Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech. *in Proc. Interspeech*, Pittsburgh, PA, pp.1395-1398.
- Narula, S. C., & Wellington, J. F. (1982). The minimum sum of absolute errors regression: A state of the art survey. *International Statistical Review/Revue Internationale de Statistique*, 50, 317-326.
- Nefian, A. V., Liang, L., Pi, X., Liu, X., & Murphy, K. (2002). Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP Journal on Advances in Signal Processing*, 2002(11), 1274-1288.
- Neti, C., Potamianos, G., Luetttin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., & Zhou, J. (2000). Audio-visual speech recognition *Final Workshop 2000 Report*. Baltimore, MD: Center for Language and Speech Processing. The Johns Hopkins University.
- Nicholls, G. H., & Ling, D. (1982). Cued Speech and the reception of spoken language. *Journal of Speech, Language and Hearing Research*, 25(2), 262.
- Nievergelt, Y. (1994). Total least squares: State-of-the-art regression in numerical analysis. *SIAM review*, 36(2), 258-264.
- Périer, O., & De Temmerman, P. (1987). L'enfant à audition déficiente: aspects médicaux, éducatifs, sociologiques et psychologiques. *Acta Oto-rhynolaryngologica Belgica*, 41, 129-420.
- Papandreou, G., Katsamanis, A., Pitsikalis, V., & Maragos, P. (2009). Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3), 423-435.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11), 559-572.
- Picone, J. W. (1993). Signal modeling techniques in speech recognition. *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol.81(9), pp.1215-1247.
- Potamianos, G., Graf, H. P., & Cosatto, E. (1998). An image transform approach for HMM based automatic lipreading. *in Proc. International Conference on Image Processing*, Vol.1, pp.173-177.
- Potamianos, G., Neti, C., Luetttin, J., & Matthews, I. (2012). Audiovisual automatic speech recognition. In G. Bailly, P. Perrier & E. Vatikiotis-Bateson (Eds.), *Audiovisual speech processing*: Cambridge University, pp. 193-247.
- Rao, C. R., Toutenburg, H., Fieger, A., Heumann, C., Nittner, T., & Scheid, S. (1999). *Linear models: least squares and alternatives* (2 ed.). New York: Springer-Verlag.

- Rao, R., Mersereau, R., & Chen, T. (1997). Using HMMs in audio-to-visual conversion. *in Proc. IEEE Workshop on Multimedia Signal Processing*, Princeton, NJ, pp.19-24.
- Reisberg, D., Mclean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*. London, U.K: Lawrence Erlbaum, pp. 97-113.
- Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1), 72-83.
- Richmond, K. (2009). Preliminary inversion mapping results with a new EMA corpus. *in Proc. Interspeech 2009*, Brighton, UK, pp.2835-2838.
- Richmond, K., King, S., & Taylor, P. (2003). Modelling the uncertainty in recovering articulation from acoustics. *Computer Speech & Language*, 17(2), 153-172.
- Schroeter, J., & Sondhi, M. M. (1994). Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(1), 133-150.
- Shannon, M., Zen, H., & Byrne, W. (2013). Autoregressive Models for Statistical Parametric Speech Synthesis. *IEEE Trans. Audio Speech Language Process.*, 21(3), 587-589.
- Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*: Cambridge university press.
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3), 199-222.
- Stone, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *The Annals of Statistics*, 3, 267-284.
- Stylianou, Y., Cappé, O., & Moulines, E. (1998). Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 6(2), 131-142.
- Su, Q., & Silsbee, P. L. (1996). Robust audiovisual integration using semicontinuous Hidden Markov Models. *in Proc. International Conference on Spoken Language*, Philadelphia, PA, pp.42-45.
- Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26, 212-215.
- Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica*, 36(4-5), 314-331.
- Swindel, B. F. (1981). Geometry of ridge regression illustrated. *The American Statistician*, 35(1), 12-15.
- Thomaz, C. E., Feitosa, R. Q., & Veiga, A. (2000). Separate-group covariance estimation with insufficient data for object recognition. *in Proc. Fifth All-Ukrainian International Conference*, pp.21-24.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.
- Toda, T., Black, A. W., & Tokuda, K. (2004). Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis. *in Proc. Fifth ISCA Workshop on Speech Synthesis*, Pittsburgh, USA, pp.31-36.

- Toda, T., Black, A. W., & Tokuda, K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Speech and Audio Processing*, 15(8), 2222-2235.
- Toda, T., Black, A. W., & Tokuda, K. (2008). Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, 50(3), 215-227.
- Tofallis, C. (2009). Least squares percentage regression. *Modern Applied Statistical Methods*, 7(2), 368-631.
- Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, Vol.3, pp.1315-1318.
- Toutios, A., & Margaritis, K. (2005). A support vector approach to the acoustic-to-articulatory mapping. in *Proc. Interspeech 2005*, pp.3221-3224.
- Uchanski, R. M., Delhorne, L., Dix, A., Braida, L., Reed, C., & Durlach, N. (1994). Automatic speech recognition to aid the hearing impaired: prospects for the automatic generation of cued speech. *Rehabilitation Research and Development*, 31(1), 20.
- Uto, Y., Nankaku, Y., Toda, T., Lee, A., & Tokuda, K. (2006). Voice conversion based on mixtures of factor analyzers. in *Proc. Interspeech 2006*, Pittsburgh, USA, pp.2278-2281.
- Webb, A. R. (2011). Finite Mixture Models. In Andrew R. Webb & K. D. Copey (Eds.), *Statistical pattern recognition*. New York: Oxford Univ.Press, 3 ed., pp. 51-53.
- Woodward, M. F., & Barber, C. G. (1960). Phoneme perception in lipreading. *Speech, Language and Hearing Research*, 3(3), 212.
- Yamamoto, E., Nakamura, S., & Shikano, K. (1998). Lip movement synthesis from speech based on hidden Markov models. *Speech Communication*, 26(1), 105-115.
- Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1), 23-43.
- Yuille, A. L., Hallinan, P. W., & Cohen, D. S. (1992). Feature extraction from faces using deformable templates. *computer vision*, 8(2), 99-111.
- Zen, H., Gales, M. J., Nankaku, Y., & Tokuda, K. (2012). Product of experts for statistical parametric speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 20(3), 794-805.
- Zen, H., Nankaku, Y., & Tokuda, K. (2011). Continuous Stochastic Feature Mapping Based on Trajectory HMMs. *IEEE Transactions on Speech and Audio Processing*, 19(2), 417-430.
- Zen, H., Tokuda, K., & Kitamura, T. (2004). An introduction of trajectory model into HMM-based speech synthesis. in *Proc. Fifth ISCA Workshop on Speech Synthesis*, pp.191-196.
- Zhang, L. (2009). *Modelling speech dynamics with trajectory-HMMs*. (Ph.D. thesis), University of Edinburgh.
- Zhang, Y., Levinson, S., & Huang, T. (2000). Speaker independent audio-visual speech recognition. in *Proc. International Conference on Multimedia and Expo.*, New York, pp.1073-1076.

- Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. *in Proc. International Conference on Machine Learning*, pp.912–919.
- Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands. *Journal of the Acoustical Society of America*, 33, 248-249.

# Appendice A

## Résumé en français de la thèse

### Introduction

Le sujet de cette thèse est la communication vocale pour les sourds qui sont oralement éduqués. La parole est concernée ici dans ses dimensions multimodales et dans le contexte d'un traitement automatique. En effet, l'avantage de l'information visuelle pour la perception de la parole (appelée «lecture labiale») est largement reconnu. A partir des travaux précurseurs de Sumbly et Pollack (1954), puis ceux de Summerfield (1979) à ceux de Benoit et al. (1991) dans lesquels la langue française est concernée, il est bien établi que l'information visuelle issue du visage du locuteur est utilisée pour améliorer la perception de la parole dans un environnement bruyant. En outre, même dans le contexte d'une parole clairement auditive, la vision reste importante: les expérimentations d'ombrage ont montré, par exemple, que les temps de réaction ont été réduits d'un facteur moyen de 7,5% en cas de stimuli audiovisuels en comparaison avec la présentation simplement auditive (Reisberg et al., 1987). L'effet McGurk démontre la capacité d'intégrer les informations auditives et visuelles même si les deux modalités ne sont pas congruentes (McGurk et al., 1976). Comme nous présentons ici, les gens ayant une audition normale ont des compétences en lecture labiale (Cotton, 1935; Dodd, 1977). Cependant, il a été démontré que les performances initiales - c'est à dire sans formation spécifique - varient grandement d'un individu à l'autre. Bernstein et al. (2000) ont comparé les performances de 96 personnes ayant une audition normale avec 72 personnes profondément sourdes. Les auteurs ont observé des performances très variables entre les individus des deux groupes, mais ont clairement montré que les meilleurs lecteurs labiaux étaient les sourds.

Cependant, même avec la bonne performance de lecture labiale, la parole ne peut pas être complètement perçue sans la connaissance de son contexte sémantique. Les meilleurs lecteurs labiaux atteignent très rarement la perfection. Seulement 41,9% à 59,1% des 10 voyelles différentes sont reconnus dans un [hVg] contexte (Montgomery et al., 1983) et 32% quant aux mots faiblement prédits (Nicholls et al., 1982). La raison principale à cela est liée à l'ambiguïté du motif visuel. Toutefois, dans la mesure où les personnes sourdes oralement éduquées sont concernés, l'acte de lecture labiale reste la modalité principale de perception de la parole. Cela a conduit Cornett (1967) à développer le système de Cued Speech (CS) comme un complément à l'information labiale.

