A Multimodal Sensor Fusion Architecture for Audio-Visual Speech Recognition

by

Mustapha A. Makkook

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2007

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

A key requirement for developing any innovative system in a computing environment is to integrate a sufficiently friendly interface with the average end user. Accurate design of such a user-centered interface, however, means more than just the ergonomics of the panels and displays. It also requires that designers precisely define what information to use and how, where, and when to use it. Recent advances in user-centered design of computing systems have suggested that multimodal integration can provide different types and levels of intelligence to the user interface. The work of this thesis aims at improving speech recognition-based interfaces by making use of the visual modality conveyed by the movements of the lips.

Designing a good visual front end is a major part of this framework. For this purpose, this work derives the optical flow fields for consecutive frames of people speaking. Independent Component Analysis (ICA) is then used to derive basis flow fields. The coefficients of these basis fields comprise the visual features of interest. It is shown that using ICA on optical flow fields yields better classification results than the traditional approaches based on Principal Component Analysis (PCA). In fact, ICA can capture higher order statistics that are needed to understand the motion of the mouth. This is due to the fact that lips movement is complex in its nature, as it involves large image velocities, self occlusion (due to the appearance and disappearance of the teeth) and a lot of non-rigidity.

Another issue that is of great interest to audio-visual speech recognition systems designers is the integration (fusion) of the audio and visual information into an automatic speech recognizer. For this purpose, a reliability-driven sensor fusion scheme is developed. A statistical approach is developed to account for the dynamic changes in reliability. This is done in two steps. The first step derives suitable statistical reliability measures for the individual information streams. These measures are based on the dispersion of the N-best hypotheses of the individual stream classifiers. The second step finds an optimal mapping between the reliability measures and the stream weights that maximizes the conditional likelihood. For this purpose, genetic algorithms are used.

The addressed issues are challenging problems and are substantial for developing an audio-visual speech recognition framework that can maximize the information gather about the words uttered and minimize the impact of noise.

Acknowledgements

Now that I look at the past two years that I spent working on my Masters degree, I remember the greatest times and the worst times. But I also remember the people who shared with me those times.

Dr. Otman Basir thanks a lot. You have been great as a supervisor as you are a great person. Thanks for your trust and for giving me the chance to prove something about myself. I should also thank Dr. Fakhri Karray, my co-supervisor, for always motivating me to work hard and for his guidance and supervision. Thanks to Dr. George Freeman and Dr. Paul Ward for reading this thesis.

Thanks to my family: Wafaa, my mom, whose coffee I miss everyday and who always stood up for what I wanted to do; Ali, my dad whom I look up to every single day, and who has taught me the greatest lessons of my life; Noha, my sister and Bilal, my brother, for being the cheerful and encouraging siblings and friends they have always been.

Thanks to Khaled and Rana Soudki, my family in Canada, for everything they did to make me feel at home. Khaled, you always kept my spirit up and believed that I could do great things, and Rana, your cooking is amazing, but your company is even better.

Thanks to my friends: Adel Fakih for the valuable discussions and for the the continuous feedback on my thesis, Kamal Mroue, Ahmad Rteil and Hatem Elbeheiry for the great times we had together. You guys are the best. Thanks to my friends and colleagues in the PAMI lab for all the great times we shared and for providing the friendly environment which I had the pleasure to work in.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The last decade has witnessed a trend towards an increasingly ubiquitous computing environment, where tiny smart devices are being integrated in mobile phones, cars, medical instruments and almost every aspect of our lives. This has been coupled by major advances in information and communication technology, with sensors, actuators, and integrated processors being connected together via high-speed networks to provide people with high quality services. In fact, today's leading-edge information-processing devices are being so integrated within the environment that it becomes necessary for people to interact with them more naturally and casually than they currently do, and in whatever context they find themselves. This applies for instance in intelligent vehicles, where an easy and immediate form of dialog between the driver and the vehicle equipment could lead to a safer and more efficient driving experience.

The most popular way to attain this natural human-machine interaction is through voice-activated controls that make use of the various speech recognition algorithms developed in literature. However, developing voice-activated controls remains challenged in noisy environments despite all the technical advances that have been developed to enhance their capabilities. Although there has been good progress in speech recognition for well-defined applications like dictation and medium vocabulary processing applications, speech recognition has not yet reached the level of performance, which allows its deployment in

embedded systems as a reliable user interface. In fact, speech recognition systems lag human speech perception even in perfectly clean acoustic environments. For speech recognition systems to be of practical use in noisy environments, such as automotive, crowded areas and simultaneous human computer discourse applications, the issue of robustness must be addressed.

This can possibly be accomplished by utilizing other sensing modalities to complement the acoustic signal of speech. As a matter of fact, in almost every context, carefully designed multimodal interfaces are shown to be more beneficial than any single-modality interface. An example in line with this strategy is to fuse visual lip movements and expressions with the acoustic signal of the speech so as to maximize information gather about the words uttered and to minimize the impact of acoustic noise. This is referred to as Audio-Visual Speech Recognition (AVSR) or Automatic Speech-Reading (ASR). In fact, understanding speech from visual information is an attractive technology that has captured the interest of researchers to improve speech recognition systems. Using a camera or infrared sensor, an audio-visual speech recognition system will be capable of providing supplementary information from the lips movement and correlating the results with input from a microphone. The idea behind using the visual modality is that both the production and perception of human speech are bimodal in nature, and thus speech recognition systems require the integration of both the visual and the acoustic modalities. The now classic *McGurk effect* is an excellent example: the auditory syllable /bi/ presented in synchrony with a videotape of a talker saying the syllable /gi/ is usually perceived as /di/, a syllable not presented to either modality. This phenomenon has important applications for theories of speech perception which must account for how and why auditory and visual signals are integrated during phonetic processing. Our primary goal in this work is to exploit this human perceptual principle of sensory integration to develop a multimodal user interface that is capable of understanding human speech.

To be able to perform in real-world human environment, the system will also have to be able to understand the user intent, an issue that is not considered within the scope of this thesis. Recognizing individual utterances and gestures can be helpful for basic commands and a restricted environment, but natural language understanding involves identifying the meaning of each gesture, as well as disambiguation between multiple meanings, since even

basic gestures can be highly ambiguous themselves.

This work aims at developing a multimodal human computer interface, which can efficiently fuse the information present in speech and visual lip gestures of the user, to form a robust and comprehensive system, capable of improving the recognition accuracy of speech utterances. This will involve the development and implementation of a multimodal framework for the intelligent feature extraction and fusion of input audio and visual signals as well as understanding the fused information.

1.1.1 Application Domains

The benefits of this work scan a wide range of areas that are capable of improving the quality of life for people who are either hearing impaired, or merely trying to cope with the technological advances of modern life. Hearing impaired people can restore by lipreading about 50% of the speech. An audio-visual speech recognition system, on the other hand, can provide them with complementary information, based on phonetic feature recognition. Moreover, this casual multimodal user-interface will present new, simpler methods to interact with mobile phones, personal digital assistants, traditional computers, and even smart homes or entertainment systems.

Perhaps the most promising application of this field of research is its implications for speech therapy. The idea here is related to those people who cannot speak due to hearing loss. Since it would seem that speech perception is a combination of the audio and visual inputs, then those people can use the visual component to learn speech. Indeed that is the project that has been started by Cole at al. [18], which utilizes a computer-generated human head that is designed to speak using extremely precise mouth movements. This computer program is implemented at the Tucker-Maxon Oral School in Portland, Oregon to teach its deaf students how to speak by seeing and mimicking the head's oral gestures.

Another application domain for AVSR is the in-vehicle environment. In a car for instance, operating a telephone or navigational system by hand may distract the driver, whereas speaking the command is much safer. Unfortunately, speech recognition in cars is a demanding task due to interference noise caused by the wind, tires and engine. Consequently, it is crucial in such an environment to use visual observation of the lips to convey speech information.

An application that is also of interest is in videophones and telephone handset cameras. Business applications such as stock and commodity trading profit from transcriptions, but making accurate transcriptions is quite difficult in noisy environments. A miniature camera and infrared source could be built into the telephone handset for video image capture for a speechreading system.

Other applications include security and identification, video games, transcription of television broadcasts for the deaf, defense applications and interactive web browsers. This work will serve as a baseline for future work to be developed in order to bring this area of research one step forward towards commercialization. In addition, most of the problems addressed in this research will benefit other areas of research, such as speaker identification and verification [42], [84], speaker localization [10], [85] and visual text-to-speech [17], [23]. Further investigation is expected to bring about better performance and to address several hardware and software implementation issues in order to come up with a robust and natural intelligent user interface.

1.2 Objectives

The primary goals of this research are to design, implement and evaluate a novel audio-visual speech recognition system that is capable of providing a reliable and user-friendly interface for smart embedded devices. This goal will be realized through the following objectives.

1. Visual feature extraction: Several approaches have been proposed in literature to address the issue of tracking lips or mouth regions and then extracting a representative set of features that are able to fully represent the visual speech information. However, we need to propose new visual processing and extraction algorithms that are well adapted and perfectly tuned to match our fusion and recognition modules.

2. Stream reliability: The reliability of an audio or video stream depends on the noise that is present in the respective stream as well as on the level of voicing and other factors. This introduces the problem of how to perform an adaptive integration of the two modalities for the task of automatic speechreading. The integration scheme

has to be able to achieve the best synergy by dynamically weighting the multimodal streams based on their reliabilities.

3. Synchronization: Perceptual studies have shown that the production of sound and the movements of the lips are not perfectly synchronous. Indeed, lips start articulation by around 120 ms before sound is produced. Consequently, there is a need to create a certain level of synchronization between acoustic and visual information during the process of audio-visual speech recognition.

4. Level of integration: Several experiments have been performed to understand whether sensory integration in man and machine occurs at an early (feature level), late (decision level), or intermediate stage. However, no model was developed to clearly depict the necessary level of integration. While some studies claim that an early feature-level fusion is capable of eliminating the need for synchronization, others argue that feature-level fusion techniques are not capable of modeling the reliability of the multi-modal streams and thus late decision-level fusion techniques should be used.

5. Robustness: The developed audio-visual speech recognition system should be able to perform equally well in different contexts and across different speakers. In addition, it should be able to achieve a level of performance that is better than that of the independent unimodal systems.

6. Visual speech modeling: A particular issue of interest is how to model visual speech in the task of automatic speech recognition. In fact, the basic units of acoustic speech, *phonemes*, are not well suited for visual speech representation because of the fact that different sounds are perceived equally in the visual domain. For instance, the phonemes /p/, /b/ and /m/ are identical in the visual domain and thus cannot be distinguished using the visual modality. Using visually distinguishable units called *visemes*, which consist of phoneme clusters, may bring a feasible solution to this problem. However, it is questionable whether visemes are able to fully represent the visual information, and whether more representative units are required.

1.3 Thesis Overview

The remaining of this thesis is structured as follows.

Chapter 2 reviews the state of the art of audio-visual speech recognition, first by dividing the system into modules, and then by describing and analyzing the various contributions to date of each module.

Chapter 3 gives an overview of the overall system architecture and principal contributions.

Chapter 4 deals with the audio-front end design. Since an extensive amount of work has been done in this area, this chapter only uses a speech model and feature extraction algorithm previously developed in literature.

Chapter 5 explains the fundamental building block developed for the visual front-end design. This involves visual speech modeling, video preprocessing, and visual feature extraction. The visual front-end system is tested on a database of mouth region images of several speakers.

Chapter 6 elaborates on the multimodal fusion framework, that combines the audio and visual features based on the reliability of their respective channels. For this purpose, a stream reliability assessment model is developed and mapped into stream weights in an optimal fashion. A multimodal recognition system is used to test the AVSR system on an audio-visual database.

Chapter 7 summarizes the contributions of this thesis and introduces the focus of future research.

Chapter 2

Background and Literature Review

The first audio-visual speech recognition system was introduced in 1984 by Petagan [63]. This system used simple image thresholding to extract binary mouth images from which a set of visual features (mouth height, width, perimeter, and area) are derived. Since then, a lot of work has been done in this domain. Most of this work has shown improved performance over single-modality systems. However, this improvement was limited by the size of used dataset and the level of the acoustic signal-to-noise ratios.

This chapter reviews the state of the art of what has been done in processing and understanding speech using multiple modalities. When building an AVSR system, four main issues must be considered: feature design and extraction, the choice of speech units, recognition, and multimodal fusion. Most of the current work in AVSR is based on methods that implement these steps sequentially and independently. Before exploring what has been done in literature for implementing these steps, we will introduce an in-depth representation of the complete physiological mechanism of speech production.

2.1 Speech Production

In order to develop practical audio-visual speech recognition systems, it is beneficial to understand the physiological mechanism of speech production. The speech waveform is an acoustic sound pressure wave which originates from the movements of the human speech production system [70]. The main components that comprise this mechanism are the lungs,

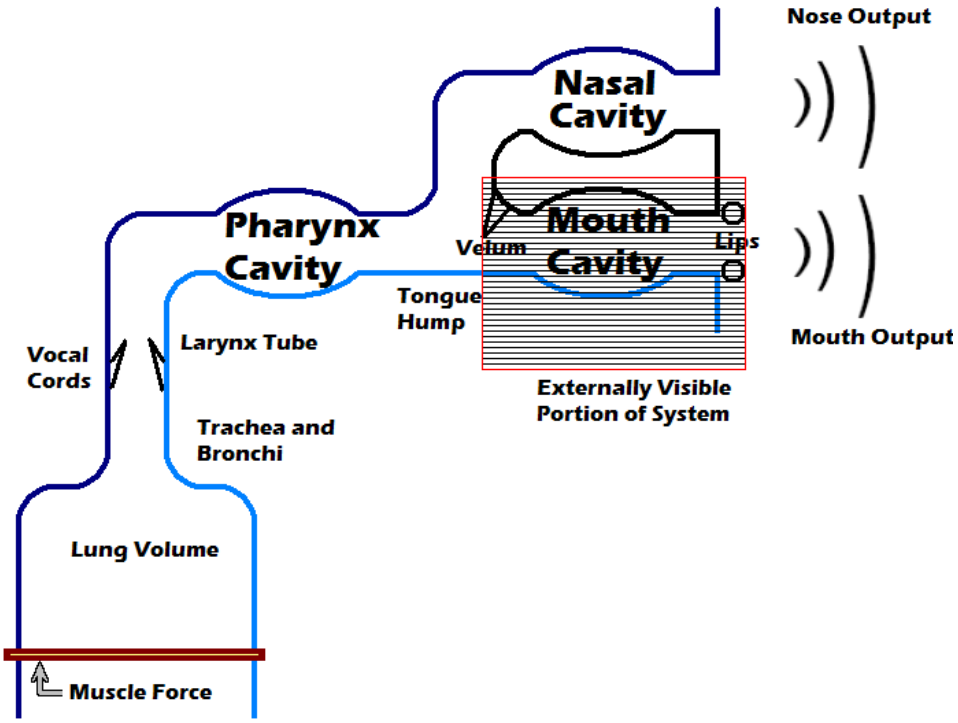

Figure 2.1: Complete physiological mechanism of speech production [70].

trachea, larynx, pharyngeal cavity (throat), oral cavity (mouth) and nasal cavity (nose). Figure 2.1 shows a representation of the physiological mechanism of speech production. The lungs and associated muscles act as the source of air exciting the vocal mechanism. The muscle force pushes air out of the lungs and through the bronchi and trachea. Speech sounds can be classified into voiced and unvoiced sounds. Voiced sounds are produced when the vocal cords are tensed incurring a vibration from the air flow. Unvoiced sounds, on the other hand, are produced by turbulent flow of air created at a constriction in the vocal tract.

In the acoustic speech modality the resultant pressure wave stemming from the mouth and nasal cavities is a combination of all parts of the speech production. This does not imply that the acoustic representation of speech is complete, as the McGurk effect still illustrates the necessity of the visual modality in speech production and perception. However, it does illustrate that, unlike the visual speech modality, the acoustic modality does have direct interaction with the entire speech production mechanism.

2.2 Feature Design and Extraction

As in any pattern recognition problem, the first major issue in automatic speechreading is feature design and extraction. This includes features derived from the audio cues as well as features derived from the visual cues. Much work has already been done in deriving audio features for audio-only systems [70], and therefore this area will not be explored in depth in this work. The important issue here is the visual front-end design.

The visual front-end stage encodes stimuli coming from the visual cues (mainly the lips) of a speaker and transforms it into a suitable representation that is compatible with that of the recognition module. However, prior to this feature extraction process, a number of preprocessing steps have to be done as shown in Figure 2.2. This involves face detection followed by ROI extraction. Then, the lips of the speaker are tracked in consecutive frames. Following these steps, and given an informative set of features, the visual front-end module can proceed with feature extraction. A number of approaches have been proposed in literature for this purpose. These approaches can be categorized as either appearance-based, shape-based, or both. We will give an overview of the preprocessing

steps (face detection, ROI extraction and lip tracking) and then proceed to talk about the visual feature extraction.

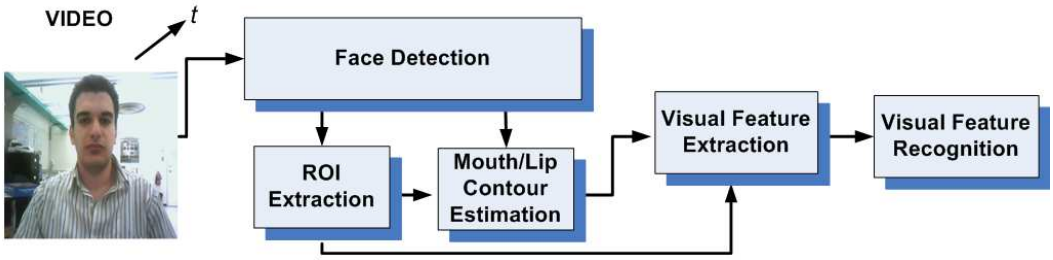

Figure 2.2: Visual speech recognition system.

2.2.1 Visual Preprocessing

Before extracting any visual features, a number of preprocessing steps are required as described in Figure 2.2. The first step is usually face detection followed by *Region Of Interest* (ROI) extraction. The ROI consists of the region of the face that contains most of the speech information. Of course there is no unique understanding of where most of the speech information is located and hence there are many interpretations of what we mean by ROI. This issue will be discussed later. However, we can now establish that the ROI depends on the type of visual data being provided to the visual speech recognition system. In case shape-based visual features are to be extracted, for instance, the additional step of lips detection might be needed using tools such as *snakes*, *templates*, and *active shape and appearance models*. Next, the mouth of the subject needs to be tracked in consecutive frames to extract relevant features that will be fed into a classifier for speech recognition.

Face and facial part detection is a classical pattern recognition problem that has captured the interest of researchers for many years. Consequently, many techniques have been developed to solve this problem. These techniques can be divided into two categories: the traditional image processing techniques use methods such as color segmentation, edge detection, image thresholding, template matching, or motion information [32] for determining the region of the face; the second category is based on statistical modeling approaches, employing neural networks for instance [74].

Among the various algorithms used for face and facial part detection, two popular approaches have been used extensively in literature for AVSR. Both of these techniques belong to the second category. The first algorithm developed by Liang et al. [45] begins with the detection of the user's face using the neural network-based approach described in [3]. Once the face region is detected, a cascade of *Support Vector Machine* (SVM) classifiers is used to locate the mouth within the lower region of the face. To adapt to scale variations, a multi-scale search in an estimated range is employed by repeatedly resampling the source region image by a constant factor. Next, two SVM filters with and without facial hair, are applied to each test pattern and its rotated versions in the image plane. The highest mouth classification score among all rotated patterns and SVM classifiers is used to determine the refined location of the mouth.

The second algorithm is the one developed by Senior [76]. Many systems have used this algorithm including Neti et al. [60]. This algorithm starts by choosing a face template size. Then it uses a face pyramid over all permissible face locations and scales in the given image to search for possible face candidates. By "permissible" we mean the face candidates that contain a relatively high proportion of skin-tone pixels. The selected face candidates are then normalized to the chosen template size, and their grayscale pixel values are placed into separate feature vectors. Each of these vectors is then given a score based on a two-class (face versus non-face) *Fisher Linear Discriminant Analysis* (LDA) [71], as well as its *Distance From Face Space* (DFFS), i.e. the face vector projection error onto a lower-dimensional space obtained by means of *Principal Component Analysis* (PCA) [12]. All candidate regions having a score that exceeds a certain threshold are returned by the algorithm as faces.

After detecting the face, an AVSR system has to locate the ROI. The ROI in an AVSR system is usually a rectangle containing the image pixels of the speaker's mouth region. As mentioned before, there is no unique understanding of where the boundaries of a mouth region lie in terms of relevant speech information. The ROI can include large parts of the lower face, such as the jaw and cheeks [66], or even the entire face [60]. It can be a disk around the mouth center [25], or a three-dimensional entity containing adjacent frame ROIs [64], which enables the system to capture dynamic speech information.

The algorithm proposed by Senior [76] for face detection (described above), also per-

forms ROI extraction and is commonly used for AVSR systems. It does that by attempting to detect a set of predefined features on the face region. These features include the lip corners and centers. Each feature location is determined by using a score that is based on prior feature location statistics, linear discriminant, and *distance from feature space*, based on the chosen feature template size. A training step is required for face detection and facial feature estimation.

2.2.2 Visual Feature Extraction

Given an input video of a person speaking, the task of any visual speech recognition system is to extract visual speech features that could be used for recognizing the uttered words. Visual features are either appearance-based, shape-based, or a combination of both.

Shape-based Features

Shape-based feature extraction is usually based on deriving features for the lip contours. These features include geometric features such as mouth height and width [2], [63], Fourier and image moment descriptors of the lip contours [34], snakes [43], statistical models of shape, such as *Active Shape Models* (ASM) [22] , or other parameters of lip-tracking models. A *snake* is an elastic curve represented by a set of control points. The control point coordinates are iteratively updated, by converging towards the local minimum of an energy function, defined on basis of curve smoothness constraints and a matching criterion to desired features of the image [43]. Such an algorithm is used for lip contour estimation in the speechreading system of Chiou and Hwang [15]. Another widely used technique for lip tracking is by means of lip *templates*, employed in the system of Chandramohan and Silsbee [11] for example. Templates constitute parametrized curves that are fitted to the desired shape by minimizing an energy function, defined similarly to snakes. *B-splines*, used by Dalton et al. [24], work similarly to the above techniques as well.

