In compliance with the
Canadian Privacy Legislation
some supporting forms
may have been removed from
this dissertation.

While these forms may be included
in the document page count,
their removal does not represent
any loss of content from the dissertation.

Lip Feature Extraction and Audio - Video Synchronization

submitted by

Chengliang Li, M. Eng.

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of

Master of Applied Science

Ottawa-Carleton Institute for Electrical and Computer Engineering
Faculty of Engineering
Department of Systems and Computer Engineering

Carleton University
Ottawa, Ontario, Canada K1S 5B6
May 2003

© Copyright 2003, Chengliang Li

National Library
of Canada

Bibliothèque nationale
du Canada

Acquisitions and
Bibliographic Services

Acquisitions et
services bibliographiques

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-612-88880-0

Our file Notre référence

ISBN: 0-612-88880-0

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

Canada

ABSTRACT

Human speech is bimodal in nature: acoustic and visual. The spectral acoustic speech is shaped by the time-varying configuration of the vocal tract. The vocal tract events are relevant to the facial motions. The dependence of acoustic and facial events on the behavior of the vocal tract implies that there is an inherent relationship between acoustic and visible aspects of speech.

In this thesis, to allow exploitation of the relationship between acoustic and visual speech, we present an automatic algorithm to extract the components of visual speech, such as lip position which includes the lip width, outer lip height and inner lip height. These features can be extracted by making use of both color and edge information, combined within a Markov random field (MRF) framework.

The extracted lip features are then used in the research work on the problem of acoustic to visual speech mapping. Linear spectrum pair (LSP) coefficients are used to parameterize the acoustic speech. LSP coefficients and the extracted lip features are then used to train a mapping model. The resulting model can be used to estimate lip features from acoustic speech. The estimation of lip features from acoustic speech gives a way to integrate acoustic and visual speech, which is very useful for speech driven face animation, audio-video synchronization, joint audio-video coding and foreign film dubbing.

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisors, Dr. Richard M. Dansereau and Dr. Rafik A. Goubran, for all of their guidance, patience and encouragement for my research.

I would like to thank all of the graduate students in the DSP lab for their helpful discussions. Also I would thank Danny Lemay and all other technicians for their help in keeping the network operational.

I would like to thank Mitel Networks, Natural Sciences and Engineering Research Council of Canada (NSERC) and Carleton University for their financial support.

Finally, I would like to thank my wife, for her constant support and encouragement during my university career.

TABLE OF CONTENTS

ABSTRACT	iii
ACKNOWLEDGEMENTS.....	iv
TABLE OF CONTENTS.....	v
LIST OF FIGURES	viii
LIST OF TABLES	x
LIST OF ABBREVIATIONS.....	xi
 CHAPTER 1 INTRODUCTION.....	 1
1.1 Overview.....	1
1.1.1 Human Speech Production.....	2
1.1.2 Human Speech Perception	3
1.1.3 Acoustic to Visual Speech Conversion.....	5
1.2 Problem Statement	6
1.3 Objective	8
1.4 Thesis Outline	9
 CHAPTER 2 BACKGROUND REVIEW.....	 11
2.1 Visual Speech Features	11
2.1.1 Global-based Visual Feature.....	12
2.1.2 Local-based Visual Feature.....	12
2.2 Visual Speech Extraction	13
2.2.1 Previous Approaches	14
2.2.1.1 Image-based Approach.....	14
2.2.1.2 Model-based Approach.....	15
2.2.1.3 Geometric Features Approach.....	17
2.2.2 Discussion on Feature Extraction Approaches	18
2.3 Acoustic to Visual Speech Conversion	20
2.3.1 Phonemes-visemes Conversion	21

2.3.2 Direct Conversion	23
2.3.2.1 Neural Network Approach.....	24
2.3.2.2 Gaussian Mixture Model and EM algorithm	24
2.4 Applications of Acoustic to Visual Speech Conversion.....	25
2.4.1 Lip Synchronization.....	25
2.4.2 Joint Audio-Video Coding.....	26
2.4.3 Speech-driven Face Animation.....	27
2.4.4 Foreign Film Dubbing.....	28
 CHAPTER 3 VISUAL FEATURE EXTRACTION.....	29
3.1 Color Image Representation.....	30
3.1.1 RGB	31
3.1.2 CMY	32
3.1.3 YIQ	32
3.1.4 HSI.....	32
3.1.5 Grayscale and Binary.....	33
3.2 Image Processing Techniques	33
3.2.1 Image Histogram.....	33
3.2.2 Edge Detection.....	34
3.2.2.1 Gradient-based Methods.....	35
3.2.2.2 Discrete Gradient Operators	36
3.2.2.3 Canny's Method	37
3.2.3 Motion Detection	39
3.2.4 Image Segmentation.....	40
3.3 Visual Speech Feature Extraction	42
3.3.1 Color Analysis	42
3.3.2 Lip Region Locating	47
3.3.3 MRF-based Lip Segmentation	51
3.3.3.1 Principle of Bayesian Segmentation.....	52
3.3.3.2 Markov Random Field.....	53
3.3.3.3 The Energy Function	54

3.3.3.4 High Confidence First (HCF) Algorithm	56
3.3.4 Lip Feature Extraction	59
3.3.4.1 Feature Points Detection.....	59
3.3.4.2 Lip Geometric Feature Extraction	62
3.4 Summary	64
 CHAPTER 4 ACOUSTIC-VISUAL SPEECH CONVERSION.....	66
4.1 Training Data Set	68
4.1.1 Visual Speech Extraction.....	68
4.1.2 Acoustic Speech Extraction	69
4.2 Mapping Model Training	74
4.2.1 Learning Criteria for Neural Networks Training.....	74
4.2.2 Optimization of Neural Network Architecture	75
4.2.2.1 Single Multi-output versus Multi Single-output Architecture.....	76
4.2.2.2 Number of Nodes in Hidden Layer	79
4.2.2.3 Linearity of the Nodes in the Hidden Layer	81
4.2.2.4 Training Algorithm.....	82
4.3 Estimation of Visual Feature from Acoustic Feature.....	84
4.4 Summary	88
 CHAPTER 5 CONCLUSIONS.....	91
REFERENCES	95

LIST OF FIGURES

Figure 1.1 The human speech production system.....	2
Figure 2.1 Phoneme-viseme conversion approach.....	22
Figure 2.2 Direct acoustic-visual speech conversion.....	24
Figure 3.1 Original image and ROI.....	42
Figure 3.2 Histograms of R/G/B components.....	43
Figure 3.3 Histograms of r/g/b components.....	44
Figure 3.4 Histogram of red hue	45
Figure 3.5 Lip feature extraction.....	47
Figure 3.6 Lip region locating.....	48
Figure 3.7 Lip segmentation by simple thresholding.....	51
Figure 3.8 Edges of the lip region in figure 3.7(a).....	52
Figure 3.9 Five-point clique system.....	54
Figure 3.10 Steps of MRF-based lip segmentation.....	58
Figure 3.11 Segmented lips overlayed on the original image.....	59
Figure 3.12 Illustration of the extracted geometric feature of the lips.....	59
Figure 3.13 (a) Tongue is separated from the lips. (b) Tongue merges with the lips	61
Figure 3.14 Examples of detected feature points	62
Figure 3.15 Visual features (W, H, h) for my utterance of English words '1' to '10'	63
Figure 4.1 LSP and LPC coefficients of the same speech.	70
Figure 4.2 Procedures of acoustic speech extraction.	71

Figure 4.3 (a) Single multi-output versus (b) Multi single-output architecture	77
Figure 4.4 Learning curves for different neural network architecture	78
Figure 4.5 Optimization of the number of nodes in the hidden layer.	80
Figure 4.6 Learning curves versus linearity of the nodes in the hidden layer.....	82
Figure 4.7 Learning curves of the same network set trained by different algorithms.....	83
Figure 4.8 Estimation results over one of the training video sequence.....	86
Figure 4.9 Estimation results over a testing video sequence.....	86
Figure 4.10 Estimation results over video sequence of utterance of 'audio'	87

LIST OF TABLES

Table 1.1 Examples of McGurk Effect	4
Table 2.1 Viseme Groups for English.....	22

LIST OF ABBREVIATIONS

ASM	Active Shape Model
AVI	Audio Video Interleave
BP	Back-propagation
CC	Cross Correlation
CGM	Conjugate Gradient Method
CIF	Common Intermediate Format
CMY	Cyan Magenta Yellow
EM	Expectation Maximization
FF	Feed-Forward
GMM	Gaussian Mixture Model
HCF	High Confidence First
HMM	Hidden Markov Model
HSI	Hue Saturation Intensity
ICM	Iterated Conditional Models
LP	Linear Prediction
LPC	Linear Prediction Coding
LSF	Line Spectral Frequency
LSP	Line Spectral Pair
MAP	Maximum <i>a posteriori</i>

MRF	Markov Random Field
MSE	Mean Square Error
NN	Neural Network
PCA	Principal Components Analysis
PCM	Pulse Code Modulation
RGB	Red Green Blue
ROI	Regions of Interest
SCG	Scaled Conjugate Gradient
TDNN	Time-Delay Neural Network
VQ	Vector Quantization

CHAPTER 1 INTRODUCTION

1.1 Overview

Human speech is bimodal in nature: acoustic and visual [1]. Speech is produced by the vibration of the vocal cord and the configuration of the vocal tract that is composed of articulatory organs. Acoustic speech often refers to the acoustic waveform produced by the speaker, which is shaped by the time-varying configuration of the vocal tract. The vocal tract events are relevant to the articulatory movements of the lips, tongue, and other facial muscles. The articulatory movements are often referred to as visual speech. The dependence of acoustic and articulatory events on the behavior of the vocal tract implies that there is an inherent relationship between the acoustic and visible aspects of speech [1]. This fact results in the existence of a conversion from acoustic to visual speech. So if one of them is known, the other can be estimated with a certain degree of accuracy. In my thesis, I focus on the visual speech extraction and estimation of visual speech from acoustic speech. These techniques are very useful for speech driven face animation, audio-video synchronization, joint audio-video coding and foreign film dubbing, etc.

1.1.1 Human Speech Production

Figure 1.1 shows the middle section of the human speech production system [2]. The main components which determine the speech waveform are the lungs, trachea, larynx, vocal cords, pharyngeal cavity, mouth and nose.

Speech production is based on the basic mechanisms of phonation, related to the vibration of the vocal cords, and of vocal articulation, related to the time-varying geometry of the vocal tract responsible for the phonemic structure of speech. As the diaphragm is forced up, air is pushed out from the lungs into the trachea, glottis (the gap between the two vocal cords), and larynx before reaching the upper part of the vocal tract. The vocal tract begins at the vocal cord constriction at the top of the trachea and ends at the lips. The cross-sectional area of the vocal tract is non-uniform and is determined by the position of the articulators which comprise the lips, jaw, tongue and palate. The vocal tract, in fact, can be varied in shape by moving the jaw, tongue, lips, and palate.

Figure 1.1 The human speech production system

From the human speech production discussed above, we can see that the spectral acoustic speech is shaped by the time-varying configuration of the vocal tract. The vocal tract events are relevant to the articulatory movements. The dependence of acoustic and facial events on the behavior of the vocal tract implies that there is an inherent relationship between the acoustic and visible aspects of speech. However, even under the best circumstances, only the lips, the frontal part of the tongue, and the jaw are visible to the speechreader. So it is generally agreed that most of the visual information is contained in the lips [3]. Therefore most research on the relationship between acoustic and visual speech focuses on the relationship between acoustic features and lip features.

1.1.2 Human Speech Perception

In addition to speech production, speech perception is also bimodal. Speech perception is determined by acoustic and visual speech cues. The large influence of visual articulation on human perception of speech is demonstrated by the McGurk effect [4]. The McGurk effect establishes the bimodality by showing that, when a person is presented with conflicting audio and visual stimuli, the person may perceive a sound that does not exist in either modality. For this experiment, subjects are presented simultaneously with an acoustic recording of an utterance and a visual recording of a speaking face corresponding to a different utterance. The effect is that what the subject perceived does correspond to neither the acoustic nor the visual stimuli. For example, when a person hears an audio recording of /ba/ and sees the synchronized video of a talker saying /da/,

he perceives neither /ga/ nor /ba/. Instead, what is perceived is something close to /da/. The McGurk effect shows that the speech perceived by a person depends not only on acoustic cues, but also on visual cues such as lip movements. Some other examples of audio-visual combinations are shown in Table 1.1.

Table 1.1 Examples of McGurk Effect

Acoustic	+	Visual	→	Perceived
ba		ga		da
pa		ga		ta
ma		ga		na

Due to the bimodality in speech perception, seeing the face of the speaker is evidently of great help to speech comprehension and almost necessary. Thus, acoustic-visual speech synchronization becomes an important design factor for multimodal communication systems, such as video telephony and video conferencing. In such communication systems, when the acoustic and visual speech is out of synchronization, the listeners may not perceive the speech correctly because of the speech perception bimodality. However, we can take advantage of the speech production bimodality, estimating visual speech from acoustic speech, so that what the listener is hearing and what the listener is seeing are the same speech. For this reason, the acoustic to visual speech conversion technique has been gaining interest recently [5] [6] [7].

1.1.3 Acoustic to Visual Speech Conversion

Among the possible interactions between different media types, the interaction between audio and video is the most interesting. A recent trend in multimedia research is to convert acoustic speech to visual speech to exploit such interactions. For multimedia applications that involve person-to-person conversations, e.g., video telephony and video conferencing, such interaction is very important. Research topics on audio-visual integration include speech-driven face animation, human-computer interfaces, lip synchronization, foreign film dubbing, joint audio-video coding, and bimodal person authentication.

One key issue in acoustic to visual speech conversion is to establish a model which can convert acoustic parameters to the mouth shape parameters. In other words, given the acoustic parameters a , such as the LSP (Linear Spectrum Pairs) or cepstral coefficients, the corresponding mouth shape v can be estimated by a certain function $v = f(a)$.