CS est un système de communication visuelle qui utilise les formes de main placées dans différentes positions à proximité du visage, en combinaison avec lecture labiale naturelle de la parole, pour améliorer la perception de la parole à partir de données visuelles. Il s'agit d'un système dans lequel le locuteur, face à l'observateur, déplace sa main en correspondance avec la parole (Attina et al., 2004, pour une étude détaillée sur l'organisation temporelle du CS en langue française). Le CS améliore considérablement la perception de la parole pour les personnes dont l'audition est endommagée (Nicholls et al., 1982) concernant l'identification de syllabes anglo-américain; concernant l'identification des phrases en langue anglo-américaine, les scores se répartissent entre 78 et 97% (Uchanski et al., 1994). Par ailleurs, le CS propose aux sourds une représentation complète du système phonologique, dans la mesure où ils ont été exposés à cette méthode depuis leur jeunesse, et donc il a un impact positif sur le développement de langage (Leybaert, 2000).

Comme nous l'avons vu dans ce court résumé, la méthode CS offre un réel avantage pour la perception de la parole complète. Aujourd'hui, un des défis majeurs est le problème de la communication vocale entre les personnes ayant l'audition normale qui ne pratiquent pas le CS mais les discours vocaux et les sourds sans restes auditifs qui utilisent la lecture labiale complété par le code CS pour la perception de la parole. Pour résoudre ce problème, on peut utiliser un traducteur humain. Une autre solution est basée sur le développement de systèmes de traduction automatique. Pour cela, et dans un cadre plus général, deux sources d'information pourraient contribuer à cette

opération de traduction: (i) La connaissance a priori des contraintes phonétiques, phonologiques et linguistiques; (ii) la connaissance a priori des corrélations entre les différentes activités vocales: les activités neuronales et neuro-musculaires, les mouvements articulatoires, les paramètres aérodynamiques, la géométrie du tract vocal, la déformation du visage et le son acoustique. De différentes méthodes réalisent leur modélisation et leur fusion optimale avec les signaux d'entrée et ceux de sortie. Sur un axe classant les méthodes en fonction de leur dépendance à l'égard de la langue utilisée, on peut trouver aux deux extrémités: (i) la méthode utilisant le niveau phonétique de l'interface, combinant la reconnaissance vocale et la synthèse vocale pour tenir compte de l'organisation phonologique de la parole. Il faut noter que le processus de reconnaissance et de synthèse peut nécessiter très diverses techniques de modélisation. Si les modèles phonétiques basés sur les modèles de Markov cachés (HMM) sont à la base des principaux systèmes de reconnaissance, la synthèse basée sur la concaténation d'unités multi-paramétrées de différentes longueurs est toujours très populaire. Notez aussi l'intérêt croissant de la synthèse par les modèles de trajectoire sur la base de HMM permettant l'apprentissage joints des systèmes de reconnaissance et de synthèse (Tokuda et al., 2000; Zen et al., 2004; Zen et al., 2011), (ii) les méthodes utilisant la corrélation entre les signaux sans l'aide du niveau phonétique, mais en utilisant diverses techniques de cartographie. Ces techniques capturent les corrélations entre les échantillons d'entrée et de sortie à l'aide de la Quantification Vectorielle ou le Modèle Mélange Gaussien (Toda et al., 2004; Uto et al., 2006).

Pour le CS, la méthode classique pour convertir la parole auditive aux composants CS consiste en coupler un système de reconnaissance à un synthétiseur vocal 'Text-to-visual' (Duchnowski et al., 2000; Attina et al., 2004; Gibert et al., 2005; Beautemps et al., 2007). La liaison entre les deux systèmes nécessite au moins le niveau phonétique élevé.

Avant ce travail, aucune étude ne vise à utiliser un niveau de signal très faible. Cette thèse est une contribution à ce défi dans le cas des voyelles françaises. Une nouvelle approche basée sur la cartographie des paramètres spectraux de la parole avec les composantes visuelles constituées des paramètres CS et labiaux est proposée. Dans ce

contexte, le but du processus de cartographie consiste à fournir des paramètres visuels qui peuvent être utilisés en tant que paramètres de cible pour la synthèse de la parole visuelle.

Une autre problématique de cette thèse porte sur la cartographie (ou l'estimation) des paramètres géométriques de la lèvre par les caractéristiques de l'apparence obtenues par l'image naturelle de la région d'intérêt de la bouche (ROI) sans utiliser les artifices de la lèvre. Pour estimer les paramètres de la lèvre, deux approches sont généralement considérées comme suggère par Potamianos et al. (2012): Caractéristiques basées sur la forme et les caractéristiques basées sur l'apparence. Dans le premier cas, le contour interne et externe de la lèvre sont extraits de la vue d'image du visage. Un modèle de contour de la lèvre peut être obtenu sur le plan statistique (Luettin et al., 1996; Dupont et al., 2000) ou paramétrique (Hennecke et al., 1996; Chiou et al., 1997). Ensuite, l'ensemble des paramètres de modèle contient les informations visuelles. Dans la seconde approche, les transformations appropriées, telles que la transformation en cosinus discrète (DCT) ou l'analyse du composant principal (PCA), sont appliquées sur les pixels de l'image correspondant à la région d'intérêt (ROI) de la bouche du locuteur (Matthews et al., 1996; Gray et al., 1997). Les deux approches sont souvent combinées comme dans (Matthews, 1998), où les modèles actifs d'apparence sont construits à la fois sur les caractéristiques de la forme et de l'apparence.

## **1. Etat de l'art du CS**

Le CS est un système inventé par Dr Cornett qui utilise les codes de la main comme un complément à la lecture labiale, et vise à permettre l'accès des personnes auditivement handicapées à la langue orale. Dans le système CS, la main déplaçant à l'emplacement spécifique autour du visage est utilisée pour coder les voyelles, tandis que les formes sont utilisées pour coder les consonnes. La version finale définie par Cornett pour l'anglais américain est basée sur quatre positions de la main et huit configurations (quatre groupes de voyelles et huit groupes de consonnes). Les phonèmes de chacun de ces groupes peuvent être distingués par la forme de la lèvre tandis que les phonèmes avec la forme labiale similaire seront distingués par l'utilisation de différents codes. Par exemple, les consonnes [p], [b] et [m] (qui ont la même forme de la lèvre) sont respectivement codées par les configurations 1, 4 et 5.



Le CS a été adapté à plus de 60 langues et dialectes dans le monde entier, y compris le français. Le CS a été importé en France en 1977. L'adaptation a été appelée LPC pour souligner le fait que le LPC est entièrement basé sur la langue française.

## **1.1. Études de perception**

La valeur du LPC réside dans son efficacité dans l'amélioration de la perception de la parole. C'est la principale raison pour laquelle ce système a été inventé. L'expansion du CS dans le monde par ses adaptations aux différentes langues et son utilisation croissante dans plusieurs environnements (à la maison ou à l'école), démontre l'efficacité du système pour une bonne réception de la parole. Le site de l'ALPC a présenté la preuve concrète des parents utilisant le code LPC pour communiquer avec leurs enfants sourds et a démontré la contribution du LPC perceptif.

Sur le plan expérimental, plusieurs études ont été menées sur les différentes versions du CS dans le monde. Pour CS, dans sa version originale, nous pouvons citer les travaux de (Ling et al., 1975; Clarke et al., 1976; Nicholls et al., 1982; Uchanski et al., 1994). Pour la version française du LPC, nous pouvons trouver de telles études (Charlier et al., 1990; Alegria et al., 1999; Attina et al., 2002; Attina et al., 2004; Attina, 2005; Aboutabit, 2007; Heracleous et al., 2009) et ainsi de suite. Nous présentons certains de leurs travaux dans la section suivante.

## **1.2. Les études de production**

Aucune étude fondamentale n'a été dédiée à l'analyse de la production qualifiée de gestes CS jusqu'à l'invention de la LPC. Cornett n'a remarqué incidemment que certains groupes de consonnes devraient être retardés pour laisser suffisamment de temps à la main pour atteindre la position correcte au cours d'enquêtes technologiques (Cornett, 1967).

Le travail de pionnier dans ce domaine a été réalisé à l'Institut de la Communication Parlée (ICP) pour montrer comment le mouvement de la main co-produit sur la consonne et la voyelle en LPC. Tout d'abord, Attina et al. (2002, 2004) ont mis l'accent sur l'organisation temporelle des indications manuelles en collaboration avec le mouvement de la lèvre et le signal acoustique correspondant. Il montre un

avancement du début de mouvement de la main de 200 ms en moyenne par rapport à la réalisation du CV syllabe acoustique à partir de l'analyse d'un codeur LPC (Attina et al., 2002; Attina et al., 2004). Les résultats de l'avancement de la main ont été confirmés par l'analyse de la production de trois codeurs supplémentaires (Attina, 2005).