Appearance-based Features

However, shape-based features are often insufficient on their own and do require the integration of appearance-based features. Appearance-based features are extracted from

the entire image containing the region of interest (usually the mouth and the chin of the speaker). This results in a feature vector with a dimensionality that is quite large to allow for statistical modeling of speech classes using HMMs for example. Therefore, the dimensionality of the feature set is reduced using linear transform techniques that capture most of the speech-relevant information. These techniques include *Principal Component Analysis* (PCA), *Discrete Cosine Transform* (DCT) [69], *discrete wavelet transform* [64], *Hadamard and Haar transforms* [32] and *Linear Discriminant Analysis* (LDA) based data projection [71]. Some examples of this approach include simple gray levels [30], PCA of pixel intensities [9], transform-based compression coefficients [69], edges [2], and filters such as sieves [51].

In this thesis, we introduce an approach for visual feature extraction that is appearance-based. Therefore, we will review some of the approaches that have been proposed for appearance-based visual feature extraction, and that are related to our approach. In [9], Bregler et al. introduced the concept of "eigenlips" in regards to a similar approach from Turk and Pentland [79] for face recognition, called "eigenfaces". In this work, the principal components of the lip contours are extracted. These principal components are used to construct a set of basis lip images such that the PCA coefficients are uncorrelated. The problem with this approach is that it only captures second-order statistics and thus may not be appropriate for motion of the lips, where higher order statistics are required. In [27], Black et. al introduced the concept of constructing basis flow fields from example motions using principal component analysis. In motions such as that of the mouth, principal component analysis may not be perfectly appropriate. This is due to the fact that lips movement is complex in its nature, as it involves large image velocities, self occlusion (due to the appearance and disappearance of the teeth) and a lot of non-rigidity. In [77], Tamura et. al proposed a multimodal speech recognition method using optical flow analysis for lip images. Their method calculates two kinds of visual feature sets in each frame. The first feature set consists of variances of vertical and horizontal components of optical flow vectors. The second feature set consists of the maximum and minimum values of the integral of the optical flow. Our approach aims at extracting other features related to the optical flow, namely the coefficients of their independent components, as will be discussed in Chapter 5.

As said before there are some approaches which integrate shape-based feature with appearance-based features in a single feature vector for visual speech recognition. An example of this approach is the *Active Appearance Model* (AAM) [21]. Once the visual features are extracted, they are integrated to synchronously extracted audio features for audio-visual speech recognition.

2.3 Recognition

Whatever the final choice of representation of the visible speech gestures, the other major issue is how to recognize this information along with the information about the acoustic stream of information so that the best use can be made of the two modalities together. A number of recognition approaches have been proposed in literature for the task of audio-visual recognition. These approaches include: a simple weighted distance in visual feature space [63], artificial neural networks [9], [44], SVMs [30] and *Dynamic Bayesian Networks* (DBNs), which include *Hidden Markov Models* (HMMs). We will discuss in this section some of the most widely used approaches.

2.3.1 Dynamic Bayesian Networks (DBN)

The HMM as well as other audio-visual models used in existing AVSR systems, are special cases of dynamic Bayesian networks. DBNs are directed graphical models of stochastic processes in which the hidden states are represented in terms of individual variables or factors. A DBN is specified by a directed acyclic graph, which represents the conditional independence assumptions and the conditional probability distributions of each node [40].

Hidden Markov Models (HMM)

The HMM provides a stochastic framework that is commonly used for speech recognition. It is the most commonly used classifier in both audio-only and audio-visual speech recognition. HMMs statically model transitions between the speech classes and assume a class-dependent generative model for the observed features. Let us denote the set of speech classes by C, and the l_s-dimensional feature vector in stream s at time t by $\mathbf{o}_t^{(s)} \in R^{l_s}$,

where $s = a$ if we are referring to the audio stream and $s = v$ if we are referring to the video stream. We will use this notation throughout this thesis. The HMM assumes a sequence of hidden states that are sampled according to the transition probability parameter vector $\mathbf{a}_s = [Pr[c'|c''], c', c'' \in C]$. These states subsequently emit the observed features with class-conditional probabilities $P(\mathbf{o}_t^{(s)}|c)$, $c \in C$. In most of the work done in this area, the HMMs are assumed to have a continuous observation probability density, modeled as a mixture of Gaussian densities:

$$Pr[\mathbf{o}_t^{(s)}|c] = \sum_{k=1}^{K_{s,c}} w_{s,c,k} \mathcal{N}(\mathbf{o}_t^{(s)}; m_{s,c,k}, s_{s,c,k}), \tag{2.1}$$

where the $K_{s,c}$ mixture weights $w_{s,c,k}$ are positive and add to one, and $\mathcal{N}(\mathbf{o}; \mathbf{m}, \mathbf{s})$ is the l-variate normal distribution with mean $\mathbf{m}$ and a diagonal covariance matrix $\mathbf{s}$.

As will be discussed in Section 2.4, there are two principal models for audio-visual fusion: feature fusion and decision fusion. Feature fusion models combine acoustic and visual features into a single feature vector and transmit them directly to a single, bimodal classifier. For these models, a regular left-to-right HMM [70] is used. On the other hand, in decision fusion systems, two parallel, unimodal classification systems are employed and the results from each are fed forward for fusion and final decision making, for example, on a probabilistic basis. For these kinds of systems, conventional HMM recognizers are useless because they assume asynchrony of the visual and acoustic data (which is not always the case). Therefore, other models have been used. Some of the most successful decision fusion models include the *Multi-Stream HMM* (MSHMM), the *Product HMM* (PHMM), the *Independent HMM* (IHMM), the *Factorial HMM* (FHMM) and the *Coupled HMM* (CHMM). The multi-stream HMM [47] assumes that the audio and video sequences are state synchronous, but unlike the HMM for feature fusion, allows the likelihood of the audio and visual observation sequences to be computed independently. Although more flexible than the HMM, the multi-stream HMM cannot accurately describe the state synchrony of the audio-visual speech. The audio-visual multi-stream product HMM [26], on the other hand, extends the previous model by representing each hidden state of the multi-stream HMM as a pair of one audio and one visual state. Due to its structure, the multi-stream product HMM allows for audio-visual state asynchrony, controlled through the state transition matrix of the model, and forces the audio and video streams to be in synchrony

at the model boundaries (phone level or word level). The audio-visual streams can also be modeled using two independent HMMs [60], one for audio and one for visual features. This model extends the level of asynchrony between the audio and visual states of the previous models, but fails to preserve the natural dependency over time of the acoustic and visual features of speech.

In [57], Nefian et al. used two novel models for audio-visual speech recognition: the factorial HMM (FHMM) and the coupled HMM (CHMM). The FHMM has not been widely used and therefore we will not talk about it here. The CHMM, which was introduced in [8], models multiple interacting processes while maintaining the Markov condition that each state must depend only on the prior state. Unlike the independent HMM used for audio-visual speech recognition, the CHMM can capture the interactions between the audio and video streams through the transition probabilities between the backbone nodes. The CHMM can model the audio-visual state asynchrony and preserve the natural audio-visual dependencies over time. This technique has been used in many systems including that of Potamianos et al. [67].

In general, HMMs are very beneficial and widely used. However, there are some problems associated with them which drive certain researchers to use other recognition modules. For instance, it is often hard to determine the right HMM state complexity, which means that one must search through the model space for the proper number of states. Furthermore, HMMs make certain assumptions that often do not hold, for instance, that the features are uncorrelated. They also often fail to capture co-articulation during speech production.

Other Models

Other models that are based on DBNs have been proposed in literature. In [75] for example, Saenko et al. work on the task of phrase recognition, and they propose a DBN with loosely couple streams of articulatory features, where the observation model is a Gaussian mixture over the feature classifier outputs. This approach can capture many of the effects of co-articulation during speech production. In fact, it is shown that this approach can better account for variation introduced by speech which was spoken more quickly than that in the training data. Another model that uses DBNs has been developed by Gowdy et al.

[31]. The advantage of this approach is that it can handle two or more streams, and can account for the reliability of these streams.

2.3.2 Neural Networks (NN)

In contrast to HMMs, *Neural Networks* (NN) make only few assumptions about the underlying data and thus they can be generalized to large classes using sufficiently large training data. However, training is slow and asynchrony modeling is difficult to achieve. One such approach is proposed by Meier et al. [52]. This approach uses a *Multiple State-Time Delayed Neural Network* (MS-TDNN) for recognition of the audio-visual speech task. Combining visual and acoustic data is done on the phonetic layer or on lower levels. Another approach that uses neural networks is developed by Yuhas et al. [83]. This work uses layered feed-forward networks. The image of the speaker's mouth is presented in the bottom layer of units, which then passes the signal to a layer of hidden units, that in turn projects this signal to an output layer. As a signal travels from unit to unit, it is multiplied by a weight that resembles the reliability of the signal.

2.4 Multimodal Fusion

2.4.1 Model Architectures

Multimodal fusion is a very important research area that relies on measuring a set of complementary features from multiple sensors or modalities and combining these features in an "intelligent" way that maximizes information gather and minimizes the impact of noise coming from the individual sensors. In AVSR, the issue of multimodal fusion has received a lot of attention, as it aims to combine the multiple speech informative streams into a multimodal classifier that can achieve better classification results than the audio- and visual-only classifiers. The first issue to be addressed in fusion of audio-visual speech is where the fusion of the data takes place. Cognitive studies have suggested four architectures for the combination of audio and visual modalities:

- Direct Identification (DI), in which acoustic and visual data are combined and transmitted directly to a single, bimodal classifier.

- Separate Identification (SI), in which two parallel, unimodal classification systems are employed and the results from each are fed forward for fusion and final decision making.

- Dominant Recoding (DR), in which auditory processing is supposed to be dominant. Visual data is recoded into the dominant modality. Each modality thus generates a representation appropriate to the dominant modality, such as a tract transfer function. The two estimates are then fused and fed forward to a classifier.

- Motor-space Recoding (MR), in which both inputs are projected and recoded into an amodal (not audio or video) common space, such as that of articulatory configurations, and the two representations are fused and passed to the classifier.

In literature, the most widely used model architectures are the direct identification model (also referred to as feature fusion) and the separate identification model (also referred to as decision fusion). For this purpose, in this section we will only discuss the work that has been done for these two models.

Feature Level Fusion

The first architecture integrates data on the feature level. This is referred to in literature as feature fusion, direct identification (DI), or early integration. In this case, the audio and visual features are used simultaneously and equally for classification using a bimodal classifier. Feature-level fusion algorithms train this classifier on the concatenated vector of audio and visual features or any appropriate transformation of it. Examples of feature fusion methods include plain feature concatenation [2] and hierarchical discriminant feature extraction [60].

Concatenative feature fusion is the simplest fusion technique. Given time-synchronous audio and visual feature vectors $\mathbf{o}_t^{(A)}$ and $\mathbf{o}_t^{(V)}$, with dimensionalities D_A and D_V respectively, this method generates a concatenated audio-visual feature vector at every time instance t:

$$\mathbf{o}_t^{(AV)} = [\mathbf{o}_t^{(A)T}, \mathbf{o}_t^{(V)T}]^T \in R^D, \tag{2.2}$$

where $D = D_A + D_V$ is the dimensionality of the combined feature vector. This can result in a very large feature vector which may not be suitable for representing the underlying data. That is where discriminative feature fusion can be used.

Hierarchical discriminant feature fusion projects the concatenated feature vector $\mathbf{o}_t^{(AV)}$ into a lower-dimensional feature vector. In [60], this is done using *Linear Discriminant Analysis* (LDA) projection on the concatenated vector, while seeking the best discrimination among the speech classes of interest. LDA is followed by a *Maximum Likelihood Linear Transform* (MLLT) rotation of the feature vector to improve maximum-likelihood data modeling using the Gaussian mixture emission probability densities of Equation 2.1. This approach is hierarchical because it is found on two stages (LDA followed by MLLT). It is therefore referred to as hierarchical LDA or HiLDA. The resulting audio-visual feature vector is:

$$\mathbf{o}_t^{(HiLDA)} = \mathbf{P}_{MLLT}^{(AV)} \mathbf{P}_{LDA}^{(AV)} \mathbf{o}_t^{(V)}. \tag{2.3}$$

Decision Level Fusion

The problem with feature fusion techniques is that they provide no way of capturing the reliability of the individual streams of information. Reliability in audio-visual integration is an important issue because factors such as noise, face occlusions and volume of the speaker's sound can lead to a certain modality being more "trustworthy" than the other. Decision-level fusion (also called separate identification (SI) or late integration), on the other hand, uses the two outputs of the audio and visual classifiers to combine the two modalities. This framework provides a mechanism for capturing the reliability of each modality, by borrowing from classifier combination literature [39].

A number of classifier combination methods have been used in the AVSR literature. One such technique is to use a cascade of fusion modules, some of which using only rank-order classifier information about the speech classes of interest [63]. However, the most popular approach in this regard uses a parallel architecture, adaptive combination weights, and class measurement level information [67]. This corresponds to finding the most likely speech class using linear combination in the log-likelihood domain of the two single-modality classifier decisions. This combination in the log-likelihood domain can be done at several levels, and every level corresponds to a certain recognition model as described in the previous section.

In the case where single-stream HMMs, with the same set of speech classes (states), are used for both audio- and visual-only classification, we consider this likelihood combination to be at a frame (HMM state) level, and it is modeled by means of a multistream HMM. The state-dependent emission of the audio-visual observation vector $\mathbf{o}_t^{(AV)}$ is thus governed by:

$$P(\mathbf{o}_t^{(AV)}|c) = P(\mathbf{o}_t^{(A)}|c)^{\lambda_A} P(\mathbf{o}_t^{(AV)}|c)^{\lambda_V}, \tag{2.4}$$

for all HMM states $c \in C$.

Since we are dealing with continuous speech recognition, where a sequence of classes should be estimated (a number of phones or words), there should be a certain way of accounting for the temporal differences between the two streams (stream asynchrony). This asynchrony is observed in [9] to be up to the order of 120 ms. Decision fusion at a state or frame level is not good enough because the states are probably not in synchrony. For this reason, decision should be done at a later stage. One approach could be to wait until the end of an utterance and then fuse the decisions about the different streams based on their log-likelihoods. One such technique is the one applied in [28], which uses the discriminative model combination technique. A third way to approach decision fusion, is to fuse at an intermediate stage. Such a scheme is typically implemented by means of the product HMM [81], or the coupled HMM [8], as discussed in the previous section. So, in conclusion, to allow asynchrony between the audio and visual streams, it is required to integrate at a late or intermediate stage. This allows for the modeling of the audio and visual streams temporal properties.

2.4.2 Stream Reliability

For multimodal fusion to be of practical use in noisy environments, such as automotive, crowded areas and simultaneous human-computer discourse applications, the issue of observation reliability must be addressed. Consider a sensor that at times gives an erroneous output and therefore has a certain level of unreliability associated with it. Then, when the recognition system receives information from this sensor, it should assume with a certain probability that this information might be incorrect. What sensor fusion does is integrate information from a different sensor (which has its own probability of error as well) in an

effort to ramify any classification errors. Since one sensor might have a higher probability of error than another sensor (due to higher noise elements, sensor malfunction, etc.), there should be a way of estimating the reliability of every sensor and giving greater confidence in the classification phase to the sensor with higher reliability. This constitutes an important aspect of the multimodal sensor fusion paradigm in AVSR. Here, the system autonomously gathers observations from its multiple sensors adapting itself to environmental changes and sensor malfunction. For this purpose, a common approach in many stream integration methods is to use stream weights that operate as exponents to each stream's probability density. Such stream weights have been applied in AVSR using HMMs, DBNs, ANNs, and other classifiers. These weights are estimated from some reliability measures of the individual streams. Consequently, this problem can be formulated as two main tasks: the first task is to derive a suitable reliability measure for the individual information streams, and the second task is to find an optimal mapping between the reliability measures and the stream weights that maximizes information gather.

Signal-based Approaches

The most popular approach to reliability estimation is that developed in [68] by Potamiaonos and Potamianos. This algorithm uses a multi-stream HMM for AVSR recognition as described in the previous section. The resulting classification result should thus satisfy:

$$c = \underset{c}{\operatorname{argmax}}\, P(c|\mathbf{o}_t^{(A)})^{\lambda_A} P(c|\mathbf{o}_t^{(V)})^{\lambda_V}. \tag{2.5}$$

λ_A and λ_V are the audio and video exponents respectively, which model the reliability of each stream, and they satisfy:

$$0 \leq \lambda_A, \lambda_V \leq 1 \tag{2.6}$$

$$\text{and} \quad \lambda_A + \lambda_V = 1. \tag{2.7}$$

The Generalized Probabilistic Descent (GPD) algorithm is used to estimate the stream exponents early during the training phase, and then the estimated stream exponents are fixed for a particular audio-visual environment and database. The problem with this approach is that it cannot model the rapid changes in observation conditions over time. For

example, possible noise bursts, face occlusion, or other face tracking failures can greatly change the reliability of the affected stream, and thus the estimated weights might not correctly reflect the reliability of each of the signals. Thus, a major concern here is to derive a reliability assessment model that can adjust the stream weights dynamically during recognition.

Neti et al. [59] and Glotin et al. [28] proposed a dynamic technique for reliability estimation based on the degree of voicing present in the audio stream averaged over the entire utterance such that $0 \leq \lambda_A = \textit{degree of voicing} \leq 1$ and $\lambda_V = 1 - \lambda_A$. Using dynamic weights demonstrated an improved performance over statically-weighted schemes in noisy environments. However, there are two main problems with this approach. The first problem, as shown by the results of this work, is that in clean acoustic environments, some of the late fusion techniques were outperformed by the early fusion. The second problem with this system arises when there are sudden noise bursts such as a sudden loud noise. In this case, this sudden noise burst will affect the overall voicing average and thus the estimated stream weights will not resemble the actual reliability of the modalities. Using the median instead of the mean might help, but still the performance would degrade drastically at extra noisy levels.

Statistical Approaches

The problem with the above-mentioned approaches is that they only consider the noise in the audio channel and hence do not allow the modeling of the possible variations in the visual stream reliability. This is due to the fact that estimating the noise in the visual signal is quite hard. Therefore, statistical indicators of classifier confidence on the stream data can be used as shown in [67]. These indicators capture the reliability of each stream at a local frame level and have the advantage of not depending on the properties of the underlying signal which means that they can capture the reliability of the visual classifier as well as the audio classifier in the same manner. Some algorithms have been proposed in literature to make use of this statistical approach of reliability modeling.

Adjoudani and Benoit [2] used a certainty factor to differentially weight each subsystem. The advantage of this certainty factor weighting scheme is that it not based on the level of acoustic noise within the signal, but rather on the dispersion of the N-best hypotheses

in each modality. This is justified by the fact that large differences in probabilities equate to greater certainty, close probabilities to less certainty. This dispersion value is based on the variance of the output classifier:

$$\sigma^2 = \frac{1}{N-1}\sum_{n=1}^{N}(R_n - \mu)^2, \tag{2.8}$$

where R_n is the n^{th} output of the classifier. The variances of the N-best hypotheses of every stream are calculated over all noise levels and a mapping is established between these values and the stream weights as follows:

$$\lambda_A = \frac{\sigma_A}{\sigma_A + \sigma_V}, \tag{2.9}$$

for the acoustic weight and a similar mapping for the visual weight.

Potamianos and Neti [65] also use a similar dispersion method that uses Gaussian Mixture Model (GMM) to classify speech classes. Their method uses an N-best dispersion method that is formulated as the difference between each pair of the n^{th}-best hypotheses, given by:

$$\frac{2}{N(N-1)}\sum_{n=1}^{N}\sum_{n'=n+1}^{N}(R_n - R_{n'}), \tag{2.10}$$

where $N \geq 2$ and R_n is equal to the n^{th}-best hypothesis. The choice of the value of N that results in the best weighting scheme was found to be 4 by both Adjoudani and Benoit [2] and Potamianos and Neti [65]. Another dispersion measure that was also used by Potamianos and Neti, called the N-best likelihood ratio average is calculated as the difference against the best hypothesis:

$$\frac{1}{N-1}\sum_{n=2}^{N}(R_1 - R_n), \tag{2.11}$$

where R_1 to R_N are sorted in descending order, and thus R_1 is the best hypothesis. Comparing the above mentioned approaches, the one using dispersion as a reliability measure proved to be the best, yielding a phoneme accuracy of 55.19%, followed by the one using the ratio average which achieved a 55.05% accuracy.

Another interesting approach to be discussed in this section is the one developed by Heckmann et al. [36] which uses a hybrid ANN/HMM AVSR system with the NNs providing the aposteriori probabilities for the HMM. The HMM in turn models the phonemes and words and is used for recognition. Heckman et al. develop a method for stream weighting that they call *Geometric Weighting*. First they calculate a value c that reflects an estimate of the SNR of the acoustic signal. Then, they use this value to estimate the stream weights. Detecting the most probable phoneme is found by a conditional probability that is augmented by the geometric weights. Their weighting scheme uses a similar idea as dispersion that exploits the distribution of the a posteriori probabilities at the output of the MLP, but based on the calculated entropy:

$$H = -\frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{N} P(H_{n,k}|x_{A,k}) log_2 P(H_{n,k}|x_{A,k}), \tag{2.12}$$

where N is the number of phonemes and K is the number of frames. The idea behind using entropy as a reliability measure is that high entropy corresponds to an even spread implying high ambiguity and low reliability. The mapping between c and H was established using an empirical analysis of the values (optimization process). Results on the word error rate show an improved performance using this technique until an SNR value of -6 dB (high noise level) where it starts to perform than worse the visual-only system. When comparing the entropy confidence measure to the voicing index and the dispersion methods, they showed that the entropy based approach gave the best results.

The last approach that we are going to discuss is of particular importance because it combines both feature and decision fusion in one system. This system, developed by Rogozan [72] first uses a HMM to produce a hypothesis based on a concatenated audio-visual feature vector (feature fusion). Then, another system (a HMM or a ANN) refines the result using the visual observations. The system then uses a confidence measure based on the dispersion of the N-best hypotheses, to fuse the results of the two subsystems (decision fusion).

In this thesis, we attempt to solve the above-mentioned problems by introducing a statistical reliability assessment model that also takes into consideration the previous behavior of the classifier with respect to changes in observation conditions, and we propose a method for exponent estimation based on these indicators.

Chapter 3

Architecture and Principal Contributions

With all the improvements over audio-only systems, developing an audio-visual speech recognition system introduces new challenges and difficulties that should be tackled. Figure 3 depicts the overall system architecture developed in this work to solve these problems. The system starts by acquiring the audio speech signal of a speaker through a microphone as well as the video frames of the speaker's face by means of a camera. The audio and visual streams are then ready for analysis at a signal level.