A number of approaches have been proposed for the task of converting acoustic speech to mouth shape parameters. The problem can be solved from two different perspectives, according to the acoustic speech analysis being used.

The first way is at the phoneme level. It stresses that speech is a linguistic entity. The speech is first segmented into a sequence of phonemes, and then each phoneme is mapped to the corresponding viseme. Viseme is the basic unit of mouth movements [20], which forms the smallest visibly distinguishable unit of visual speech. This scheme could

be implemented using a complete speech recognizer followed by a lookup table [8]. The advantage of this approach is that the acoustic speech signal is explored to the full extent, so that all the context information is utilized and co-articulations are completely incorporated. Therefore, this approach provides the most precise speech analysis. However, this approach has a certain amount of computational overhead because to achieve acoustic-to-visual speech conversion one does not really need to recognize the spoken words or sentences. In addition, the construction of the lookup table that maps phonemes to visemes is non-trivial.

The other way is direct conversion at the frame level. It concentrates on speech being a physical phenomenon. Since there is a physical relationship between the shape of the vocal tract and the sound that is produced, there exists a functional relationship between the acoustic speech parameters, typically LSP coefficients or LPC (Linear Prediction Coding) cepstrum, and the visual parameters set. The conversion problem becomes one of finding the best functional approximation, given sets of training data. There are many algorithms that can be modified to perform this task. These approaches include neural networks [10], and hidden Markov models with Gaussian mixtures [6].

1.2 Problem Statement

Because human speech perception is bimodal, audio-video synchronization is one of the most important issues in audio-video communication systems. In some video conferencing systems, the microphone array is used to locate the approximate speaker

location and then a camera is turned to put the speaker in its field of view. It possible that there are other people in the camera's field of view, and these people could be confused for the speaker if they are somehow moving their mouth (i.e., chewing gum). In such a scenario, it is very hard for the system to decide which person is the actual speaker.

One solution for this problem can be implemented as follows: first, the lip features of each person are extracted. Next a set of lip features can be estimated from the acoustic speech. Then the estimated lip features are compared with each of the extracted lip features over a period of time. The extracted features close to the estimated features should belong to the actual speaker. Thus the actual speaker is detected and the camera can be zoomed to get the speaker in the center of the screenshot.

A good acoustic-visual speech conversion model is crucial for such a scenario. Acoustic-visual speech conversion models are also useful for other scenarios, such as speech-driven face animation, human-computer interface, foreign film dubbing and joint audio-video coding. Because of its practical use, the acoustic to visual speech conversion technique has been gaining interest and deserves further investigation.

A visual speech feature extraction algorithm is also of importance for the audio-video synchronization scenario above. It is generally agreed that most of the visual information is contained in the lips [3], so visual speech feature extraction essentially is the extraction of the geometric lip features. Lip feature extraction is a difficult problem due to the large appearance variability across different people and due to appearance variability during speech production. Different illumination, different face complexion and different poses of the face cause further difficulties in the image analysis. To overcome these difficulties,

automatic lip feature extraction systems have been considered by many researchers. This problem is another emphasis in my thesis.

1.3 Objective

The objectives of the thesis are as follows:

- Develop and investigate the algorithm to locate the approximate lip region from a color image sequence.
- Develop and investigate the algorithm to segment the exact lip from the resulting lip region and to extract geometric lip features from the segmented lip.
- Investigate and train an acoustic-visual speech conversion model using training data set.
- Estimate visual speech from acoustic speech using the conversion model and evaluate the performance of the conversion model, by comparing the estimated visual features with the actual visual features qualitatively and quantitatively.

The main contributions of the thesis are the approach used to extract the visual speech and the methods used to investigate the relationship between the acoustic and visual speech. It breaks the boundary between the acoustic speech feature space and visual speech feature space. It is useful in a large number of multimedia application scenarios.

1.4 Thesis Outline

The rest of the thesis is organized as follows:

Chapter 2 reviews different previous approaches of lip feature extraction. These approaches include an image-based approach, a model based approach and a geometric feature segmentation based approach. During the introduction of each approach, their advantages and disadvantages are also discussed. We also review different acoustic to visual speech conversion techniques and their applications in different multimedia scenarios in the remaining part of this chapter.

Chapter 3 concentrates on our proposed lip feature extraction algorithm – three-stage MRF (Markov Random Field) based lip segmentation. In this chapter, we first introduce some useful image processing techniques which are used in our proposed approach. In the first stage of our approach, the approximate lip region is located using color and motion information. In the second stage, the exact lip area is segmented from the resulting lip region, based on a Markov random field (MRF) framework combining color and edge information. The geometric lip features, such as the lip width (W), inner height (h) and outer lip height (H) are then derived from the segmented lip area.

Chapter 4 investigates the relationship between the acoustic and visual speech features. We use neural networks to model their relationship. To obtain such a model, some known acoustic and visual speech features are used to train the model. In my thesis, the visual speech features include the lip width (W), inner height (h) and outer lip height (H), which are extracted using our approach. The acoustic speech feature is represented

by LSP coefficients, which are useful because they are strongly related to the speech formants, which are determined by the vocal tract shape. The optimal model architecture is obtained by empirical experiments in terms of MSE (Mean Square Error) criterion. Then the performance of the resulting model is evaluated by comparing the estimated and actual lip feature using some other data set than the training data set.

Chapter 5 concludes the thesis and proposes some suggestions for further studies.

CHAPTER 2 BACKGROUND REVIEW

In this chapter, we will review different approaches for lip feature extraction as found in the literature. These approaches include the image-based approaches, model based approaches and geometric feature segmentation based approaches. During the introduction of each approach, their advantages and disadvantages are also discussed. We also review different acoustic to visual speech conversion techniques and their applications in different multimedia scenarios in the remaining part of this chapter.

2.1 Visual Speech Features

The speech waveform is an acoustic sound pressure wave which originates from movements of the human speech production system. The main components which determine the speech waveform are the lungs, trachea, vocal cord, pharyngeal cavity, mouth and nose.

The articulators which influence the speech production are the vocal cords, palate, tongue, teeth, lips, jaw and air pressure. The articulators move to different positions to produce different speech sounds. The movements of the articulators are referred to as visual speech. It is generally agreed that most visual speech information is contained in the lips [3]. Thus, visual speech analysis mainly focuses on lip feature extraction. The

choice for a visual representation of lip movement has led to different approaches. At one extreme, the entire image of the talker's mouth is used as a feature. With other approaches, only a small set of parameters describing the relevant information of the lip movement is used.

2.1.1 Global-based Visual Feature

Global image approaches are based on transformations with minimal information reduction. They are usually data intensive. Often, the intensity image containing the mouth is either used directly or after some pre-processing. Advantages of this method are that very little data is disregarded, since it is not known, for example what lip information is truly important (protrusion, position, rounding, teeth, tongue, dynamics). A disadvantage is that global-based approaches are usually very sensitive to lighting conditions, viewing angle and orientation. They are also computationally expensive, which usually rules them out of real-time lip tracking applications. Works in [17] [18] [19] [21] use global-based visual features to improve speech recognition by integrating the acoustic-visual speech features.

2.1.2 Local-based Visual Feature

Local-based visual feature refers to specific facial feature points. This type of features is mainly parameterized descriptions of the lip contours or geometric lip features, which are assumed to contain valuable visual speech information.

The main advantage of this approach is its potential ability to represent the visual speech in a lower dimensionality than the holistic approach. Generally, approaches, including the deformable model approaches and geometric feature approaches, are often used to extract local-based visual features. Recently, most researchers in this area have used local-based visual features to represent the visual speech [7] [10] [13] [15] [25] [26] [27] [28].

2.2 Visual Speech Extraction

In any acoustic-visual speech integration system, it is required that some sort of features be extracted from the image sequence to provide the desired information. Visual speech feature extraction and modeling has become an important issue in potential audio-video applications. It is also a difficult problem due to the large appearance variability across different persons and due to appearance variability during speech production. Different illumination and different poses of the face cause further difficulties in the image analysis. These difficulties arising from such variability, however, are sometimes circumvented by taking some kinds of measures, such as marking the subjects' lips with color or reflective markers [7][15], recording the lip movements with a head mounted camera [16], using one subject only for the experiments, hand segmenting the lip region, or using very controlled lighting conditions [10] [13].

Motivated by psychological studies, several researchers have developed automatic visual speech feature extraction algorithms. While these systems have validated the

benefit of visual speech information, there are still many discussions about how to determine which visual features are important for acoustic-visual speech integration applications, how to represent them, and how to extract them automatically in a robust manner. Although some approaches circumvent these difficulties by applying some constraints, there are also many approaches to solve the feature extraction problems without the use of any constraint.

2.2.1 Previous Approaches

The problem of visual feature extraction has been studied extensively. The choice for a visual representation of lip movements has led to different approaches. At one extreme, the entire images of the talker's mouth are used as a feature. This is the image-based approach. With other approaches, only a small set of parameters describing the relevant information of the lip movement is used. These approaches include model-based approaches and geometric feature-based approaches. The main approaches for extracting visual speech information from image sequences can be grouped into the image based, model based and geometric feature based approaches.

2.2.1.1 Image-based Approach

In the image-based approach, the whole image containing the mouth area is used as a feature either directly [17] [18], or after some preprocessing such as a principal components analysis [19] or vector quantization [21]. Recently, more sophisticated data

preprocessing has been used, such as a linear discriminant analysis projection and maximum likelihood linear transform feature rotation [22]. The advantage of the image-based approach is that no information is lost, but it is hard to determine the relevant features in the image. A common shortcoming of this approach is that it tends to be very sensitive to changes in illumination, position, camera distance, rotation and speaker [21].

2.2.1.2 Model-based Approach

In the model based approach, a model of the visible speech articulators, usually the lip contour, is built and its configuration is described by a small set of parameters. The advantage of this method is that important features can be represented in a low dimensional space and can be made invariant to translation, scaling, rotation and lighting. A disadvantage is that the particular model used may not consider all relevant speech information. The main difficulty in the model-based approach is the design of the model topology and the design of a robust algorithm which accurately maps the model to the image.

The application of **deformable templates** [25] for lip-tracking has been described in [26] [27]. The outline of the lips is modeled by a set of polynomials, which are matched onto the outline of the lips, represented by the image gradient. Deformable templates provide both a parameterized model and an energy function. The model may be simple, such as modeling the outer contour of the mouth by two intersecting parabolas [27]. More complex models can also be built [26]. These models have different numbers of parameters. The simplest two-parabola model may be specified by four parameters while

a complex mouth model may contain more than ten parameters. In order to extract the visual speech feature or lip contour, the model then deforms and moves over the entire image, in an attempt to minimize a certain energy function. Since the deformation of deformable templates is constrained by the initial choice of polynomials, they are often not able to resolve fine contour details. An image search is performed by fitting the template to image gradients, assuming strong edges at the lip contours. This assumption is, however, often violated, as the gradient along the contour is dependent on the speaker, illumination, reflection, facial hair, visibility of teeth, and mouth opening. A similar approach, but based on color information, has been presented in [28].

Kass et al. [29] have described active contour models, so-called "snakes", for lip tracking. These are able to resolve fine contour details but shape constraints are difficult to incorporate and one has to compromise between the degree of elasticity and the ability to resolve fine contour details. Snakes allow one to parameterize a closed contour. When the total energy is minimized, the optimal snake location can be found. Another modified version of snakes, together with the use of color information, has been proposed by Chiou [30].

Another deformable model for locating objects whose shape can vary (e.g., faces or internal organs) is to use **active shape models (ASM)** [31]. This method relies on building a statistical model of shape variation from examples in a training set. Each example object is represented using a fixed number of landmark points $(x_i, y_i)(i=1, 2, \dots, N_p)$, each of which marks a particular point on the object. The

training examples are aligned into a common coordinate frame and each is then represented by a $2N_p$ element vector $\mathbf{x} = (x_1, y_1, \dots, x_{N_p}, y_{N_p})^T$. The ASM matches the model points to any valid new image using an iterative technique which is a variant on the expectation maximization (EM) algorithm. A search is made around the current position of each point to find a point nearby which best matches a model of the texture expected at the landmark. The parameters of the shape model controlling the point positions are then updated to move the model points closer to the points found in the image. Any valid shape can then be approximated by adding a reduced subset of these models to the mean shape. This allows valid lip shapes to be represented in a compact, statistically derived shape space by transforms of scale, rotation and translation.

2.2.1.3 Geometric Features Approach

The geometric feature based approach assumes that certain measures such as the height or width of the mouth opening are important features. Some works extract the lip geometric features by painting the lips of the talker to facilitate feature extraction [7] [10] [13] [15]. Other works extract the lip geometric features automatically by certain image processing techniques, such as segmentation and morphological operations [34]. The advantage of the method is that the features are represented in a compact form and that many image processing techniques are available for use.

Adriano V. Barbosa *et al.* [7] described the relationship between acoustic and visual speech where visual features were extracted by measurements on highly reflective dots

placed around the speaker's face. A similar feature extraction method was used in [15]. Other researchers [10] [13] have highlighted the lips with color to facilitate visual processing and extraction of geometric features [33]. Approaches using markers or lipstick mainly serve to determine the usefulness of visual features but they cannot be applied to practical applications. Other researchers segment the lip area from the speaker's face by certain image processing techniques and then measure the geometric visual features from the segmented lip areas [34].

Petajan [11] was probably the first researcher to develop a speechreading system by the geometric based approach. The system was based on geometric features of the mouth opening, like height, width and area. A simple thresholding technique was used to find the mouth opening. A similar feature extraction method was later used by Goldschen [32], who developed a continuous speechreading system.

2.2.2 Discussion on Feature Extraction Approaches

As we discuss above, the image-based approaches contain much information redundancy and are sensitive to changes in illumination, position, camera distance, rotation and speaker, so they are seldom used by researchers currently.