### **1.3. Études des systèmes automatiques**

Dans le système Autocuer Cornett (Cornett, 1988), les indices sont définis par la reconnaissance de son de la parole prononcée et sont affichés dans les groupes de LEDs sur les lunettes portées par le lecteur de la parole. L'ensemble du processus implique un retard de 150 à 200 ms pour l'affichage de l'indice, par rapport au moment de production du son correspondant. Ce système, conçu pour des mots isolés, atteint 82% d'identification correcte. Dans le système de génération automatique de CS développé par (Duchnowski et al., 2000) pour l'anglais américain, les indices sont présentés avec l'aide des mains préenregistrées, et les règles de coordination temporelle avec le son sont proposées. Ce système utilise un système de reconnaissance phonétique de la parole audio pour obtenir une liste des téléphones qui sont ensuite convertis en un flux temporel de codes d'indice. Les indices appropriés sont visuellement affichés en superposant les formes de la main sur le signal vidéo du visage du locuteur.

Récemment, de nombreux travaux ont focalisé sur la traduction automatique du langage visuel, y compris les caractéristiques de la forme des lèvres et les gestes de la main en parole acoustique. Les caractéristiques visuelles fusionnées par le modèle multi-flux HMM ont été utilisées pour réaliser la reconnaissance de la voyelle ou de la consonne française (Aboutabit, 2007; Heracleous et al., 2009). La méthode classique pour convertir le discours audio en discours visuel consiste de coupler un système de reconnaissance à un synthétiseur vocal 'text-to-visuel' (Duchnowski et al., 2000; Attina et al., 2004; Gibert et al., 2005; Beautemps et al., 2007). La liaison entre les deux systèmes nécessite au moins un haut niveau lexical.

## **2. Matériel de la parole et du CS**

### **2.1. L'enregistrement de base de données**

Les données proviennent d'un enregistrement vidéo d'un locuteur prononçant et codant en LPC (CS), un ensemble de 50 mots isolés français. Les mots sont constitués de 32 chiffres (de 0 à 31), 12 mois et 6 autres mots qui sont ordinaire. Chaque mot a été présenté une fois sur un écran placé devant le haut-parleur, dans un ordre aléatoire. Le corpus a été prononcé 10 fois. Le locuteur est un locuteur natif féminin du français diplômé en CS. L'enregistrement a été fait dans une cabine insonorisée et le taux d'enregistrement d'image vidéo a été réglé à 25 images par seconde. Le locuteur est assis devant un micro et une caméra reliée à un enregistreur Betacam. Les marques d'intérêt ont été placés entre les sourcils et à l'extrémité des doigts pour faciliter l'extraction des coordonnées utilisées en tant que paramètres LPC de la main. En outre, un papier carré a été enregistré pour la conversion pixel-à-centimètre. L'enregistrement vidéo a été réalisé sous le format PAL, ainsi sauvegardé comme des images numériques RGB Bitmap constituées des demi-frames entrelacées de la vidéo (respectivement les lignes paires et impaires). Chaque image a été désentrelacée en deux demi-frames et les lignes manquantes de la demi-frame chacun ont été remplis par interpolation linéaire, pour obtenir deux images complètes de-entrelacées correspondant à deux enregistrements séparés de 20 ms.

### **2.2. Extraction des lèvres et de la main caractéristiques visuelles**

Avant de choisir manuellement les coordonnées sur les images correspondantes, le signal audio est utilisé pour localiser les échantillons des voyelles. Après les trames vidéo correspondantes aux voyelles sont choisies, les coordonnées du contour intérieur de la lèvre ont été sélectionnées manuellement pour en extraire les caractéristiques de la lèvre dans l'étape suivante. Dans notre travail, nous avons dénommé respectivement la largeur, la hauteur et la surface interne de la lèvre A, B et S comme les paramètres de la lèvre inclus dans les caractéristiques visuelles du contour des lèvres. Selon Lallouache (1991), les paramètres de la lèvre ont été calculés sur la base des points manuellement sélectionnés sur le contour intérieur de la lèvre. La méthode de Lallouache (1991) fonctionne bien dans la plupart de temps.

Mais quand la hauteur de la lèvre, c'est-à-dire le paramètre de lèvre B, est très faible, il y a un problème avec cette méthode. Par exemple, dans le cas de prononcer les voyelles [ø, u], la hauteur du contour de la lèvre interne est très faible et une partie des valeurs d'extraction correspondantes au paramètre B sont égales à zéro. Cela est dû à la difficulté de sélectionner manuellement suffisamment de points sur le contour intérieur de la lèvre pour l'approximation parabolique lorsque la hauteur de la lèvre est faible. Afin de surmonter ce problème, nous avons utilisé une méthode basée sur les pixels de la région d'intérêt de la lèvre interne (ROI) au lieu des points sélectionnés manuellement pour calculer le paramètre B. Nous avons d'abord utilisé les points sélectionnés manuellement pour tracer la lèvre interne ROI et ensuite utilisé l'ellipse pour s'adapter à la forme du contour interne de la lèvre pour calculer les paramètres de la lèvre (Leung et al., 2004).

Contrairement à la forme de la lèvre qui est normalement synchronisée avec la parole acoustique, la position de la main définie dans le CS n'est pas toujours synchronisée avec la parole acoustique. Cela est dû à la raison que la parole est produite directement par les articulateurs y compris la lèvre plutôt que la position de la main ou la forme qui sont utilisées dans le CS. Par conséquent, nous avons besoin de vérifier si la position de la main correspond à la voyelle dans le cadre de l'image sélectionnée avant d'en extraire la position de la main.

La position de la main est déterminée par la position du bout du doigt. Le point marqué en bleu au milieu des sourcils a été défini comme le point d'origine du nouveau système de coordonnées. L'axe X est la coordonnée horizontale et l'axe Y est la coordonnée verticale. Ainsi, l'emplacement du bout des doigts par rapport à l'origine a été défini comme la position de la main dans ce travail.

Il faut noter que nous avons sélectionné le doigt du milieu comme le doigt de guidage s'il apparaissait. Sinon, l'index est choisi comme le doigt de guidage quand plus de deux doigts apparaissent dans l'image pour constituer la forme de la main indiquant les consonnes dans le CS.

### 2.3. Extraction de paramètres spectraux

Dans ce travail, l'enregistrement sonore a été numérisé à 44100 Hz et ré-échantillonné à 16000 Hz. Trois types de paramètres spectraux du signal de parole ont été utilisés: les valeurs des formants, les MFCCs (Mel-Frequency Cepstral Coefficients) et les LSP (paires de raies spectrales). Un de nos principaux objectifs était de déterminer les paramètres qui peuvent donner une meilleure performance lorsqu'ils sont utilisés pour prédire les lèvres et les paramètres de la main. Comme nous le savons, les valeurs de formants sont les plus pertinentes pour décrire les voyelles et ils ont présenté une relativement faible variance d'un locuteur à l'autre. Cependant, le petit nombre (2 à 4) de valeurs de formants limitera les performances d'estimation. Au lieu de cela, les paramètres MFCC et LSP offriront de l'information spectrale suffisante pour la modélisation des estimations.

Les MFCCs ont été calculés à partir d'un spectre de parole à court terme mis à l'échelle (en dB) par une DCT (transformée en cosinus discrète) sur la base d'une fenêtre Hamming de 32 ms centré au  $t_{0i}$ . De cette façon, les MFCCs décrivent bien le spectre de parole perçu dans l'oreille. Les MFCCs sont prouvés plus efficaces que les autres paramètres spectraux pour les systèmes de reconnaissance vocale et les systèmes d'identification de locuteur (Davis et al., 1980).

Comme les formants sont dérivés des coefficients LSP dans la pratique, nous pouvons utiliser l'enveloppe spectrale obtenue pour vérifier les valeurs aberrantes des formants et retirer le cadre anormale dans le corpus pour le rendre propre.

### 2.4. Structure de la base de données

Dans ce travail, nous nous sommes concentrés sur les voyelles. Ainsi, la base de données est composée de 1 371 occurrences des 10 voyelles françaises (tableau 2.1). Chaque trame comprend des paramètres de la lèvre, comme la largeur de la lèvre (A), la hauteur de la lèvre (B) et la surface de la lèvre (S), la position des mains en coordonnées X et Y, les paramètres spectraux comme les quatre formants, les 16 MFCCs, les 16 LSP et l'étiquette phonétique correspondante.

Tableau 2.1 Liste des dix voyelles françaises avec leur apparition

| Voyelles   | [i] | [e] | [ɛ] | [a] | [y] | [ø] | [œ] | [ɔ] | [o] | [u] |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Apparition | 236 | 255 | 231 | 168 | 37  | 80  | 137 | 83  | 40  | 104 |

### 3. Méthodes de cartographie

#### 3.1. Méthode multi-linéaire (MLR) sur la base de l'APC régression

Un moyen direct d'évaluer l'interrelation entre les différents groupes de données collectées est d'utiliser les estimateurs linéaires pour mesurer à quel point un groupe de données peut être déterminé à partir d'un autre (Yehia et al., 1998). Dans ce travail, la méthode MLR de régression PCA (Principal Analysis Component) a été utilisée pour mettre en correspondance les paramètres spectraux, les paramètres de la lèvre et les positions des mains. Le PCA est utilisé pour éliminer la corrélation entre les paramètres spectraux conduisant à des estimations instables et potentiellement trompeuses de l'équation de régression (Jolliffe, 1986). La méthode MLR PCR (Principal Components Regression) peut être divisé en trois étapes: (1) La première étape est de lancer une analyse en composants principaux sur la table des variables explicatives (2) La deuxième étape consiste à exécuter un OSL (Ordinary Least Squares Regression) sur les composants sélectionnés: les facteurs qui sont les plus corrélés à la variable dépendante seront sélectionnés. (3) Enfin, les paramètres du modèle sont calculés selon le critère d'erreur de carré minimal.

