Visual Speech And Speaker Recognition

Juergen Luettin
Department of Computer Science
University of Sheffield

Dissertation submitted to the University of Sheffield
for the degree of Doctor of Philosophy

May 1997

Summary

This thesis presents a learning based approach to speech recognition and person recognition from image sequences.

An appearance based model of the articulators is learned from example images and is used to locate, track, and recover visual speech features. A major difficulty in model based approaches is to develop a scheme which is general enough to account for the large appearance variability of objects but which does not lack in specificity. The method described here decomposes the lip shape and the intensities in the mouth region into weighted sums of basis shapes and basis intensities, respectively, using a Karhunen-Loéve expansion. The intensities deform with the shape model to provide shape independent intensity information. This information is used in image search, which is based on a similarity measure between the model and the image.

Visual speech features can be recovered from the tracking results and represent shape and intensity information. A speechreading (lip-reading) system is presented which models these features by Gaussian distributions and their temporal dependencies by hidden Markov models. The models are trained using the EM-algorithm and speech recognition is performed based on maximum posterior probability classification.

It is shown that, besides speech information, the recovered model parameters also contain person dependent information and a novel method for person recognition is presented which is based on these features. Talking persons are represented by spatio-temporal models which describe the appearance of the articulators and their temporal changes during speech production. Two different topologies for speaker models are described: Gaussian mixture models and hidden Markov models.

The proposed methods were evaluated for lip localisation, lip tracking, speech recognition, and speaker recognition on an isolated digit database of 12 subjects, and on a continuous digit database of 37 subjects. The techniques were found to achieve good performance for all tasks listed above. For an isolated digit recognition task, the speechreading system outperformed previously reported systems and performed slightly better than untrained human speechreaders.

Acknowledgements

First, I would like to thank my supervisor Neil Thacker. He has been a great source of inspiration and has provided the right balance of suggestions, criticism, and freedom. I also would like to thank Steve Renals for taking over the role of supervisor during the last year of my research and for his support during that time.

Many thanks go to Neil Thacker and Tony Lacey for providing the outstanding computer vision research tool TINA and for their technical support which made my work much easier. I also have to thank Neil Powell and Georg Thimm for ensuring the smooth running of the computing resources and especially Georg Thimm for his help with various kinds of software.

Special thanks go to Gilbert Maitre for providing a wonderful working atmosphere at the Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP) in Switzerland, where the final year of this research was carried out. It is hard to imagine a nicer place to write a thesis than at IDIAP, where the social interaction of the members as well as skiing, trekking, and hiking provide the right balance for successful research.

I am grateful to those people who critically read drafts of this thesis – Neil Thacker, Steve Renals, Hervé Bourlard and Gerard Chollet. I would also like to thank my examiners Martin Cooke and David Hogg for critical reading of the thesis and for their valuable comments.

Finally, I would like to thank the German Academic Exchange Service (DAAD) for funding the first year of my research and the University of Sheffield for providing the subsequent scholarship. Part of this work was performed at IDIAP under the European ACTS-M2VTS project and I thank the Swiss Office for Science and Education (SOSE) for their financial support of this project.

Contents

List of Figures

List of Tables

Glossary

ABM	Appearance Based Model
AI	Articulation Index
AOI	Area Of Interest
ASM	Active Shape Model
ASR	Automatic Speech Recognition
AVSR	Audio-Visual Speech Recognition
CDHMM	Continuous Density Hidden Markov Model
DHMM	Discrete Hidden Markov Model
DSM	Downhill Simplex Method
DTW	Dynamic Time Warping
EI	Early Integration
EM	Expectation-Maximisation
FLMP	Fuzzy Locical Model of Perception
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
K-L	Karhunen-Loéve
LI	Late Integration
LP	Linear Prediction
MAP	Maximum A Posteriori
MFCC	Mel Frequency Cepstrum Coefficients
ML	Maximum Likelihood
PCA	Principal Component Analysis
PDM	Point Distribution Model
s.d.	standard deviation
SNR	Sound Noise Ratio
TD	Text Dependent
TI	Text Independent
VOT	Voice-Onset Time
VPAM	Vision provides Place, Acoustic provides Manner of articulation
VSR	Visual Speech Recognition

Notation

$\mathbf{O} = \mathbf{o}(1), \mathbf{o}(2), \ldots, \mathbf{o}(T)$: sequence of observation vectors.