Human perceptual studies [35] show that more visual speech information is contained within the inner lip contours. However, several model-based approaches are not able to represent the principal deformations of the lip contours well and often only consider the outer lip contour. Although active shape models can extract the inner and outer lip

contour, they need to train initial shapes beforehand using some known training data set. For all three model-based approaches, – deformable templates, "snakes", and active shape models, – an image search is performed by fitting a model to the edges of the lips, where intensity values are commonly used. The difficulty with these approaches usually arises when the contrast is poor along the lip contours, which occurs quite often under normal lighting conditions. In particular, edges on the lower lip are hard to distinguish because of shading and reflection. These algorithms are usually hard to extend to various lighting conditions, people with different skin colors, or people with facial hair. In addition, the teeth and tongue are not easy to detect using intensity-only information. The skin-lip and lip-teeth edges are highly confusable if only intensity information is presented.

An obvious way of overcoming the inherent limitations of the intensity-based approach is to use color, which can greatly simplify lip identification and extraction. Lip feature extraction using color information has gained interest with the increasing processing power and storage of hardware, making color image analysis more affordable.

Edge information provides additional useful information in lip feature extraction. It can be applied to the lip region to refine the lip boundary and to extract the lip features.

In chapter 3, I will introduce our approach to locate the lip region and extract the geometric lip features, in which, the lip region is located using color and motion information. Based on the located lip region, a segmentation operation combining both color and edge information within a Markov random field (MRF) framework is

performed to segment the lip from its surrounding face pixels. The geometric lip features can be extracted from the segmented lips.

2.3 Acoustic to Visual Speech Conversion

From the human speech production discussed in chapter 1, we can see that there is an inherent relationship between the acoustic and visible speech. It is generally agreed that most of the visual information is contained in the lips, so most research on the relationship between acoustic and visual speech focuses on the relationship between acoustic features and lip features.

A recent trend in multimedia research is to convert acoustic speech to visual speech features. For multimedia applications that involve person-to-person conversation, e.g., video telephony and video conferencing, such interaction is especially important. Audio-visual speech conversion is very useful in speech-driven face animation, human-computer interfaces, lip synchronization, foreign film dubbing, joint audio-video coding, and bimodal person authentication.

The purpose of acoustic to visual speech conversion is to establish a model which can convert acoustic parameters to the mouth shape parameters with certain accuracy. A number of approaches have been proposed for the task of converting acoustic speech to mouth shape parameters. According to the acoustic speech analysis being used, the problem can be solved from two different perspectives: mapping in the phoneme level and mapping in the frame level.

2.3.1 Phonemes-visemes Conversion

Most languages, including English, can be described in terms of a set of distinctive sounds, or phonemes. A phoneme is a member of the set of the smallest units of speech that serve to distinguish one utterance from another in a language or dialect. In particular, for American English there are about 42 phonemes including vowels, semi-vowels and consonants. The international standard method to represent phonemes is International Phonetic Alphabet [36].

Similarly, in the visual domain, the basic unit of mouth movements is called a viseme. A viseme therefore is the smallest visibly distinguishable unit of speech. Strictly speaking, instead of a still image, a viseme can be a sequence of several images that capture the movements of the mouth. This is especially true for some vowels. There are many acoustic sounds that are visually ambiguous. These sounds are grouped into the same class that represents a viseme. There is therefore a many-to-one mapping between phonemes and visemes. For example, the /p/, /b/, and /m/ phonemes are all produced by a closed mouth shape and are visually indistinguishable. Therefore they form one viseme group. It is commonly agreed that the visemes for English can be grouped into 14 distinct groups as in Table 2.1 [37].

Table 2.1 Viseme Groups for English

1	p, b, m	8	n, l
2	f, v	9	r
3	th, dh	10	A
4	t, d	11	E
5	k, g	12	I
6	sh, zh	13	O
7	s, z	14	U

The phoneme-visemes mapping approach stresses that speech is a linguistic entity. This approach contains three stages, as shown in figure 2.1. In the first stage, the acoustics features must be classified into one of a number of phonemes. The second stage maps each phoneme into a corresponding viseme. The third stage looks up the viseme in the database to get the corresponding lip movements of that viseme. Therefore the lip movements trajectories are output, given a sequence of acoustic speech signal.

Figure 2.1 Phoneme-viseme conversion approach. Speech is converted to lip movements through the intermediate stage of phoneme classification

The advantage of this approach is that the acoustic speech signal is explored to the full extent, so that all the context information is utilized and co-articulations are completely incorporated. However, this approach has a certain amount of computation overhead because, to achieve acoustic-to-visual mapping, one does not really need to recognize the spoken words or sentences. In addition, the construction of the lookup table that maps phonemes to visemes is non-trivial. Another shortcoming of the classification-based method is that it does not produce a continuous mapping, but rather produces a distinct number of output levels. This often leads to a staircase-like reproduction of the output.

2.3.2 Direct Conversion

The other way is direct mapping at the frame level, as shown in figure 2.2. It concentrates on speech being a physical phenomenon. Since there is a physical relationship between the shape of the vocal tract and the sound that is produced, there exists a functional relationship between the speech parameters, typically LSP coefficients or LPC cepstrum, and the visual parameters. The mapping problem becomes one of finding the best functional approximation, given sets of training data. There are many algorithms that can be modified to perform this task. These approaches include neural networks [10] [13] and hidden Markov models with Gaussian mixtures [6].

Figure 2.2 Direct acoustic-visual speech conversion. Speech is converted to lip movements without the intermediate stage of phoneme classification.

2.3.2.1 Neural Network Approach

Neural networks can be used to convert acoustic parameters into visual parameters. In the training phase, input patterns and target output patterns are presented to the network, and an algorithm called back propagation (BP) 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.

2.3.2.2 Gaussian Mixture Model and EM algorithm

Except for neural networks, Gaussian mixture models (GMM) can be used to model the relationship too. Let the acoustic parameters with the visual parameters compose the joint feature vector $O = [a^T, v^T]^T$. The probability distribution of the audio-visual vectors O is modeled using GMM. The GMM is a weighted sum of K Gaussian functions, where K is the dimension of joint acoustic-visual feature vector of a frame.

The GMM can be trained with the Expectation-Maximization (EM) algorithm using the training data set. After the model is initialized, the EM algorithm iterates to update the model by an updating algorithm, the old model parameter is replaced with the new parameter, and converges to the maximum.

After the GMM is trained with the training data set, we can use the model to map the audio feature to the visual feature.

2.4 Applications of Acoustic to Visual Speech Conversion

As we discussed in the previous section, the visual speech can be estimated from acoustic speech since there is a relationship between them. Once the boundary between acoustic and visual speech is broken down, the conversion technique can be useful in a large number of applications.

2.4.1 Lip Synchronization

Lip synchronization is useful for video telephony and video conferencing applications. There has been much research that shows the importance of combined audiovisual conversion to improve bimodal perceptual quality of video-conferencing systems [38], [39]. A typical problem that impairs such perceptual quality in video conferencing is that the sound and the lip movements are not synchronized due to either bandwidth constraints, network delays, or both. One solution is to extract information from the audio

signal to determine the corresponding mouth movements and then process the speaker's mouth image accordingly to improve lip synchronization [8].

In applications for very low bit-rate videophone coding, synthetic images of the speaker's mouth – generated from the articulatory estimates provided by speech analysis – can be interleaved with the actual images, thus increasing the frame rate at the decoder. In such a scheme, a deformable wire-frame model is adapted to the mouth of an original frame and is then animated by means of the articulatory estimates provided by the visual speech estimator. The animated textured mouth is therefore suitably pasted onto the original image, thus generating interpolated/extrapolated synthetic images.

2.4.2 Joint Audio-Video Coding

For many years, audio coding and video coding were studied independently. Recently, there has been a trend of research on joint audio-video coding [8] [9] [40] [41]. One example is to exploit the correlation between the audio and video signals in a predictive coding manner. Predictive coding of video has traditionally used information from previous video frames to help construct an estimate of the current frame. The difference between the original and estimated signals can then be transmitted to allow the receiver to reconstruct the original video frame. This method has proven extremely useful for removing the temporal redundancy in video. Similarly, the prediction can be done in a cross-modal manner to explore cross-modal redundancy. The basic idea is that there is information in the acoustic signal that can be used to help predict what the video signal

should look like. Since the acoustic data is also transmitted, the receiver is able to reconstruct the video with very little side information.

2.4.3 Speech-driven Face Animation

In this direction, researchers have tried to produce visual speech from auditory speech, i.e., to generate speech-driven talking heads. The major applications of this technique include: human-computer interfaces, computer-aided instruction, cartoon animation, video games, and multimedia telephony for the hearing-impaired.

Two approaches to generating talking head images are the flipbook method [42] and the wireframe model (two-dimensional or three-dimensional) [43]. In the flipbook approach, a number of mouth images of a person, called key frames, are captured and stored. Each image represents a particular mouth shape. Then, according to the speech, the corresponding mouth images are "flipped" one by one to display and form the animation. The flipbook method typically generates jerky motion during key frame transition, when the number of key frames is limited.

In [43], a three-dimensional wireframe facial model was used to synthesize the facial expressions, particularly the lip motion. The lip parameters used therein form an eight-dimensional vector that includes the position of the upper lip, the position of the chin, etc. These parameters can be derived either by text input or by speech input. In the case of text input, the input text is then analyzed into a sequence of phonemes and the corresponding lip feature vectors are used to drive the facial model to produce the lip motion. In the case of speech input, the lip feature points are driven by a classifier that

derives lip parameters from the input acoustic speech. In this work, facial features other than lip movements, such as eyebrows and eyelids, were controlled randomly to make the facial motion look more realistic.

2.4.4 Foreign Film Dubbing

One main problem in dubbed foreign films is the loss of lip synchronization. To solve this problem, we can analyze the speech or the image of the dubber and then modify the dubbee's mouth accordingly using image-interpolation techniques. The same technique can be also used to facilitate the animation of lip movements of cartoon characters, which has been a timeconsuming process in cartoon animation.

CHAPTER 3 VISUAL FEATURE EXTRACTION

In this chapter, a method of automatic visual speech feature extraction is described. It is generally agreed that most of the visual information is contained in the lips. So visual speech feature extraction is a problem of lip feature extraction. In my thesis, the lip features include the width of the lip contour, the height of the inner contour and the height of the outer contour. The algorithm first reliably locates the mouth region by using hue and motion information from a color video sequence of a speaker's frontal face. Based on the rough lip region, the algorithm subsequently segments the exact lip area from its surroundings by making use of both color and edge information, combined within a Markov random field framework. From the segmented exact lip area, the geometric lip features that define the lip movements are detected and the relevant visual speech parameters are derived. These visual speech parameters can be used to investigate the relationship between the acoustic and visual speech features in the following chapters.

We applied the feature extraction algorithm to some video sequences in the Clemson University audio-video database [47] and some free demo video sequences in the XM2VTS database [48]. These two databases include head-shoulder full frontal face color video sequences of a person talking. The feature extraction algorithm works well for most of the sample sequences in the two data sets. In a few cases, a few pixels of inaccuracy are observed. The limitation of the color-based feature extraction occurs when

the lip color and its surrounding skin color are very close to each other, which exists in a small percentage of the population. In these cases, the extraction of the key points on the upper and lower lips becomes unstable. These errors can be controlled by applying some constraint on the geometric features. I also recorded a few sequences of my frontal face view as AVI files. The feature extraction algorithm works well for these video sequences.

This chapter begins with the introduction of the color image representation and some image processing techniques that are used in my feature extraction algorithm. The proposed feature extraction algorithm is then presented in detail. Finally a conclusion and discussion on the results are given.

3.1 Color Image Representation

An image is a two-dimensional signal, typically a brightness function of two spatial variables. Common examples are photographs, still frames of video, radar and sonar images and x-rays. A digital image results from sampling of the spatial domain and a quantization of the brightness values. The sampling produces a finite 2-D array of values uniformly distributed over the field of view, while the quantization restricts the sample values to a finite integer range. These necessary operations convert real-world analog sensory data to a form suitable for computer processing and storage.

There are several models used to describe the tristimulus color scheme: RGB, CMY, YIQ, and HSI. Each model is derived for specific purposes and has certain advantages

over the others. Converting between the different models is generally done by a relatively simple mapping.

3.1.1 RGB

A color digital image is typically represented by a triplet of values, one for each of the color channels, as in the frequently used RGB color scheme. The letters R, G, and B stand for red, green, and blue. The individual color values are almost universally 8-bit values, resulting in a total of 3 bytes per pixel. This yields a three-fold increase in the storage requirements for color versus monochrome images. The three primary colors may be used to synthesize any one of 2^{24} or approximately 16 million colors.

In this model, the primary colors are red, green, and blue. It is an additive model, in which colors are produced by adding components, with white having all colors present and black being the absence of any color. This is the model used for active displays such as television and computer screens. The RGB model is usually represented by a unit cube with one corner located at the origin of a three-dimensional color coordinate system, the axes being labeled R, G, and B, and having a range of values $[0, 255]$. The origin $(0, 0, 0)$ is considered black and the diagonally opposite corner $(255, 255, 255)$ is called white. The line joining black to white represents a gray scale and has equal components of R, G, and B.

There is another RGB representation called normalized RGB, denoted by rgb , in which r is defined by $R/(R+B+G)$, g by $G/(R+B+G)$ and b by $B/(R+B+G)$.

3.1.2 CMY

Cyan, magenta, and yellow are the secondary colors with respect to the primary colors of red, green, and blue. However, in this subtractive model, they are the primary colors, while red, green and blue are the secondaries. In this model, colors are formed by subtraction, where adding different pigments causes various colors not to be reflected and thus not to be seen. Here, white is the absence of colors, and black is the sum of all of them. This is generally the model used for printing.

3.1.3 YIQ

This model was designed to separate chrominance from luminance. This was a requirement in the early days of color television programs when black-and-white television sets were still in use. The Y-channel contains luminance information. It gives all information required by a black-and-white television sets. While I and Q channels carried the color information for color television sets. A color television set would take these three channels, Y, I, and Q, and map the information back to R, G, and B levels for display on a screen.