3.1 Acoustic Front End

The analysis of the audio speech signal has been extensively investigated in the speech recognition community and much work has been done in order to mitigate the speech classification errors. Therefore, audio speech recognition will not be improved within the scope of this work. Instead, we will extract audio features that have been proven in literature to be robust to noise. The predominant feature type used in speech recognition is the the Mel Frequency Cepstral Coefficients (MFCCs). These features are inspired by properties of human auditory system. The derivation of the MFCCs and the audio signal analysis are reviewed in Chapter 4.

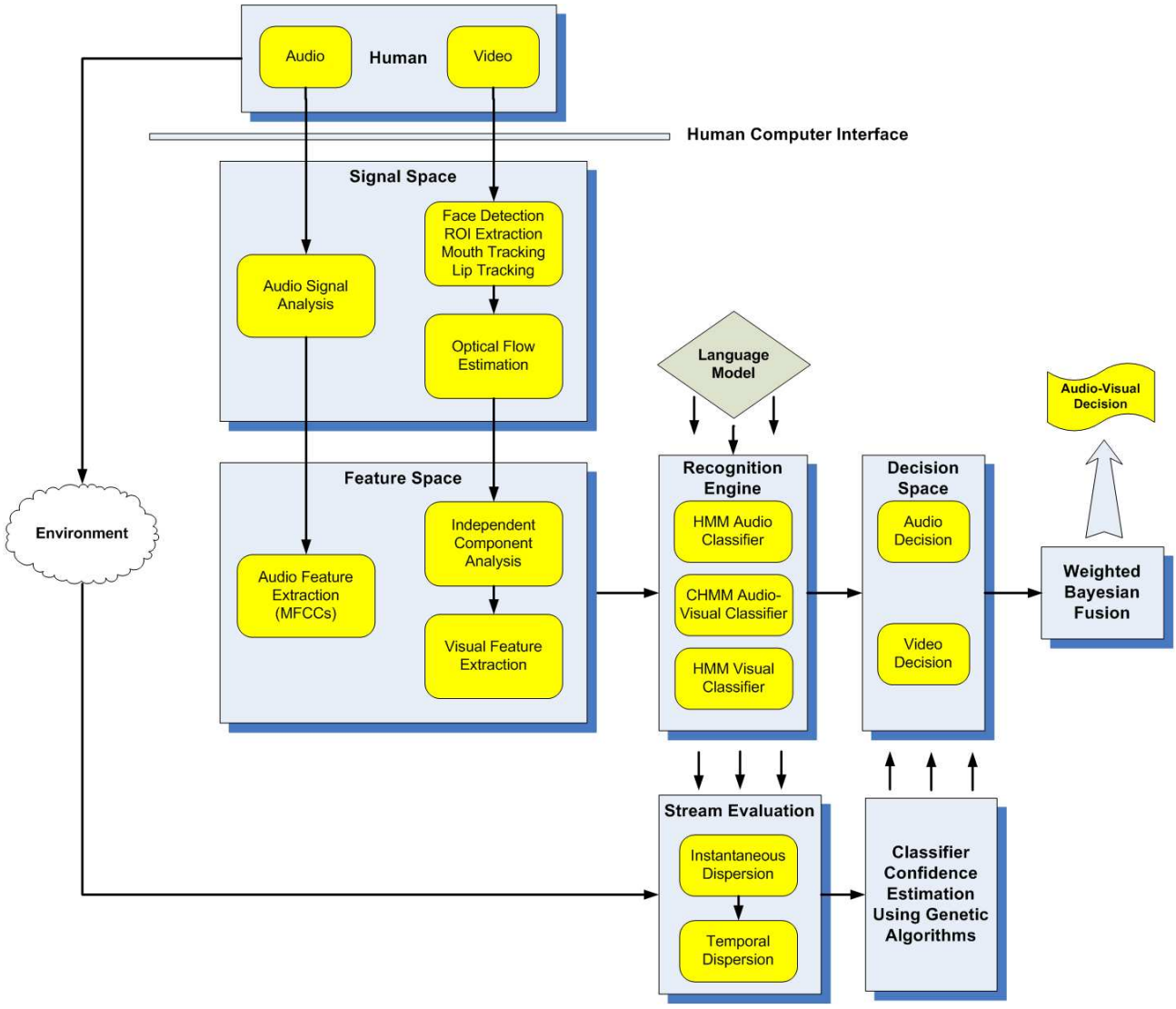

Figure 3.1: Overall system architecture.

3.2 Visual Front End

The first major issue in audio-visual speech recognition is the visual front-end design. Given an input video of a person speaking, the task of the visual front-end system is to extract visual features from the lips and to recognize these features as well-defined units of visual speech or visemes. As discussed in Chapter 2, the first step that should be performed for this purpose is face detection. Following this step, the system should extract the ROI which consists of the region of the face that contains most of the speech information (the mouth and the jaw). Next, the mouth of the subject is tracked in consecutive frames to extract relevant features, where a Kalman filter, for instance, can be used. In Chapter 2, we investigated some efficient approaches to the above-mentioned preprocessing steps. These approaches have been widely used in literature due to their robustness to various contexts and databases. Therefore, for the scope of this thesis, no contribution will be done for these steps.

A major contribution of this thesis is in the visual feature extraction module. In this thesis, we develop a novel approach for lips reading using two main steps. In the first step, the motion of the mouth is estimated using a robust *optical flow* technique. The goal of optical flow estimation is to compute an approximation to the motion field (or the 2D pixel velocities) from time-varying image intensity. In the second step, *Independent Component Analysis* (ICA), as proposed by Bell and Sejnowski [5], derives a set of basis flow fields for the frames generated by a word utterance. Consequently, every word is expressed as a linear combination of these basis flow fields, and the coefficients of this linear combination comprise the reduced feature set. ICA is a technique for finding a transformation in which transformed components are as statistically independent as possible. Typically associated with the *Blind Source Separation* (BSS) problem, ICA has also been used to separate *Electroencephalogram* (EEG) signals, *Functional Magnetic Resonance Imaging* (fMRI) signals, and for face detection.

Once features become available from the visual front end, we can proceed with automatic recognition of the spoken utterances by combining them to synchronously extracted acoustic features for audio-visual speech recognition. Visual speech modeling is required in this process, its two central aspects being the choice of speech classes that are assumed to generate the observed features, and the statistical modeling of this generation process.

Both issues are important, as they are also embedded into the design of audio-visual fusion. For visual speech modeling, an AVSR system could use word, viseme, or sub-viseme speech classes, which are based on articulatory features of the lips movement [75]. While the latter provides more information, it complicates audio-visual integration because the speech classes in the audio and visual modalities are no longer identical. The architecture developed in this thesis uses viseme classes.

3.3 Bimodal Sensor Fusion

The framework will involve an intelligent fusion mechanism. The fusion can take place at the feature-level, decision level or both. As described in Chapter 2, feature-level fusion algorithms train a single classifier on the concatenated vector of audio and visual features or any appropriate transformation of it. Decision-level fusion, on the other hand, uses the two outputs of the audio and visual classifiers to combine the two modalities. In the framework of this thesis, a decision-level Bayesian fusion scheme is deployed. For decision-level fusion, the two modalities or data streams should be weighted based on their reliability. The combination weights should be dynamic, in the sense that they should adapt to the varying levels of noise, voicing and other channel effects. To achieve this, certain reliability measures have to be defined. These measures have to depend on the varying levels of noise and channel effects. However, they cannot be trained for deterministic models and thus have to assume no prior knowledge about the channel. Moreover, efficient techniques should be developed to avoid the problem of having to estimate the channel characteristics including noise. This boils down to an optimization problem, where an optimal weighting scheme should be generated based on an optimal reliability measure.

Estimating stream reliability is a complex issue. For instance, in a vehicle environment, the wind, tires and engine all introduce noise to the microphone. Simply estimating the noise in the channel (or assuming prior knowledge of the noise) is not enough, because the noise makes the speakers increase their vocal effort. This change in vocal production makes the voice sound different, further reducing the accuracy of speech recognition. Consequently, although the reliability of the channel is degraded because of the noise and because the voice sounds different, it has been enhanced by the increased vocal effort. This

makes the problem of stream reliability a difficult problem and thus gives rise to the need for robust statistical methods.

A stream reliability assessment model is developed in this thesis, which is also considered one of the major contributions of this work. The approach works in two steps. In the first step, two reliability measures are developed to evaluate the performance of the individual streams based on observing statistically the respective classifier behavior. The developed reliability measures, which we call *instantaneous dispersion* and *temporal dispersion* are based on the dispersion of the a posteriori probabilities of the unimodal classifiers. They take into consideration both the current as well as the past performance of the classifier in order to create a behavioral profile of each modality. In the second step, an optimal mapping from the derived stream indicators to the stream weight measures is developed using genetic algorithms. This approach is superior to previous approaches because it is dynamic, easy to implement and considers an arbitrary number of streams.

Another issue with multimodal fusion is synchronization. This problem is coupled with the issue of the level of integration. Indeed, it is important to determine whether the fusion module should integrate at an early (phone level), intermediate (word level), or late (utterance level) stage in order not to lose synchronicity across modalities. In this work, the audio-visual recognition engine is implemented by means of a coupled HMM. The coupled HMM allows for asynchrony modeling between the audio and the visual streams by allowing the multimodal streams to interact. The concept of a CHMM is elaborated in more detail in Chapter 6. In addition to this audio-visual classifier, an audio HMM classifier and a visual HMM classifier are implemented on the audio and visual feature vectors respectively in order to compare the multimodal system with the individual unimodal systems.

The current popular approaches to fusion involve fuzzy based systems, Hidden Markov Models (HMMs), Bayesian networks, rule-based systems and evidential reasoning methods. Many variants and modified algorithms based on the above and a few other methods have been developed for different scenarios. This research investigates current approaches, and seeks to develop an improved fusion algorithm, which can efficiently perform its job in the target environment. The fusion algorithm is a weighted Bayesian fusion scheme that is designed with the following requirements:

1. Robust performance even under high noise levels, like that of an in-vehicle environ-

ment. In addition, the performance of the system should not degrade due to excessive noise in one of the modalities. Thus noise suppression is an important stage of the fusion process.

2. Temporal fusion of the different modalities. As has been mentioned in Section 1.2, the two modalities may not be synchronous in time. Thus the system has to be able to handle the delay between the two modalities, by means of different synchronization techniques.

3. Ability to provide output even if one modality is inefficient, faulty or even non-functional. The system should be able to provide an output based on the information present in the other modality and the previously identified speech sequence.

Chapter 4

Audio Front-End Design

Audio speech recognition has been extensively addressed in literature. Since the main goal of this work is to improve on speech recognition by making use of the visual modality, it is important to elaborate on the audio modality design and implementation before describing its integration with the visual front-end. For this purpose, this chapter addresses two main issues that are to be considered in the audio front-end design. The first issue is the audio speech modeling and the second issue is the audio feature extraction.

4.1 Audio Speech Modeling

4.1.1 The Concept of a Phoneme

For the acoustic stream, the basic units of speech are the *phones*. A phone is an acoustic realization of a *phoneme*, a theoretical unit for describing how speech conveys linguistic meaning. The acoustic realization of a phoneme depends on the speaker, the word context, and so forth. The variations in the pronunciation of the same phoneme are called *allophones*. In literature, people usually use the words phone and phoneme interchangeably. In this work, we are going to deal with speech recognition in English, so we will only use the phonemes developed for the English language. Usually there are about 10-15 vowels or vowel-like phones and 20-25 consonants. The most popular computer-based phonetic alphabet in American English is ARPABET which consists of 48 phones [14]. The publicly

available Carnegie Mellon University dictionary [1] can be used to to transcribe the English words into its phonetic transcription. This dictionary uses a subset of the ARPABET which consists of 39 phones. For instance, for the word "three", this transcription would give "TH-R-IY". Table 4.1 shows the phonemes of the English language.

4.2 Audio Features Extraction

The predominant feature type used in speech recognition is the Mel Frequency Cepstral Coefficients (MFCCs). Figure 4.1 show a block diagram of the MFCC computation. These features are derived after applying a Fourier transform based filterbank designed to give approximately equal resolution on a mel-scale. The mel-scale is inspired by properties of human auditory system. The emphasis is on a better sensitivity at lower frequencies to the expense of lower sensitivity to high frequencies.

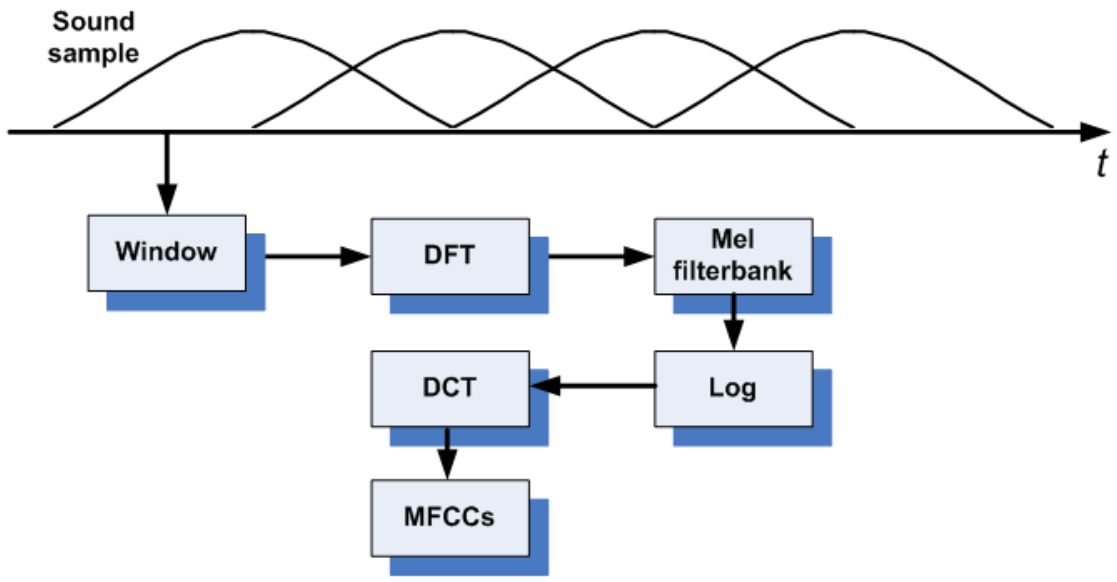

Figure 4.1: Block diagram of MFCC computation.

The filters in the filterbank are triangular. The filtering results in multiplying each discrete Fourier (DT) magnitude with the filter gain. Afterwards all values per filter are

Table 4.1: Phonemes in the English language.

Phone Classes	Phonemes	Example Words	Phone Classes	Phonemes	Example Words
Stops	/B/	bee	Front	/AE/	bat
	/D/	day		/EH/	bet
	/G/	gay		/IH/	bit
	/K/	key		/IY/	beet
	/P/	pea	Mid	/AA/	bott
	/T/	tea		/AH/	but
Affricates	/CH/	choke		/AO/	bought
	/JH/	joke		/ER/	bird
Fricatives	/DH/	then	Back	/UH/	book
	/F/	fin		/UW/	boot
	/S/	sea	Diphthongs	/AW/	bout
	/SH/	she		/AY/	bite
	/TH/	thin		/EY/	bait
	/V/	van		/OW/	boat
	/Z/	zone		/OY/	boy
	/ZH/	azure	Silence	/H#/	
Nasals	/M/	mom			
	/N/	noon			
	/NG/	sing			
Glides	/L/	lay			
	/R/	ray			
	/W/	way			
	/Y/	yacht			
Whisper	/HH/	hay			

added (accumulated). Typically before the accumulation, power values (squared magnitudes) can be computed. The number of triangular filters depend on the speech frequency band [82]. After applying the filter bank, we get filterbank coefficients. To get the log-filterbank coefficients, a logarithm is taken of the filterbank coefficients. This step also mimics human logarithmic perception to the increase in speech volume. Finally, to derive the MFCCs the log-filterbank coefficients are transformed into the cepstral domain, using discrete cosine transform. Usually, the low-order 12 or 15 MFCCs are taken as final speech features, as they describe in a compressed form the spectral envelope of the short-term speech spectrum. MFCCs are shown to perform better in noisy conditions than spectral representations (e.g. filterbank coefficients) [20]. However, MFCCs also get contaminated by additive and convolutional noise. If the convolutional noise component is constant or slowly varying, it appears as a bias in the time evolution of the MFCCs values. Cepstral mean normalization can be used with MFCCs to remove constant convolutional noise. Such noise may result from the transfer function of the microphone or the transmission channel through which speech is communicated. Applying cepstral mean normalization results in features robust to convolutional noise that does not change rapidly over time. In that way features can become much less sensitive to different microphone equipment. The steps to construct MFCC features are as follows:

1. Pre-Emphasis:
   The following FIR pre-emphasis filter is applied to the input waveform:

$$y[n] = x[n] - \alpha x[n-1]. \tag{4.1}$$

   α is provided by the user or set to the default value. If $\alpha = 0$, then this step is skipped. In addition, the appropriate sample of the input is stored as a history value for use during the next round of processing.

2. Windowing:
   The frame is multiplied by the following Hamming window:

$$w[n] = 0.54 - 0.46\cos(\frac{2\pi N}{N-1}), \tag{4.2}$$

   where N is the length of the frame.

3. Power Spectrum:
   The power spectrum of the frame is computed by performing a DFT of length specified by the user, and then computing its magnitude squared:

$$S[k] = (real(X[k]))^2 + (imag(X[k]))^2. \tag{4.3}$$

4. Mel Spectrum:
   The mel spectrum of the power spectrum is computed by multiplying the power spectrum of each of the triangular mel weighting filters and integrating the result:

$$\tilde{S}[l] = \sum_{k=0}^{N/2} S[k] M_l[k] \qquad l = 0, 1, ..., L-1, \tag{4.4}$$

   where N is the length of the DFT, and L is the total number of triangular mel weighting filters.

5. Mel Cepstrum:
   A DCT is applied to the natural logarithm of the mel spectrum to obtain the mel cepstrum:

$$c[n] = \sum_{i=0}^{L-1} \ln(\tilde{S}[i]) \cos(\frac{\pi n}{2L}(2i+1)) \qquad n = 0, 1, ..., C-1, \tag{4.5}$$

   where C is the number of cepstral coefficients.

Delta MFCC

Also, delta MFCC coefficients can be calculated using the following equation:

$$\Delta c[n] = c[n+1] - c[n]. \tag{4.6}$$

Delta delta MFCC

Moreover, to obtain delta-delta coefficients, the following equation can be used:

$$\Delta\Delta c[n] = \Delta c[n+1] - \Delta c[n]. \tag{4.7}$$

The database used in this work for both the unimodal and bimodal recognition is the Tulips1 database. For this database, the mel-cepstrum coefficients are already extracted from the speech signal. The first 12 cepstral coefficients, 12 delta-cepstral coefficients, 1 logpower and 1 delta log-power are used. Audio-only recognition is done using HMMs.

Chapter 5

Visual Front-End Design

As discussed in Chapter 3, the first major issue to address in audio-visual speech recognition is the visual feature design and representation scheme. This involves the extraction of visual features from the lips and mouth movement and recognizing these features as well-defined units of visual speech. This requires robust face detection, as well as location estimation and tracking of the speaker's mouth or lips, followed by visual feature extraction. Consequently, the problem is reduced to a pattern recognition formulation where it is important to reduce the dimensionality of the feature vector into a representative feature set.

5.1 Problem Definition

Let $\mathbf{x}_i, i = 1, 2, ..., P$ be the set of patterns corresponding to feature vectors describing the mouth shape. The feature vector $\mathbf{x}_i$ can be related to a low level representation of the mouth image like the gray levels from a rectangular image region containing the mouth. It can also comprise geometric parameters like the mouth height, width, perimeter, etc., or the coefficients of a linear transformation of the mouth image. All the feature vectors from the set have the same number of components M. We also denote the pattern classes by $C_j, j = 1, 2, ..., Q$, where Q is the total number of classes. We apply this pattern recognition formulation to our visual front-end design in the following way:

- Each unknown pattern represents the optical flow derived between two consecutive

frames of the speaker's mouth region at a particular time instant.

- Each class label represents one viseme.

As mentioned before, it is important to reduce the dimensionality of the feature vector into a representative feature set. Several unsupervised statistical methods, such as principal component analysis (PCA) [12], have been proposed in literature to achieve this goal. These methods, while deriving reduced feature sets, are only sensitive to second-order statistics. The problem is that second-order statistics capture the amplitude spectrum of images but not their phase spectrum. The high-order statistics, on the other hand, capture the phase spectrum. For natural images, such as those of the lips, the phase spectrum, not the power spectrum, contains structural information that drives human perception. Consequently, there is a need for a better solution that is well suited for representing lips movement.

In this chapter, a novel approach for lips reading using two main steps is established. In the first step, the motion of the mouth is estimated using a robust optical flow technique. The optical flow fields are derived between the consecutive ROI images of a speaker's image sequence. These optical flow estimates provide valuable information about the motion of the mouth from time-varying image intensity. In the second step, independent component analysis, as proposed by Bell and Sejnowski [5] is used to derive a set of basis flow fields for the frames generated by a word utterance. Each speech class is then expressed as a linear combination of these basis flow fields, and the coefficients of this linear combination comprise the reduced feature set. ICA is a technique for finding a transformation in which the transformed components are as statistically independent from each other as possible. Typically associated with the blind source separation (BSS) problem, ICA has also been used to separate EEG signals, fMRI signals, and for face recognition.

In Chapter 2, we mentioned that we are going to develop a feature extraction model that is appearance-based, and we discussed some of the main approaches that have been proposed in literature for this purpose. In particular, we addressed some drawbacks of the used approaches, and came up with a few observations:

- Lips movement is complex in its nature, as it involves large image velocities, self occlusion (due to the appearance and disappearance of the teeth) and a lot of non-rigidity. As a result, such kind of motion is characterized by higher-order dependencies.

- PCA-based coefficients are insufficient on their own because they only capture second-order statistics and thus may not be appropriate for motion of the lips, where higher order statistics are required.
- For complex objects that are characterized by higher-order dependencies, a representation using independent components provides more knowledge.

For these reasons, ICA may be more appropriate for analyzing the motion of the mouth. Figure 5.1 shows an overview of the visual front-end design.

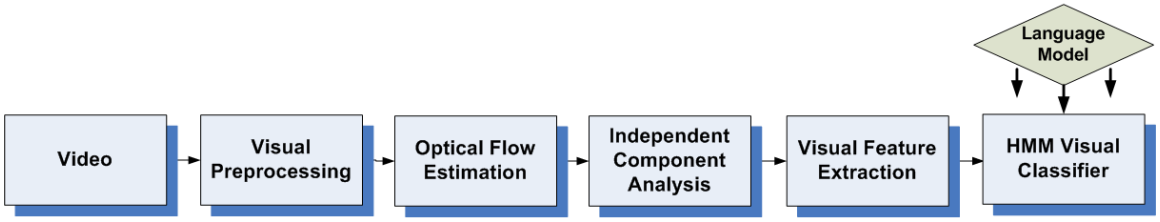

Figure 5.1: Overview of the visual front-end system.