$\mathbf{O}^a$: acoustic observation sequence.

$\mathbf{O}^v$: visual observation sequence.

$\mathbf{o}(\tau) = [o_1(\tau), o_2(\tau), \ldots, o_n(\tau)]^T$: observation vector at time τ.

$\Delta\mathbf{o}(\tau)$: delta observation vector.

$\widehat{\Lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_M\}$: set of possible speech unit HMMs.

$\mathbf{s} \ = \ \{s_1, s_2, \ldots, s_N\}$: set of HMM states.

$\mathbf{A} = \{a_{ij}\}$: matrix of state transition probabilities from state i to state j.

$\mathbf{B}$: vector of observation probabilities. $b_j(\mathbf{o}(\tau))$ associated with each emitting state j.

π: vector with the initial state probabilities π_i of entering the model at state i.

c_{im}: mixture weight for state i and mixture m.

$\mathcal{N}(\mathbf{o}(\tau), \mu, \Sigma)$: multivariate Gaussian with mean μ and covariance matrix Σ.

$P(\lambda|\mathbf{O})$: posterior probability of model λ given the observation $\mathbf{O}$.

$P(\mathbf{O}|\Lambda)$: likelihood of observation $\mathbf{O}$ given model λ.

$P(\lambda)$: prior probability of model λ.

$P(\mathbf{O})$: prior probability of the observation sequence $\mathbf{O}$.

$\alpha_i(t)$: forward probability.

$\beta_i(t)$: backward probability.

$\phi_j(t)$: partial maximum likelihood of observing the observation sequence up to time t and being in state j.

$\mathcal{H}(\Lambda, \hat{\Lambda})$: auxiliary function of the old parameter set Λ and the new set $\hat{\Lambda}$, used in the EM algorithm.

$L_{jm}(\tau)$: probability of being in state j and mixture m at time τ.

$\mathbf{v}_i = (x_0, y_0, x_1, y_1, \ldots, x_{N_s-1}, y_{N_s-1})^T$: shape vector where (x_j, y_j) are the coordinates of the j^{th} point $(j = 0 \ldots N-1)$.

$\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_T)$: matrix of the T most dominant column eigenvectors $\mathbf{p}_t$.

$\mathbf{b} = (b_1, b_2, \ldots, b_t)$: vector of eigenvector weights or model parameters.

$\tilde{\mathbf{b}} = (\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_t)$: normalised model parameters.

$\mathbf{h}_i = (\mathbf{g}_{i0}, \mathbf{g}_{i1}, \ldots, \mathbf{g}_{iN_s-1})^T$: global intensity vector for image i

g_{ij}: local intensity vector for image i at the model point j.

$M(s, \theta)[\mathbf{x}]$: performs a rotation by θ and a scaling by s of $\mathbf{x}$.

$\Omega = (t_x, t_y, s, \theta, \mathbf{b_s}, \mathbf{b}_i)$: lip model with translation t_x and t_y, scale s, angle θ, shape weights b_s and intensity weights b_i.

E: cost between the image and the model, used in image search.

E_g: negative sum of gradient magnitudes.

$E_{n,m}$: residual error for the modes n to m.

E_i^2: sum of square errors.

$f(t_x, t_y, s, \theta, \mathbf{b}_s)$: image search function.

I_i: irradiance or brightness.

I_r: radiance or illumination.

$\hat{W}_i = \{w_{i1}, w_{i2}, \ldots, w_{iN}\}$: set of word models for person i.