3.1.4 HSI

In this model, as in the YIQ model, luminance or intensity (I) is decoupled from the color information, which is described by a hue channel and a saturation channel. Hue and saturation of colors respond closely to the way humans perceive color. Hue is an attribute

associated with the dominant wavelength in a mixture of light waves. Saturation tells the amount of white light mixed with a hue. HSI color space is very useful in digital image processing. On the one hand, it is because human vision system looks at the objects in this way, i.e., it recognizes the hue and saturation. On the other hand, I is decoupled from the other two parameters in an image.

3.1.5 Grayscale and Binary

A **grayscale** image is an image consisting solely of shades of gray (with a maximum of 256 different shades). Therefore, grayscale images are small files, compared to **RGB** files, and they can convey more detail than indexed color images.

Binary image is the simplest and smallest bitmap file that can exist. A binary consists only of white or black pixels. Such images are extremely simple, but can be quite expressive.

3.2 Image Processing Techniques

3.2.1 Image Histogram

An important concept common for many image processing operations is the histogram, which is a count of the luminance values occurring in the image. Image histograms are important in many image processing applications, most notably compression and thresholding.

The histogram of a monochrome $M \times N$ image is simply a count (or relative frequency, if normalized) of the various gray levels in the image. The histogram of level i is given by

$$h_i = \frac{N_i}{MN} \quad (3-1)$$

where N_i is a count of the number of pixels in the image with value i . The calculation of an image histogram is therefore relatively simple. Typically, an array of length 256 one-element for each of the possible pixel values, is initialized to 0. Next, as the image is scanned and a pixel of value i is encountered, the array location at index i is incremented by 1. In the end, the value at position i represents the number of occurrences of the luminance value i in the image. If we think of a pixel as a random variable then the relative frequency is a measure of the probability of a pixel having a particular value. The histogram is therefore equivalent to a probability density function. Alternatively, the image histogram may be viewed as a luminance density function, since the histogram defines the relative area occupied by a given luminance value.

It is straightforward to extend the definition of a histogram to color images. Since each channel may be considered a monochrome image, an RGB-color image would be represented by three histograms.

3.2.2 Edge Detection

Edge detection is a problem of fundamental importance in image analysis. It is well known that edges are important to mammalian vision. In typical images, edges

characterize object boundaries and are therefore useful for segmentation, registration, and identification of objects in a scene. Some sort of edge detection capability is present in the visual systems of a wide variety of creatures, so it is obviously useful in their abilities to perceive their surroundings.

This section covers the basics of gradient edge detection methods in some detail. Following it, we also describe several important and useful edge algorithms based on that approach.

3.2.2.1 Gradient-based Methods

The core of gradient edge detection is the gradient operator ∇ . In continuous form, it is applied to a continuous-space image $f_c(x, y)$,

$$\nabla f_c(x, y) = \frac{\partial f_c(x, y)}{\partial x} i_x + \frac{\partial f_c(x, y)}{\partial y} i_y \quad (3-2)$$

where i_x and i_y are the unit vectors in the x and y directions, The first derivative assumes a local maximum at an edge.

The gradient magnitude is defined as

$$|\nabla f_c(x, y)| = \sqrt{\left(\frac{\partial f_c(x, y)}{\partial x}\right)^2 + \left(\frac{\partial f_c(x, y)}{\partial y}\right)^2} \quad (3-3)$$

The most commonly used method of producing edge segments or contour consists of two stages: thresholding and thinning. In the thresholding stage, the gradient magnitude at every point is compared to a predefined threshold value. All points satisfying the criterion are classified as candidate edge points.

The set of candidate edge points tends to form strips, which have positive width. Since the desire is usually for zero-width boundary segments or contours to describe the edges, a subsequent processing stage is needed to thin the strips to the final edge contours. Edge contours derived from continuous-space images should have zero width because any local maxima of $|\nabla f_c(x, y)|$, along a line segment that crosses the edge, cannot be adjacent points. For the case of discrete-space images, the nonzero pixel size imposes a minimum practical edge width.

3.2.2.2 Discrete Gradient Operators

Let the discrete space representation of $f_c(x, y)$ be $f(n_1, n_2)$, with n_1 describing the horizontal position and n_2 describing the vertical. The gradient estimate is

$$\hat{\nabla}f(n_1, n_2) = f_1(n_1, n_2)\mathbf{i}_{n_1} + f_2(n_1, n_2)\mathbf{i}_{n_2} \quad (3-4)$$

where

$$f_1(n_1, n_2) = f(n_1, n_2) * h_1(n_1, n_2) \quad (3-5)$$

$$f_2(n_1, n_2) = f(n_1, n_2) * h_2(n_1, n_2) \quad (3-6)$$

Two filters are necessary because the gradient requires the computation of an orthogonal pair of directional derivatives. The gradient magnitude can then be computed as follows:

$$|\hat{\nabla}f(n_1, n_2)| = \sqrt{f_1^2(n_1, n_2) + f_2^2(n_1, n_2)} \quad (3-7)$$

There are many possible derivative-approximation filters for use in gradient estimation. These filters include the Roberts operator, Prewitt operator, Sobel operator and Frei-Chen operator.

Roberts operator:

$$h_1(n_1, n_2) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad h_2(n_1, n_2) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (3-8)$$

Prewitt operator:

$$h_1(n_1, n_2) = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}, \quad h_2(n_1, n_2) = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} \quad (3-9)$$

Sobel operator:

$$h_1(n_1, n_2) = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad h_2(n_1, n_2) = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (3-10)$$

Frei-Chen operator:

$$h_1(n_1, n_2) = \begin{bmatrix} -1 & 0 & 1 \\ -\sqrt{2} & 0 & \sqrt{2} \\ -1 & 0 & 1 \end{bmatrix}, \quad h_2(n_1, n_2) = \begin{bmatrix} 1 & \sqrt{2} & 1 \\ 0 & 0 & 0 \\ -1 & -\sqrt{2} & -1 \end{bmatrix} \quad (3-11)$$

3.2.2.3 Canny's Method

Canny's method [59] uses the concepts of both the first and second derivatives in a very effective manner. His method is a classic application of the gradient approach to edge detection in the presence of additive white Gaussian noise. The method has three

simultaneous goals: low rate of detection errors, good edge localization, and only a single detection response per edge.

Canny's approach begins by smoothing the image with a Gaussian filter:

$$g(i, j) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{i^2 + j^2}{2\sigma^2}\right) \quad (3-12)$$

At each pixel, Eq. (3-7) is used to estimate the gradient direction. From a set of prepared edge detection filter masks having various orientations, the one oriented nearest to the gradient direction for the targeted pixel is then chosen. When applied to the Gaussian-smoothed image, this filter produces an estimate of gradient magnitude at that pixel. Next, the goal is to suppress non-maxima of the gradient magnitude by testing a 3×3 neighborhood, comparing the magnitude at the center pixel with those at interpolated positions to either side along the gradient direction.

The pixels that survive to point are candidates for the edge map. To produce an edge map from these candidate pixels, Canny applies thresholding by gradient magnitude in an adaptive manner. An estimate of the noise in the image determines the values of a pair of thresholds, with the upper threshold typically two or three times that of the lower. A candidate edge segment is included in the output edge map if at least one of its pixels has a gradient magnitude exceeding the upper threshold, but pixels not meeting the lower threshold are excluded. This action helps reduce the problem of broken edge contours while improving the ability to reject noise.

3.2.3 Motion Detection

A video sequence is a much richer source of visual information than a still image. This is primarily due to the capture of motion; while a single image provides a snapshot of a scene, a sequence of images registers the dynamics in it. The registered motion is a very strong cue for human vision; we can easily recognize objects as soon as they move even if they are inconspicuous when still. Motion is equally important for video processing and compression for two reasons. First, motion carries a lot of information about spatiotemporal relationships between image objects. This information can be used in such applications as traffic monitoring or security surveillance, for example, to identify objects entering or leaving the scene or objects that just moved. Secondly, image properties, such as intensity or color, have a very high correlation in the direction of motion, i.e., they do not change significantly when tracked in the image (the color of a car does not change as the car moves across the image). This can be used for the removal of temporal video redundancy; in an ideal situation, only the first image and the subsequent motion have to be transmitted.

It is essential to realize that the motion of image points is not perceived directly, but rather through intensity changes. However, such intensity changes over time may be also induced by camera noise or illumination changes. Moreover, object motion itself may induce small intensity variations or even none at all. The latter will happen in the rare case of exactly constant luminance and color within the object. Clearly, motion detection from time-varying images is not as easy as it may seem initially.

3.2.4 Image Segmentation

Image segmentation process is one that partitions a digital image into disjointed regions. A region is a connected set of pixels – a set in which all the pixels are adjacent or touching. Segmentation divides the image into regions that hopefully correspond to structural units in the scene or distinguish objects of interest. Segmentation is often described by analogy to visual processes as a foreground-background separation, implying that the selection procedure concentrates on a single kind of feature and discards the rest.

Thresholding is a particularly useful region-approach technique for scenes containing solid objects resting upon a contrasting background. It is computationally simple. When using a threshold rule for image segmentation, one assigns all pixels at or above the threshold gray level to the object. All pixels with gray level below the threshold fall outside the objects

Thresholding works well if the objects of interest have uniform interior gray level and rest upon a background of different but uniform gray level. If the objects differ from their background by some property other than gray level, one can first use an operation that converts that property to gray level. Then gray-level thresholding can segment the processed image.

In the simplest implementation of segmentation by thresholding, the value of the thresholding gray level is held constant throughout the image. If the background gray level is reasonably constant throughout, and if the objects all have approximately equal

contrast above the background, then a fixed global threshold will usually work well, provided that the threshold gray level is properly selected.

In many cases, the background gray level is not constant, and the contrast of objects varies within the image. In such cases, a threshold that works well in one area of the image might work poorly in other areas. In these cases, it is convenient to use a threshold gray level that is a slowly varying function of position in the image. This approach is called adaptive thresholding.

Unless the object in the image has extremely steep sides, the exact value of the threshold can have considerable effect on the boundary position and overall size of the extracted object. This means that subsequent size measurements are sensitive to the threshold. For this reason, an optimal thresholding method can be used. But determining an 'optimal' threshold is problematic since there is usually variation within scenes.

To overcome the problem of threshold-based segmentation, recently Bayesian segmentation has been gaining interest. The image is used for an initial characterization and then an optimal segmentation is inferred using a defined random field. The segmentation procedure is performed by a decision algorithm that maximizes the *a posteriori* probability. The procedure is iterated until it converges to a stable solution. The advantage of Bayesian segmentation is that the decision algorithm can incorporate more information of the image in addition to the gray-level information. As a result, it's more flexible and robust than the thresholding approach [52].

3.3 Visual Speech Feature Extraction

3.3.1 Color Analysis

The RGB color model is most widely used in computer vision because color CRTs use red, green, and blue phosphors to create the desired color. However, its inability to separate the luminance and chromatic components of a color hinders the effectiveness of color in image recognition. Previous studies [45] have shown that even though different people have different skin colors, the major difference lies in the intensity rather than the color itself. To separate the chromatic and luminance components, various transformed color spaces can be employed, such as the normalized RGB space (which we denote as *rgb* in the following discussion).

Figure 3.1 Original image and ROI. (a) Original image, (b) Manually extracted lip ROI.

The choice of an appropriate color space is of great importance for successful feature extraction. To analyze the statistics of each color model, we build histograms of the three

color components in each color space by discretizing the image colors and counting the number of times each discrete color occurs in the image. We construct histograms for the entire image and for the manually extracted lip regions of interest (ROI) bounded within the contour, as shown in figure 3.1.

Typical histograms of the color components in the RGB, rgb color spaces are shown in figures 3.2 and 3.3, where two cases are given: (a) those for the entire image and, (b) those for the extracted lip region only.

Figure 3.2 Histograms of R/G/B components. (a) Entire image. (b) Lip ROI.

(a)

(b)

Figure 3.3 Histograms of r/g/b components. (a) Entire image. (b) Lip ROI.

Based on the histograms obtained from video sequences taken under various test conditions and for different test subjects, we can make the following observations:

i) Color distributions of the R/G/B components in the whole image and in the lip ROI (figure 3.2) are wide-spread in the range of possible values, since they contain luminance component. Therefore, the RGB color space is not suitable for lip identification and is discarded in the following analysis,

ii) The color components (r, g, b) exhibit peaks in the histograms of the lip region. This indicates that the color distribution of the lip region is narrow and implies that the color for the lip region is fairly uniform. However, the color histogram of (r, g, b) in the lip region still more or less overlaps with that of the whole image (figure 3.3). For example, for the b components in (a), a half of all pixels concentrate around 0.25, for the b components in (b), almost all pixels concentrate around 0.25. This shows that rgb also does not have high discriminative power for lip detection.

By intuition, the lip region is usually redder than the face region, even though different people have different skin colors. A red hue can be defined by $H = R/(R+G)$, where R and G denote the red and green components of a pixel. Since the red hue value of lips lies at the high end of the red hue, the lips appear to be at the brightest region in the red hue image. The histograms of the red hue H are shown in figure 3.4, where two cases are given: (a) those for the entire image and, (b) those for the extracted lip region only.

Figure 3.4 Histogram of red hue. (a) Entire image. (b) Lip ROI

Based on the histograms obtained from video sequences taken under various test conditions and for different test subjects, we can make the following observations:

i) The red hue value of lips lies at the high end, the values of most lip pixels are greater than 0.60. This indicates that the lips appear to be at the brightest region in the red hue image.

ii) The color histogram of the red hue in the lip region has little overlaps with that of the whole image, since the values of most pixels in the whole image are smaller than 0.6, nevertheless, the values of most lip pixels are greater than 0.6 (figure 3.4). This shows that the red hue is therefore suitable for lip object identification.