#### 3.2. Méthode sur la base du modèle de mélanges gaussiens (GMM)

Les méthodes statistiques telles que la quantification vectorielle (VQ), le réseau de neurones, et le GMM ont été fréquemment utilisées pour mettre en correspondance la parole acoustique et les caractéristiques visuelles de la reconnaissance de la parole audio-visuelle, l'interface homme-machine et l'inversion acoustique-à-articulatoire, etc .

La VQ est initialement utilisée pour la compression de données, et puis utilisée pour la modélisation acoustique discrète depuis le début des années 1980. Plus tard, la méthode est largement utilisée pour la cartographie des caractéristiques de la parole

aux caractéristiques visuelles dans la parole audio-visuelle (Morishima et al., 1991). Dans l'inversion acoustique-à-articulatoire, la VQ est également utilisée pour créer un dictionnaire de paires de paramètres articulatoires-acoustiques quantifiés (Schroeter et al., 1994; Hogden et al., 1996). La méthode VQ est simple, mais elle génère des erreurs de quantification indésirables.

Les réseaux de neurones peuvent aussi être utilisés pour convertir les paramètres acoustiques aux paramètres visuels. Un réseau neuronal basé sur un réseau neuronal à trois couches est utilisée pour la conversion de voix à image (Morishima et al., 1991). Le réseau de densité mixte (MDN) a prouvé son efficacité pour la cartographie d'inversion articulatoire (Richmond et al., 2003; Richmond, 2009). Le ANN est implémenté dans une inversion audiovisuel-à-articulatoire (Kjellström et al., 2009).

L'efficacité de GMM a été illustrée dans de nombreux domaines de recherche de la parole, tels que la reconnaissance de la parole. Le GMM est souvent utilisé comme le cadre de la densité probabiliste, l'identification de locuteur (Reynolds et al., 1995), la conversion de la parole (Kain et al., 1998; Stylianou et al., 1998; Chen et al., 2003; Toda et al., 2007) et bien sûr dans le domaine du discours audiovisuel (Huang et al., 1998) et la cartographie d'inversion acoustique-à-articulatoire (Nakamura et al., 2006; Toda et al., 2008; Zen et al., 2011).

La procédure principale de la méthode de cartographie à base du GMM est (Figure 3.1):

(1) La fonction de densité de probabilité conjointe des séquences de vecteur caractéristique de la source et de la cible est modélisée par un GMM qui est formé par EM. Il est prouvé que l'utilisation de la source conjointe et des vecteurs de cible plutôt que les vecteurs de source (Stylianou et al., 1998) seulement pour former le GMM est plus robuste pour les petits montants, puisque la densité conjointe devrait conduire à un regroupement plus judicieux pour le problème de régression (Kain et al., 1998).

(2) La fonction de densité de probabilité conditionnelle d'une séquence de vecteur caractéristique de la cible dans le cas où une séquence de vecteur caractéristique de la source est estimée à partir de la fonction de densité de probabilité conjointe.

(3) La séquence de vecteur caractéristique mappé de la cible est déterminée avec les approches Bayésiennes pour minimiser son erreur quadratique moyenne (MMSE) ou maximiser sa probabilité a posteriori (MAP) basée sur la fonction de densité de probabilité conditionnelle estimée.

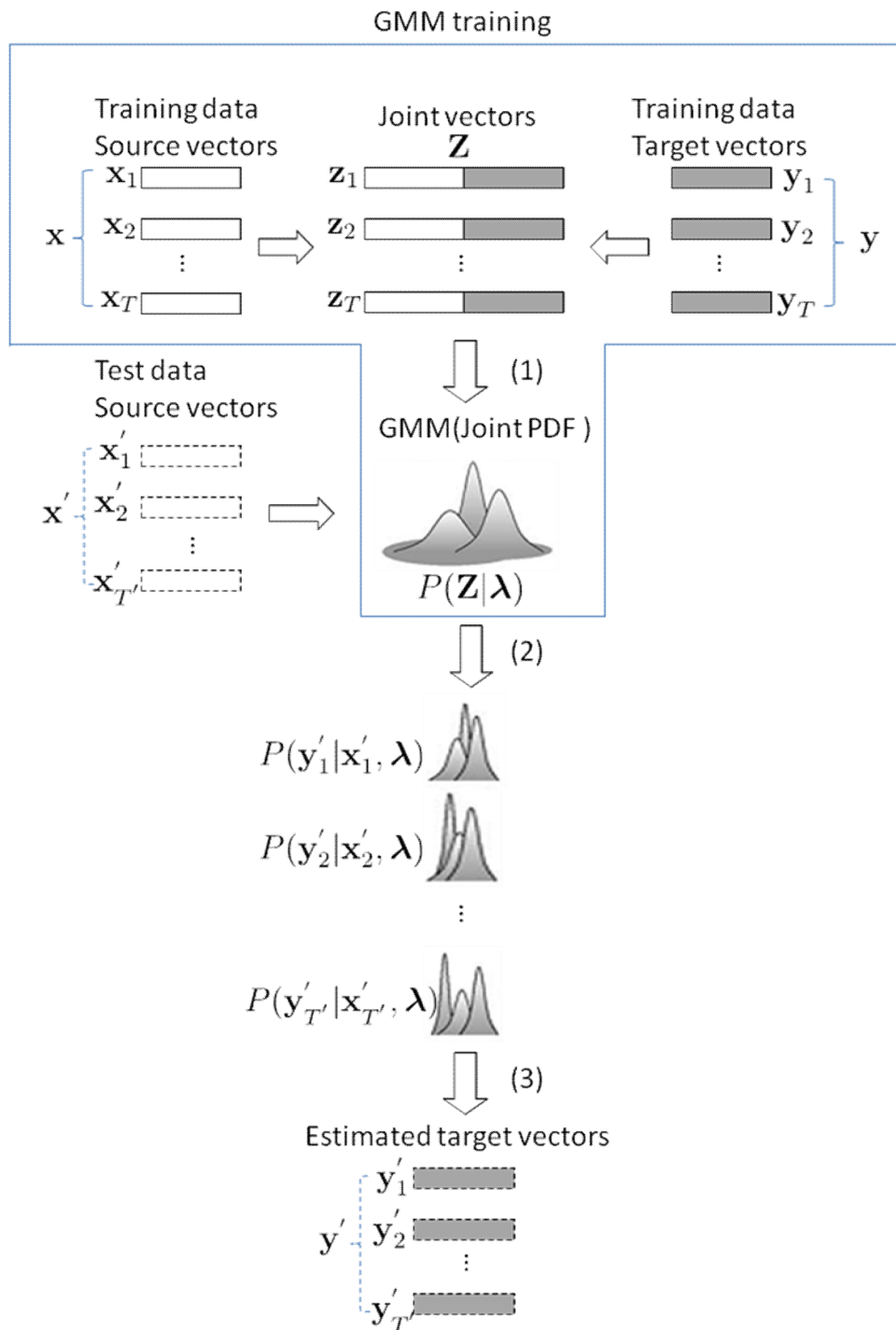


Figure 3.1: La procédure de la méthode de cartographie à base du GMM.



## **4. Cartographie entre la parole et le CS: l'approche linéaire**

L'objectif de ce travail est d'associer la caractéristique acoustique représentée par les paramètres spectraux des signaux de parole à la forme de la lèvre et les coordonnées de la main qui constituent le CS dans le cas des voyelles françaises. Au début de ce travail, la modélisation MLR est utilisée pour prédire la forme de la lèvre et les coordonnées de la main. La validation croisée, parfois appelée l'estimation en rotation, a été utilisée comme la méthode d'évaluation de cette expérimentation. Cette méthode pourrait estimer le niveau de précision auquel d'un modèle prédictif se produira en pratique, surtout lorsque d'autres échantillons sont coûteuses ou difficiles à collecter (Kohavi, 1995). Nous avons utilisé la variance résiduelle et l'erreur quadratique moyenne (RMSE) comme les critères d'évolution pour évaluer l'efficacité et la précision de la prédiction.

### **4.1. L'évaluation de la prédiction des caractéristiques de la lèvre**

Quatre types de paramètres spectraux, le Formant, MFCCs, LSP, et le mélange du MFCC et LSP ont été utilisés pour prédire les paramètres de la lèvre. Tout d'abord, la variance résiduelle et l'erreur quadratique moyenne diminue avec l'incrément du nombre de prédicteurs utilisés dérivés des paramètres spectraux. La variance résiduelle reste élevée en utilisant les formants. Cela est probablement dû à un manque de dimensions. Les MFCCs permettent une diminution plus rapide tandis que les coefficients LSP atteignent une variance résiduelle inférieure. La prédiction basée sur la concaténation des MFCCs et du LSP a les avantages de la propriété de diminution rapide des MFCCs et le bas résiduels du LSP. Ce mélange de MFCC et LSP est donc considéré comme les meilleurs paramètres pour cette prédiction. La forte corrélation linéaire entre les formes de la lèvre et le spectre acoustique donne de bons résultats d'estimation des paramètres de la lèvre (voir le tableau 4.1, tableau 4.2).