Chapter 1

Introduction

Human speech perception is inherently a multi-modal process, which involves the analysis of the uttered acoustic signal and which includes higher level knowledge sources such as grammar, semantics, and pragmatics. One information source which is mainly used in the presence of acoustic noise is lipreading or so-called speechreading[1]. It is well known that seeing the talker's face in addition to audition can improve speech intelligibility, particularly in noisy environments.

Automatic speech recognition (ASR) by machine has been an active research area for several decades, but in spite of the enormous efforts, the performance of current ASR system is far from the performance achieved by humans. Most state-of-the-art ASR systems make use of the acoustic speech signal only and ignore the visual speech cues. They are therefore susceptible to acoustic noise, and essentially all real-world applications are subject to some kind of noise. Although several audio-visual speech recognition (AVSR) systems have been proposed in the literature, most systems were dependent on constrained visual environments which prevents their use for real-world applications. The aim of the speechreading system described here is to perform speaker independent speech recognition for a large variety of speakers and without the use of visual aids.

Whereas speech recognition is concerned with the problem of finding an estimate of the word sequence given the speech signal, the task of speaker recognition is to recognise the person that uttered the phrase. In analogy to the audio-visual perception of speech, human identification of persons is a multi-modal process too, of which the analysis of the face and of the voice are two of the most dominant modalities. In the recognition of persons by machine, these two modalities have usually been treated separately, with visual face recognition and acoustic speaker recognition being two of the main approaches. The performance of a unimodal system can usually be increased by combining the information of several classifiers and recently, a few studies have been performed on the integration of face recognition and speaker recognition system scores. In these approaches, face recognition was based on a static face image, which was recorded before the person was speaking.

[1] Lipreading is the perception of speech purely based on observing the talkers lip movements. Speechreading is the visual perception of speech which also includes observation of facial and manual gestures. Audio-visual speech perception is the perception of speech by combining speechreading with audition.

Here, a new modality for person identification is introduced, which uses the temporal image sequence of the talking person for identification.

The aim of the work presented in this thesis can be divided into three tasks:

- Robust lip localisation, lip tracking, and feature extraction in grey-level image sequences of various subjects.
- Speech recognition purely based on visual spatio-temporal information.
- Speaker recognition purely based on visual spatio-temporal information.

Although the emphasis of the work is on visual speech and speaker recognition, the problem of audio-visual speech and speaker recognition will also be addressed briefly.