However, red hue information alone is not sufficient to guarantee the accuracy of the visual speech feature extraction. For example, difficulties can result when a person has a very red face, or he is wearing a red shirt or tie, or the background is lip-colored. To overcome such difficulties, motion information combined with color information can be used to increase the robustness of locating the lip region. Based on the located lip region, the lip boundary can then be refined by a MRF-based segmentation operation combining both color and edge information. From the segmented exact lip area, the geometric lip features that define the lip movements are detected and the relevant visual speech parameters are derived. Figure 3.5 shows the procedures involved in the proposed three-stage visual feature extraction algorithm. I will address these steps in detail in the following sections.

Figure 3.5 Lip feature extraction

3.3.2 Lip Region Locating

In the first stage of the visual analysis, the speaker's mouth in the video sequence is located. We utilize red hue and motion information for this purpose. Given an RGB image of the frontal view of a talker, as shown in figure 3.6(a), a red hue color image can be derived (figure 3.6(b)) by color transformation ($H = R / (R+G)$) and then a histogram equalization operation applied to the image of H . Histogram equalization operation extends the distribution of the red hue values from a narrow range to a wider range. Since the red hue value lies at the high end, the lips appear to be the brightest region. To locate the rough lip region, it requires that H exceed a certain preset value. For locating the lips, we label a pixel as a lip pixel if $H(i, j) > H_0$, where H_0 is set to 0.6, which is decided by empirical experiments on different Asian and white people in our data set. The accuracy of this threshold value is not very critical, and it is proved to generalize well for other talkers. The resulting binary image is shown in figure 3.6 (c).

Lots of noise is caused by the non-lip red blobs in the image, for example, when there are distracting red blobs in the clothing, or if the person has a ruddy complexion. In this case, we exploit motion cues to increase the robustness of detecting the lips region.

Figure 3.6 Lip region locating. (a) Original RGB color image. (b) Red hue image. (c) Binary red hue image after thresholding. (d) Accumulated motion image. (e) Binary motion image after thresholding. (f) Result from And-operation of (c) and (e). (g) Original image with the identified lip region.

To eliminate distractions of other lip-colored blobs, we also use motion information to increase the robustness of locating the lip region. A difference images between subsequent frames can be built and then summed up over a series of frames. Defining $F_k(i, j)$ as the corresponding gray-scale image of frame k , the motion difference is computed as

$$\delta_k(i, j) = |F_k(i, j) - F_{k-1}(i, j)| \quad (3.13)$$

To detect the moving lips, an accumulated motion image (AMI) is built over a series of frames. The accumulated motion image is defined as follows:

$$\begin{aligned} AMI_1(i, j) &= 0 \\ AMI_k(i, j) &= AMI_{k-1}(i, j) + \delta_k(i, j), \quad k = 2, 3, \dots \end{aligned} \quad (3.14)$$

An example of an accumulated difference image is shown in figure 3.6(d). To separate the moving lips from the background, we use a thresholding operation. The threshold is applied to the entire image, where threshold T is derived by using Otsu's method [49]. This operation separates the moving objects from the background. The binary image is shown in figure 3.6(e) with moving objects being highlighted. When this is combined with the binary image from the red hue thresholding, shown in figure 3.6(c), the binary image figure 3.6(f) is obtained by combining the two binary images using an AND operation. Based on the resulting image, we extract the lip region from its surroundings by finding the largest connected region. The identified lip area is shown as a white bounding box in figure 3.6(g). It should be noted, however, that it is assumed here that the video sequence contains the frontal view of a speaker without significant

head motion. If there is significant head motion frequently, the head tracking algorithm [63] can be applied first before the lip region locating. By taking this measure, the algorithm can still locate the lip region reliably for the video sequence with significant head motion.

The effectiveness of this rather simple algorithm lies in the fact that the hue color is very efficient in identifying the lips due to its color constancy and high discriminative power. Furthermore, the algorithm can reliably locate the mouth region by using both color and motion information from a color video sequence of a talker's frontal view. Since we take advantage of motion information, the lip locating algorithm won't mistake the red blobs in faces as lips. On the other hand, the algorithm won't mistake the blinking eyes as lips since we use color information meanwhile.

The computation of the accumulated motion image described above needs to be done only once for the first image of the sequence. For the succeeding frames, we estimate the lip region using the identical AMI image and the corresponding red hue image of that frame.

When applying this lip region locating algorithm to the video sequence we have, the rough lip region is located very well for all images available. It shows that the proposed lip region locating algorithm is feasible. The robustness and the good quality of the lip region locating allow the MRF-based segmentation to segment the exact lip area more reliably and robustly.

3.3.3 MRF-based Lip Segmentation

In some previous work, after the approximate lip region is located, a thresholding operation is applied to the red hue image of the lip region (shown in figure 3.7 (a)). The exact lip area is obtained after the thresholding operation. Figure 3.7 (b) (c) and (d) are the results of the thresholding operation with threshold $T = 0.55$, 0.70 and 0.80 respectively.

Figure 3.7 Lip segmentation by simple thresholding: (a) Red hue image of lip region. (b) (c) (d) The segmentation results after a thresholding operation with threshold $T = 0.55$, 0.70 and 0.80 respectively.

From the results above, we can see that the threshold T is crucial to the performance of the lip segmentation. It's hard to determine a threshold T to guarantee the performance, because of the difference of the light luminance and face complexion. In order that the lip

segmentation is more robust and reliable, I used a MRF technique combining color and edge information to segment the actual lip area from the approximate lip region.

In addition to the color information, edges characterize object boundaries and provide additional useful information. We perform edge detection by using a Canny detector [59] on the gray image of the approximate lip region. In the Canny detector, the input ed hue image H is convolved with the first derivative of a Gaussian function $g(i, j) = e^{-\frac{i^2 + j^2}{2}}$ to obtain an image with enhanced edges. Figure 3.8 shows the edges of the lip region from figure 3.7(a).

Figure 3.8 Edges of the lip region in figure 3.7(a)

To combine the edge and hue color information, I chose to use the machinery of the Markov random field, which has been shown to be suitable for the problem of image segmentation.

3.3.3.1 Principle of Bayesian Segmentation

Bayesian segmentation algorithm performs a maximum *a posteriori* (MAP) estimation of the unknown partition. Let y be the observed gray scale image (for example, figure 3.7(a))

and y_s , the intensity of the pixel at location s . The Bayesian [52] segmented image will be denoted by x , where $x_s = 0/1$ means that the pixel at location s belongs to face/lip area. We then want to maximize the *a posteriori* probability function $p(x|y)$ to obtain the MAP estimate of x . Thus, we want to find the most likely segmentation x given the gray scale image y . From Bayes' theorem

$$p(x | y) \propto p(y | x)p(x) \quad (3-15)$$

where $p(x)$ is the *a priori* probability of the region process and $p(y|x)$ is the conditional probability of the observed images given the segmentation.

3.3.3.2 Markov Random Field

An MRF is a probabilistic model defined over a lattice of sites. The sites are related to each other through a neighborhood. In an MRF, the sites on $\mathfrak{R}$ are related to one another via a neighborhood system, which is defined as $N = \{N_i, i \in \mathfrak{R}\}$, where N_i is the set of sites neighboring i , $i \notin N_i$ and $i \in N_j \Leftrightarrow j \in N_i$. A random field X said to be an MRF on $\mathfrak{R}$ with respect to a neighborhood system N if and only if $P(x) > 0$, $\forall x \in X$ and $P(x_i | x_{\mathfrak{R}-(i)}) = P(x_i | x_{N_i})$. In MRFs, only neighboring sites have interaction with each other. It was originally used to study the thermodynamic properties of interacting particle systems. Now the MRF is often used as a *priori* model in Bayesian formulation to enforce image segmentation.

Mathematically, in an MRF, the *a priori* probability of the region process of x is described or defined by a Gibbs distribution [60]:

$$p(x) \propto e^{-V(x)} \quad (3-16)$$

Here $V(x)$, is a sum of neighboring-interaction terms, called clique potentials [60], i.e.,

$$V(x) = \sum_{c \in C} V_c(x) \quad (3-17)$$

where C is a clique, and $V_c(\cdot)$ is one clique potential.

3.3.3.3 The Energy Function

In my work, I use the clique potential as defined in [61]

$$V(x) = \sum_s \left[\sum_{q \in C_s} V_{cs}(s, q) + \sum_{q \in C_s} V_{ce}(s, q) \right] \quad (3-18)$$

with C_s being the set of all neighbors of s . In my thesis, a 5-point neighborhood system, shown in figure 3.9, is used.

Figure 3.9 Five-point clique system.

The spatial components of the energy function is given by

$$V_{cs}(s, q) = \begin{cases} -1, & \text{if } x_s = x_q \\ +1 & \text{otherwise} \end{cases} \quad (3-19)$$

The proposed spatial component $V_{cs}(s, q)$ is intuitive and says that if a pixel and its neighbor have the same label, they are more likely to belong to the same region, and if a

pixel and its neighbor have different label, they are more likely to belong to different regions.

Further, we obtain a set, Σ , of all edge pixels by applying an edge detector to the image y . This can be done by a Canny operator [59]. The edge component is defined by

$$V_{CE}(s, q) = \begin{cases} +1, & \text{if } s \notin \Sigma, q \in \Sigma, t \notin \Sigma, x_s = x_t \\ -1, & \text{if } s \notin \Sigma, q \in \Sigma, t \notin \Sigma, x_s \neq x_t \\ 0, & \text{otherwise} \end{cases} \quad (3-20)$$

where $t = s + 2(q - s)$. The proposed edge component $V_{CE}(s, q)$ is intuitive and says that if there is an edge element between two non-edge pixels, they are more likely to belong to different regions. The potential $V_{CE}(s, q)$ is zero if s or t is an edge pixel, because an edge pixel can belong to either side.

Finally, we model the observed image y , given x as a constant region mean (i.e., lip region mean or face region mean) plus additive white Gaussian noise with variance $\sigma_{x_s}^2$ as in [50]

$$p(y | x) \propto \prod_s \exp\left(-\frac{(y_s - u_{x_s})^2}{2\sigma_{x_s}^2}\right) \quad (3-21)$$

where u_{x_s} , $\sigma_{x_s}^2$ are the mean and variance of lip region or face region. They are obtained by using Otsu's methods [49] based on the histogram.

By combining the above equations (3-15), (3-16), (3-18) and (3-21), the *a posteriori* probability function $p(x|y)$ can be derived as

$$\begin{aligned}
p(x | y) &\propto \prod_s \exp\left(-\frac{(y_s - u_{x_s})^2}{2\sigma_{x_s}^2}\right) \exp\left[-\sum_s \left(\sum_{q \in C_s} V_{CS}(s, q) + \sum_{q \in C_s} V_{CE}(s, q)\right)\right] \\
&= \exp\left[-\sum_s (y_s - u_{x_s})^2 / 2\sigma_{x_s}^2 - \sum_s \left[\sum_{q \in C_s} V_{CS}(s, q) + \sum_{q \in C_s} V_{CE}(s, q)\right]\right] \quad (3-22) \\
&= \exp(-E(x))
\end{aligned}$$

Here $E(x)$, often called as energy function, is defined by

$$E(x) = \sum_s \exp\left(\frac{(y_s - u_{x_s})^2}{2\sigma_{x_s}^2}\right) + \sum_s \left[\sum_{q \in C_s} V_{CS}(s, q) + \sum_{q \in C_s} V_{CE}(s, q)\right] \quad (3-23)$$

In order to maximize the *a posteriori* probability function $p(x|y)$, the energy function $E(x)$ has to be minimized.

3.3.3.4 High Confidence First (HCF) Algorithm

A method to minimize (3.23) has to be selected. For the optimization strategy, a stochastic relaxation technique, such as simulated annealing, can be used to find the globally optimal interpretation for the image. However, an exhaustive search for a global optimum imposes a large computational burden. Therefore, alternative estimates have been suggested, including the Monte Carlo method, mean field technique, iterated conditional models (ICM), and high confidence first (HCF) algorithm [51][52]. I chose to use the HCF, because it is deterministic, computationally attractive, and achieves good performance [51]. The HCF algorithm minimizes the energy function in equation. 3-23 as follows.

Initially, all pixels in figure 3.7(a) are uncommitted and the special label -1 (uncommitted) is assigned to them. All pixels are divided into two segments (lip segment, labeled by 1 and face segment labeled, by 0) by a thresholding operation. The threshold T is chosen so that the number of the lip pixels is the same as that of face pixels.

The labels of the pixels are then updated in order of decreasing the energy function. Pixels are allowed to change their label during the process, but once committed they cannot become uncommitted again. Before giving the definition of confidence, the local energy at s is defined

$$E_s(i) = \frac{(y_s - u_i)^2}{2\sigma_i^2} + \sum_{q \in C_s} V_{CS}(s, q) + \sum_{q \in C_s} V_{CE}(s, q) \quad (3-24)$$

$E_s(i)$ is the local energy at s for $x_s = i$, $i = 0/1$. Some neighbors q of s might be uncommitted and for those q , $V_{CS}(s, q)$ and $V_{CE}(s, q)$ are set to zero because they must not influence s . Further, the local energy for uncommitted pixels is set to $E_s(-1) = 0$.

The confidence $c_s(x)$ of a pixel is now given by

$$c_s(x) = \max\{E_s(i) - E_s(l)\} \quad i, l \in \{0, 1\} \quad (3-25)$$

A positive value of $c_s(x)$ indicates that a more stable (lower energy) segmentation will result if the label at s is changed from $x_s = i$ (0/1) to $x_s = l$ (1/0). The larger $c_s(x)$, the more confident the label at s is about to change.

The algorithm stops when no pixel s with $c_s(x) > 0$ is left. The pixels with label 1 are considered part of the lips and those with label 0 are part of the rest of the face. Figure 3.10 shows the steps of MRF-based lip segmentation

Figure 3.10 Steps of MRF-based lip segmentation. (a) Modified red hue image of lip region. (b) Initial lip segmentation. (c) Edges of the lip region. (d) Lip segmentation after MRF & HCF algorithms.

The teeth sometimes may be labeled as lips, since the red hue value of the tooth also lies at the high end because the color of teeth is close to white. To eliminate these errors, our algorithm will look at the extracted lips for the pixels whose red hue value H is close to 0.5 and eliminates those pixels from the extracted lips.