Tableau 4.1: La variance résiduelle (RVAR) de la prédiction des caractéristiques de la lèvre en formant les données avec les prédicteurs issus de différents paramètres spectraux.

| RVAR                       | Nombre de prédicteurs | A   | B   | S   |
|----------------------------|-----------------------|-----|-----|-----|
| Formant(Formant1,Formant2) | 2                     | 42% | 37% | 43% |
| Formant(Formant1-Formant4) | 4                     | 42% | 36% | 42% |
| MFCC                       | 16                    | 27% | 26% | 28% |
| LSP                        | 16                    | 18% | 18% | 18% |
| MFCC+LSP                   | 16                    | 16% | 14% | 15% |

Tableau 4.2: Erreur quadratique moyenne (RMSE) de la prédiction des caractéristiques de la lèvre en formant les données avec les prédicteurs issus de différents paramètres spectraux.

| RMSE(mm)                   | Nombre de prédicteurs | A    | B    | S(mm <sup>2</sup> ) |
|----------------------------|-----------------------|------|------|---------------------|
| Formant(Formant1,Formant2) | 2                     | 8,70 | 2,53 | 11,17               |
| Formant(Formant1-Formant4) | 4                     | 8,69 | 2,52 | 11,12               |
| MFCC                       | 16                    | 6,99 | 2,12 | 9,03                |
| LSP                        | 16                    | 5,74 | 1,76 | 7,31                |
| MFCC+LSP                   | 16                    | 5,31 | 1,55 | 6,67                |

## 4.2. L'évaluation de la prédiction de la position de la main

Néanmoins, dans le cas de la main, l'approche MLR directe affiche une mauvaise performance. Cela est dû aux positions de la main qui sont utilisées pour lever l'ambiguïté sur la similitude des formes de la lèvre dans le CS n'ayant effectivement aucun rapport avec le signal de la parole. La variance résiduelle finale des coordonnées prédites X et Y de la main est respectivement 39% et 29% avec les 16 meilleurs prédicteurs dérivés d'un mélange de MFCC et LSP.

## 4.3. Prédiction de la position de la main dans l'espace intermédiaire

A partir de la mauvaise performance de l'approche MLR directe pour estimer les positions de la main, nous pouvons constater que l'espace des positions de la main a une structure différente de la topologie que l'espace acoustique. Par exemple, pour les voyelles acoustiquement proche (par exemple [i] et [e]), leurs positions

correspondantes dans le plan XY est probablement très loin, bien au contraire, les deux voyelles loin dans l'espace acoustique peuvent partager la même position de la main, comme les voyelles [a] et [o]. C'est probablement la raison pour laquelle l'approche MLR directe obtient une mauvaise performance pour estimer la position de la main. Afin d'établir une structure de topologie similaire de l'espace acoustique, l'espace intermédiaire est introduit où les positions de la main sont relocalisées en fonction de la distribution des voyelles dans l'espace de formants F1-F2. La performance de l'estimation de la position de la main dans l'espace intermédiaire améliore de manière significative (voir le tableau 4.3). Cette amélioration prouve que le manque de structure de topologie similaire entre l'espace acoustique et l'espace de la position de la main est la raison qui explique la mauvaise performance de l'approche linéaire pour estimer la position de la main. Cependant, il est nécessaire de reconfigurer les positions estimatives de la main de l'espace intermédiaire à l'espace d'origine. Les méthodes de classification ont été introduites pour classer la position de la main dans l'espace intermédiaire pour faire correspondre les valeurs de traduction par lesquelles les coordonnées de la main dans l'espace intermédiaire permettent de reconfigurer l'espace d'origine. Les classificateurs Bayes LDA et QDA ont été évalués tous les deux, en utilisant la matrice de covariance arbitraire au lieu de la matrice de covariance identique, et la performance du QDA est légèrement supérieure à la LDA. Mais une petite déviation causée par la mauvaise classification dans l'espace intermédiaire entraînera probablement un très grand écart dans l'espace d'origine au cours du processus de reconfiguration. Par conséquent, les erreurs de la méthode de classification n'ont pas réussi à atteindre la bonne performance d'estimation dans l'espace d'origine (voir le tableau 4.3). Afin de libérer la contrainte linéaire du modèle MLR et de surmonter la propriété binaire de la méthode de classification, nous avons besoin d'une approche qui a à la fois les propriétés de classification et de cartographie. Ainsi, nous nous tournons vers l'approche de cartographie à base du GMM qui peut répondre à ces exigences.

Tableau 4.3: Comparaison de la RVAR moyenne dans le cas de la prédiction par l'approche MLR dans l'espace d'origine ( $\overline{\text{RVAR}}_0$ ), dans l'espace intermédiaire ( $\overline{\text{RVAR}}_1$ ) et reconfiguration des résultats de l'espace intermédiaire à l'espace d'origine avec le classificateur QDA ( $\overline{\text{RVAR}}_2$ ).

| X coordinate | $\overline{\text{RVAR}}_0$ | $\overline{\text{RVAR}}_1$ | $\overline{\text{RVAR}}_2$ |
|--------------|----------------------------|----------------------------|----------------------------|
| Training     | 39%                        | 13%                        | 26%                        |
| Test         | 43%                        | 14%                        | 31%                        |

| Y coordinate | $\overline{\text{RVAR}}_0$ | $\overline{\text{RVAR}}_1$ | $\overline{\text{RVAR}}_2$ |
|--------------|----------------------------|----------------------------|----------------------------|
| Training     | 29%                        | 7%                         | 20%                        |
| Test         | 31%                        | 8%                         | 22%                        |

## 5. Cartographie de la parole au CS: l'approche GMM

L'approche de la cartographie basée sur GMM dont l'efficacité a été illustrée par les applications de la conversion de la parole et de l'inversion acoustique-à-articulatoire possède des propriétés de régression et de classification qui nous inspirent à l'employer dans notre travail. La 5-fois validation croisée, la variance résiduelle et le RMSE sont utilisés pour évaluer la méthode de cartographie à base du GMM.

### 5.1. Les méthodes de formation du GMM

Il existe de nombreuses méthodes pour former le GMM. Dans ce travail, nous avons utilisé trois méthodes: la procédure supervisée, sans surveillance et semi-supervisée séparément au niveau de la théorie d'apprentissage de la machine. Lorsque les paramètres de GMM qui sont les vecteurs de moyennes  $\mu_m^{(x)}$  et  $\mu_m^{(y)}$ , les matrices de covariance  $\Sigma_m^{(xx)}$ ,  $\Sigma_m^{(yy)}$ ,  $\Sigma_m^{(xy)}$ ,  $\Sigma_m^{(yx)}$  et les poids de mélange de chaque gaussienne  $\alpha_m$ , le processus de régression MMSE et MAP sont séparément implémentés sur les différents GMMs.

La méthode de formation supervisée est mis en œuvre de deux façons différentes: la première est basée sur les caractéristiques visuelles de la cible, 3 visèmes de la lèvre par exemple, pour estimer les paramètres de la lèvre, ou 5 positions de la main définies dans le CS pour estimer les positions de la main. Le second moyen est basé

sur la caractéristique acoustique, comme les 10 voyelles, qui peut être utilisée pour estimer les paramètres de la lèvre ou les positions de la main. C'est-à-dire, la source conjointe et les vecteurs cibles sont regroupés en 3 groupes en fonction des visèmes de la lèvre, ou 5 groupes en fonction des positions de la main, ou 10 groupes en fonction des différents types de voyelles par l'étiquette phonétique individuelle. Chaque groupe correspond à une gaussienne et les paramètres de gaussienne sont estimés sur la base des groupes.

La méthode non supervisée est initialisée par K-moyenne et utilise l'algorithme EM (Expectation-maximisation) pour estimer automatiquement les paramètres du GMM. La méthode de formation non supervisée est une procédure sans la connaissance a priori telle que les informations d'étiquette. Ainsi, la méthode de formation non supervisée est supérieure pour trouver les composants cachés du GMM parmi les données.

Afin d'inclure les avantages des deux méthodes de formation et d'atténuer les inconvénients, la méthode de formation semi-supervisée est proposée. Dans la méthode semi-supervisée, les vecteurs communs sont classés en plusieurs grands groupes par la méthode supervisée au préalable, puis les gaussiennes au sein de chaque groupe sont précisément formées par l'EM.

## **5.2. Évaluation de l'approche de GMM pour estimer les paramètres de la lèvre et les positions de la main**

L'approche de cartographie à base du GMM basée sur le GMM formé par la méthode supervisée montre une grande efficacité et une bonne robustesse bénéficiant de l'information phonétique a priori en comparaison avec le GMM formé par l'EM. La méthode de formation EM peut être plus affectée par les aberrantes due à la dépendance des données elles-mêmes. Mais la méthode de formation supervisée est également limitée en exigeant des connaissances a priori telles que les informations d'étiquette de chaque individu. Alors que la méthode de formation non supervisée (i.e. algorithme EM) surmonte le problème sans l'obtention des connaissances a priori à l'avance, les composants latents parmi les données de formation peuvent être automatiquement explorés, et d'assez bons résultats peuvent être obtenus, comme

indiqué dans notre travail. Mais les inconvénients de la méthode de formation EM comprennent la faible efficacité (en utilisant beaucoup plus de composants) pour atteindre les résultats comparables à ceux de la méthode de formation supervisée et le processus de formation relativement plus compliqué. La méthode de formation semi-supervisée ne montre pas d'amélioration apparente en comparaison avec les deux autres méthodes, cela est dû à la matrice de covariance commune est utilisé dans le processus de cartographie plutôt que la matrice de covariance arbitraire.