5.2 Visual Speech Modeling

Prior to the feature extraction stage, visual speech modeling is required, its two central aspects being the choice of speech classes that are assumed to generate the observed features, and the statistical modeling of this generation process. Both issues are important, as they are also embedded into the design of audio-visual fusion.

5.2.1 The Concept of Viseme

As in the acoustic space, one can define the basic unit of speech in the visual space, the *viseme*. The concept of viseme is usually defined in accordance with the mouth shape and mouth movements. An example where the concept of viseme is related to the mouth dynamics is the viseme /AO/ which represents the movement of the mouth from a position close to /A/ to a position close to /O/. In such a case, to represent a viseme, one

should develop a method for representing the video sequence, further complicating the video processing stage. Fortunately, most of the visemes can be represented by stationary mouth images. The benefit of using such a representation is that it can be mapped directly to the acoustic speech units, which makes the integration process pretty easy. However, in some scenarios, the use of visemes may not be representative of the speech content and therefore there is a need for using sub-visemic speech classes, which are based on articulatory features of the lips movement [75]. While this technique provides more information, it complicates audio-visual integration because the speech classes in the audio and visual modalities are no longer identical. Therefore, for the scope of this thesis, we will use visemes as visual speech units.

To be able to design the visual front-end, it is desirable to define for each phoneme its corresponding viseme. This enables us to integrate the visual speech recognition system into existing acoustic-only systems. Unfortunately, speech production involves invisible articulatory organs, which renders the mapping of phonemes to visemes into many-to-one. Consequently, there are phonemes that cannot be distinguished in the visual domain. For example, the phonemes /P/, /B/, and /M/ are all produced with a closed mouth and cannot be distinguished visually, so they will be represented by the same viseme. It is important also to consider the effect of the dual of the allophone, where the same viseme can be realized differently in the visual domain due to the speaker variability and the context. Table 5.1 shows the most commonly used visemes for the English consonants in literature [13]. Unlike the phonemes, there is no viseme set that is commonly used by all researchers.

5.3 The Basic Structure of the HMM for Visual Speech Recognition

In order to develop a visual model that can be easily integrated with existing audio speech recognition systems, we will use HMMs for recognition. At this stage, it is crucial to define the basic structure of the HMM developed for viseme-based visual speech recognition, so that we can understand the general framework in which our proposed visual features will be integrated. When we introduced visemes in the previous section, we only showed which

Table 5.1: The most commonly used visemes for the English consonants [13].

Viseme Group Index	Corresponding Consonants
1	/F/, /V/
2	/TH/, /DH/
3	/S/, /Z/
4	/SH/, /ZH/
5	/P/, /B/, /M/
6	/W/
7	/R/
8	/G/, /K/, /N/, /T/, /D/, /Y/
9	/L/

visemes can be present in the pronunciation of a certain word and not their duration. Let $T_i, i = 1, 2, ..., S$ denote the duration of the i^{th} viseme in a word model of S visemes. Let T be the duration of the video sequence that results from the pronunciation of this word. The duration of a viseme or a word is defined here to be the number of optical flow fields per viseme or word, which is equal to the number of frames minus one. Representation using optical flow will be discussed later.

For the purpose of aligning the video sequence of duration T with the visemic model of S visemes, we create a temporal Viterbi lattice [82] containing as many states as the optical flow fields in the video sequence, that is, T. An example of this Viterbi lattice is shown in Figure 5.2 for the representation of the word "four". A similar Viterbi lattice is derived for every visemic model $w_d, d = 1, 2, ..., D$, where D is the total number of visemic models. Each node in this lattice generates an observation that belongs to a certain class at each time instant. Let $l_k = 1, 2, ..., Q$ be the class label that the observation o_k generated at time instant k belongs to. Let us denote the emission probability of that observation by $b_{l_k}(o_k)$. Each solid line between any two nodes in the lattice represents a transition probability between two states. Denote by $a_{l_k,l_{k+1}}$ the transition probability from the node corresponding to the class l_k at time instant k to the node corresponding to the class l_{k+1}

at time instant $k+1$. The class labels l_k and l_{k+1} may or may not be different.

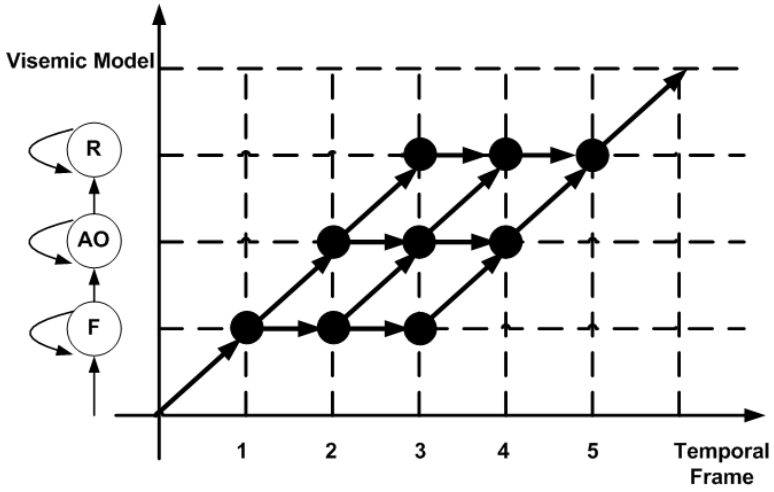

Figure 5.2: A Viterbi lattice for the word "four" having duration of 5.

Having a video sequence of T frames for a word and a Viterbi lattice for each visemic word model $w_d, d = 1, 2, ..., D$, we can compute the probability that the visemic word model w_d is realized, following a path l in the Viterbi lattice as:

$$p_{d,l} = \prod_{k=1}^{T} b_{l_k}(o_k) \prod_{k=1}^{T-1} a_{l_k, l_{k+1}}. \tag{5.1}$$

The probability that the visemic word model w_d is realized can be computed by:

$$p_d = \max_{l=1}^{L} p_l, \tag{5.2}$$

where L is the number of all possible paths in the lattice. Among the words that can be realized following any possible path in any of the D Viterbi lattices, the word described by the model whose probability $p_d, d = 1, 2, ..., D$, is maximum is finally recognized. In the visual speech recognition approach discussed in this chapter, the emission probability $b_{l_k}(o_k)$ is given by the corresponding HMM, HMM_k. Here we assume equal transition probabilities $a_{l_k, l_{k+1}}$ between any two states. Therefore, it is sufficient to take into account only the probabilities $b_{l_k}(o_k), k = 1, 2, ..., T$, in the computation of the path probabilities

$p_{d,l}$ which gives the simplified equation:

$$p_{d,l} = \prod_{k=1}^{T} b_{l_k}(o_k). \tag{5.3}$$

5.4 Visual Preprocessing

For the scope of this work, we will use the Tulips1 database. For the visual part of this database, the mouth regions are already extracted, which means that we do not have to perform face detection and ROI extraction. However, a number of image preprocessing steps need to be performed for these regions as described in [33] and shown in Figure 5.3.

First, the contour of the outer lips is tracked using point distribution models, a data-driven technique based on analysis of the grey-level statistics around lip contours. The mouth images are then normalized for translation and rotation. This is accomplished by first padding the image on all sides with 25 rows or columns of zeros, and modulating the images in the spatial frequency domain. The images are then symmetrized with respect to the vertical axis going through the center of the lips. This makes the final representation more robust to horizontal changes in illumination. The images are then cropped to a size 36×50 and their intensity is normalized using logistic gain control. This in turn produces a grayscale region of interest that is robust against varying lighting conditions.

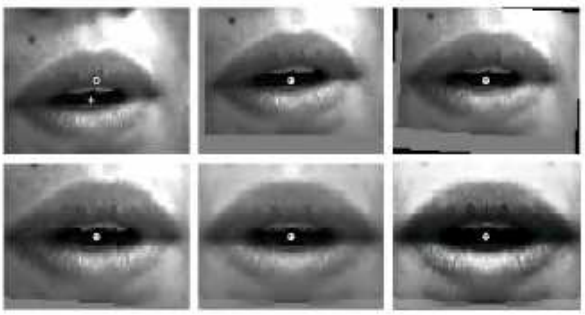

Figure 5.3: Visual preprocessing steps.

5.5 Optical Flow of Mouth Motion

Once the images are preprocessed, optical flow is derived between consecutive images of mouth regions, in order to understand the motion of the mouth from time-varying image intensity. Optical flow in computer vision is a concept used to measure the motion of objects within a visual representation. In camera-oriented coordinates each point on a 3D surface moves along a 3D path $\mathbf{X}(t)$. When projected onto the image plane each point produces a 2D path $\mathbf{x}(t) \equiv (x(t), y(t))^T$, the instantaneous direction of which is the velocity $d\mathbf{x}(t)/dt$. The 2D brightness velocities for all visible surface points is often referred to as the *2D motion field*. The goal of *optical flow* estimation is to compute an approximation to the motion field from time-varying image intensity. Typically the optical flow between two consecutive frames is represented as vectors originating or terminating at pixels in a digital image sequence. Figure 5.4 shows the flow field estimated for two consecutive images of a person speaking.

5.5.1 Optical Flow Estimation

Optical flow estimation can be classified into three types [4]: intensity-based differential methods, frequency-based filtering methods and correlation-based methods. The method we use for optical flow estimation is a differential technique based on [7]. This is a regularization technique, where a robust statistics method is employed to alter the traditional objective function to be minimized. This makes it feasible to reject outliers and obtain a more accurate optical flow field.

Let $I(x, y, t)$ be the image brightness, or a filtered version of the image brightness, at a point (x, y) at time t. The data conservation constraint can be expressed in terms of the standard *brightness constancy assumption* as follows:

$$I(x, y, t) = I(x + u\delta t, y + v\delta t, t + \delta t), \tag{5.4}$$

where (u, v) is the horizontal and vertical image velocity at a point and δt is small. This is equivalent to stating that the image value at time t, at a point (x, y), is the same as the value in a later image at a location offset by the optical flow.

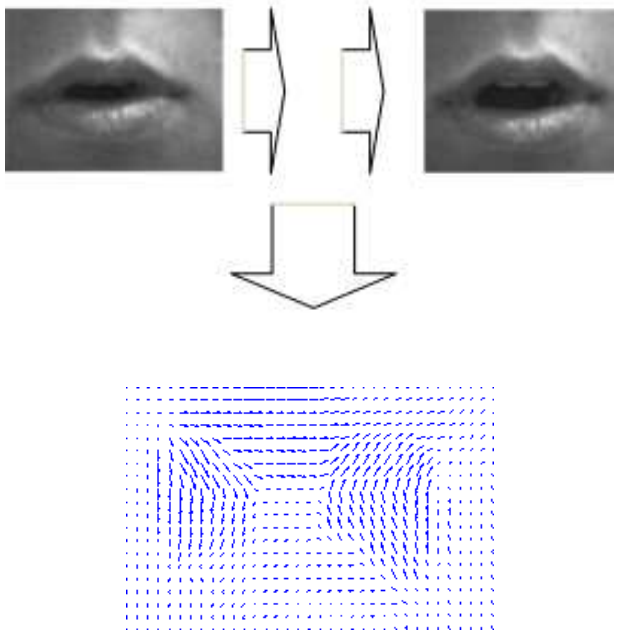

Figure 5.4: Optical flow of two consecutive images.

A Taylor series expansion of the right side of Equation 5.4 and discarding the terms higher than first-order leads to the following equation:

$$I_x u + I_y v + I_t = 0, \tag{5.5}$$

where $I_x = \frac{\partial I}{\partial x}$, $I_y = \frac{\partial I}{\partial y}$, and $I_t = \frac{\partial I}{\partial t}$. For convenience, Equation 5.5 is rewritten as:

$$(\nabla I)^T \mathbf{u} + I_t = 0, \tag{5.6}$$

where ∇I denotes the local brightness gradient vector, and $\mathbf{u} = [u, v]^T$ denotes the flow vector. Equation 5.6 is known as the brightness constancy constraint equation and is not sufficient to recover $\mathbf{u}$ since it is one equation with two unknowns u and v. This equation constrains u and v to lie on a line as shown in Figure 5.5.1.

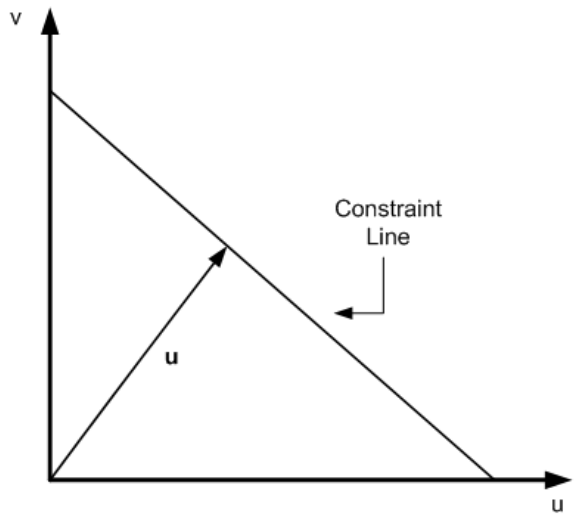

Figure 5.5: Brightness constancy constraint equation.

In order to provide other constraints, many methods have been used in literature. One common approach to constrain $\mathbf{u}$ is to use gradient constraints from nearby pixels assuming they share the same 2D velocity. The central idea here is to derive the constraints in a neighboring region and minimize them using least squares. Least squares is used here because of the pooling of constraints over some spatial neighborhood R. While R should

be large enough to constrain the solution, the larger the region of integration the more likely it is to contain multiple motions with competing constraints. The optical flow is consequently derived by minimizing the following quantity:

$$E(u,v) = \sum_{(x,y)\in R} [I_x(x,y,t)u + I_y(x,y,t)v + I_t(x,y,t)]^2. \tag{5.7}$$

Using the notation of Equation 5.6, Equation 5.7 is rewritten as:

$$E(\mathbf{u}) = \sum_R [(\nabla I)^T \mathbf{u} + I_t]^2. \tag{5.8}$$

Other methods assume global smoothness through regularization approaches and are represented by the classic method of Horn and Shunk [38]. The method works by minimizing the following function over all the image:

$$E(\mathbf{u}) = \sum_R [((\nabla I)^T \mathbf{u} + I_t]^2 + \lambda(u_x^2 + u_y^2 + v_x^2 + v_y^2)), \tag{5.9}$$

where λ is a regularization parameter, $u_x = \frac{\partial u}{\partial x}$, $u_y = \frac{\partial u}{\partial y}$, $v_x = \frac{\partial v}{\partial x}$, and $v_y = \frac{\partial v}{\partial y}$.

However, the smoothness assumption is violated if two or more motions are present in a neighborhood. In this case, one set of constraints will be consistent with one of the motions while the other set of constraints will be consistent with the other motion. If we are considering one of the motions, the constraints for the other motion will appear as errors referred to as *outliers*. Least-squares estimation do not produce good results in the presence of outliers. For this purpose, Black and Anandan [7] propose a robustification technique to deal with the sensitivity of the least-squares estimator to measurement outliers. The objective function for their regularization approach, with a gradient-based data term, becomes:

$$E(\mathbf{u}) = \sum_R \rho((\nabla I)^T \mathbf{u} + I_t, \sigma) + \lambda \sum_R [\rho(u_x, \sigma) + \rho(v_x, \sigma)], \tag{5.10}$$

where

$$\rho(x,\sigma) = \log(1 + \frac{1}{2}(\frac{x}{\sigma})^2) \tag{5.11}$$

is the *Lorentzian* ρ-function, and σ is a scale parameter.

5.5.2 Representation in Terms of Basis Flow Fields

Since the optical flow is computed without extracting the speakers lip contours and location, robust visual features can be obtained for lip movements. This is actually the prime advantage of using optical flow visual features. The optical flow field, $\mathbf{u}(\mathbf{x};\mathbf{c})$, over positions $\mathbf{x} = (x, y)$ can be written as a weighted sum of *basis flow fields*:

$$\mathbf{u}(\mathbf{x};\mathbf{c}) = \sum_{j=1}^{n} c_j \mathbf{b}_j(\mathbf{x}), \tag{5.12}$$

where $\mathbf{b}j(\mathbf{x})_{j=1,...,n}$ is the basis set and $\mathbf{c} = (c_1, ..., c_n)$ is the vector containing the scalar coefficients. Figure 5.6 shows how an optical flow field is represented by a linear combination of a set of basis flow fields. A translational model, such as that we're dealing with here, requires two basis flow fields, encoding horizontal and vertical translation. Consequently, determining the optical flow for consecutive frames becomes a problem of estimating the coefficients $\mathbf{c}$ of the basis flow fields. In [27], the principal components of a training set of flow fields were used as the basis. In this work, we extract the independent components of a training set and use them as the basis in order to derive a linear motion model. This motion model will then be used to recognize linguistic events.

5.6 Independent Component Analysis

Having obtained the optical flow fields of consecutive mouth regions, the next step is to derive coefficient values of a linear transformation of the optical flow fields. Independent component analysis is used for this purpose due to the fact that it is superior to other transformation techniques when it comes to representing natural images such as that if the mouth. Independent component analysis of multivariate data aims at finding a transformation in which the transformed components are as statistically independent from each other as possible. The concept of ICA is usually coupled with the concept of Blind Source Separation (BSS), in which it is desired to separate multiple mixed signals based on the fact that the sources of these signals are statistically independent (also known as the cocktail party problem).

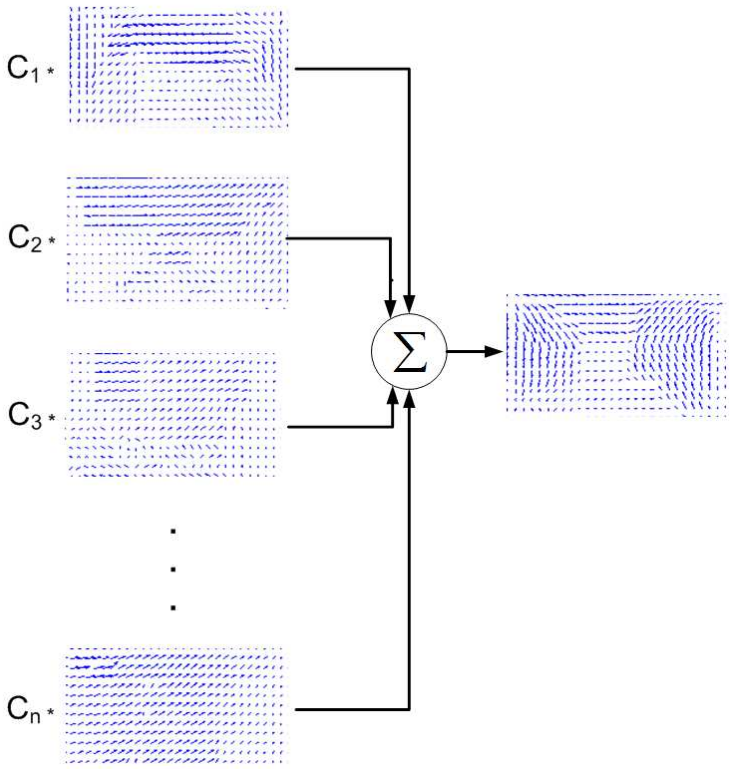

Figure 5.6: Representation of optical flow by basis flow fields.

5.6.1 ICA Derivation

A number of algorithms have been proposed in literature for performing ICA [5], [19], [16]. In this work we will use ICA as proposed by Bell and Sejnowski [5] to derive our motion model. This approach employs the principle of optimal information flow in neural networks with sigmoidal transfer functions, by minimizing output joint mutual information (infomax principle) as described below.

Let $\mathbf{X}$ be an n-dimensional random vector representing a distribution of inputs in a certain context. Let W be an $n \times n$ invertible matrix. Let $\mathbf{U} = W\mathbf{X}$ be an n-dimensional vector of linear combinations of inputs. $\mathbf{Y} = f(\mathbf{U})$ is an n-dimensional random variable, representing the outputs of n neurons. Each component of $f = (f_1, ..., f_n)$ is typically the logistic function:

$$f_i(u) = \frac{1}{1 + e^{-u}}. \tag{5.13}$$

The goal in Bell and Sejnowski's algorithm is to maximize the mutual information between the environment $\mathbf{X}$ and the output $\mathbf{Y}$ of the neural network. This is achieved by performing gradient ascent on the entropy of the output with respect to the weight matrix W. The gradient update rule for W is shown in equation 5.14:

$$\Delta W \propto \nabla_W H(\mathbf{Y}) = (W^T)^{-1} + E(\mathbf{Y}'\mathbf{X}^T), \tag{5.14}$$

where $H(\mathbf{Y})$ is the output joint mutual information, $\mathbf{Y}'_i = f''_i(\mathbf{U}_i)/f'_i(\mathbf{U}_i)$ is the ratio between the second derivative and the first derivative of f, E stands for expected value, and $\nabla_W H(\mathbf{Y})$ is the gradient of the entropy in matrix form. In the case of the sigmoidal transfer function in Equation 5.13, we have $\mathbf{Y}'_i = (1 - 2\mathbf{Y_i})$, and therefore Equation 5.14 reduces to:

$$\Delta W \propto \nabla_W H(\mathbf{Y}) = (W^T)^{-1} + (1 - 2\mathbf{Y_i})\mathbf{X}^T. \tag{5.15}$$

When the cumulative distribution function (CDF) of the independent components is aligned with the transfer function f (up to a scaling and rotation), maximizing output joint entropy becomes equivalent to minimizing the mutual information between the individual outputs. This in turn causes the individual outputs to be more statistically independent. The algorithm is speeded up by including a "sphering" step prior to learning [6]. The row

means of $\mathbf{X}$ are subtracted, and then $\mathbf{X}$ is passed through the whitening matrix W_z, which is twice the inverse square root of the covariance matrix:

$$W_z = 2 * (Cov(\mathbf{X}))^{(-1/2)}. \tag{5.16}$$

This removes the first and the second-order statistics of the data; both the mean and covariances are set to zero and the variances are equalized. When the inputs to ICA are the "sphered" data, the full transform matrix W_I is the product of the sphering matrix and the matrix learned by ICA:

$$W_I = WW_z. \tag{5.17}$$

MacKay [49] and Pearlmutter [62] showed that the ICA algorithm converges to the Maximum Likelihood estimate of W^{-1} for the following model of the data:

$$\mathbf{X} = W^{-1}\mathbf{S} \tag{5.18}$$

where $\mathbf{S} = (\mathbf{S}_1, ..., \mathbf{S}_n)'$ is a vector of independent random variables, called the sources, with cumulative distributions equal to f_i. This means that the W^{-1}, the inverse of the weight matrix in Bell and Sejnowski's algorithm, can be interpreted as the source mixing matrix and the $\mathbf{U} = W\mathbf{X}$ variables can be interpreted as the maximum-likelihood estimates of the sources that generated the data.