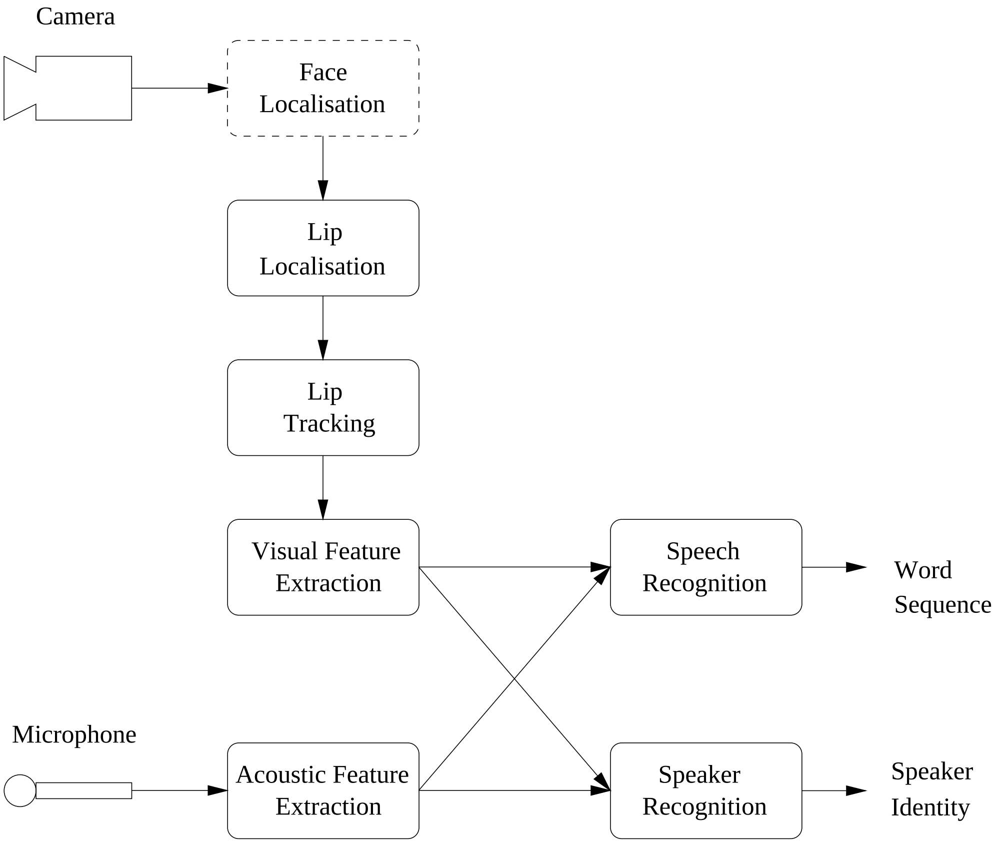


Figure 1.1: Overview of the complete system for audio-visual speech and speaker recognition. In this work, all components of the system are considered except the module for face localisation.

Figure 1.1 displays the structure of an audio-visual speech and speaker recognition system. The image and the speech of the talking person are recorded by a camera and a microphone, respectively. Visual and acoustic features are extracted from these signals and serve as input to a speech or speaker recognition module. The face localisation module will not be considered in this work. All experiments were performed on two audio-visual databases: the Tulips1 database, which contains recordings of digits, spoken in isolation by 12 persons, and the M2VTS database, which contains recordings of continuously spoken digits from 37 persons.

1.1 Feature Extraction

Both visual speech and speaker recognition systems require some sort of features to be extracted from the image sequence which provide the desired information. In the case of speech recognition, the features should contain information about the different speech units spoken, while in the case of speaker recognition, they should provide information about the identity of the person. Facial speech feature extraction and modelling has become an important issue in both, automatic speech processing and automatic face processing. Potential applications include audio-visual speech recognition [143], recognition of talking persons [110], lip synchronisation [34], speech-driven talking heads [140], and speech-based image coding [134].

Facial feature extraction is a difficult problem due to the large appearance variability across different persons and due to appearance variability during speech production. Different illumination and different pose of the face cause further difficulties in the image analysis. For a real-world application, whether it is in a car, an office or a factory, the system has to be able to deal with these kinds of image variability. The robustness of algorithms to such variability, however, has received only little attention in the research community, and was often circumvented by marking the subjects lips with a reflective marker, by using very controlled lighting conditions for the recordings, or by restricting the experiments to one subject only.

Whereas the extraction of acoustic speech features is fairly established, it is not well known which visual speech features are important for speechreading and the investigation of different feature extraction methods is still subject to much ongoing research. In this work, important visual features are determined from psychological and physiological studies, then an appearance based model aiming to represent these features is developed. The emphasis of the method is to provide a robust feature extraction method which performs well for different subjects and without the use of visual aids.

1.2 Visual Speech Recognition

There has been much progress in automatic speech recognition over the past decade and state of the art systems perform very well in controlled lab environments. However, once these systems are applied to real-world environments, where background noise or cross-talk are present, their performance degrades significantly [78]. Such application environments are for example an office, car, factory or aircraft and essentially all of these are subject to some interfering noise. Much research effort in automatic speech recognition has therefore been directed towards systems for noisy speech environments and the robustness of speech recognition systems has been identified as one of the biggest challenges in future research [41].

Noise sources can be divided into environmental noise, speech related noise, altered speaking manner, and channel/microphone related noise. Environmental noise can be either continuous, such as fans and engines, or non-continuous, such as cars passing, telephone ringing, and competing speech. Speech related noise is dependent on the speech signal and the room acoustics and can be caused by reverberation and reflection. The effect of altered speaking manner includes the Lombard effect, which describes how speakers alter their speech due to stress or background noise, and other effects like differing speaking rate, pitch, stress and breath. Finally, channel/microphone related noise can be due to the use of different microphones, different filter characteristics, and limited frequency bandwidth, which can all alter the speech signal.