L'erreur quadratique minimale (MMSE) et le critère de régression probabilité a posteriori maximale (MAP) ne montrent pas de grandes différences dans n'importe quel modèle de cartographie. Dans notre travail, nous indiquons également que lorsque le nombre de la gaussienne de GMM est égal à 1, l'approche de cartographie à base du GMM est en effet égale à l'approche MLR. Il a été démontré dans les paramètres de la lèvre mais aussi dans l'estimation de la position de la main. L'approche de cartographie MLR est en effet un cas particulier de la méthode de cartographie à base de GMM avec une seule gaussienne.

*Tableau 5.1: Le RVAR optimal des paramètres de la lèvre estimés par 16 prédicteurs spectraux basés sur l'approche de cartographie à base de GMM sur les données de test selon les critères du MMSE.*

| (%) | MLR | Supervised GMM (3 components) | Supervised GMM (10 components) | Unsupervised GMM | Semi-supervised GMM (35 comp.) |
|-----|-----|-------------------------------|--------------------------------|------------------|--------------------------------|
| A   | 17% | 9%                            | 9%                             | 7% (38 comp.)    | 7%                             |
| B   | 12% | 9%                            | 12%                            | 9% (40 comp.)    | 9%                             |
| S   | 14% | 8%                            | 10%                            | 8% (40 comp.)    | 8%                             |

*Tableau 5.2: Le RVAR optimal des positions de la main estimée par 16 prédicteurs spectraux basé sur l'approche de cartographie à base de GMM sur les données de test selon les critères du MMSE.*

| (%) | MLR direct | MLR indirect | Supervised GMM (5 comp.) | Supervised GMM (10 comp.) | Unsupervised GMM | Semi-supervised GMM |
|-----|------------|--------------|--------------------------|---------------------------|------------------|---------------------|
| X   | 42.57%     | 30.67%       | 7.65%                    | 13.97%                    | 7.98% (60 comp.) | 5.07% (19 comp.)    |
| Y   | 30.59%     | 21.88%       | 3.78%                    | 7.73%                     | 5.48% (54 comp.) | 3.78% (5 comp.)     |

## 5.4 Discussion sur les approches utilisées pour estimer la position de la main

Nous avons utilisé une interpolation linéaire entre les voyelles [a] et [i] dans l'espace acoustique afin de comparer les différentes approches de cartographie telles que la méthode de cartographie à base de MLR, GMM et la méthode de classification à base de GMM (voir Figure 5.1, Figure 5.2). La méthode MLR produit un résultat linéaire continu avec une grande diversion en particulier dans le cas de l'estimation de la position de la main car il n'y a pas de relation entre l'espace acoustique et l'espace de la position de la main. La méthode de classification à base de GMM obtient les réponses d'étape avec les valeurs moyennes de chaque groupe. La méthode de classification permet d'éliminer la variance de la source de données au sein du groupe. En raison de la caractéristique de classification, cette méthode diminue la variance résiduelle par rapport à la MLR. L'approche de cartographie à base de GMM présente les propriétés de classification et de régression. Au sein de chaque phase de stabilité, la régression locale permet de projeter localement la variance des données de source. Cependant, la régression locale a une signification différente dans le cas de l'estimation des paramètres de la lèvre ou de l'estimation de positions de la main. Dans le cas de l'estimation des paramètres de la lèvre, la régression locale a la même tendance que l'approche MLR car les paramètres de la lèvre ont une forte corrélation linéaire avec les paramètres spectraux. Dans le cas de l'estimation de la position de la main, la régression locale ne varie pas en fonction du MLR même dans la direction opposée, puisque le mouvement de la main n'a pas de relation avec les paramètres spectraux. En outre, la transition est facile entre les phases stables de l'approche de cartographie à base de GMM, ce qui est différent de la transition binaire dans la méthode de classification. A partir de la transition, nous pouvons voir qu'il est la propriété de partitionnement de classe qui permet à l'approche de cartographie à base de GMM de diminuer de manière significative la variance résiduelle par rapport à l'approche MLR.

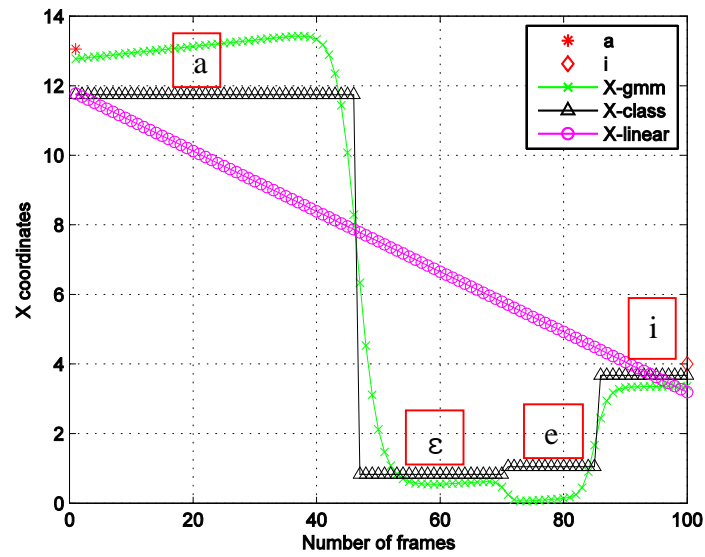


Figure 5.1 La transition dynamique de coordonnées  $X$  de la position de la main par interpolation entre la voyelle [a] et [i].

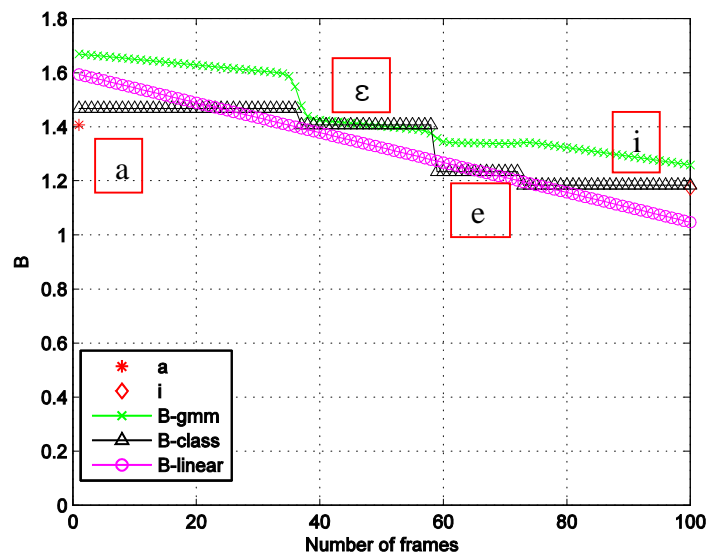


Figure 5.2 La transition du paramètre  $B$  de la lèvre par interpolation entre la voyelle [a] et [i].



## **6. L'estimation de caractéristiques de lèvres à partir de l'image naturelle de la région d'intérêt (ROI) de la bouche**

Cette étude est une contribution dans le domaine de traitement de la parole visuelle. Il se concentre sur l'extraction automatique des caractéristiques de la lèvre en discours à partir de l'image naturelle de la lèvre. La méthode est basée sur la prédiction directe de ces caractéristiques à partir des prédicteurs dérivant d'une transformation appropriée des pixels de la région d'intérêt de la lèvre. La transformation est réalisée par une Transformée 2-D en Cosinus Discrète (DCT) en combinaison avec une analyse du composant principal appliquée à un sous-ensemble des coefficients DCT. Les approches de cartographie à base de MLR et GMM ont été utilisées séparément pour prédire les caractéristiques géométriques de la lèvre en parole. Dans le cas de l'approche MLR, la variance du paramètre B de la lèvre par estimation peut atteindre 95% avec 17 prédicteurs issus du DCT correspondant à environ 3% du total du DCT. Mais cette approche a des limites lorsque les données à être prédites sont proches de zéro. Toutefois, les valeurs estimées données par l'approche de cartographie à base de GMM sont apparemment plus proches de la cible. En outre, l'approche de cartographie à base de GMM est aussi plus efficace que le MLR, comme il n'utilise que 8 prédicteurs issus du DCT correspondant à environ 1% du total des DCTs pour converger en contraste avec les 17 dans le cas du MLR. Par ailleurs, le GMM qui est composé de 3 composants a la meilleure performance en comparaison avec le GMM avec 4 ou 5 composants. L'approche de cartographie à base de GMM est également affectée par le problème de sur-correspondance causé par l'utilisation de trop de prédicteurs. Toutefois, le nombre de prédicteurs rendant les valeurs estimées à la divergence est bien supérieur que le nombre de prédicteurs nécessaires pour la convergence.

## **7. Conclusion et perspectives**

Dans cette thèse, nous nous concentrons sur le problème de la cartographie à partir des paramètres spectraux acoustiques de la parole aux paramètres CS et les paramètres de la lèvre (la largeur de la lèvre, la hauteur de la lèvre et la surface de la lèvre) concernant les voyelles en français. Cette étude se situe dans le cadre général de la conversion du CS à la parole.