A few examples of extracted lips are demonstrated in figure 3.11. We observe that the highlighted pixels match the true lip area fairly well.

Figure 3.11 Segmented lips overlaid on the original image.

3.3.4 Lip Feature Extraction

Based on the segmented lip image, we are able to extract the key feature points on the lips (figure 3.12). We detect four feature points along the vertical lip line — the upper/lower outer/inner lip and two feature points along the horizontal lip line.

Figure 3.12 Illustration of the extracted geometric feature of the lips

3.3.4.1 Feature Points Detection

Firstly, the pixels in the segmented lip area (e.g., figure 3.10 (d)) are summed for each vertical stripe to form a horizontal profile, by which the derivatives are used to find the x

coordinates of P_1 and P_2 . The y coordinates of P_1 and P_2 are determined by searching the vertical line crossing the x coordinates of P_1 and P_2 .

Then the edges of the segmented lip area (e.g. figure 3.10 (d)) are generated by a Canny edge detector. Searching along the line perpendicular with $\overline{P_1P_2}$ from the top and bottom, the two intersections between the edge and $\perp \overline{P_1P_2}$ are P_3 and P_6 . Similarly, we can determine the coordinates of P_4 and P_5 .

To increase the accuracy of the identified feature points, we incorporate intensity gradient information. If the gradient of the detected point is below a preset value, we start searching for the largest gradient in its vicinity, and replace the old one with it.

During the lip feature point detection, difficulties may be caused by the presence of tongue and teeth. The presence of the tongue may cause difficulty for the detection of the coordinates for P_4 and P_5 since the red hue of the lip and tongue is very similar. This problem can be demonstrated in two cases, as shown in figure 3.13. In the first case, the tongue is separated from the lips by the teeth. Tongue detection is trivial in this case. In the second case however, the tongue merges with the upper lip or lower lip or both. From the segmented image, we have a lip closure case. Here, we use the gradient of the intensity to detect the inner upper/lower lip. In this case, we search for intensity gradient values along the vertical lip line. If the gradients of two points exceeding a preset value are found, they are identified as upper/lower inner lip. The parameter for the tongue is represented by the total number of lip-color pixels within the inner lip contour.

Figure 3.13 (a) Tongue is separated from the lips. (b) Tongue merges with the lips

The teeth can also cause trouble to the lip feature point detection, but this problem is easy to solve since their red hue H values are distinctly different from the hue of the lips. This is a big advantage compared with gray-level-based approaches that may confuse skin-lip and lip-teeth edges.

The limitation of the color-based feature extraction for the coordinates of P_4 and P_5 occurs when the tongue merges with the lips. In these cases, the extraction of the key points on the upper and lower lips becomes unstable. These errors can be controlled by using the geometric constraint method. In this method, we assume that the variation of all parameters between successive frames is within a limited range.

Since the evaluation of feature extraction methods is often subjective, it is common that a direct and quantitative evaluation is impossible. In our experiments, we evaluate the accuracy of the derived visual features by a set of video sequences with about 4500 frames of 5 different people. We verify the computed results by visual inspection of the original images. The results show that the computed feature sets have an accuracy of 98%

for images with tongue absence and 91% for images with tongue presence. A few examples of extracted feature points are shown in figure 3-14. Most failure occurs in those frames in which there is significant reflective flare at or around the feature points.

Figure 3.14 Examples of detected feature points

3.3.4.2 Lip Geometric Feature Extraction

Based on the extracted key feature points, we can derive the geometric dimensions of the lips. The following features are used in my thesis: Lip width (W), the height of the outer lip contour (H) and the height of the inner lip contour (h). An illustration of the geometry is shown in figure 3.12. Let the lip width of the first frame be 1, all other feature parameters are represented as the ratio to this value. Figure 3.15 shows these three feature parameters when I speak English words from 'one' to 'ten'.

Figure 3.15 Visual features (W , H , h) for my utterance of English words '1' to '10'

We applied the feature extraction algorithm on some video sequences. These video sequences include head-shoulder full frontal face shoots of a person talking. Test subjects have various skin complexions with no particular lipstick. The feature extraction algorithm works well for most of the sample sequences, which includes about 4500 frames of images of 5 different people. The overall accuracy is about 98% for images with tongue absence and 91% for images with tongue presence.

3.4 Summary

This chapter begins with the introduction of color image representation and some image processing techniques that are used in the visual feature extraction algorithm. The proposed feature extraction algorithm is then presented in detail.

In this chapter, we described a method of automatic visual speech feature extraction, which consists of three stages – lip region locating, MRF-based lip segmentation and lip geometric feature extraction. In the first stage, the mouth region is reliably located by using hue and motion information from a color video sequence of a speaker's frontal view. The algorithm can reliably locate the mouth region. Since we take advantage of motion information, the lip locating algorithm won't mistake the red blobs in faces as lips. On the other hand, the algorithm won't mistake the blinking eyes as lips since we use color information. Since the region of interest (ROI) of the following stages is a small region rather than a whole image, the processing time of the lip segmentation is much shorter than that of segmenting the lip from the whole image.

In the lip segmentation stage, based on the approximate lip region, the algorithm subsequently segments the lips from the rest of the face by making use of both color and edge information, combined within a Markov random field framework. Because we make use of both color and edge information under an MRF framework, the segmented lip matches the actual lip much better than that of some segmentation approaches based on a simple thresholding operation.

From the segmented exact lip area, the geometric lip features that define the lip movements are detected and the relevant visual speech parameters are derived. These visual speech parameters can be used to investigate the relationship between the acoustic and visual speech features in the following chapters.

We applied the feature extraction algorithm on the Clemson University audio- video database [47] and some free demo video sequences in the XM2VTS database [48], as well as some video sequences recorded by myself. These two databases include head-shoulder full frontal face color video sequences of a person talking. Test subjects have various skin complexions with no particular lipstick. The feature extraction algorithm works well for most of the sample sequences in the two data sets. In a few cases, a few pixels of inaccuracy are observed. The limitation of the color-based feature extraction occurs when the lip color and its surrounding skin color are very close to each other, which exists in a small percentage of the population. In these cases, the extraction of the key points on the upper and lower lips becomes unstable. These errors can be controlled by using the geometric constraint methods.

Since the evaluation of lip feature extraction methods is often subjective, it is common that a direct and quantitative evaluation is impossible. We verify the computed results by visual inspection of the original images. The results show that the accuracy is 98% for images with tongue absence and 91% for images with tongue presence. Most failure occurs in those frames in which there is significant reflective flare at or around the feature points.

CHAPTER 4

ACOUSTIC-VISUAL SPEECH CONVERSION

As we discuss in chapter 2, the speech waveform is a sound pressure wave which originates from movements of the human speech production system. The main components which determine the speech waveform are the lungs, trachea, throat, pharyngeal cavity, mouth and nose. The articulators move to different positions to produce different speech sounds. During speech production, air is forced from the lungs through the trachea into the pharynx or nasal cavity. The acoustic waveform is often referred to as acoustic speech and the movements of the articulators are referred to as visual speech. It is generally agreed that most of the visual information is contained in the lips [1]. Thus most research on relationship between acoustic and visual speech focuses on the relationship between acoustic features and lip features.

A recent trend in multimedia research is to convert acoustic speech to visual speech to exploit their relationship. The objective of this research is to find a mapping model capable of modeling the relation between acoustic speech and visual feature parameters. For this purpose, it is assumed that this relation can be described by a function $v_m = f(a_m)$. There are many techniques, such as neural networks and Gaussian mixture models that

are often used for converting human speech into visual parameters of lip movements. For multimedia applications that involve person to person conversation, e.g., video telephony and video conferencing, this relationship is especially important. Research topics along the direction of audio-visual conversion include speech-driven face animation, human-computer interface, lip synchronization, foreign film dubbing and joint audio-video coding.

We will use neural networks to model the relationship between the acoustic-visual speech features. In the training phase, input patterns and target 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 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.

In this chapter, we first discuss the acquisition of training data set. Then we present the design choice to obtain an optimal mapping model with the training data set available. The design choice includes these following aspects: multi-output versus single-output architecture, the number of nodes in the hidden layers, the training algorithm to train the network, as well as the linearity of the nodes in the hidden layer. Finally, the performance of the mapping model is given by comparing the actual and estimated visual speech from acoustic speech.

4.1 Training Data Set

To train our mapping model, we record a video sequence while I say the English words from 1 through 10, each 10 times. The distance between the camera and me is maintained constant and the room illumination is also kept constant. The position is almost frontal view with no rotation and no significant head motion.

The mono acoustic speech is sampled at 12kHz and represented with a precision of 16 bits per sample.

At the same time, the corresponding video images of the speaker's front view face are captured at 30 frames per second. The spatial resolution of the image was Audio Video Interleave (AVI) with 320×240 pixels, 24 bits per pixel. The silence portions in the speech and their corresponding video frames are removed.

4.1.1 Visual Speech Extraction

The geometric dimensions of the lips can be derived using the extraction algorithm discussed in chapter 3. The following features are used in my thesis: Lip width (W), the height of the outer lip contour (H) and the height of the inner lip contour (h) as illustrated in figure 3.12. Let the lip width of the first frame be 1; all other feature parameters are presented as the ratio to this value.

In our work, each frame m of the image is parameterized as the vector

$$v_m = [W, H, h]^T. \quad (4-1)$$

These vectors are then grouped in the matrix

$$V = [v_1, v_2, \dots, v_M]. \quad (4-2)$$

4.1.2 Acoustic Speech Extraction

There are different sets of parameters representing the spectral information of acoustic speech, such as LPC (Linear Predictive Coding), LSP (Line Spectrum Pairs) and cepstrum parameters.

Linear Predictive Coding (LPC) is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate. It provides extremely accurate estimates of speech features, and is relatively efficient for computation.

LPC starts with the assumption that the speech signal is produced by a linear system in short period of time. So LPC analysis is done on short chunks of the speech signal, which are called frames. Usually frames of less than 20 ms give good performance.

Line spectrum pair (LSP) was first introduced by Itakura [57] as an alternative kind of LPC spectral representation. LSP parameters have both well-behaved dynamic range and filter stability preservation properties, and can be used to encode LPC spectral information more efficiently than other parameters [11]. The LPC, on the other hand, is not bounded to a well-behaved dynamic range. The LPC and LSP coefficients of the same speech are shown in figure 4.1. It's very clear that the LSP coefficients have better-behaved dynamic range than LPC coefficients. Furthermore, LSP coefficients have a

good property in that they are strongly related to the speech formants, which are determined by the vocal tract shape [7]. The vocal tract motion, in turn, is primarily responsible for the facial motion during speech.

Figure 4.1 LSP and LPC coefficients of the same speech.

Figure 4.2 Procedures of acoustic speech extraction.

Figure 4.2 shows the steps of acoustic speech feature extraction; in my thesis, the acoustic speech is parameterized into LSP coefficients as follows:

1. Pre-emphasis

The speech signal, $s(n)$, is filtered with a first-order FIR filter to spectrally flatten the signal. We used one of the most widely used pre-emphasis filters

$$H(z) = 1 - az^{-1} \quad (4-3)$$

where $a = 15/16$.

2. Frame Blocking

The pre-emphasized speech signal is blocked into frames of N samples, with no overlap between adjacent frames. Since the acoustic signal is acquired at a rate of 12,000 samples per second, we choose a frame length of 16.67 ms, yielding a rate of 60 frames/s. The frame rate is not important. We choose 60 frames/s in order that it matches the video frame rate.

3. Windowing

Each individual frame is windowed to minimize the signal discontinuities at the borders of each frame. If the window is defined as $w[n]$, $0 < n < N-1$, then the windowed signal is

$$\tilde{s}(n) = s(n)w(n) \quad (4-4)$$

where $0 < n < N-1$. We used a Hamming window, a typical window used for the autocorrelation method of LPC.

4. LPC Parameters

P th-order LPC analysis starts with the assumption that the current samples can be represented by a linear combination of P past samples.

$$\tilde{s}(n) = \sum_{k=1}^P a_k \tilde{s}(n-k) \quad (4-5)$$

Given a frame of speech samples, the LPC coefficients can be derived by Durbin algorithm [62], a classic autocorrelation method as

$$a = R^{-1}r \quad (4-6)$$

where,

$$r = [r_n(1), r_n(2), \dots, r_n(P)]^T \quad (4-7)$$

$$R = \begin{bmatrix} r_n(0) & r_n(1) & r_n(2) & \dots & r_n(P-1) \\ r_n(1) & r_n(0) & r_n(1) & \dots & r_n(P-2) \\ r_n(2) & r_n(1) & r_n(0) & \dots & r_n(P-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_n(P-1) & r_n(P-2) & r_n(P-3) & \dots & r_n(0) \end{bmatrix} \quad (4-8)$$

$$r_n(i-k) = \sum_{n=0}^{N-1-(i-k)} \tilde{s}(n) \tilde{s}(n+i-k) \quad (4-9)$$

5. LPC Parameter Conversion to LSP

Given a specific P th order LPC coefficients $1, a_1, a_2, \dots, a_p$, an LPC polynomial is given by

$$H(z) \stackrel{\text{def}}{=} \frac{1}{A_p(z)} = \frac{1}{1 + \sum_{p=1}^P a_p z^{-p}} \quad (4-10)$$

The LSFs (Line Spectral Frequencies) are the poles of the polynomial in equation (4-10). Let $w_1, w_2, \dots, w_p$ denote these LSFs. LSPs are the cosine values of $w_1, w_2, \dots, w_p$.

In our work, the acoustic speech is parameterized into LSP coefficients as follows: the acoustic signal, acquired at a rate of 12,000 samples per second, is analyzed using a

frame length of 16.67 ms, yielding a rate of 60 frames/s. LPC analysis of order $P = 13$ is applied to each frame. The LPC coefficients are then converted into LSP coefficients.