5.6.2 ICA Compared to PCA

As stated before, the use of ICA is proposed in this work due to the fact that ICA is superior to PCA and can thus can yield better performance than the traditional approaches based on PCA. In fact, PCA is a special case of ICA which uses Gaussian source models. For PCA, the mixing matrix W^{-1} is unidentifiable, which means that there is an infinite number of equally good ML solutions for W^{-1}. In fact, for PCA, the rows of W are the eigenvectors of the covariance matrix of the data. Therefore, PCA deals with second-order statistics and only captures the amplitude spectrum of the mouth motion not the phase spectrum. However, for natural images such as that of the mouth, structural information is contained in the phase spectrum, and the amplitude spectrum is not enough. The high-order statistics, obtained through ICA, capture the phase spectrum [6]. Since PCA is only

sensitive to the power spectrum of images, it may not be the best way for representing natural images.

Another drawback of PCA is that it assumes Gaussian sources, which may be inadequate when the true sources are non-Gaussian. In fact, it has been empirically observed that many natural images, including speech, natural images, and EEG are better described as linear combinations of sources with long tailed distributions [5] instead of Gaussian distributions. These sources are called *high-kurtosis*, *sparse*, or *super-Gaussian*. When the sources are modeled by high-kurtosis models, ICA is better than PCA for the following reasons: 1) It provides a better probabilistic model of the data. 2) It uniquely identifies the mixing matrix W^{-1}. 3) It finds a not-necessarily orthogonal basis which may represent the data better than PCA in the presence of noise. 4) It is sensitive to high-order statistics, not just the covariance matrix.

5.7 Data Collection and Processing

5.7.1 The Tulips1 Database

Our goal in this work is two-fold: to *understand* the motion of the lips, and to *recognize* the motion of the lips. In order to achieve the first goal, we need to derive the basis flow fields from a training set. The training set should be representative in the sense that it should include people from different ages, genders and appearances (beard or no beard for example), uttering a defined set of words at approximately equal rates. For this purpose we use the *Tulips1* database compiled by Movellan [53]. This is a small, publicly available database of 12 subjects (9 males and 3 females), pronouncing the first four digits in English two times in repetition. The audio part is sampled at 11127 Hz with 8 bits per sample. The video part consists of 934 gray scale lip images of size 100×75, sampled at the rate of 30 frames per second. Subjects are undergraduate students from the Cognitive Science Program at the University of California, San Diego. Although the number of words is small, this database is challenging due to the differences in illumination conditions, ethnicity, and gender of the subjects.

For the purpose of our work, the visual speech recognizer was tested using the leave-

one-out testing strategy for the 12 subjects in the Tulips1 database. This implies training the visual speech recognizer 12 times, each time using only 11 subjects for training and leaving the 12th out for testing.

5.7.2 Processing Using The Proposed Approach

For the purpose of training our system, we collect the video sequences of the 12 subjects in the database. Depending on the approach used to model the spoken words in the visual domain, visual speech recognition systems can be based on either a word-oriented model, a viseme-oriented model, or a sub-viseme-oriented model. In this work, we develop a viseme-oriented model, where each of the visemes is treated as a separate unit of visual speech and is thus processed independently. Consequently, we obtain a training set that is made up of the mouth regions for consecutive frames. Figure 5.7 shows an overall diagram of our approach.

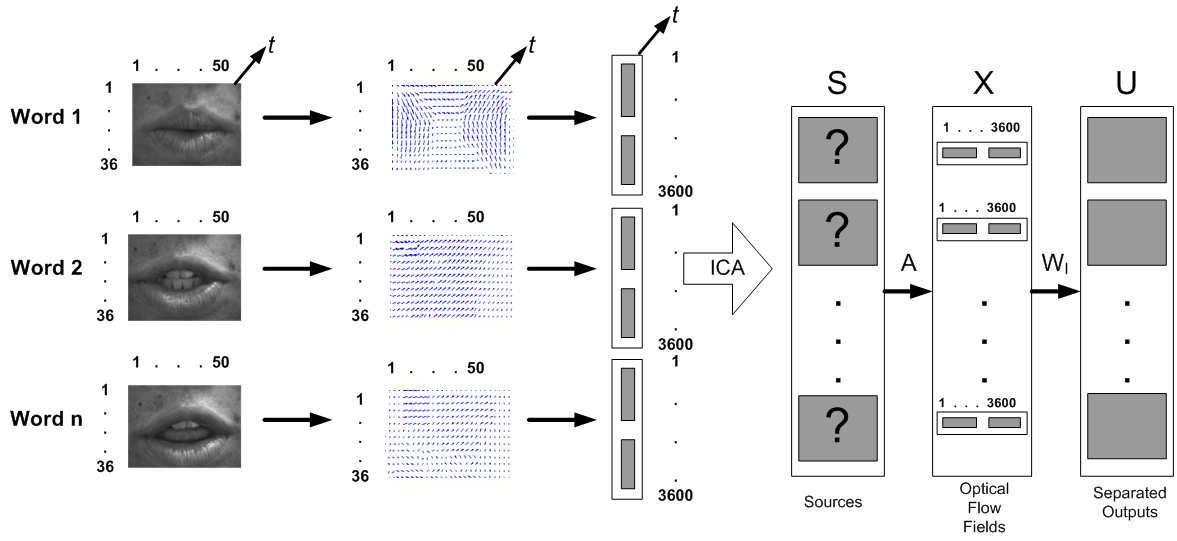

Figure 5.7: Processing using proposed approach.

Given the sequence of mouth regions, the images are normalized to 36×50 (see Section 5.4) in size and fed into the visual feature extraction module. In the first place, the input mouth region frames are mapped into optical flow fields derived between consecutive

frames. Since training is done offline, we can afford to use a computationally expensive optical flow technique. The method used for optical flow estimation is based on [7] as discussed in Section 5.5.1.

Next, we use the derived optical flow fields as our training ensemble. Each word consists of a variable number of consecutive frames and thus a variable number of optical flow fields. For 36×50 images, each image has $36 \times 50 = 1800$ pixels, and each flow field contains $2 \times 1800 = 3600$ quantities (i.e. the horizontal and vertical elements of the flow at each pixel). We place these 3600 values into a vector by scanning the horizontal components of the flow followed by the vertical components. Since we have 11 subjects (training set) and each subject utters 4 words, this gives us $11 \times 4 = 44$ vectors that become the rows of a 44×3600 matrix X. Each row of X represents a word (from "one" to "four") per person. We subtract the mean from each row so that each word has zero mean.

After deriving X, ICA is used to find a matrix W such that the rows of $U = WX$ are as statistically independent as possible. From W we can find W_I as shown in Section 5.6.1, and the mixing matrix $A \equiv W_I^{-1}$. The source optical flow fields estimated by U are then used as the basis flow fields for representing the motion of the mouth. The coordinates for these basis flow fields are contained in the rows of the mixing matrix A. Consequently, A has 44 rows, and each row of A is the feature vector for each of the corresponding words in the training sequence. The dimensionality of the input is equal to the number of ICs found by the ICA algorithm (44 in this case). Figure 5.8 shows how ICA is performed on the optical flow fields.

Prior to performing ICA, we perform PCA on the matrix X and use the first 12 principal components as the input to our ICA module. The first 12 PCs account for over 95% of the variance in the optical flow fields. PCA captures the second order statistics (i.e. the power spectrum), but not the higher order statistics (i.e. the phase spectrum). This does not mean that the higher order relationships are lost by performing PCA. These relationships still exist in the data but are not separated until we perform ICA. ICA is then performed on the first 12 principal components and this results in a matrix of independent source optical flow fields in the rows of U.

The following steps summarize the approach followed:

- Let P be the matrix containing the first 12 principal components in its rows.

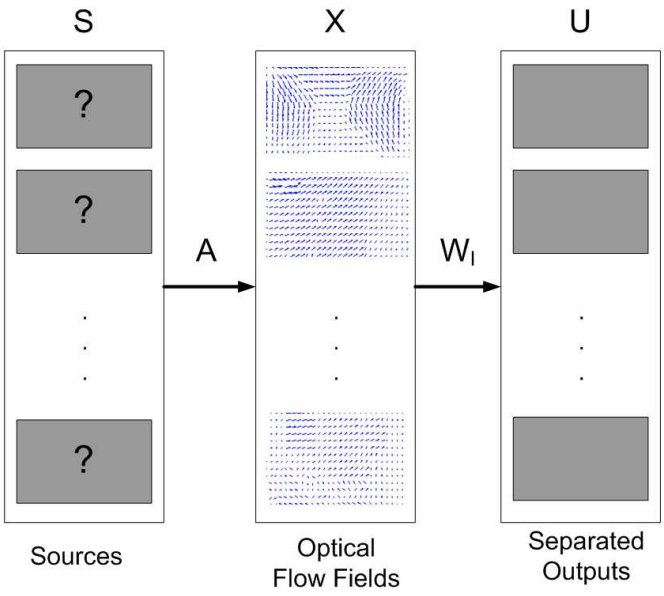

Figure 5.8: ICA for optical flow fields.

- The PC representation of X is defined as $R = XP^T$.
- A minimum squared error approximation X is obtained by $\widehat{X} = RP$.
- The ICA algorithm produces $W_I = WW_z$ such that:

$$W_I P = U. \tag{5.19}$$

- This implies that:

$$\widehat{X} = RP = RW_I^{-1}U. \tag{5.20}$$

- This means that the coefficients of the basis flow fields are contained in RW_I^{-1}. Therefore, if we denote the coefficient matrix by F, then $F = RW_I^{-1}$.

Figure 5.9 shows the first 5 IC's of the optical flow fields.

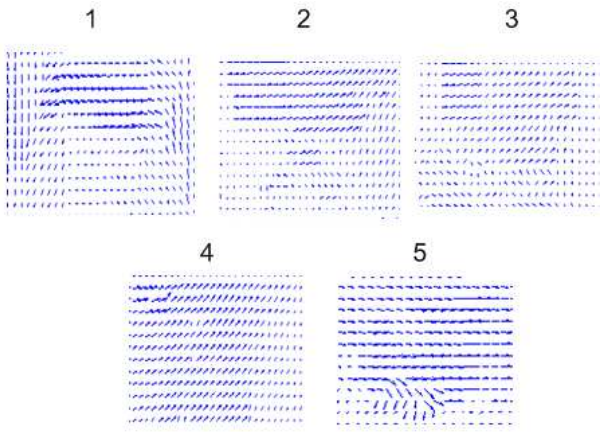

Figure 5.9: The first five derived basis flow fields.

5.8 Training the Motion Model

In order to evaluate the performance of our motion model, we will use the coefficients of the basis flow fields derived from training sequence as the training set. The recovered coefficients from the testing sequence will be used for evaluating the recognition task. For training, we used 4 streams of image sequences for every speaker, which correspond to the 4 words of the training set. As indicated before, training is performed for 11 users in every training set.

The first 12 eigenvectors (those with the largest eigenvalues) were used to estimate the motion of the mouth. This means that we obtain 12 basis images and consequently 12 coefficients. The learning rate was initialized at 0.0005 and annealed down to 0.0001. After training we obtain a 44×12 matrix F containing the coefficient values for the basis flow fields of the training set. Next, we perform the same steps on the testing set (which comprises one speaker in every one of the 12 tests uttering the 4 words) and obtain a 4×12 matrix F_{test} containing the coefficient values for the testing set. Next, the coefficient values are derived over a number of frames and a coefficient trajectory is found for every speaker in the training set. The coefficient trajectories are then averaged over the 11 users. The dimensionality of our feature vector at every frame level is then equal to 12. This feature vector is used to derive the viseme class label and then decide on the word class as dictated

by the Viterbi lattice (described in Section 5.3). The following subsections demonstrate the experimental setup that is performed to show the efficiency of the developed model.

5.9 Experimental Results

In order to assess the efficiency of our feature extraction approach, we perform recognition using the leave-one-out principle described earlier. Since we have 12 subjects and 4 word classes repeated 2 times by each speaker in the Tulips1 database, this means that we have 96 word recognition tests. In this section, we present our experimental results and compare them to others reported in literature for the same experiment on the Tulips1 database. As a performance measure, we use the Word Recognition Rate (WRR) averaged over all the recognition tasks.

Figure 5.10 shows the first and sixth frame of four subjects uttering the word "one". It also shows the optical flow fields derived at these frames (bottom of Figure 5.10). If the proposed model is correctly capturing the motion of the lips, then the estimated coefficients of each speaker should be the similar. For this purpose, the first four coefficients (c_1, c_2, c_3 and c_4) are plotted at the top of Figure 5.10 over 15 frames. In general, the plots appear to be highly correlated for different speakers saying the same word. This correlation is high for some coefficients (like coefficient c_1) and low for other coefficients (like coefficient c_4). However, there appears to be a trend of coefficient temporal variation for the same word across different speakers, which demonstrates the efficiency of our proposed model.

To further illustrate that the derived coefficients contain structural information about the speech classes, the coefficient values are plotted over 15 frames for both the training and testing data. Figure 5.11 shows the the first 5 coefficient values obtained by averaging the coefficient trajectories of the four words over the training set (11 subjects). Figure 5.12, on the other hand, shows the first 5 coefficient trajectories of the four words for the speaker being tested. At each time instant there is a vector of coefficients that represents the flow field of the ROI. Note that the trajectories for the coefficients of a certain word are similar for both the training set and the testing set.

Speech is usually characterized by large, rapidly changing motions, which can be difficult to estimate, especially across multiple speakers. Consequently, our model, having

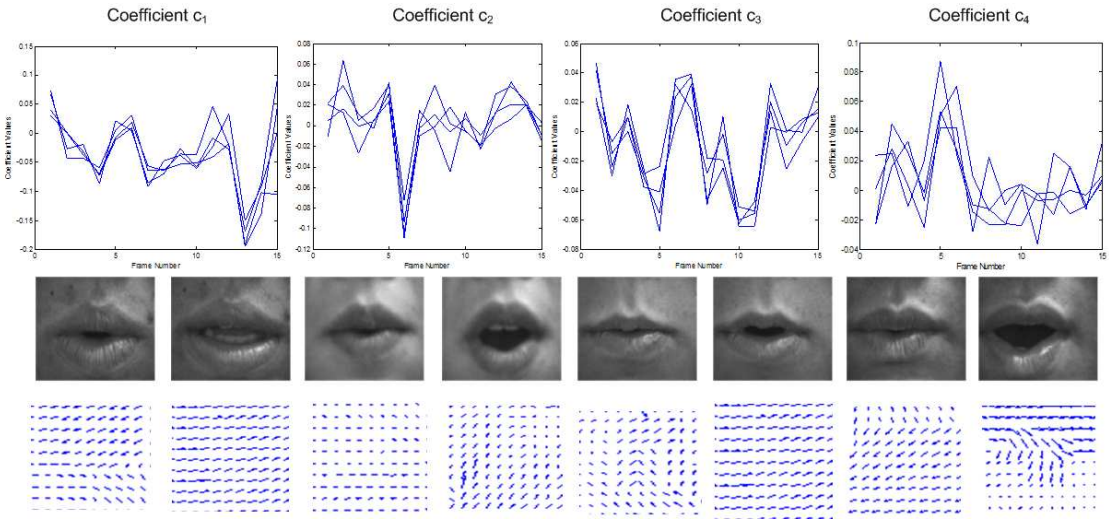

Figure 5.10: Four coefficients for four subjects uttering the word "one".

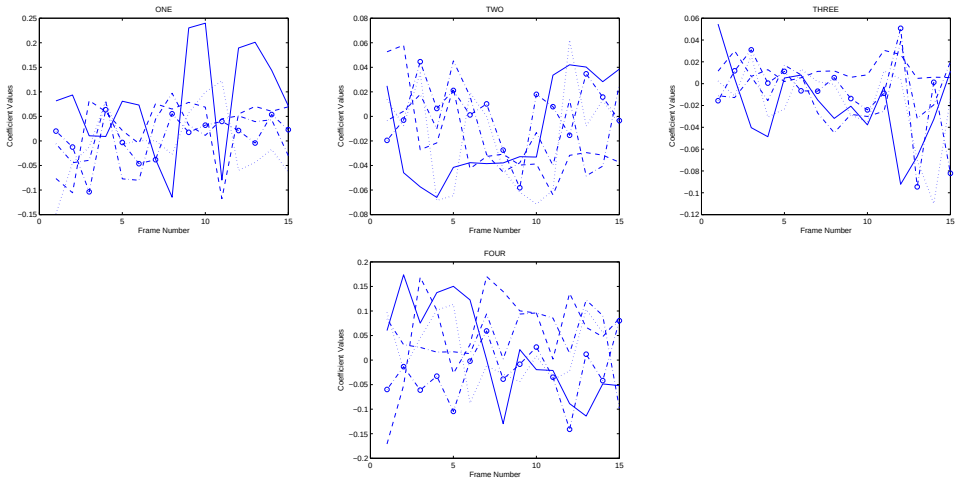

Figure 5.11: Training set: 5 coefficient values over 15 frames.

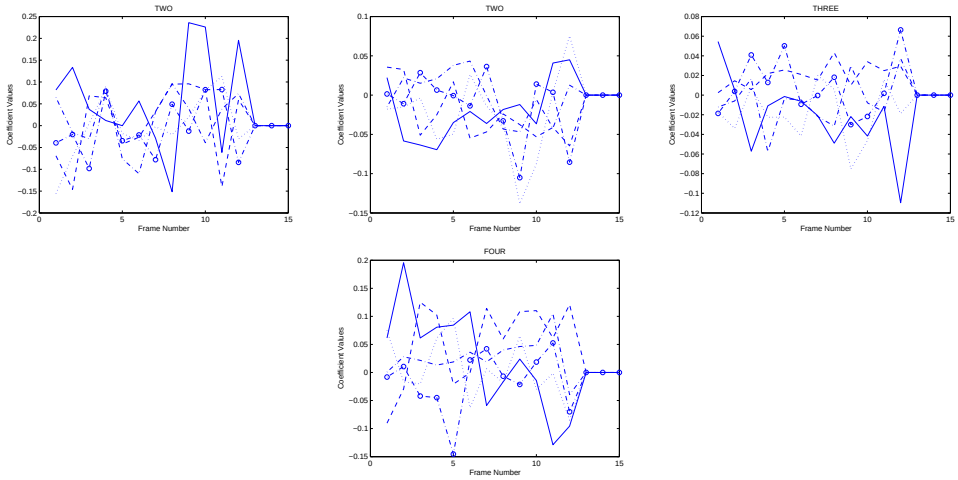

Figure 5.12: Testing set: 5 coefficient values over 15 frames.

achieved this high level of correlation, is expected to have performed well for the task of speech recognition. In order to evaluate and quantify this performance, we have to perform classification. We will first discuss our experimental protocol.

5.9.1 Experimental Protocol

We start the design of the visual speech recognizer with the definition of the viseme classes for the first four digits in English. We first obtain the phonetic transcriptions of the first four digits in English using the CMU pronunciation dictionary [1]:

"one" → "W-AH-N"
"two" → "T-UW"
"three" → "TH-R-IY"
"four" → "F-AO-R". We then define the viseme classes so that

- A viseme class includes as few phonemes as possible
- We have as few different visual realizations of the same viseme as possible.

The definition of viseme classes was based in the visual examination of the video part from the Tulips1 database as shown in [30]. The clustering of the different mouth images

into viseme classes was done manually on the base of visual similarity of these images. Accordingly, we obtain the viseme classes described in Table 5.2 and the phoneme-to-viseme mapping given in Table 5.3.

Table 5.2: Viseme classes for the Tulips1 database [53].

Viseme Group Index	Symbolic Notation	Viseme Description
1	(W)	Small-rounded open mouth state
2	(AO)	Larger-rounded open mouth state
3	(WAO)	Medium-rounded open mouth state
4	(AH)	Medium ellipsoidal mouth state
5	(N)	Medium open, not rounded, mouth state; teeth visible
6	(T)	Medium open, not rounded, mouth state; teeth and tongue visible
7	(TH)	Medium open, not rounded mouth state
8	(IY)	Longitudinal open mouth state
9	(F)	Almost closed mouth state; upper teeth visible; lower lip moved inside

5.9.2 Classification Results

Given the dataset, we estimate the optical flow fields and perform independent component analysis as described earlier. The reduced feature vectors obtained from the independent components of the optical flow fields are then used for classification. For this purpose, each of the 4 words is fed into an HMM for recognition.

Table 5.4 shows the WRR per subject, obtained by the proposed approach, and compared to other approaches that have worked on the same database. These other approaches are not explained in this work, but their results are provided for comparison.

This table shows that the minimum WRR attained in the proposed approach is 75% and the maximum WRR is 100%, a result that is generally acceptable in speech recogni-

Table 5.3: Phoneme-to-viseme mapping for the Tulips1 database [53].

Viseme Group Index	Corresponding Phonemes
1,2, or 3 (depending on the speaker's pronunciation)	/W/, /UW/, /AO/
1 or 3 (depending on the speaker's pronunciation)	/R/
4	/AH/
5	/N/
6	/T/
7	/TH/
8 or 4 (depending on the speaker's pronunciation)	/IY/
9	/F/

Table 5.4: WRR per subject in the Tulips1 database.

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Accuracy [%]	87.5	75	100	87.5	100	100	100	100	75	87.5	87.5	100
Accuracy [%] [30]	100	75	100	100	87.5	100	87.5	100	100	62.5	87.5	87.5
Accuracy [%] [48]	100	87.5	87.5	75	100	100	75	100	100	75	87.5	87.5

tion. However, this performance metric does not give a clear understanding of the overall enhancement introduced by the proposed approach.

Table 5.5, on the other hand, shows the confusion matrix between the words actually uttered and the words recognized by the HMM classifier. This confusion matrix is compared to the average human confusion matrix [53] in Table 5.6. It is shown here that there is a relation between the two matrices. For example, both of the human confusion matrix and the confusion matrix of the proposed approach show higher accuracy in recognizing the words "two" and "four" than the words "one" and "three".

Table 5.5: Confusion matrix for visual word recognition.

	One	Two	Three	Four
One	91.67%	0%	0%	8.33%
Two	0%	95.83%	4.17%	0%
Three	8.33%	4.17%	87.5%	0%
Four	0%	4.17%	0%	95.83%

Table 5.6: Average human confusion matrix.