En ce qui concerne les formes acoustique-à-lèvre, le problème peut être partiellement considéré comme un problème de cartographie acoustique-à-articulatoire (A à A). Mais en termes de la position acoustique-à-main, il y a un nouveau problème. Comme il n'existe pas de relation entre la position de la main et les paramètres acoustiques du signal de la parole, nous avons d'abord appliqué l'approche MLR afin de connaître la limite en termes de la performance et aussi de comprendre pourquoi la cartographie linéaire ne peut pas offrir une précision suffisante d'estimation. Pour utiliser les paramètres spectraux de la parole pour estimer les paramètres de la lèvre ou de la main, nous avons d'abord appliqué l'analyse en composantes principales (PCA) sur les paramètres spectraux de la parole afin de surmonter le problème de corrélation entre les paramètres spectraux qui mènent aux estimations instables et potentiellement trompeuses de l'équation de régression (Jolliffe, 2002) et de réduire le nombre de régresseurs dans le modèle MLR surtout lorsque le groupe est large. Une bonne performance peut être obtenue en utilisant les 16 prédicteurs dérivés du mélange de coefficients MFCC et LSP qui a la meilleure performance en comparaison avec les autres paramètres spectraux, comme les formants, MFCC ou LSP. En utilisant l'approche MLR, les paramètres de la lèvre peuvent être estimés avec une précision favorable. Mais pour l'estimation de la position de la main, la performance est médiocre.

Les raisons pour cette mauvaise performance sont principalement: 1). Il n'y a aucune relation entre la position de la main et les paramètres acoustiques du signal de la parole; 2). Les positions de la main pour chaque voyelle sont spécialement conçues pour éliminer l'ambiguïté dans la lecture labiale. Une conséquence directe est le suivant: pour deux voyelles "acoustique proche" (par exemple, [i] et [e]), leur position correspondante dans l'espace de la main sont très loin! Cependant, deux voyelles loin dans l'espace acoustique peuvent partager la même position de la main, comme pour les voyelles [a] et [o]. Après plusieurs études de simulation, il a été démontré que l'espace de la main et de l'espace du spectre acoustique ont une structure de topologie totalement différente qui cause la mauvaise performance de l'approche MLR pour estimer la position de la main.

Pour résoudre ce problème, nous avons mis en place un espace intermédiaire pour estimer la position des mains, dans lequel la position de la main de chaque voyelle est déplacée suivant la règle: leur situation relative doit être en cohérence avec leur position "acoustique". Par conséquent, la position redistribuée de la main a une structure topologique similaire avec l'espace de spectre acoustique, par exemple le "triangle acoustique" composé par F1-F2 dans l'espace de formant.

Cela prouve que la performance de l'approche MLR pour estimer la position de la main s'est améliorée dans l'espace intermédiaire bénéficiant de la structure de topologie similaire. Cependant, le problème de la classification pendant le processus de reconfiguration n'a pas réussi à obtenir la bonne performance dans l'espace d'origine.

Tous ces résultats montrent que même si un certain niveau de linéarité est conservé, c'est aux dépens d'une forte variance de l'estimateur linéaire. Lorsque nous utilisons l'espace intermédiaire, le problème de la variance résiduelle élevée est résolu. Mais la nécessité d'utiliser une reconfiguration provoque de grandes erreurs de classification. Afin de continuer à libérer les contraintes de linéarité, nous avons introduit la méthode de cartographie basée sur le mélange de gaussiennes (GMM). C'est exactement notre cas.

La méthode GMM nécessite une formation plus complexe que l'approche MLR. Dans ce travail, le GMM a été formé par les trois méthodes dans la vue de la théorie d'apprentissage de la machine: la méthode supervisée, non supervisée et semi-supervisée de formation. La méthode de formation supervisée est efficace, simple et robuste. Mais elle est également limitée par l'acquisition d'une connaissance a priori avant la formation. La méthode de formation non supervisée (comme l'EM) résout le problème dans la méthode de formation supervisée en regroupant les données automatiquement. Toutefois, la vitesse de convergence de la méthode EM est lente comme le regroupement est entièrement dépendant des données et que les valeurs aberrantes affectent probablement la convergence. La méthode de formation semi-supervisée est proposée visant à inclure les avantages des méthodes de formation supervisée et non supervisée. La méthode semi-supervisée ne montrent pas d'amélioration évidente par rapport aux autres méthodes en raison de la matrice de

covariance commune qui est utilisée et qui affaiblit la capacité individuelle d'ajustement de chaque composant. Lorsque les paramètres GMM sont fixes, les méthodes de régression d'erreur quadratique minimale (MMSE) et de probabilité a posteriori maximale (MAP) ont été utilisées séparément pour estimer les cibles. Le MMSE et MAP n'ont pas montré de grandes différences pour estimer les paramètres de la lèvre ou de la position de la main.

Dans notre travail, il indique aussi que l'approche de cartographie MLR est effectivement le cas particulier de l'approche à base de GMM avec une seule gaussienne. En augmentant le nombre de composants du GMM, la performance d'estimation de l'approche de cartographie à base de GMM peut améliorer considérablement sa performance de cartographie en comparaison avec l'approche MLR.

Nous avons aussi recherché l'extraction automatique des caractéristiques géométriques de la lèvre à partir de l'apparence de la lèvre qui est obtenue par la transformation DCT combinée avec le PCA de la bouche du locuteur et les images naturelles du ROI. Les approches de cartographie à base de MLR et GMM ont été séparément utilisées pour prédire les caractéristiques géométriques de la lèvre durant la parole. Par l'approche MLR, nous pouvons utiliser seulement 2% du total des DCTs pour expliquer 95% de la variance en termes du paramètre B de la lèvre, tandis que l'approche de cartographie à base de GMM améliore d'avantage la performance d'estimation au cas où les données de cible sont proches de zéro. En outre, l'approche de cartographie à base de GMM a un rendement plus élevé pour l'estimation des paramètres de la lèvre en comparaison avec l'approche MLR. Cependant, l'approche de cartographie à base de GMM introduit le problème de sur-correspondance lorsque la dimension du GMM est trop élevée.

Dans le futur, l'approche de cartographie à base de GMM devrait être étendue à la situation continue. Dans notre travail, nous avons déjà une discussion préliminaire sur la performance de différentes approches de cartographie de la situation continue réalisée par l'interpolation linéaire dans l'espace de spectre acoustique. A partir des transitions obtenues par l'interpolation linéaire, nous pouvons voir clairement les différentes propriétés de l'approche MLR, l'approche de cartographie à base de

GMM et l'approche de classification à base de GMM. L'approche MLR produit toujours un résultat linéaire continu mais la performance dépend de la corrélation linéaire entre la source et la cible. L'approche de classification à base de GMM obtient une réponse d'étape qui localise les données dans le centre de chaque groupe de GMM. La méthode de classification ne peut pas projeter la variance de données de source au sein du groupe et supprime ainsi la propriété continue de la source de données. L'approche de cartographie à base de GMM bénéficiant à la fois des propriétés de régression linéaire et de celles de classification peut améliorer les performances d'estimation par les propriétés de classification lorsque la corrélation linéaire entre les données de source et les données de cible est faible. Cette approche peut aussi projeter la variance de données de base au niveau local par les propriétés de régression linéaire. D'ailleurs, la transition entre les phrases de l'approche de cartographie à base de GMM est continue en pondérant les contributions de toutes les gaussiennes contrairement à la méthode de classification binaire. Ainsi, il est raisonnable d'étendre l'approche de cartographie à base de GMM sur le problème de l'estimation continue comme le problème d'estimation continue du CS qui comprend les voyelles et les consonnes. Par ailleurs, nous pouvons aussi explorer de nouvelles approches pour estimer les composants CS dans le contexte continu, par exemple les approches proposées récemment par (Zen et al., 2011; Zen et al., 2012; Shannon et al., 2013) sur la base de la trajectoire GMM ou la trajectoire HMM dans l'inversion acoustique-à-articulatoire, dans le domaine de la synthèse vocale, qui examine la trajectoire des trames voisines au lieu de la trame unique dans le processus d'estimation.

Mots clés: LPC; mapping de la parole acoustique vers LPC; modèle linéaire; GMM; MMSE; MAP.