Thus, each frame m of speech is represented as a $P = 13$ -dimensional vector of LSP coefficients

$$a_m = [a_{1m}, a_{2m}, \dots, a_{Pm}]^T. \quad (4-11)$$

These vectors are then grouped in the matrix

$$A = [a_1, a_2, \dots, a_M]. \quad (4-12)$$

4.2 Mapping Model Training

After the parameterized acoustic and visual features are obtained, they can be used to train neural networks for conversion of acoustic features to visual features.

4.2.1 Learning Criteria for Neural Networks Training

While training a neural network, it is usually required to define a valid criterion to assess the achieved level of quality. Often the mean squared error (MSE) and the cross-correlation (CC) coefficients are used for this purpose.

Let u_t and d_t denote the neural network output and target vectors of pattern t , the following costs can be defined:

- MSE (Mean Square Error)

$$\text{MSE} = \frac{1}{N_{Pat}} \sum_{t=1}^{N_{Pat}} (u_t - d_t)^2 \quad (4-13)$$

This figure, despite being the most commonly used, turns out not to be reliable enough for those specific applications when u_i and d_i are very small. In such an cases, even if the difference between u_i and d_i is significant to d_i , the difference is still very small, so the MSE is small. However, its value is guaranteed to be monotonically decreasing during training time, so it's often used as the criterion during training.

- Cross-Correlation Coefficient

$$r = \frac{E\{(u - \bar{u})(d - \bar{d})\}}{\sigma_u \sigma_d} \quad (4-14)$$

In contrast to mean square error, the cross-correlation coefficient measures the similarity between the two sequences with invariance to the scale factor and to the particular distribution of the sequence samples. So r overcomes the shortcomings of MSE. However, its value is not guaranteed to be monotonically increasing at learning time [13], so it's not often used as the criterion during training, but it can be used to measure the performance after training.

4.2.2 Optimization of Neural Network Architecture

The optimization design choice lies in selecting a suitable topology for the network. The number of nodes in hidden 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.

The parameterization of the neural network architecture to optimize the estimation performances is evidently a very complex task, both in consideration of the many parameters of the network and of the large dimension of the training set.

It's impossible to optimize every aspect of neural network architecture, so I will focus on discussion of the design options in these four aspects: (1) multi-output versus single-output architecture, (2) number of nodes in hidden layer, (3) the linearity of the nodes in the hidden layer, as well as (4) the training algorithm to train the network.

4.2.2.1 Single Multi-output versus Multi Single-output Architecture

The first thing that must be decided is choosing the architecture of a single multi-output neural network (figure 4.3 (a)) or a multiple single-output neural networks (figure 4.3 (b)).

In single multi-output neural network structure, the input is the acoustic speech feature. The output is a vector of 3 lip features. In particular, as shown in figure 4.3 (a), the involved parameters concern the lip width, the outer lip height and the inner lip height.

In multiple single-output neural networks, the input of each neural network is the acoustic speech feature. The output of each neural network, as shown in figure 4.3 (b), is the lip width, the outer lip height and the inner lip height correspondingly. In figure 4.3 (b), NN_1, NN_2 and NN_3 are single-output neural networks. Each neural network has the same hidden layer structure as that in Figure 4.3 (a)

(a)

(b)

Figure 4.3 (a) Single multi-output versus (b) Multi single-output architecture

(a)

(b)

Figure 4.4 Learning curves for different neural network architecture. (a) Single multi-output architecture. (b) Multi single-output architecture

In order to compare which architecture is more suitable for acoustic-visual speech mapping for our training data, we train a multi-output neural network and three independent single-output neural networks using the same training data set. The convergence for each output of each structure is shown in figure 4.4.

Comparing the two figures in figure 4.4, we can see that the MSE of each output in single-output architecture converges more quickly and steadily than that in multi-output architecture does. This result is not surprising, since the weights are updated to minimize the MSE for all outputs, however, the weights are updated to minimize the MSE for only one output. It's easier for MSE to converge more quickly in a multi single-output architecture than in a multi-output one.

The experiments with different initial conditions have the same results as discussed above, so we can conclude that the multi single-output architecture is more suitable for modeling the relationship between the acoustic-visual speech features in our training data set.

4.2.2.2 Number of Nodes in Hidden Layer

Based on the results of the previous section, three single-output neural networks will be used to model the relationship between the acoustic and visual speech. The optimal number of nodes in the hidden layer has to be decided empirically.

Several neural network configurations are investigated with the constraint of keeping their complexity constant and limiting their architecture to a one-hidden-layer

architecture. The MSE reduction is analyzed when varying the number of nodes in the hidden layer. Figure 4.5 provides an example of the optimization procedure results for MSE minimization with the lip width W as the target output.

Figure 4.5 Optimization of the number of nodes in the hidden layer with minimization of the MSE. In this case, the architecture has been fixed to single-output, one-hidden-layer. This result is obtained by using the lip width W as the target output. Similar results can be obtained by using the parameters H or h as the target output.

From the figure above, we can see that the greater the number of nodes in the hidden layer, the better the performance the neural network can achieve. However, when the number of nodes in the hidden layer is greater than 30, the performance improvement is not significant as the number increases, but the computation load and the training time

increase dramatically. Forty nodes in the hidden layer can achieve a balance between performance improvement and the training time consumed.

Similar results are obtained by using the lip parameters H or h as the target output. Thus the number of nodes in the hidden layer is set to 40 in all three independent neural networks in the following discussion.

4.2.2.3 Linearity of the Nodes in the Hidden Layer

Another design option which affects the performance of neural networks is the linearity of the nodes in the hidden layer. To investigate this problem, two sets of neural networks are trained using our training data, one set with linear hidden layer, the other one with nonlinear hidden layer. The configuration of each neural network uses one hidden layer with 40 nodes. These neural networks are trained using our training data. The learning curves are demonstrated in figure4.6

(a)

(b)

Figure 4.6 Learning curves versus linearity of the nodes in the hidden layer. (a) nonlinear hidden layer (b) linear hidden layer

From the results above, the performance of the nonlinear model is much better than that of linear model. This fact reveals that the relationship between acoustic speech and corresponding visual features is not strictly linear, but intrinsically nonlinear.

4.2.2.4 Training Algorithm

Several adaptive learning algorithms for feed-forward neural networks have been discovered. Backpropagation (BP) is one of them. Many of these algorithms are based on the gradient descent algorithm. They usually have a poor convergence rate, since they only make use of the first order information of the training data set. At the same time, many fast learning algorithms have been introduced. Scaled conjugate gradient algorithm

(SCG) [58] is one among them. SCG benchmarked against the performance of the standard backpropagation algorithm. SCG yields a speed-up of at least an order of magnitude relative to BP. The algorithm is based upon a class of optimization techniques well known in numerical analysis as the conjugate gradient methods (CGM). SCG uses second order information from the neural network.

(a)

(b)

Figure 4.7 Learning curves of the same network set trained by different algorithms. (a) scaled conjugate gradient algorithm, (b) standard backpropagation algorithm.

Figure 4.7 lists the learning curves of the same neural network set trained by SCG and BP. The one trained by SCG converges much faster.

Through the empirical experiments above, far from being exhaustive, a neural network architecture with good performance can be decided. In my thesis, three independent feed-forward neural networks are used to convert the vector of LSP coefficients to each of the three components (H , h , W) of the visual vectors. Neural networks with one nonlinear hidden layer are used. The number of nodes in the hidden layer is 40. Each neural network is trained using the scaled conjugated gradient algorithm [58] using the training data set.

4.3 Estimation of Visual Feature from Acoustic Feature

After they are trained, these neural networks can be used to estimate the visual speech from the acoustic speech.

We will use some video sequences to test the performance of the conversion model. Given a video sequence, we can measure the visual speech features by the lip feature extraction algorithm presented in chapter 3. These visual features obtained are referred to as 'actual features'.

To estimate the visual parameters from the acoustic speech, first the silence portions in the speech are detected; since there is no information to predict the acoustical parameters, we simply set the visual parameters as the mouth is closed.

For the non-silence portions in the speech, we extract the LSP coefficients of each frame. Then these coefficients are used as the input to the corresponding neural networks and the outputs of the neural networks are the estimated visual features, which are referred to as 'estimated features'.

We use three video sequences to test the performance of the conversion model. The first one is the utterance of the English words from 1 through 10, which is one of the 10 video sequences in training set. The second one is also the utterance of the English words from 1 through 10, but it is not used to train the conversion model. The third one is the utterance of the English word 'audio'.

The actual visual features and estimated visual features of these three testing video sequences are illustrated in figure 4.8, figure 4.9 and figure 4.10 respectively. Similar results can be found in our work in [64]. The actual visual features are plotted in blue lines. The estimated visual features from the acoustic speech are plotted in red lines. In each figure, the curves of the lip width, outer lip height and inner lip height are placed in order. And the correlation coefficients between the measured and estimated parameters for each visual feature are labeled besides the corresponding curve.

Figure 4.8 Estimation results over one of the training video sequence.

Figure 4.9 Estimation results over a testing video sequence of utterance of '1' through '10', which is not used to train the model.

Figure 4.10 Estimation results over video sequence of utterance of 'audio'.

From the results in figure 4.8 and figure 4.9, we see that the estimated visual features match the actual ones well not only with qualitative observation but also with quantitative evaluation. The results in figure 4.8 are not surprising, since it is a portion of the training set. The results in figure 4.9 are quite interesting, since the data for figure 4.9 are not used to train the model, and the speaking speed of the video sequence in figure 4.9 is different from that in figure 4.8. However, the results in figure 4.9 are still good. The global correlation coefficient is roughly 0.90. This shows that the model is capable of estimating the visual speech from acoustic speech of some syllables if the training data of those syllables are used to train the model.

Although the results in figure 4.10 are not very good, they are understandable, since the syllables in 'audio' are different from all syllables in '1' through '10' of the training

set. The estimation models have not established a mapping relationship between the acoustic and visual features for those syllables in 'audio'. Thus it's not surprising that the estimation performance in figure 4.10 is not good.

Based on the results above, we can make the following observations: i) Neural networks are capable of estimating visual speech features from acoustic speech features, if they are training using some known training data sets; ii) In order to obtain a general model, we should use as many as sentences, containing as many as syllables, or even the same sentences spoken at different speaking speed and mood to train the estimation model.

Estimation of visual speech features from acoustic speech features breaks the boundary between the acoustic speech feature space and visual speech feature space. It gives a way to integrate audio and video information in a large number of multimedia application scenarios, such as in speech-driven face animation, lip synchronization, foreign film dubbing and joint audio-video coding.

4.4 Summary

In this chapter, we exploit the relationship between acoustic and visual speech. The objective here is to find a mapping model capable of modeling the relation between acoustic speech and visual feature parameters. There are many techniques, such as neural networks and Gaussian mixture models, which are often used for converting the human speech into visual parameters of lip movements. Audio-visual speech conversion is of

great importance in multimedia applications involving person to person conversation, such as speech-driven face animation, human-computer interface, lip synchronization, foreign film dubbing, joint audio-video coding, and bimodal person authentication.

The neural network is a tool capable of modeling the relationship between the acoustic-visual speech features. Neural networks have to be trained using some known input patterns and target output patterns, before they can be used to convert acoustic parameters into visual parameters. So the acquisition of training data set is presented at the beginning of this chapter.

An important issue related to neural networks is selecting a suitable topology for the network. Four aspects of the optimal neural network architecture, including (1) multi-output versus single-output architecture, (2) number of nodes in hidden layer, (3) the linearity of the nodes in the hidden layer, as well as (4) the training algorithm to train the network, are discussed. These problems can be decided empirically by experiments over the training data set.

Based on the empirical experiments, the optimal neural network architecture is decided. In my work, three independent feed-forward neural networks are used to convert the vector of the LSP coefficients to each of the three components (H , h , W) of the visual vectors. Neural networks with one nonlinear hidden layer are used. The number of nodes in the hidden layer is 40. Each neural network is trained using the scaled conjugated gradient algorithm using the training data set.

At the end of this chapter, the performance of the model is tested using a testing video sequence. The estimated features match the actual ones very well. The global correlation coefficient between the actual and estimated features is roughly 0.90.

Estimation of visual speech features from acoustic speech features breaks the boundary between the acoustic speech feature space and visual speech feature space. It gives a way to integrate audio and video information in a lot of multimedia applications, such as in speech-driven face animation, human-computer interfaces, lip synchronization, foreign film dubbing and joint audio-video coding.

CHAPTER 5 CONCLUSIONS

Human speech is bimodal in nature. Such bimodality has two aspects, the production and the perception [1]. Acoustic speech often refers to the acoustic waveform produced by the speaker, visual speech refers to the movements of the lips, tongue, and other facial muscles of the speaker. There is an inherent relationship between the acoustic and visible speech. This relationship is very useful in many multimedia applications.

In this thesis, with the purpose of exploiting the relationship between acoustic and visual speech, we present an automatic algorithm to extract the visual speech. This method consists of three stages – lip region locating, MRF-based lip segmentation and lip geometric feature extraction. In the first stage, the mouth region is located by using hue and motion information from a color video sequence of a speaker's frontal face. Since we take advantage of motion information, the lip locating algorithm won't mistake the red blobs in faces as lips. On the other hand, the algorithm won't mistake the blinking eyes as lips, since we use color information. Since the region of interest (ROI) of the following stage is a small region rather than a whole image, the processing time of the lip segmentation is much shorter than that of segmenting the lip from the whole image.

In the lip segmentation stage, based on the approximate lip region, the algorithm subsequently segments the lip from its surroundings by making use of both color and edge information, combined within a Markov random field framework. Because we make use of both color and edge information under an MRF framework, the segmented lip matches the actual lip much better than that of some segmentation approaches based on a simple thresholding operation.

From the segmented exact lip area, the geometric lip features that define the lip movements are extracted and the relevant visual speech parameters are derived. These visual speech parameters can be used to investigate the relationship between the acoustic and visual speech features.