	One	Two	Three	Four
One	89.36%	0.46%	8.33%	1.85%
Two	1.39%	98.61%	0%	0%
Three	9.25%	3.24%	85.64%	1.87%
Four	4.17%	0.46%	1.85%	93.52%

Table 5.7 shows the overall WRR for all subjects, along with the WRR confidence intervals, in comparison to those obtained by other approaches. The WRR of our approach was calculated by averaging the diagonal values of the confusion matrix.

From examining these tables, we can see that our system achieves an overall WRR of 92.7%. This exceeds the WRR of the best rate reported in [33] and in [56] by 1%. It

Table 5.7: Overall recognition rate compared to other methods.

Method	WRR	Confidence Interval
Our Method	92.7%	[87.5%, 95.83%]
Global PCA [33]	79.2%	[70%, 86.1%]
Global ICA [33]	74%	[64.4%, 81.7%]
Blocked filter bank PCA/ICA [33]	85.4%	[76.9%, 91.1%]
Unblocked filter bank PCA/ICA [33]	91.7%	[84.4%, 95.7%]
Diffusion network shape+intensity [56]	91.7%	[84.4%, 95.7%]
Delta features/SVM classifier [30]	90.6%	[83.1%, 94.7%]
Delta features/HMMs [53]	89.93%	[82.3%, 94.5%]

should also be mentioned that the minimum value of the confidence interval as found by our approach is 87.5% which is larger than the best value obtained in [33] and in [56] by 3.1%. In addition, it is to be noted that the methods reported in [33] require a lot of local processing, by the use of a bank of linear shift invariant filters with unblocked selection whose response filters are ICA or PCA kernels of very small size (12×12 pixels). On the other hand, our approach is much simpler to implement.

5.9.3 Quality of the Chosen Eigenvectors

We have stated before that the reason for choosing 12 basis flow fields is that the first 12 eigenvectors of the source fields (those corresponding to the highest eigenvalues), account for 95% of the variance. We show here why the first 12 basis flow fields are enough for representing the motion of the mouth.

As discussed in [27], in order to observe the quality of the chosen basis fields, we measure the fraction of the variance of the training set that is accounted for by the first n components:

$$Q(n) = \frac{(\sum_{j=1}^{n} \lambda_j^2)}{(\sum_{j=1}^{p} \lambda_j^2)}, \tag{5.21}$$

where p is the total number of eigenvalues. As $Q(n)$ approaches 1, the first n eigenvectors will account for a large portion of the variance. Consequently, these eigenvectors will act as a good representation because eigenvalues λ_j are very small for $j > n$. On the other hand, if $Q(n)$ is relatively small for a given n, then the first n eigenvectors are not enough to represent the model of interest. This means that the dimensionality of the model depends on how fast $Q(n)$ increases with respect to n, being very large for instance if $Q(n)$ increases slowly. Performing this calculation, we find that $Q(12) = 0.95$ which means that the first 12 eigenvectors account for 95% of the variance. This is why we choose to use the first 12 principal components in the PCA stage prior to the ICA calculation. This means that we have coefficient values per class at each time frame.

5.10 Chapter Summary and Discussion

Visual speech is complex in its nature as it includes a lot of non-rigidity, self-occlusion and high image velocities. Consequently, much of the information is contained in the higher order statistics (the phase spectrum), with a little knowledge provided by second-order statistics (the amplitude spectrum). The fact that PCA derives only second order statistics makes it an inefficient tool for our application. For this purpose, we have used Bell and Sejnowski's ICA, which is based on the principal of optimal information flow in sigmoidal neurons by minimizing output joint mutual information. PCA is a special case of ICA in which the source models are Gaussian. The assumption that sources are Gaussian is not necessarily true, especially for natural signals such as speech, EEG and natural images. Consequently, ICA provides a better probabilistic model of the data and derives a basis set that can reconstruct the data better than PCA in the presence of noise. In addition, it uniquely identifies the mixing matrix A and captures higher-order statistics that are not provided by the covariance matrix.

The purpose of the work of this chapter was two-fold: to *understand* the motion of the lips, and to *recognize* the motion of the lips. We proposed a novel approach for achieving these goals. Each word from the Tulips1 database was represented by a sequence of frames, and the optical flow was derived between every pair of consecutive frames. ICA was then performed on the optical flow fields to find a basis set that is statistically independent.

The temporal trajectory of the model coefficients provided a rich description of the visual speech information. Next, we used the coefficients of these basis images to recognize units of speech (defined here as the visemes of the words "one" to "four" uttered by subjects of the Tulips1 database).

For recognition we first created a temporal Viterbi lattice containing as many states (N) as the optical flow frames in the video sequence. We then used an N-state HMM based on this Viterbi lattice to classify the visemes and the corresponding words. We found that the classification results improved by 1% as compared to the best performing state of the art system. We also showed that the first 5 eigenvectors (those corresponding to the largest eigenvalues) of the input flow fields account for 95% of the variance in the data. This means that the chosen 12 basis optical flow fields are highly representative of the lips motion model.

Chapter 6

Multimodal Fusion

Having presented the design methodology for the visual and audio feature design and extraction, this chapter develops a method of modality fusion that is based on reliability. In order to achieve this purpose, several issues should be addressed. The first issue is where the fusion of the data takes place. As discussed in Chapter 2, several architectures have been developed in literature to tackle this issue. Feature fusion integrates data on the feature level, where audio and visual features are used simultaneously and equally to identify the corresponding speech unit. Decision fusion, on the other hand, takes place after the independent identification of each stream and is thus an integration of identification results. Other methods, such as Motor Recoding, Dominant Recoding and hybrid methods also exist. However, the most widely used integration techniques use either a feature-based or a decision-based fusion architecture.

For our recognition experiments we exclusively followed a Separate Integration (decision fusion) architecture because different comparisons showed superior performance of the SI compared to the other fusion architectures [78], [73]. Once the level of integration is chosen, the next issue to tackle is how the fusion of the identification results takes place. The identification results in our case are the a posteriori probabilities of the observation vectors. The quality of this estimate is related to the match of the training and testing conditions. Since the training data was all recorded in a clean environment, the reliability if the testing set depends solely on the noise present in the test condition and not on any other conditions (such as sensor malfunction). In order to account for the dynamic changes

in reliability, a weighting of the audio and visual probabilities is desirable. For this purpose, we propose in this chapter two stream reliability indicators based on the dispersion of the a posteriori probabilities of the observation vectors. These reliability indicators are then mapped into stream weights using the genetic algorithm, in such a way that maximizes the conditional likelihood. Figure 6.1 shows an overall diagram of our fusion system. Note that although the work in this chapter focuses only on two modalities (the audio and the video), all the derivations are done with respect to a random number of input modalities. This facilitates the integration of our fusion module in future architectures that may use more than two streams of information.

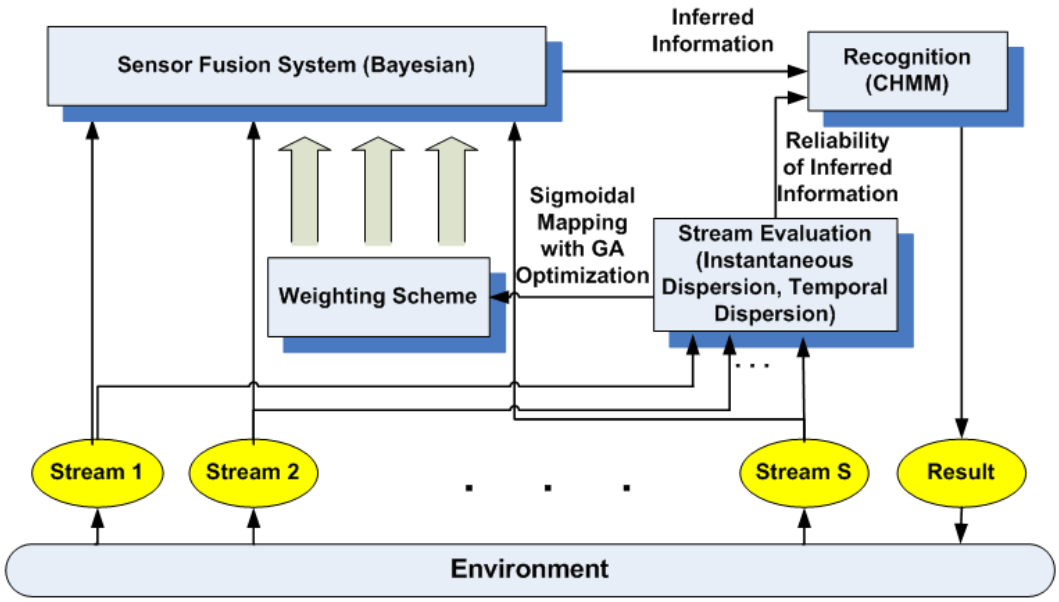

Figure 6.1: Overview of the multimodal fusion system.

6.1 Bayesian Fusion

In pattern classification problems, we measure a property (feature) of a pattern instance and try to decide to which of M classes $c_i, i = 1, ..., M$ it should be assigned. Multimodal fusion or integration combines S complementary features, originating from a single or

multiple modalities, in order to maximize information gather and to overcome the impact of noise in each individual stream.
Let $S_k, k = 1, ..., S$, denote the information streams that we want to integrate.
Let $\mathbf{x}_s, s = 1, ..., S$, denote the feature vectors of every stream.

The simplest way to combine audio and video data is to use Bayes' rule and multiply the audio and video a posteriori probabilities. From a probabilistic perspective, this approach is valid if the audio and video data are independent. Perceptive studies have shown that in human speech perception, audio and video data are treated as class conditional independent [54], [50]. In this case, the conditional probability of the observation vector $\mathbf{x}_{1:S} = (\mathbf{x}_1, ..., \mathbf{x}_S)$ given the class label c_i is governed by the product:

$$P(\mathbf{x}_{1:S}|c_i) = P(\mathbf{x}_1, ..., \mathbf{x}_S|c_i) = \prod_{s=1}^{S} P(\mathbf{x}_s|c_i). \tag{6.1}$$

Using Bayes' rule, we get the desired a posteriori probability of the class given the features:

$$P(c_i|\mathbf{x}_{1:S}) = \frac{\prod_{s=1}^{S} P(c_i|\mathbf{x}_s)}{P(c_i)} . \frac{\prod_{s=1}^{S} P(\mathbf{x}_s)}{P(\mathbf{x}_{1:S})}. \tag{6.2}$$

By replacing the probabilities P by estimates $\hat{P}$, we get a representation of the *Bayesian Fusion (BF)*:

$$\hat{P}_{BF}(c_i|\mathbf{x}_{1:S}) = \frac{\prod_{s=1}^{S} P(c_i|\mathbf{x}_s)}{P(c_i)} . \eta, \tag{6.3}$$

where the terms independent of the actual class are replaced by the normalization factor η:

$$\eta = \frac{1}{\sum_{j=1}^{M} \frac{\prod_{s=1}^{S} P(c_j|\mathbf{x}_s)}{P(c_i)}}, \tag{6.4}$$

where M is the number of classes. This probability can then be used in classification by making use of the *Maximum A Posteriori (MAP)* rule:

$$\hat{c} = \underset{c_i \in C}{\operatorname{argmax}} \hat{P}_{BF}(c_i|\mathbf{x}_{1:S}). \tag{6.5}$$

It is to be noted here that the counterpart of the Bayesian Fusion model in the area of human perception is the *Fuzzy Logical Model of Perception (FLMP)*.

6.2 Weighted Bayesian Fusion

The standard Bayesian Fusion approach does not deal with varying reliability levels of the input streams. In order to improve classification performance, several authors have introduced stream weights $\{\lambda_1, \lambda_2, \lambda_3, ..., \lambda_S\}$ as exponents in Equation 6.3, resulting in the modified score:

$$\hat{P}_{WBF}(c_i|\mathbf{x}_{1:S}) = \frac{\prod_{s=1}^{S} P(\mathbf{x}_s|c_i)^{\lambda_s}}{\sum_{j=1}^{M}\prod_{s=1}^{S} P(\mathbf{x}_s|c_j)^{\lambda_s}}. \tag{6.6}$$

Notice that Equation 6.6 corresponds to a linear combination in the log-likelihood domain; however, it does not represent a probability distribution in general, and will consequently be referred to as a score. Such schemes have been motivated by potential differences in reliability among different information streams, and larger weights are assigned to information streams with better classification performance. Using such weighting mechanisms has experimentally been proven beneficial for feature integration in both intra-modal and inter-modal scenarios.

In order to determine the weights $\{\lambda_1, \lambda_2, \lambda_3, ..., \lambda_S\}$, we first need to define reliability measures for the individual streams. These reliability measures should reflect the quality of the observation conditions by considering statistical information conveyed in both prior and current classification results. The second step is to find an optimal mapping between these reliability indicators and the stream weights $\{\lambda_1, \lambda_2, \lambda_3, ..., \lambda_S\}$. The only constraints that this problem has are that the weights should be positive and should add up to 1.

$$\sum_{s=1}^{S} \lambda_s = 1, \tag{6.7}$$

$$\lambda_s \geq 0 \tag{6.8}$$

6.3 Reliability of Sensor Information

In this section, we provide an adaptive reliability assessment model for the different modalities. The stream reliability variables are calculated using a frame level evaluation process prior to the recognition process. Since the discriminative powers of the audio and visual signals have a temporally correlated variation which may be altered by dramatic noise

bursts, a reliability weighting scheme should take into account the previously observed behavior of the classifier. For this purpose, we propose an adaptive stream reliability measure that is based on the idea of dispersion.

As discussed in Chapter 2, the first work to ever introduce the dispersion as a measure of reliability in audio-visual speech recognition systems was that developed by Adjoudani and Benoit [2]. Since then, this idea has been further developed by other researchers. In this work, a dispersion measure developed by Potaminaos and Neti [65] is used. This measure uses an N-best dispersion method that is formulated as the difference between each pair of n^{th}-best hypotheses, and it is given by:

$$L = \frac{2}{N(N-1)} \sum_{n=1}^{N} \sum_{n'=n+1}^{N} (R_n - R_{n'}), \tag{6.9}$$

where $N \geq 2$ and R_n is equal to the n^{th}-best hypothesis. Dispersion measures provide a good estimate of stream confidence, as a large difference in classifier outputs reflects a greater confidence. Lucey et al. [46] have theoretically proven that dispersion approximately reflects the the cepstral shrinkage effect induced by additive noise.

6.3.1 Instantaneous Dispersion

The first reliability measure that we use is the instantaneous dispersion based on a local frame measurement. It is defined as:

$$L_{s,t} = L(log(P(\mathbf{x}_{s,t}|c_{s,t}))) = \frac{2}{N(N-1)} \sum_{n=1}^{N} \sum_{n'=n+1}^{N} \log \frac{P(\mathbf{x}_{s,t}|c_{s,t,n})}{P(\mathbf{x}_{s,t}|c_{s,t,n'})}, \tag{6.10}$$

where L(.) is the N-best log-likelihood dispersion function defined in Equation 6.9 and $P(\mathbf{x}_{s,t}|c_{s,t})$ is the observation emission probability generated by an HMM-based classifier. Here we choose $N = 4$ because both Adjoudani and Benoit [2] and Potamianos and Neti [65] have found that an N-best of 4 has been the most successful. This reliability measure is calculated for a particular instance of time t, which shows that it is only capable of capturing the reliability of a given stream at a certain time. However, it does not capture information about the stream from previous time or previous states. This missing information contains valuable knowledge about the reliability of a system.

It is to be noted here that the audio and visual speech signals have different discriminative powers depending on their classes. For instance, the audio channel may have a difficulty discriminating between the phonemes /m/ and /n/; however, these classes are easily distinguishable using the visual cues because /m/ takes a closed mouth shape while /n/ takes an open mouth shape. On the other hand, the classes /p/ and /m/ are not easily distinguished in the visual domain.

6.3.2 Temporal Dispersion

The instantaneous dispersion measure evaluates the stream reliability at a frame level. It is evident from the above discussion that in clean speech, the instantaneous dispersion should be low for speech classes that are perceived similar to each other and high for classes that are highly discriminative, and thus it can be implied that instantaneous dispersion is a good measure of a stream's reliability. However, this assumption is not necessarily valid in highly corrupted speech. In this case, the noise will cause the dispersions to vary rapidly. Therefore, it is hard to judge whether the dispersion changes come from the varying discriminative powers of the recognizer or from the ambient noise presented in the multimodal streams. For this purpose, the instantaneous dispersion is not sufficient to assess the stream reliability and there is a need for a temporal reliability measure. The idea of a temporal reliability is similar to that of a local reliability measure, but rather than being solely based on the current output of the system, a weighting based on temporal reliability would take into account the previously observed behavior of the classifier.

The results of the audio and visual streams should be combined dynamically according to their relative reliability. For this purpose we propose a second confidence measure that is based on the instantaneous dispersion measure. We define this weight function $R_{s,t}$ (with backward-looking time step Δt) as:

$$R_{s,t} = \sum_{n=1}^{q} \rho(n) L_{s,t-n\Delta t} + \epsilon_t, \tag{6.11}$$

where q is the window length (chosen here to be equal to 5), and ϵ_t is a white noise process with zero mean and variance σ^2. The $\rho(n), n = 1, ..., q$, in the range 0 to 1, are parameters that correspond to how rapidly past performance will be discounted. This factor $\rho(n)$

is largely responsible for determining the dynamic weights $R_{s,t}$, where recent performance should be weighted highly (large $\rho(n)$) and past performance should be gradually forgotten (low $\rho(n)$). Here the problem boils down to determining the proper values of $\rho(n)$. This value could be set to a constant raised to the power n, which implies that as n increases (by going more into the past), the weighting of the dispersion values decreases. However, this is not very appropriate in environments with a large noise variation. In order to demonstrate this idea, consider the case where at a certain frame instance t_1, a sudden noise burst occurs, and then this noise disappears at the following frame instance $t_2 = t_1 + \Delta t$. This means that for calculating the temporal dispersion at t_2, the instantaneous dispersion at t_1 should not be weighted highly. This is because, to a certain extent, the two dispersions are not correlated in this case. Consequently, the calculation of $\rho(n)$ should depend on the correlation between the dispersion values.

For this purpose, we note from Equation 6.11 that this temporal dispersion is formulated as an autoregressive model. Autoregressive (AR) models are among the most commonly used statistical models for time series modeling. The parameters $\rho(n)$ are thus calculated using the *Yule-Walker* equations [35]:

$$r_L(m) = \sum_{n=1}^{q} \rho(n) r_L(m-n) + \sigma_\epsilon^2 \delta_m, \tag{6.12}$$

where $m = 0, ..., q$, yielding $q+1$ equations, and $r_L(m)$ is the autocorrelation function of the instantaneous dispersion series. σ_ϵ is the standard deviation of the input noise process, and δ_m is the Kronecker delta function. Because the last part of the equation is non-zero only if $m = 0$, these equations are usually solved by representing them as a matrix for $m > 0$, thus getting the following equation:

$$\begin{bmatrix} r_L(1) \\ r_L(2) \\ . \\ . \\ . \\ r_L(q) \end{bmatrix} = \begin{bmatrix} r_L(0) & r_L(-1) & . & . & . & r_L(-q+1) \\ r_L(1) & r_L(0) & . & . & . & r_L(-q+2) \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ r_L(q-1) & r_L(q-2) & . & . & . & r_L(0) \end{bmatrix} \begin{bmatrix} \rho(1) \\ \rho(2) \\ . \\ . \\ . \\ \rho(q) \end{bmatrix}. \tag{6.13}$$

Using the property that $r_L(n) = r_L(-n)$ for real processes, the *Yule-Walker* equations

become:

$$\begin{bmatrix} r_L(1) \\ r_L(2) \\ . \\ . \\ . \\ r_L(q) \end{bmatrix} = \begin{bmatrix} r_L(0) & r_L(1) & . & . & . & r_L(q-1) \\ r_L(1) & r_L(0) & . & . & . & r_L(q-2) \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ r_L(q-1) & r_L(q-2) & . & . & . & r_L(0) \end{bmatrix} \begin{bmatrix} \rho(1) \\ \rho(2) \\ . \\ . \\ . \\ \rho(q) \end{bmatrix}, \tag{6.14}$$

and we solve for all ρ. For $m = 0$ we have:

$$r_L(0) = \sum_{n=1}^{q} \rho(n) r_L(-n) + \sigma_\epsilon^2, \tag{6.15}$$

which allows us to solve σ_ϵ^2. Note that since the *Yule-Walker* equations are linear in the coefficients $\rho(n)$, it is a simple matter to find the coefficients $\rho(n)$ from the autocorrelation sequence $r_L(n)$.

6.4 Stream Weight Optimization

The next step is to find a mapping between the reliability measures (L and R derived in Equations 6.10 and 6.11) and the stream weights $(\lambda_1, \lambda_2, \lambda_3, ..., \lambda_S)$. We use a sigmoid function for this purpose, as chosen by [67], due to the fact that it is monotonic, smooth and bounded between zero and one. First we define the reliability vector $d_t = [d_{1,t}, d_{2,t}, d_{3,t}, ..., d_{2S,t}] = [R_{1,t}, R_{2,t}, ..., R_{S,t}, L_{1,t}, L_{2,t}, ..., L_{S,t}]$.

Then, the mapping is defined as:

$$\lambda_s = \frac{1}{1 + exp(-\sum_{i=1}^{2S} w_{s,i} d_{i,t})}, \tag{6.16}$$

where $\mathbf{W}_s = [w_{s,1}, w_{s,2}, w_{s,3}, ..., w_{s,2S}]$ is the vector of the sigmoid parameters for stream s. As discussed before, several optimization techniques were proposed in literature to find the optimal parameters $\mathbf{W}_s$ from which we can derive the optimal weights $(\lambda_1, \lambda_2, \lambda_3, ..., \lambda_S)$. Since we have S streams and $2S$ sigmoid parameters per stream, then we have $S \times 2S = 2S^2$ sigmoid parameters that we need to optimize. This is a nonlinear optimization problem

with a large number of variables. Consequently, we propose a method based on genetic algorithms (GA) to solve this problem and show how we use this method to determine the optimal set of weights. We now give a brief introduction of genetic algorithms and then move on to show how a genetic algorithm model is developed for our problem.