# Bibliographie

- Aboutabit, N. (2007). *Reconnaissance de la Langue Française Parlée Complétée (LPC): décodage phonétique des gestes main-lèvres*. (Ph.D thesis), Institut National Polytechnique de Grenoble-INPG.
- Alegria, J., Charlier, B. L., & Mattys, S. (1999). The role of lip-reading and cued speech in the processing of phonological information in French-educated deaf children. *European Journal of Cognitive Psychology*, 11(4), 451-472.
- Attina, V. (2005). *La Langue française Parlée Complétée: production et perception*. (Ph.D thesis), Institut National Polytechnique de Grenoble-INPG.
- Attina, V., Beautemps, D., & Cathiard, M.-A. (2002). Coordination of hand and orofacial movements for CV sequences in French Cued Speech. in *Proc. Seventh International Conference on Spoken Language Processing*, Denver, Colorado, USA, pp.1945-1948.
- Attina, V., Beautemps, D., Cathiard, M.-A., & Odisio, M. (2004). A pilot study of temporal organization in Cued Speech production of French syllables: rules for a Cued Speech synthesizer. *Speech Communication*, 44(1), 197-214.
- Beautemps, D., Girin, L., Aboutabit, N., Bailly, G., Besacier, L., Breton, G., Burger, T., Caplier, A., Cathiard, M.-A., Chêne, D., Clarke, J., Elisei, F., Govokhina, O., Jutten, C., Le, V.-B., Marthouret, M., Mancini, S., Mathieu, Y., Perret, P., Rivet, B., Sacher, P., Savariaux, C., Schmerber, S., Sérignat, J.-F., Tribout, M., & Vidal, S. (2007). Telma: Telephony for the hearing-impaired people. From models to user tests. in *Proc. ASSISTH'2007*, pp.201-208.
- Charlier, B., Hage, C., Alegria, J., & Périer, O. (1990). Evaluation d'une pratique prolongée du LPC sur la compréhension de la parole par l'enfant atteint de déficience auditive. *Glossa*, 22, 28-39.
- Chen, Y., Chu, M., Chang, E., Liu, J., & Liu, R. (2003). Voice conversion with smoothed GMM and MAP adaptation. in *Proc. Eurospeech-2003*, Geneva, Switzerland, pp.2413-2416.
- Chiou, G. I., & Hwang, J.-N. (1997). Lipreading from color video. *IEEE Transactions on Image Processing*, 6(8), 1192-1195.
- Clarke, B. R., & Ling, D. (1976). The Effects of Using Cued Speech: A Follow-Up Study. *Volta Review*, 78(1), 23-34.
- Cornett, R. O. (1967). Cued speech. *American Annals of the Deaf*, 112, 3-13.
- Cornett, R. O. (1988). Cued speech, manual complement to lipreading, for visual reception of spoken language. Principles, practice and prospects for automation. *Acta Oto-Rhino-Laryngologica Belgica*, 42(3), 375.
- Cotton, J. C. (1935). Normal "visual hearing.". *Science*, 82(2138), 592-593.
- Dodd, B. (1977). The role of vision in the perception of speech. *Perception*, 6(1), 31-40.
- Duchnowski, P., Lum, D. S., Krause, J. C., Sexton, M. G., Bratakos, M. S., & Braid, L. D. (2000). Development of speechreading supplements based on automatic speech recognition. *IEEE Transactions on Biomedical Engineering*, 47(4), 487-496.
- Dupont, S., & Luetin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3), 141-151.

- Gibert, G., Bailly, G., Beutemps, D., Elisei, F., & Brun, R. (2005). Analysis and synthesis of the three-dimensional movements of the head, face, and hand of a speaker using cued speech. *The Journal of the Acoustical Society of America*, 118, 1144.
- Gray, M. S., Movellan, J. R., & Sejnowski, T. S. (1997). Dynamic features for visual speechreading: A systematic comparison. in *Proc. 3rd Joint Symposium on Neural Computation.*, La Jolla, CA, Vol.6, pp.222-230
- Hennecke, M. E., Stork, D. G., & Venkatesh Prasad, K. (1996). Visionary speech: Looking ahead to practical speechreading systems. In D. Stork & M. Hennecke (Eds.), *Speechreading by humans and machines: NATO ASI series, series F: Computer and systems science*. Berlin: Springer, Vol. 150, pp. 331-350.
- Heracleous, P., Aboutabit, N., & Beutemps, D. (2009). Lip shape and hand position fusion for automatic vowel recognition in cued speech for french. *IEEE Signal Processing Letters*, 16(5), 339-342.
- Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P., & Saltzman, E. (1996). Accurate recovery of articulator positions from acoustics: New conclusions based on human data. *The Journal of the Acoustical Society of America*, 100, 1819.
- Huang, F. J., & Chen, T. (1998). Real-time lip-synch face animation driven by human voice. in *Proc. IEEE Second Workshop on Multimedia Signal Processing*, Los Angeles, CA, pp.352-357.
- Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.
- Kain, A., & Macon, M. W. (1998). Spectral voice conversion for text-to-speech synthesis. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, WA, Vol.1, pp.285-288.
- Kjellström, H., & Engwall, O. (2009). Audiovisual-to-articulatory inversion. *Speech Communication*, 51(3), 195-209.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. in *Proc. International joint Conference on artificial intelligence*, Vol.14, pp.1137-1145.
- Leung, S.-H., Wang, S.-L., & Lau, W.-H. (2004). Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. *IEEE Transactions on Image Processing*, 13(1), 51-62.
- Leybaert, J. (2000). Phonology acquired through the eyes and spelling in deaf children. *Journal of experimental child psychology*, 75(4), 291-318.
- Ling, D., & Clarke, B. R. (1975). Cued Speech: An Evaluative Study. *American Annals of the Deaf*, 120(5), 480-488.
- Luetten, J., Thacker, N. A., & Beet, S. W. (1996). Visual speech recognition using active shape models and hidden Markov models. in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol.2, pp.817-820.
- Matthews, I. (1998). *Features for audio-visual speech recognition*. (Ph.D thesis), University of East Anglia.
- Matthews, I., Bangham, J. A., & Cox, S. (1996). Audiovisual speech recognition using multiscale nonlinear image decomposition. in *Proc. Fourth International Conference on Spoken Language*, Philadelphia, PA, Vol.1, pp.38-41.

- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746-748.
- Montgomery, A. A., & Jackson, P. L. (1983). Physical characteristics of the lips underlying vowel lipreading performance. *The Journal of the Acoustical Society of America*, 73, 2134.
- Morishima, S., & Harashima, H. (1991). A media conversion from speech to facial image for intelligent man-machine interface. *IEEE Journal on Selected Areas in Communications*, 9(4), 594-600.
- Nakamura, K., Toda, T., Nankaku, Y., & Tokuda, K. (2006). On the use of phonetic information for mapping from articulatory movements to vocal tract spectrum. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, pp.93-96.
- Nicholls, G. H., & Ling, D. (1982). Cued Speech and the reception of spoken language. *Journal of Speech, Language and Hearing Research*, 25(2), 262.
- Potamianos, G., Neti, C., Luetin, J., Matthews, I., Bailly, G., Vatikiotis-Bateson, E., Beautemps, D., Attina, V., Badin, P., Ben, Y., Bernstein, L., Beskow, J., Bregler, C., Brooke, N. M., Bruce, V., Burnham, D., Campbell, R., Cathiard, M.-A., Clark, R., Cohen, M. M., Ezzat, T., Geiger, G., Laboissière, R., Karen, L., Loevenbruck, H., Macsweeney, M., Massaro, D. W., Munhall, K., Poggio, T. A., Remez, R. E., Révère, L., Savariaux, C., Schwartz, J.-L., Scott, S. D., Sekiyama, K., Slaney, M., Tabain, M., Vatikiotis-Bateson, E., & Vilain, V. (2012). Audiovisual automatic speech recognition. In G. Bailly, P. Perrier & E. Vatikiotis-Bateson (Eds.), *Audiovisual speech processing*: Cambridge University, pp. 193-247.
- Reisberg, D., Mclean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*. London, U.K: Lawrence Erlbaum, pp. 97-113.
- Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1), 72-83.
- Richmond, K. (2009). Preliminary inversion mapping results with a new EMA corpus. in *Proc. Interspeech 2009*, Brighton, UK, pp.2835-2838.
- Richmond, K., King, S., & Taylor, P. (2003). Modelling the uncertainty in recovering articulation from acoustics. *Computer Speech & Language*, 17(2), 153-172.
- Schroeter, J., & Sondhi, M. M. (1994). Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(1), 133-150.
- Shannon, M., Zen, H., & Byrne, W. (2013). Autoregressive Models for Statistical Parametric Speech Synthesis. *IEEE Trans. Audio Speech Language Process.*, 21(3), 587-589.
- Stylianou, Y., Cappé, O., & Moulines, E. (1998). Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 6(2), 131-142.
- Toda, T., Black, A. W., & Tokuda, K. (2004). Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis. in *Proc. Fifth ISCA Workshop on Speech Synthesis*, Pittsburgh, USA, pp.31-36.



- Toda, T., Black, A. W., & Tokuda, K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Speech and Audio Processing*, 15(8), 2222-2235.
- Toda, T., Black, A. W., & Tokuda, K. (2008). Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, 50(3), 215-227.
- Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, Vol.3, pp.1315-1318.
- Uchanski, R. M., Delhorne, L., Dix, A., Braida, L., Reed, C., & Durlach, N. (1994). Automatic speech recognition to aid the hearing impaired: prospects for the automatic generation of cued speech. *Rehabilitation Research and Development*, 31(1), 20.
- Uto, Y., Nankaku, Y., Toda, T., Lee, A., & Tokuda, K. (2006). Voice conversion based on mixtures of factor analyzers. in *Proc. Interspeech 2006*, Pittsburgh, USA, pp.2278-2281.
- Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1), 23-43.
- Zen, H., Gales, M. J., Nankaku, Y., & Tokuda, K. (2012). Product of experts for statistical parametric speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 20(3), 794-805.
- Zen, H., Nankaku, Y., & Tokuda, K. (2011). Continuous Stochastic Feature Mapping Based on Trajectory HMMs. *IEEE Transactions on Speech and Audio Processing*, 19(2), 417-430.
- Zen, H., Tokuda, K., & Kitamura, T. (2004). An introduction of trajectory model into HMM-based speech synthesis. in *Proc. Fifth ISCA Workshop on Speech Synthesis*, pp.191-196.