Inaccuracy is observed when the lip color and its surrounding skin color are very close to each other. In these cases, the extraction of the key points on the upper and lower lips becomes unstable. These errors can be controlled by using the geometric constraint methods.

Currently, there is no direct and quantitative evaluation for lip feature extraction approaches. Usually, the computed results are verified by visual inspection of the original images.

Using the lip features extracted by the algorithm proposed, we can exploit the relationship between acoustic and visual speech. The objective here is to find a mapping model capable of modeling the relation between acoustic speech and visual feature parameters.

Neural networks are often used to model such a relationship between acoustic and visual speech. An important issue related to neural networks is selecting a suitable topology. By numerous empirical experiments, an optimal architecture is chosen. Each neural network is trained using the scaled conjugated gradient algorithm using the training data set.

After the model is obtained, the performance of the model is tested using a testing video sequence. The estimated features match the actual ones very well. The global correlation coefficient between the actual and estimated features is roughly 0.90.

The research on acoustic to visual speech conversion breaks down the boundary between acoustic and visual speech and facilitates some audio-video integrated applications such as speech-driven face animation, human-computer interface, lip synchronization, foreign film dubbing and joint audio-video coding.

During the research, several ideas have arisen but not yet been implemented, due to time limitations. These ideas are summarized as follows:

1. In our visual speech extraction algorithm, the color information used is the red hue $H = R/(R+G)$. Other color information, such as R/G and standard hue and saturation, also has discriminating power to separate the lip and face pixels. This can be used as a substitute for the red hue.
2. In our visual speech extraction algorithm, at the lip region locating stage, it is assumed that the video sequence contains the frontal view of a speaker without

significant head motion. This assumption will not be necessary if an algorithm of head tracking is applied first.

3. The model of the acoustic to visual speech conversion we use is three independent three-layer feed forward neural networks. Time delay neural networks have an advantage over standard feed forward neural networks for audio-to-visual conversion. A major advantage is that time-delay neural networks operate on not only the current speech frame, but also its neighbours. Therefore, the estimated lip features can incorporate information from neighbouring acoustic information.
4. The approach we use for audio-to-visual conversion belongs to the category of direct conversion. The approach of phoneme-to-viseme conversion also deserves a try.

REFERENCES

- [1] Tsuhan Chen, R.R. Rao, "Audio-visual Interaction in Multimedia Communication," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 179 -182, 1997.
- [2] J. L. Flanagan, "Speech Analysis, Synthesis, and Perception," *Kommunikation und Kyhernetik in Einzeldarstellungen*. 2nd ed., vol. 3, Springer Verlag, Berlin, 1972.
- [3] Tsuhan Chen, "Audiovisual Speech Processing," *IEEE Signal Processing Magazine*, vol. 18, pp. 9 -21, Jan. 2001.
- [4] H. McGurk, J. MacDonald, "Hearing Lips and Seeing Voices," *Nature*, pp. 746-748, Dec. 1976.
- [5] Fu Jie Huang, Tsuhan Chen, "Real-time Lip-synch Face Animation Driven by Human Voice," *IEEE Workshop on Multimedia Signal Processing*, Los Angeles, California, Dec. 1998.
- [6] S. Nakamura, E. Yamamoto, K. Shikano, "Speech-to-Lip Movement Synthesis Maximizing Audio-Visual Joint Probability Based on EM Algorithm," *IEEE Second Workshop on Multimedia Signal Processing*, pp. 53 -58, 1998.
- [7] A.V. Barbosa, H.C. Yehia, "Measuring the Relation between Speech Acoustics and 2D Facial Motion," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp.181-184, 2001.
- [8] T. Chen, H. P. Graf, K. Wang, "Lip-synchronization Using Speech-assisted Video Processing," *IEEE Signal Processing Letters*, vol. 2, no. 4, pp. 57-59, April 1995.

- [9] S. Morishima, K. Aizawa, H. Harashima, "An Intelligent Facial Image Coding Driven by Speech and Phoneme," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 1795, Glasgow, UK, 1989.
- [10] F. Lavagetto, "Converting Speech into Lip Movements: A Multimedia Telephone for Hard of Hearing People," *IEEE Trans. On Rehabilitation Engineering*, vol. 3, no. 1, pp. 1-14, March 1995.
- [11] E. D. Petajan, *Automatic Lipreading to Enhance Speech Recognition*, PhD thesis, University of Illinois at Urbana-Champaign, 1984.
- [12] J. Luetin, N. A. Thacker, S. W. Beet, "Learning to Recognize Talking Faces," *Proceedings of the International Conference on Pattern Recognition*, vol. 4, pp. 55-59, 1996.
- [13] Fabio Lavagetto, "Time-Delay Neural Networks for Estimating Lip Movements From Speech Analysis: A Useful Tool in Audio-Video Synchronization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 5, pp. 786-800, Oct. 1997.
- [14] Z. Melek, L. Akarun, "Automated Lip Synchronized Speech Driven Facial Animation," *IEEE International Conference on Multimedia and Expo*, vol. 2, pp. 623 -626, 2000.
- [15] E. Vatikiotis-Bateson, H. C. Yehia, "Estimation and Generalization of Multimodal Speech Production," *IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing*, vol. 1, pp. 23 -32, 2000.
- [16] A. Caplier, "Lip Detection and Tracking," *Image Analysis and Processing, Proceedings of 11th International Conference on*, pp. 8 -13, Sept. 2001.
- [17] B. P. Yuhas, M. H. Goldstein, T. J. Sejnowski, "Integration of Acoustic and Visual Speech Signals Using Neural Networks," *IEEE Communication Magazine*, vol. 27, pp. 65-71, 1989.

- [18] C. Bregler, H. Hild, S. Manke, A. Waibel, "Improved Connected Letter Recognition by Lipreading," *Proc. IEEE International Conference on Acoustic Speech and Signal Processing*, pp. 557-560, 1993.
- [19] C. Bregler, Y. Konig, "Eigenlips for Robust Speech Recognition," *Proc. IEEE International Conference on Acoustic Speech and Signal Processing*, pp. 669-672, 1994.
- [20] E. Owens, B. Blazek, "Visemes Observed by Hearing-impaired and Normal Hearing Adult Viewers," *J. Speech Hearing Res.*, vol. 28, pp. 381-393, Sept. 1985.
- [21] P. L. Silsbee, A. C. Bovik, "Computer Lipreading for Improved Accuracy in Automatic Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 337-351, 1996.
- [22] G. Potamianos, J. Luettin, C. Neti, "Hierarchical Discriminant Features for Audio-visual LVCSR," *Proc. IEEE International Conference on Acoustic Speech and Signal Processing*, pp. 165-168, 2001.
- [23] K. Mase, A. Pentland, "Automatic Lipreading by Optical Flow Analysis," *Systems and Computers*, vol. 22, no. 6, Japan, 1991.
- [24] M. W. Mak, W. G. Allen, "Lip-tracking System Based on Morphological Processing and Block Matching Techniques," *Signal Processing: Image Communication*, vol. 6, no. 4, pp. 335-348, Aug. 1994.
- [25] A. L. Yuille, P. Hallinan, D. S. Cohen, "Feature Extraction from Faces Using Deformable Templates," *International Journal of Computer Vision*, vol. 8, pp. 99-112, 1992.
- [26] Alan W.C. Liew, S.H. Leung, W.H. Lau, "Lip Contour Extraction Using a Deformable Model", *Proc. International Conference on Image Processing*, vol. 2, pp. 255-258, 2000.

- [27] R. Rao, R. Mersereau, "State-Embedded Deformable Templates," *International Conference on Image Processing*, vol. 1, Washington, D.C, Oct. 1995.
- [28] T. Coianiz, L. Torresani, B. Capril. "2D Deformable Models for Visual Speech Analysis," *Speechreading by Humans and Machines*, vol. 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pp. 391-398, Berlin, 1996.
- [29] M. Kass, A. Witkin, Terzopoulos, "Snakes: Active Contour Models," *International Journal of Computer Vision*, pp. 321-331, 1988.
- [30] Greg I. Chiou, Jenq-Neng Hwang. "Lipreading by Using Snakes, Principal Component Analysis, and Hidden Markov Models to Recognize Color Motion Video," *University of Washington*.
- [31] T.F. Cootes, D. Cooper, C.J. Taylor, J. Graham, "Active Shape Models - Their Training and Application," *Computer Vision and Image Understanding*. vol. 61, no. 1, pp. 38-59, Jan. 1995.
- [32] A. J. Goldschen, *Continuous Automatic Speech Recognition by Lipreading*, PhD thesis, *George Washington University*, Washington, D. C., 1993.
- [33] R. Kaucic, B. Dalton, A. Blake. "Real-time Lip Tracking for Audio-visual Speech Recognition Applications". *Proceedings of the European Conference on Computer-Vision*, vol. 2, pp. 376-386. Cambridge, 1996.
- [34] Wen-Nung Lie, Hung-Chih Hsieh, "Lips Detection by Morphological Image Processing," *Proc. Fourth International Conference on Signal Processing*, vol. 2, pp. 1084 -1087, 1998.
- [35] A. Q. Summerfield, "Lipreading and Audio-visual Speech Perception," *Philosophical Transactions of the Royal Society of London, Series B*, vol. 335, pp. 71-78, 1992.
- [36] <http://www.arts.gla.ac.uk/IPA/ipa.html>.

- [37] B. Dodd, R. Campbell, "Hearing by Eye: The Psychology of Lipreading," *Lawrence Erlbaum Associates*, London, 1987.
- [38] S. Voran, S. Wolf, "Proposed Framework for Subjective Audiovisual Testing," *ANSI Working Group T1A1*, vol. 5, pp. 93-151, Nov. 1993.
- [39] Bellcore, "Experimental Combined Audio/video Subjective Test Method," *ITU-T Study Group 12*, Feb. 1994.
- [40] D. Shah, S. Marshall, "Multi-modality Coding System for Videophone Application," *International Workshop on Image Analysis and Synthesis in Image Coding*, Berlin, Germany, Oct. 1994.
- [41] R. Rao, T. Chen, "Cross-modal Predictive Coding," *Symposium on Multimedia Communication and Video Coding*, New York, Oct. 1995.
- [42] K. Waters, T. M. Levergood, "DECface: an Automatic Lip-synchronization Algorithm for Synthetic Faces," DEC Cambridge Research Lab, Technical Report, September 1993.
- [43] P. Griffin, H. Noot, "The FERSA Project for Lip-sync Animation," Department of Interactive systems CWI, Kruislaan 413, 1098 SJ Amsterdam, Netherlands, Technical report, 1993.
- [44] Tobias Öhman, Giampiero Salvi, "Using HMMs and ANNs for Mapping Acoustic to Visual Speech," in *TMH-QPSR* vol. 1, no.2, 1999.
- [45] C. Garcia, G. Tziritas, "Face Detection Using Quantized Skin Color Regions Merging and Wavelet Packet Analysis," *IEEE Transactions on Multimedia*, vol. 1, no. 3, pp. 264-277, 1999.
- [46] M. Lievin, F. Luthon, "Lip Features Automatic Extraction," *Proc. International Conference on Image Processing*, vol. 3, pp. 168 -172, 1998.
- [47] <http://ece.clemson.edu/speech/cuave.htm>.

- [48] <http://www.ee.surrey.ac.uk/research/vssp/xm2vtsdb>.
- [49] N. Otsu, "A Threshold Selection Method from Gray Level Histograms," *IEEE Transaction on Systems, Man, and Cybernetics*, vol.9, no. 1, pp. 62-66, 1979.
- [50] T.N. Pappas, "An Adaptive Clustering Algorithm for Image Segmentation," *IEEE Trans. on Signal Processing*, vol. 40, no. 4, pp. 901-914, April 1992.
- [51] P. Chou, C. Brown, R. Raman, "A Confidence-based Approach to the Labeling Problem," *Proc. IEEE Workshop on Computer Vision*, pp. 51-56, 1987.
- [52] P.B. Chou, C.M. Brown, "The Theory and Practice of Bayesian Image Labeling," *Int. Journals of Computer Vision*, vol. 4, pp. 185-210, 1990.
- [53] J. Marroquin, S. Mitter, T. Poggio, "Computer Vision," *Journal of the American Statistical Association*, vol. 82, pp. 76-89, 1987.
- [54] D. Geiger, F. Girosi, "Parallel and Deterministic Algorithms from MRF's: Surface Reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 5, pp. 401-412, 1991.
- [55] J. Besag, "On the Statistical Analysis of Dirty Pictures," *Journal of the Royal Statistical Society B*, vol. 48, no. 3, pp. 259-302, 1986.
- [56] Fang Zheng, Zhanjiang Song, Ling Li, et al. "The Distance Measure for Line Spectrum Pairs Applied to Speech Recognition," *Int. Conf. on Spoken Language Processing*, vol. 3, pp. 1123--1126, Sydney, Nov. 30, 1998.
- [57] Itakura F., "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals," *J. Acoust. Soc. Am.*, 1975.
- [58] Martin F. Møller, "A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning," *Neural Networks*, vol. 6, pp. 525-533, 1993.

- [59] J. Canny, "A Computational Approach to Edge Detection," *IEEE Trans. On pattern analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679-698, Nov. 1986.
- [60] J. Besag, "Spatial Interaction and the Statistical Analysis of Lattice System," *Journal Royal Statistic Soc. B*, vol. 36, no. 2, pp. 192-236, 1974.
- [61] S. Geman, D. Geman, "Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721-741, 1984.
- [62] Lawrence Rabiner, Biing-Hwang Juang, "Fundamental of Speech Recognition," *Prentice Hall International Inc.*, 1993.
- [63] Han Zhao, *Talker Detection for Multimedia Conferencing*, M.A.Sc. Thesis, Carleton University, Ottawa, Canada, December 2002.
- [64] Chengliang Li, Richard Dansereau, Rafik Goubran, "Acoustic Speech to Lip Feature Mapping for Multimedia Applications," accepted by *IEEE 3rd International Symposium on Image and Signal Processing and Analysis*, Rome, Italy, Sept. 18-20, 2003.