6.4.1 Introduction to Genetic Algorithms

Linear programming and dynamic programming techniques often fail (or reach local optima) in solving NP-hard problems with large number of variables and non-linear objective functions. For this purpose, researchers have proposed evolutionary algorithms that mimic the biological evolution of species, and that are capable of reaching near-optimum solutions. GAs are the most popular of these systems and they comprise a family of stochastic search methods that are based on improving fitness through evolution [37]. A solution to a given problem is represented in the form of a string, called a "chromosome", consisting of a set of elements, called "genes", that hold a set of values for the optimization variables [29]. The fitness of each chromosome is determined by evaluating it against an objective function. To simulate the natural "survival of the fittest" process, the best chromosomes exchange information, through crossover or mutation, to produce offspring chromosomes. The offspring solutions are then evaluated and used to evolve the population if they provide better solutions than their parents. Usually, the process is continued for a large number of generations to obtain a best-fit (near-optimum) solution. The building blocks of any GA algorithm are the GA operators. The simplest form of genetic algorithm involves three types of operators: selection, crossover and mutation.

6.4.2 Problem Modeling Using Genetic Algorithms

Having established a formal definition of genetic algorithms, we now describe the way a GA model is formulated for our problem. Constructing a genetic algorithm starts with selecting the way in which candidate solutions are encoded. Then a certain technique for each operator (i.e. selection, crossover, and mutation) is chosen. The following subsections describe our model.

Encoding Encoding is a central factor in the success of a genetic algorithm. As most applications in GA, a fixed-length, fixed-order bit string (binary encoding) is selected to encode the candidate solution. In our case, the objective function of the genetic algorithm is to find the optimum weights for the S streams such that the system reliability is maximized.

In our GA model, we have S streams $\{S_1, S_2, S_3, ..., S_S\}$ which require $2S^2$ parameters $\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2 ..., \mathbf{W}_S\}$. Each parameter is encoded by k bits. In this case the chromosome will contain $2S^2 \times k$ bits. Each parameter has 2^k possible values. This will determine the resolution of the weight values, which is in this case $1/2^k$.

Selection and Fitness Function This "selection" operator selects chromosomes in the population for reproduction. The fitter the chromosome, the more times it is likely to be selected to reproduce.

The objective function of the genetic algorithm used in this experiment is to optimize the system reliability by adjusting the weights. This requires a fitness function that assigns a fitness to each chromosome in the current population. The fitness function should be selected such that the fitness of a chromosome will reflect how well that chromosome enhances the system reliability. For this purpose we choose our objective function to be the maximum conditional likelihood (MCL) estimates of parameters $\mathbf{W}$ over the training set. Given an observation vector $\mathbf{x}_{1:S}$, we first represent the conditional likelihood of class c_i by:

$$\hat{P}_{WBF}(c_i|\mathbf{x}_{1:S}) = \frac{\prod_{s=1}^{S} P(\mathbf{x}_s|c_i)^{\lambda_s}}{\sum_{j=1}^{M}\prod_{s=1}^{S} P(\mathbf{x}_s|c_j)^{\lambda_s}}. \tag{6.17}$$

We then seek the parameters $\mathbf{W}$ over a time interval T in the training set as:

$$\hat{\mathbf{W}} = \underset{W}{\operatorname{argmax}} \sum_{t\in T} \log P_{WBF}(c_{i,t}|\mathbf{x}_{1:S}), \tag{6.18}$$

where $c_{i,t}$ is the class c_i at time t. Equation 6.18 is our objective function, subject to the constraints:

$$\sum_{s=1}^{S} \lambda_s = 1, \tag{6.19}$$

$$\lambda_s \geq 0. \tag{6.20}$$

The purpose of the selection process is to emphasize the fitter individuals in the population hoping that their offspring will in turn have even higher fitness. The selection process has to be balanced with variation from crossover and mutation (the exploitation/exploration balance); i.e., too-strong selection will result in suboptimal highly fit individual who may take over the population, reducing the diversity needed for further change and progress; too-weak selection will result in too-slow evolution.

A common selection method in GA's is fitness-proportionate selection, in which the number of times an individual is expected to reproduce is equal to its fitness divided by the average fitness of the population, (this is equivalent to what biologists call "viability selection").

A simple method of implementing fitness-proportionate selection is the "roulette-wheel sampling", which is conceptually equivalent to giving each individual a slice of a circular roulette wheel equal in area to the individual's fitness. The roulette wheel is spun, the ball comes to rest on one wedge-shaped slice, and the corresponding individual is selected.

The fitness proportionate selection with elitism is used in this experiment. Elitism is an addition to many selection methods that forces the GA to retain some number of the best individuals at each generation. Such individuals can be lost if they are not selected to reproduce, or if they are destroyed by a crossover or a mutation. In this experiment only one elite (the best individual) is kept. The rest of the selection process is a normal roulette wheel sampling selection.

Crossover This operator randomly chooses a locus and exchanges the subsequences before and after that locus between two chromosomes to create two offspring. There are many variants of crossover found in the GA literature such as one point crossover, two point crossover, parametrized uniform crossover, etc. In this experiment, two point crossover with probability of 0.7 is selected.

Mutation This operator randomly flips some of the bits in a chromosome. Mutation can occur at each bit position in a string with some probability, usually very small. In this experiment, we set the mutation rate to 0.007.

For the genetic algorithm modeling, the "Evolver" tool was used. This is implemented as a macros in Microsoft Excel and has proven to be an efficient and easy-to-use tool for

GA applications.

6.5 The Recognition System

It is well known that the visual speech activity precedes the audio signal by as much as 120 ms [9], which is close to the average duration of a phoneme. As discussed in Chapter 2, the multi-stream HMM enforces state synchrony between the audio and visual streams. It is therefore important to relax the assumption of state synchronous integration, and instead allow some degree of asynchrony between the audio and visual streams. Many recognition models, such as the product HMM, the factorial HMM and the coupled HMM have been proposed for this purpose. In this work, the coupled HMM is used and is discussed next.

6.5.1 The Coupled Hidden Markov Model

The coupled HMM (CHMM) [58] is a dynamic Bayesian network (DBN) that allows the backbone nodes to interact, and at the same time to have their own observations. In the past, CHMM have been used to model hand gestures [8], the interaction between speech and hand gestures [61], or audio-visual speech [58]. A CHMM can be seen as a collection of HMMs, one for each modality stream, where the hidden backbone nodes at time t for each HMM are conditioned by the backbone nodes at time $t-1$ for all related HMMs. Figure 6.2 shows a continuous two-stream CHMM used for our audio-visual speech recognition system. The squares represent the hidden discrete nodes, while the circles describe the continuous observable nodes. Unlike the independent HMM used for audio-visual data, the CHMM can capture the interactions between audio and video streams through the transition probabilities between the backbone nodes. The CHMM can model the audio-visual state asynchrony and preserve the natural audio-visual dependencies over time.

The training of the CHMM parameters is performed in two stages. In the first stage, the CHMM parameters are estimated for isolated phoneme-viseme pairs. These parameters are determined first using the Viterbi-based initialization described in [58], followed by the expectation-maximization (EM) algorithm [41]. In the second stage, the parameters of the CHMMs, estimated individually in the first stage, are refined through the embedded training of all CHMMs. In a way similar to the embedded training for HMMs [82], each

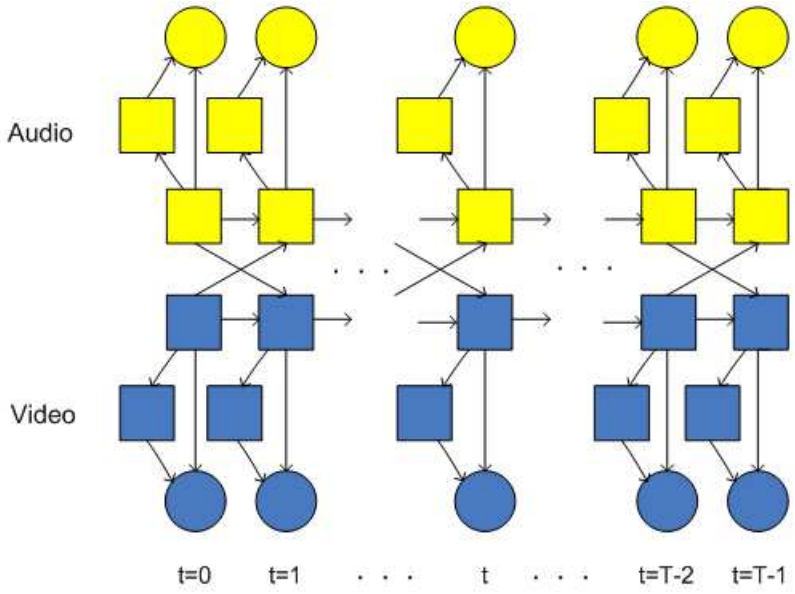

Figure 6.2: The audio-visual coupled HMM.

of the models obtained in the first stage are extended with one entry and one exit non-emitting states. The audio-visual speech recognition is carried out via a graph decoder applied to a word network consisting of all the words in the test dictionary. Each word in the network is stored as a sequence of phoneme-viseme CHMMs, and the best sequence of words is obtained through an extension of the token passing algorithm.

For the recognition task, we used the publicly available open source Audio-Visual Continuous Speech Recognition (AVCSR) toolkit developed by Intel. It is written in *C++*. Besides a CHMM-based audio-visual speech recognition decoder, two more recognition engines, an audio-only decoder (audio-only speech recognition engine) and a visual-only decoder (lip-reading engine) are integrated into this toolkit. We integrate the derived audio and video feature vectors along with the derived stream weights in an appropriate format into this toolkit. By using the CHMM and HMM recognition decoders of this toolkit, we compare the performance the different systems (as shown in the next section).

6.6 Experiments

6.6.1 Experimental Setup

In order to evaluate the efficiency of the fusion method, several experiments are conducted to show that audio-visual speech recognition can benefit from the suggested approach. This is achieved via classification experiments for the task of word classification of isolated digits. The Tulips1 database is used for all the experiments, which consists of 12 subjects, uttering the first four digits in English two times in repetition. Similar to the visual front-end experimental setup, the audio-visual speech recognizer is tested using the leave-one-out testing strategy for the 12 subjects in the Tulips1 database. This implies training the visual speech recognizer 12 times, each time using only 11 subjects for training and leaving the 12th out for testing. The audio signal is contaminated by adding 5 kinds of noise taken from the NOISEX database [80]:

- White noise
- Noise recorded in a car at 120 km/h
- Babble noise
- Two types of factory noise.

These different types of noise are mixed with the audio signal at 8 SNR levels ranging from -12dB to 30dB (clean speech). The SNR levels are: -12dB, -5dB, -3dB, 0dB, 5dB, 10dB, 20dB, and 30dB.

Mel frequency cepstral coefficients (MFCCs) are utilized as observations for the audio stream, constructing a 26-dimensional vector (derived in Chapter 4). As far as the visual front-end is concerned, a 12-dimensional visual feature vector is derived from the independent components of the optical flow fields (derived in Section 5.6.1). For the acoustic and visual modeling of the observations, 3-state left-right word Coupled HMMs with a single 5-continuous-Gaussian observation probability distribution per stream are used. The models are trained on clean data. As for the speech classes, the 9 visemes of Table 5.2 and the 10 phonemes of Table 5.3 are used.

6.6.2 Experimental Results

Figure 6.3 shows the temporal dispersions over 3 SNR levels (-5dB, 10dB and 20dB) for the audio stream of the same utterance. It is clear in general from these plots that the higher the SNR values, the higher the temporal dispersion measures are. This proves that the proposed measure can efficiently assess the reliability of the multimodal streams in a way that can capture the noise level in the channel. Figure 6.4 shows two plots. The top figure plots the instantaneous dispersions of both audio and visual streams varying over a 70-frame time window, where calculations are done at an SNR of 10dB taken from a speaker in the testing set. From this plot, it is clear that instantaneous dispersions vary dramatically over time. In most of the time, the instantaneous dispersions of the visual stream are much lower than that of the audio stream, which implies that the audio stream provides more information about the speech classes than the video stream. The bottom figure, on the other hand, plots the temporal dispersions calculated for the same speaker and same conditions. It is to be noted that the temporal dispersion plot shows more smoothness in estimating the channel reliability than the instantaneous dispersion plot. This translates into a reliability estimate that is more robust against noise bursts and that can assess classifier performance based on prior behavior.

Figure 6.5 shows the word error rate on average for the approach based on dispersion and gradient descent method (top) and our proposed method (bottom) respectively. The values of the word error rate are evaluated for the different types of noise applied on the different noise levels (babble noise, white noise, noise in a car, and two types of factory noise). It is shown that our proposed approach using the genetic algorithm optimization technique, yields a much better performance for almost all noise types and levels.

In order to more clearly see this performance improvement, Figure 6.6 presents the percentage of the correctly classified words in the isolated-digit recognition task. It shows 5 curves. The “A” curve represents the audio-only speech recognition classification accuracy shown for baseline comparison. This accuracy increases as the SNR level increases. This makes sense because the lower the noise in the channel (and the higher the SNR) the better the performance of the classification task. The “V” curve represents the video-only speech recognition classification accuracy. This value is constant because the level of noise in the audio channel does not affect the video channel. The “AV-Unweighted” curve is

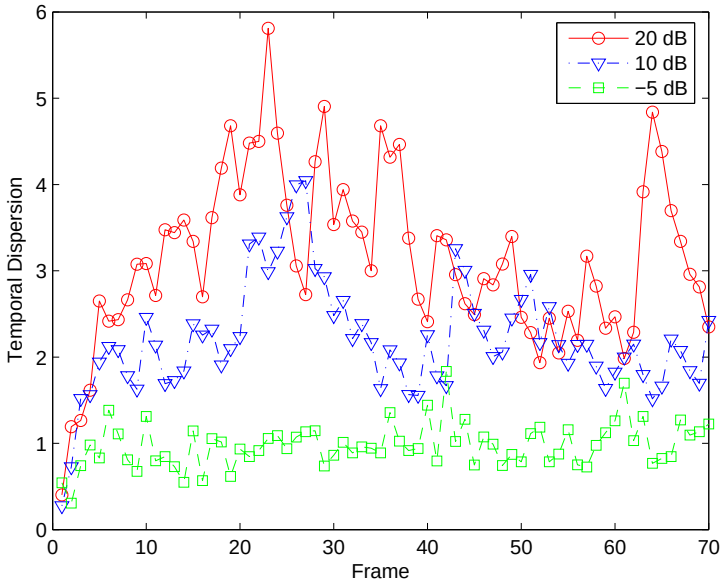

Figure 6.3: Audio temporal dispersion for 3 SNR levels.

the baseline audiovisual setup in which we use Coupled HMMs with stream weights equal to unity for both streams. For comparison, we also provide results with stream weights using the dispersion as reliability measure and the generalized gradient descent as the optimization method [67] ("AV-Dispersion" curve). The results show that our proposed approach (the "AV-Proposed Approach curve") improves AVSR performance.

One interesting result of this comparison is that the proposed approach seems particularly effective at lower SNRs. In fact, these plots show the good performance at medium to high SNR of the Bayesian fusion ("AV-Unweighted"), which does not require any weighting and hence no reliability estimation either. As can be seen in Figure 6.6, the performance of the Bayesian fusion is very close to the weighted schemes for medium and good SNR values, whereas for low SNR values the performance drastically degrades. This is due to the fact that Bayesian fusion is able to capture, in an implicit way, the varying reliability of the input streams. Decreasing the reliability of a stream results in an increase of the entropy of the corresponding classification results. Consequently, the distribution of these values flattens and in an extreme case reaches the uniform distribution when the stream

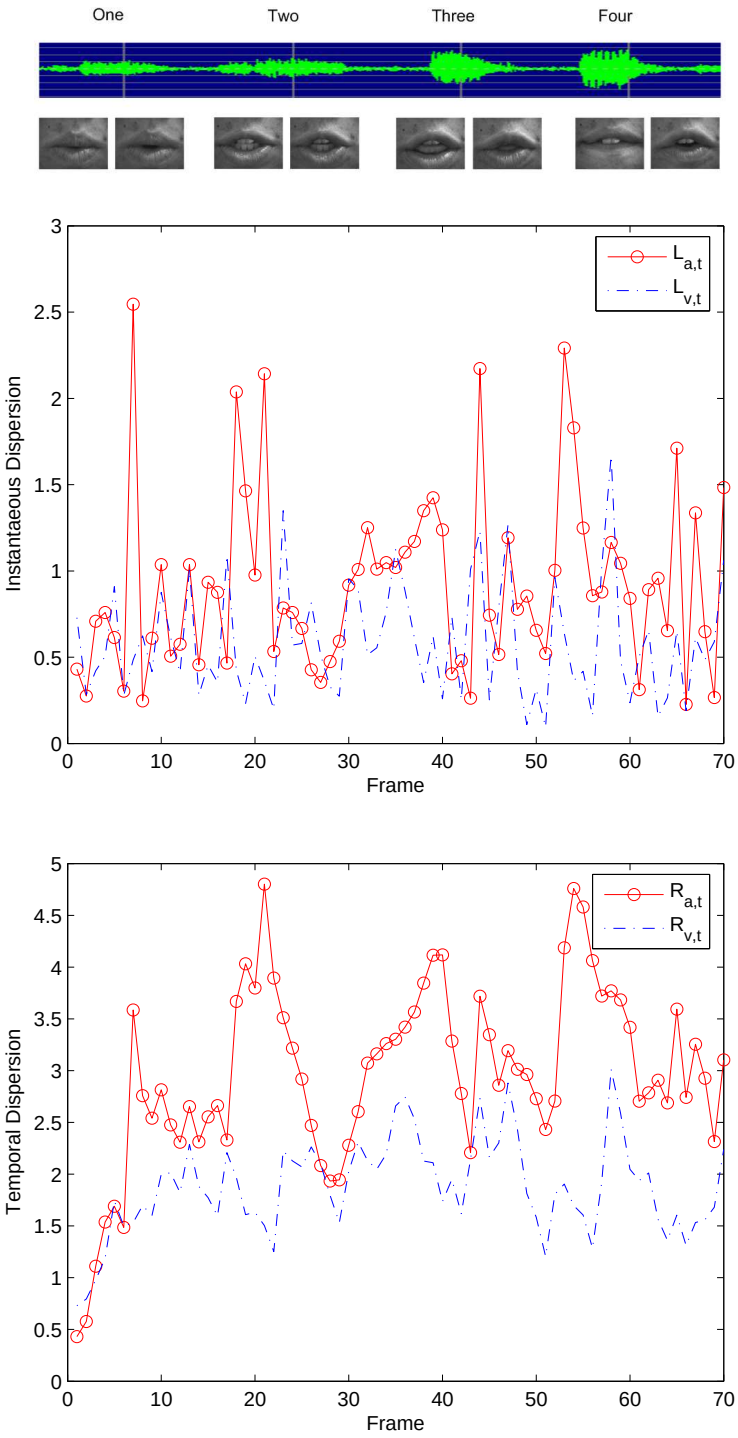

Figure 6.4: Instantaneous (top) and temporal (bottom) dispersion variation at 10dB.

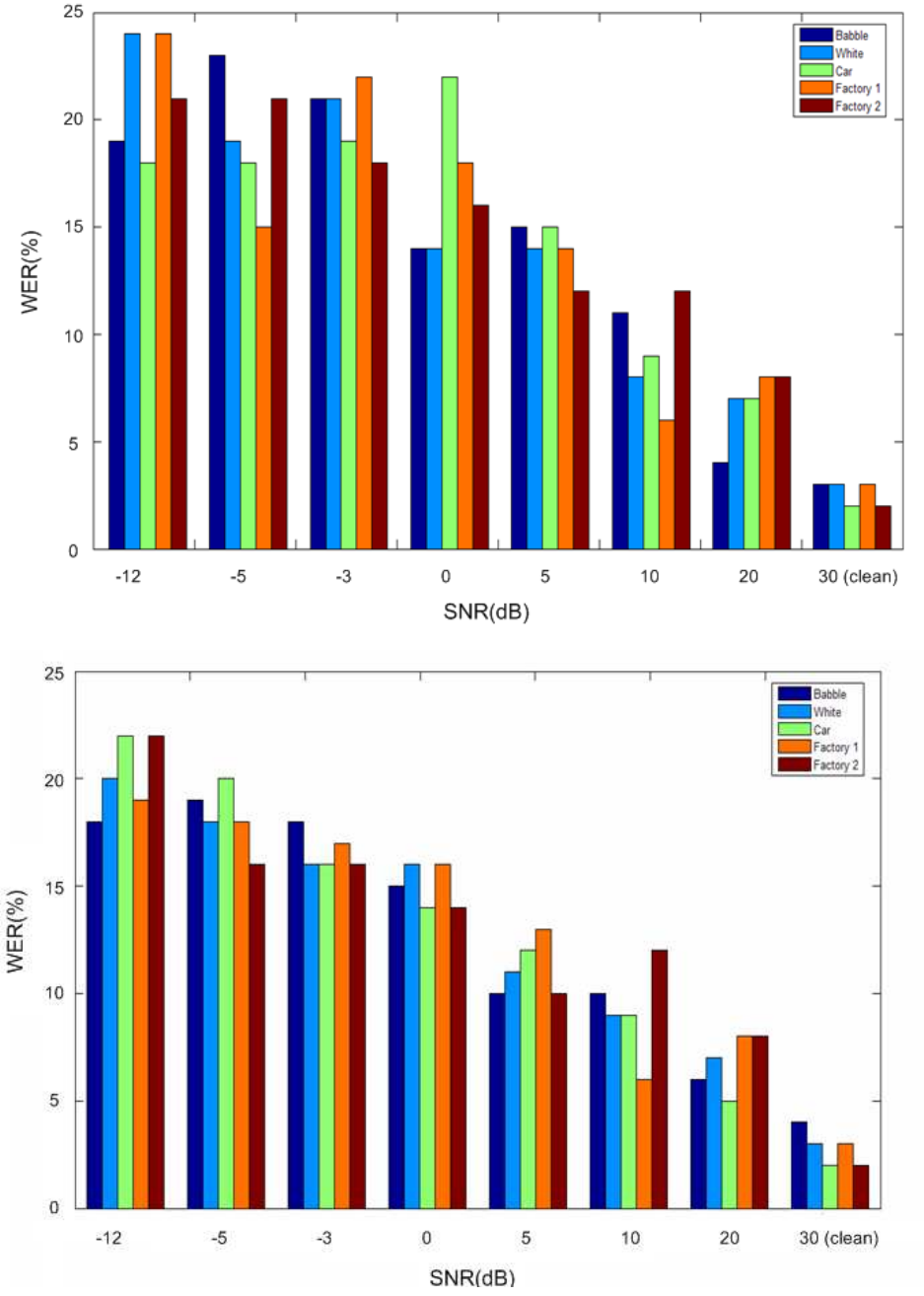

Figure 6.5: WER for dispersion-GPD (top) and proposed approach (bottom).

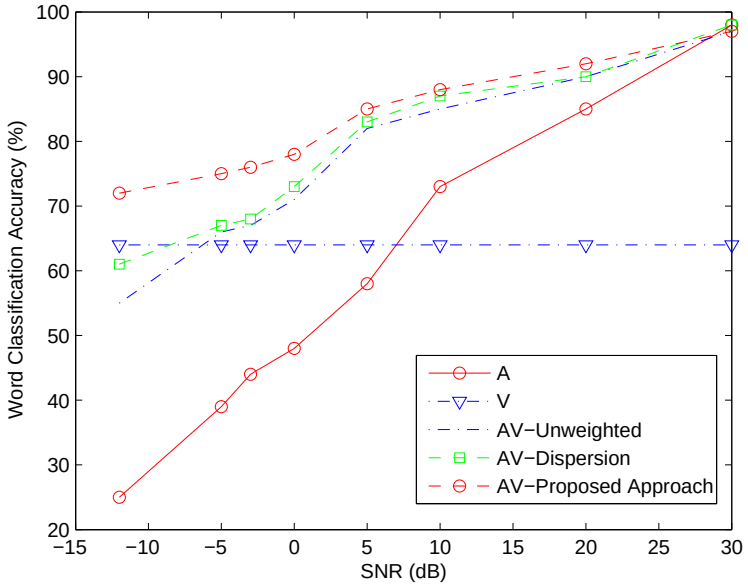

Figure 6.6: Word classification accuracy.

is totally unreliable. During fusion the uniform distribution does not interfere with the distribution of the reliable input stream as the product of the uniform distribution does not change the shape of the second distribution. Therefore, classification is not altered by the unreliable stream.

However, at very low SNR levels, the unweighted Bayesian fusion scheme performs even worse that the visual modality alone. This is referred to in literature as *catastrophic fusion* [55]. Catastrophic fusion in intelligent sensory systems happens when the accuracy of the fused outcome is less than the accuracy of both individual systems alone. Movellan and Mineiro [55], argue that most fusion systems suffer from catastrophic fusion because they make implicit assumptions and degenerate quickly when these assumptions are broken and used outside its original context. The training in this work was done for subjects in a clean environment. This means that for very low SNR values (such as -12dB), the context of training is totally different. Consequently, both the unweighted Bayesian fusion and the dispersion-based fusion suffer from catastrophic fusion. However, the proposed approach, having taken into consideration channel information within a certain interval of time, is

able to capture context information and hence does not suffer from catastrophic fusion.

It is clear now that both dispersion-based fusion and fusion using the proposed approach significantly improve AVSR performance at low SNRs, with the proposed approach being somewhat superior as it combats catastrophic fusion. For example, at -12dB SNR, the fusion method using the proposed approach produces a 72% word classification accuracy, representing a vast improvement over the audio-only rate of 25%, the unweighted fusion rate of 55%, the dispersion-based fusion rate of 61% and the video-only rate of 64%. Notice however that at the high end of the SNR range, all the recognition methods, except for the video-only, reveal a similar performance. To further illustrate quantitatively the performance of the proposed approach to fusion, we compare fusion strategies in terms of their resulting *effective SNR gain*. We measure this gain with reference to the audio-only word classification accuracy at 10dB, by considering the SNR value where the audiovisual word classification rate equals the reference audio-only word classification rate. From Figure 6.6, this SNR gain is around 10dB for both the unweighted Bayesian fusion and the dispersion-based fusion. On the other hand, classification based on the proposed approach achieves a 16dB improvement, further illustrating the efficiency of this approach.

6.7 Chapter Summary and Discussion

In this chapter, we proposed a new probabilistic reliability assessment model for multiple streams in a multimodal system. The main benefit of this assessment model is that it takes into consideration the reliability of the overall system on both a local and global level and thus is robust to sudden noise bursts. In addition, it is a model, which can be generalized for multiple information streams and multiple applications. We developed two stream reliability indicators based on the dispersion of N-best hypotheses in each modality. The first reliability indicator, the instantaneous dispersion, was simply the dispersion of the a posteriori probabilities of the observation vectors. Thus, large differences in probabilities equate to greater certainty, close probabilities to less certainty. The second reliability indicator, the temporal dispersion, was based on a linear combination of the first indicator measure over a time interval. This indicator was depicted as an autoregressive model and thus its parameters were calculated using the Yule-Walker equations.

The reliability indicators were then mapped into stream weights using the genetic algorithm, in such a way that maximized the conditional likelihood. This optimal scheme is superior to previous approaches because it is dynamic, easy to implement, and considers an arbitrary number of streams (instead of just 2 as is usually done for AVSR scenarios). Experimental results did show improvements, especially at low SNR levels, where it was able to combat catastrophic fusion. We demonstrated a significant improvement of 16dB word classification accuracy relative improvement over audio-only matched models at a 10dB SNR. Future work can extend this architecture to consider multiple streams of information on both an intramodal and intermodal level.

Chapter 7

Conclusion and Future Work

This work provided a framework for audio-visual speech recognition that can effectively maximize information gather about the words uttered and minimize the impact of noise. Two main issues that are relevant to the design of AVSR systems were addressed. The first issue is the visual front end that captures visual speech information. The second issue is the integration (fusion) of audio and visual features into the automatic speech recognizer used. Both are challenging problems, and significant research effort has been directed towards finding appropriate solutions for them.

7.1 Summary and Contributions

7.1.1 Visual Speech Modeling and Feature Extraction

This work first proposed a visual front-end system that processes visual speech information and extracts features that are representative of the speech classes. Visual speech is complex in its nature as it includes a lot of non-rigidity, self-occlusion and high image velocities. For this purpose, a motion model of the mouth movement was developed using optical flow analysis. The optical flow fields were derived between every pair of consecutive video frames. Independent component analysis was then performed on the optical flow fields to find a basis set that is statistically independent. The temporal trajectory of the model coefficients provided a rich description of the visual speech information. Next,

the coefficients of these basis flow fields were used to recognize units of speech chosen here to be the visemes.

To motivate this approach, it is important to mention that for natural images, such as the mouth, much of the speech information is contained in the higher order statistics (the phase spectrum), with a little knowledge provided by second-order statistics (the amplitude spectrum). The fact that PCA derives only second order statistics makes it an inefficient tool for our application. For this purpose, we have used Bell and Sejnowski's ICA, which is based on the principal of optimal information flow in sigmoidal neurons by minimizing output joint mutual information. PCA is a special case of ICA in which the source models are Gaussian. The assumption that sources are Gaussian is not necessarily true, especially for natural signals such as speech, EEG and natural images. Consequently, ICA provides a better probabilistic model of the data and derives a basis set that can reconstruct the data better than PCA in the presence of noise. In addition, it uniquely identifies the mixing matrix, and captures higher-order statistics that are not provided by the covariance matrix.

For recognition we first created a temporal Viterbi lattice containing as many states (N) as the optical flow frames in the video sequence. We then used an N-state HMM based on this Viterbi lattice to classify the visemes and the corresponding words. Experimental results showed good improvement (Section 5.9). We found that the classification results had a 92.7% word recognition rate showing an improvement by 1% compared to the best performing state of the art system. Moreover, the minimum value of the confidence interval as found by our approach was found to be 87.5% which is larger than the best value obtained in [33] and in [56] by 3.1%. We also showed that the first 5 eigenvectors (those corresponding to the largest eigenvalues) of the input flow fields account for 95% of the variance in the data. This means that the chosen 12 basis optical flow fields are highly representative of the lips motion model.

7.1.2 Reliability-Driven Sensor Fusion

Another major contribution of this work was in developing an assessment model that measures the reliability of the audio and visual information streams to weight the influence of the decisions in the combination. Reliability of the audio and visual streams can be obtained by measures of the signal (such as the amount of noise using SNR) or by statistical

approaches. In this work, a statistical approach was developed, which accounts for the dynamic changes in reliability. This was done in two steps. The first step derived suitable statistical reliability measures for the individual information streams. The second step found an optimal mapping between the reliability measures and the stream weights that maximized information gather.

For the purpose of the first step, this work proposed two stream reliability indicators based on the dispersion of N-best hypotheses in each modality. The first reliability indicator, the instantaneous dispersion, was simply the dispersion of the a posteriori probabilities of the observation vectors. Thus, large differences in probabilities equate to greater certainty, close probabilities to less certainty. The second reliability indicator, the temporal dispersion, was based on a linear combination of the first indicator measure over a time interval. This indicator was depicted as an autoregressive model and thus its parameters were calculated using the Yule-Walker equations. This assessment model takes into consideration the reliability of the overall system on both a local and a global level, and can thus overcome problems related to sudden noise bursts.

The reliability indicators were then mapped into stream weights using the genetic algorithm, in such a way that maximized the conditional likelihood. The proposed approach did show improvements (Section 6.6.2), especially at low SNR levels, where it was able to combat catastrophic fusion. We demonstrated a significant improvement of 16dB word classification accuracy relative improvement over audio-only matched models at a 10dB SNR. This improvement reached its maximum at -12dB, where the proposed fusion produced a 72% word classification accuracy, representing a vast improvement over the audio-only rate of 25%, the unweighted fusion rate of 55%, the dispersion-based fusion rate of 61% and the video-only rate of 64%. The improvement, however, was minimum at high SNRs yielding a 97% word classification accuracy, which is comparable to the other unimodal and bimodal approaches.

7.2 Future Work

This work clearly shows that audio-visual speech recognition is a wide area that has been explored by many researchers over the past two decades. However, issues of both practical

and research nature remain challenging, and much progress still needs to be done for capturing and integrating visual speech information.

From a practical point of view, the high quality of captured visual data, which is needed for extracting visual speech information capable of enhancing AVSR performance, is coupled with an increased cost, storage, and computer processing requirements. Furthermore, the need for a large common audio-visual corpora that is capable of capturing a wide range of contexts and speaker variabilities hinders the development of practical and robust AVSR systems.

On the research side, many substantial issues related to the design of the low-level components of AVSR systems, remain challenged and open for investigation. In the visual front end design, for example, developing mouth detection, facial feature extraction, and lips tracking algorithms that are robust to speaker, pose, lighting and environment conditions remain challenging problems. Moreover, a combined shape and appearance based three-dimensional modeling for lips tracking and visual feature extraction has not yet been addressed in the AVSR community, although such an approach would achieve the desired robustness needed for the visual front-end design. In addition, a thorough comparison between shape and appearance based features for the mouth region of interest has not been explored. Audio-visual decision fusion also has a number of issues that require further study. For example, the optimal level of integrating the audio and visual log-likelihoods, the optimal function for this integration, as well as the derivation of appropriate local estimates of the reliability of each stream into this function, are all problems that need to be tackled. Further investigation of these issues is expected to lead to improved robustness and performance of AVSR systems.

Most of the problems investigated in this research will benefit other areas of research, such as speaker identification and verification, visual text-to-speech, speech event detection, video indexing and retrieval, speech enhancement, coding, signal separation, and speaker localization. Coupled with improvements in practical and research issues, these areas will bring audio-visual speech recognition one step forward towards a commercial, natural and robust human-computer interface.

Bibliography

[1] The carnegie mellon university pronouncing dictionary v. 0.6, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[2] A. Adjoudani and C. Benoît. On the integration of auditory and visual parameters in an HMM-based ASR. *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*, pages 461–471, 1996.

[3] H. Ai, L. Liang, and G. Xu. Face detection based on template matching and support vector machines. *Proceedings of the International Conference on Image Processing*, 1, 2001.

[4] JL Barron, DJ Fleet, and SS Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.

[5] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[6] A.J. Bell and T.J. Sejnowski. The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[7] M.J. Black and P. Anandan. The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, 1996.

[8] M. Brand, N. Oliver, and A. Pentland. Coupled hidden Markov models for complex action recognition. *Computer Vision and Pattern Recognition*, pages 994–999, 1997.

[9] C. Bregler and Y. Konig. "Eigenlips" for robust speech recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP-94*, 2, 1994.

[10] U. Bub, M. Hunke, and A. Waibel. Knowing who to listen to in speech recognition: visually guided beamforming. *International Conference on Acoustics, Speech, and Signal Processing. ICASSP-95*, 1, 1995.

[11] D. Chandramohan and PL Silsbee. A multiple deformable template approach for visual speech recognition. *Proceedings of the Fourth International Conference on Spoken Language. ICSLP 96*, 1, 1996.

[12] C. Chatfield and A.J. Collins. *Introduction to multivariate analysis*. Chapman & Hall, 1980.

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

[14] T. Chen and RR Rao. Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, 86(5):837–852, 1998.

[15] GI Chiou and J.N. Hwang. Lipreading from color video. *IEEE Transactions on Image Processing*, 6(8):1192–1195, 1997.

[16] A. Cichocki, R. Unbehauen, and E. Rummert. Robust learning algorithm for blind separation of signals. *Electronics Letters*, 30(17):1386–1387, 1994.

[17] MM Cohen and DW Massaro. What can visual speech synthesis tell visual speech recognition? *Conference on Signals, Systems and Computers, 1994. Conference Record of the Twenty-Eighth Asilomar*, 1.

[18] R. Cole, J. Beskow, J. Yang, U. Meier, A. Waibel, P. Stone, A. Davis, C. Soland, G. Fortier, T. Carmell, et al. Intelligent animated agents for interactive language training. *ACM SIGCAPH Computers and the Physically Handicapped*, pages 5–10, 1998.

[19] P. Comon et al. Independent component analysis, a new concept. *Signal Processing*, 36(3):287–314, 1994.

[20] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285, 2001.

[21] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[22] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[23] E. Cosatto, G. Potamianos, and HP Graf. Audio-visual unit selection for the synthesis of photo-realistic talking-heads. *IEEE International Conference on Multimedia and Expo. ICME 2000.*, 2, 2000.

[24] B. Dalton, R. Kaucic, and A. Blake. Automatic speechreading using dynamic contours. *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*, pages 373–382, 1996.

[25] P. Duchnowski, U. Meier, and A. Waibel. See me, hear me: integrating automatic speech recognition and lipreading. To appear in proc.

[26] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151, 2000.

[27] D.J. Fleet, M.J. Black, Y. Yacoob, and A.D. Jepson. Design and use of linear models for image motion analysis. *International Journal of Computer Vision*, 36(3):171–193, 2000.

[28] H. Glotin, D. Vergyr, C. Neti, G. Potamianos, J. Luettin, and G. ICP. Weighting schemes for audio-visual fusion in speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. (ICASSP'01).*, 1, 2001.

[29] D.E. Goldberg. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1989.

[30] M. Gordan, C. Kotropoulos, and I. Pitas. A support vector machine-based dynamic network for visual speech recognition applications. *EURASIP Journal on Applied Signal Processing*, 2002(11):1248–1259, 2002.

[31] JN Gowdy, A. Subramanya, C. Bartels, and J. Bilmes. DBN based multi-stream models for audio-visual speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. (ICASSP'04).*, 1, 2004.

[32] HP Graf, E. Cosatto, and M. Potamianos. Robust recognition of faces and facial features with a multi-modal system. *IEEE International Conference on Systems, Man, and Cybernetics, 1997. 'Computational Cybernetics and Simulation'*, 3, 1997.

[33] M.S. Gray, T.J. Sejnowski, and J.R. Movellan. A comparison of image processing techniques for visual speech recognition applications. *Advances in Neural Information Processing Systems*, 13:939–945, 2001.

[34] S. Gurbuz, Z. Tufekci, E. Patterson, and JN Gowdy. Application of affine-invariant Fourier descriptors to lipreading for audio-visual speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. (ICASSP'01).*, 1, 2001.

[35] M.H. Hayes, MH Hayes, and M.H. Hayes. *Statistical digital signal processing and modeling*. John Wiley & Sons, Inc. New York, NY, USA, 1996.

[36] M. Heckmann, F. Berthommier, and K. Kroschel. A hybrid ANN/HMM audio-visual speech recognition system. *Proceedings of AVSP-2001*, 2001.

[37] J.H. Holland. *Adaptation in natural and artificial systems*. MIT Press Cambridge, MA, USA, 1992.

[38] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.

[39] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(1):4–37.

[40] F.V. Jensen. *Bayesian networks and decision graphs*. Springer.

[41] F.V. Jensen, FVV Jensen, and FV Jensen. *Introduction to Bayesian networks*. Springer-Verlag New York, Inc. Secaucus, NJ, USA, 1996.

[42] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. *Pattern Recognition Letters*, 18(9):853–858, 1997.

[43] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[44] G. Krone, B. Talle, A. Wichert, and G. Palm. Neural architectures for sensor fusion in speech recognition. *Proc. Europ. Tut. Works. Audio-Visual Speech Processing*, pages 57–60.

[45] L. Liang, X. Liu, Y. Zhao, X. Pi, and A.V. Nefian. Speaker independent audio-visual continuous speech recognition. *IEEE International Conference on Multimedia and Expo*, pages 25–28, 2002.

[46] S. Lucey, Queensland University of Technology School of Electrical, and Electronic Systems Engineering. *Audio-visual speech processing*. Queensland University of Technology, Brisbane, 2002.

[47] J. Luettin, G. Potamianos, C. Neti, and A.S. AG. Asynchronous stream modeling for large vocabulary audio-visual speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. (ICASSP'01).*, 1, 2001.

[48] J. Luettin and N.A. Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, 1997.

[49] D.J.C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. *Report, University of Cambridge, Cavendish Lab*, 1996.

[50] D.W. Massaro and D.G. Stork. Speech recognition and sensory integration. *American Scientist*, 86(3):236–244, 1998.

[51] I. Matthews, TF Cootes, JA Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):198–213, 2002.

[52] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel. Towards unrestricted lip reading. *International Journal of Pattern Recognition and Artificial Intelligence*, 14(5):571–585, 2000.

[53] J.R. Movellan. Visual speech recognition with stochastic networks. *Advances in Neural Information Processing Systems*, 7, 1995.

[54] JR Movellan and G. Chadderdon. Channel separability in the audio-visual integration of speech: a Bayesian approach. *Stork & Hennecke*, pages 473–487, 1996.

[55] J.R. Movellan and P. Mineiro. Robust sensor fusion: analysis and application to audio-visual speech recognition. *Machine Learning*, 32(2):85–100, 1998.

[56] J.R. Movellan, P. Mineiro, and R.J. Williams. Partially observable SDE models for image sequence recognition tasks. *Advances in Neural Information Processing Systems*, 13:880–886.

[57] A.V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing*, 2002(11):1274–1288, 2002.

[58] AV Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy. A coupled HMM for audio-visual speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. (ICASSP'02).*, 2, 2002.

[59] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri. Large-vocabulary audio-visual speech recognition: a summary of the Johns Hopkins Summer 2000 Workshop. *IEEE Fourth Workshop on Multimedia Signal Processing*, pages 619–624, 2001.

[60] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio-visual speech recognition. *Final Workshop 2000 Report*, 764, 2000.

[61] V.I. Pavlovic. *Dynamic Bayesian networks for information fusion with applications to human-computer interfaces*. PhD thesis, University of Illinois at Urbana-Champaign, 1999.

[62] B.A. Pearlmutter and L.C. Parra. A context-sensitive generalization of ICA. *International Conference on Neural Information Processing*, 151, 1996.

[63] ED Petajan. Automatic lipreading to enhance speech recognition. *Dissertation Abstracts International Part B: Science and Engineering*, 45(11), 1985.

[64] G. Potamianos, HP Graf, and E. Cosatto. An image transform approach for HMM based automatic lipreading. *Proceedings of the International Conference on Image Processing. ICIP 98.*, pages 173–177, 1998.

[65] G. Potamianos and C. Neti. Stream confidence estimation for audio-visual speech recognition. *Proceedings of the International Conference on Spoken Language Processing*, 3:746–749, 2000.

[66] G. Potamianos and C. Neti. Improved ROI and within frame discriminant features for lipreading. *Proceedings of the International Conference on Image Processing.*, 3, 2001.

[67] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior. Recent advances in the automatic recognition of audio-visual speech. *Proc. IEEE*, 91(9):1306–1326, 2003.

[68] G. Potamianos and A. Potamianos. Speaker adaptation for audio-visual speech recognition. *Proc. European Conference on Speech Communication and Technology EUROSPEECH*, 3:12911294.

[69] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu. A cascade image transform for speaker independent automatic speechreading. *IEEE International Conference on Multimedia and Expo. ICME 2000.*, 2, 2000.

[70] L. Rabiner and B.H. Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1993.

[71] C.R. Rao. Linear statistical inference and its applications. *New York*, 1973.

[72] A. Rogozan. Discriminative learning of visual data for audiovisual speech recognition. *International Journal on Artificial Intelligence Tools*, 8(1):43–52, 1999.

[73] A. Rogozan and P. Deleglise. Adaptive fusion of acoustic and visual sources for automatic speech recognition. *Speech Communication*, 26(1):149–161, 1998.

[74] HA Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[75] K. Saenko, T. Darrell, and J. Glass. Articulatory features for robust visual speech recognition.

[76] AW Senior. Face and feature finding for a face recognition system. *Proc. Int. Conf. Audio and Video-based Biometr. Person Authent*, pages 154–159, 1999.

[77] S. Tamura, K. Iwano, and S. Furui. A robust multi-modal speech recognition method using optical-flow analysis. *Proc. IDS02*, pages 2–4, 2002.

[78] P. Teissier, J. Robert-Ribes, J.L. Schwartz, and A. Guerin-Dugue. Comparing models for audio-visual fusion in a noisy-vowel recognition task. *IEEE Transactions on Speech and Audio Processing*, 7(6):629–642, 1999.

[79] M.A. Turk, A. Pentland, Vision, Modeling Group, Massachusetts Institute of Technology, and Media Laboratory. *Eigenfaces for recognition*. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology, 1991.

[80] A. Varga, HJM Steeneken, M. Tomlinson, and D. Jones. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. *DRA Speech Research Unit, Malvern, England, Tech. Rep*, 1992.

[81] AP Varga and RK Moore. Hidden Markov model decomposition of speech and noise. *International Conference on Acoustics, Speech, and Signal Processing. ICASSP-90.*, pages 845–848, 1990.

[82] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. The HTK book. *Cambridge University*, 1996, 1995.

[83] BP Yuhas, MH Goldstein Jr, and TJ Sejnowski. Integration of acoustic and visual speech signals using neural networks. *Communications Magazine, IEEE*, 27(11):65–71, 1989.

[84] X. Zhang, C.C. Broun, R.M. Mersereau, and M.A. Clements. Automatic speechreading with applications to human-computer interfaces. *EURASIP Journal on Applied Signal Processing*, 2002(11):1228–1247, 2002.

[85] D.N. Zotkin, R. Duraiswami, and L.S. Davis. Joint audio-visual tracking using particle filters. *EURASIP Journal on Applied Signal Processing*, 2002(11):1154–1164, 2002.