Several methods have been proposed for making ASR systems robust to some of these factors. According to Gong [78], techniques for noisy speech recognition can be classified into three categories: noise resistant features and similarity measures, speech enhancement, and speech model compensation. Although several approaches have led to robust recognition under certain noise conditions, they often require the noise pattern and level to be known at training time, or they require the noise to be stationary or to be uncorrelated with the speech signal.

Alternative methods for noise robustness are the use of close talking microphones. These are however uncomfortable to wear and for some applications, e.g. video conferencing, it is desirable that the user can move freely in the room and is not constrained to talk closely to a microphone or to wear a head set. Microphone arrays have been proposed to locate the talker by locating the loudest sound source [66]. This approach sometimes causes problems during speaking pauses, when the person moves, or during cross-talk. For such cases it would be desirable to have a system that tracks the person and then uses acoustic *beam-forming* to enhance those sounds coming from the location of the talker. Bub et al. [31] have described a system where the face of the person was located and where either the visual or the acoustic information was used for beam-forming. When competing noise was present, recognition results were shown to be better for visually guided beam-forming than for acoustic beam-forming . In comparison to acoustic beam-forming, visual beam-forming suffers in that it does not know if a given person is talking or not. Lip tracking might be useful here to improve visual beam-forming, by determining if a person is talking or not.

Another method for robust speech recognition in noise is to augment the acoustic feature representation with visual features. It is well known that humans make use of the visual modality and that their speech perception is enhanced by seeing the talker's face, especially when the acoustic speech signal is contaminated with noise [54]. In the presence of noise, the visual signal is often complementary to the acoustic signal, i.e. some phonemes which are difficult to perceive acoustically are easier to distinguish visually, and vice versa. Thus, the visual signal often provides that information which is acoustically most sensitive to noise, and can therefore increase the robustness of acoustically based recognition systems. Among the different noise sources listed above, the visual speech signal conveyed by the talkers mouth movements is unaffected by environmental noise and by speech related noise, and therefore represents an important information source for acoustic speech recognition systems.

The work presented here describes a new method for visual speech recognition (VSR). The purpose of a VSR system is to combine it with an ASR system to improve the performance of the acoustic system, particularly in noisy environments. A combined AVSR system will therefore be briefly introduced to demonstrate the benefit of visual information on contaminated speech signals. However, the focus of the work lies in the VSR system. A central issue in ASR systems is their ability to deal with speech variability like linguistic variability and speaker variability. A visual speech recognition system is not only required to deal with visual speech variability and speaker variability but also with image variability such as illumination and pose. These factors still represent some of the main problems in machine vision research. It is this variability which most previous VSR systems are not able to cope with. The aim of the work presented here will therefore be to improve the robustness of the VSR system to such variability.

1.3 Visual Speaker Recognition

A related research area to automatic speech recognition is the problem of speaker recognition. Speaker recognition and face recognition represent two of the most successful approaches to the recognition of humans by machine. Current state-of-the-art face recognition systems usually require the face images to be static and assume neutral facial expressions. The mouth is however subject to large variability due to different facial expressions and due to speech production. This often prevents the use of such systems in applications where the subject might be speaking or where the facial expression can vary such as in surveillance. Another drawback of still face recognition systems is that they might accept an impostor placing a photograph in front of the camera. It is highly desirable that systems to be used in real-world situations are able to account for facial variability.

Speaker recognition systems on the other hand can often easily be defeated by a play back of recorded voice of another person. They are furthermore subject to different noise sources as it is the case for speech recognition systems. Whereas the multi-modal nature of speech has received much attention from the speech recognition community, speaker recognition research has mainly been based on the acoustic signal and has ignored the visual cues for a person's identity inherent in the speech signal.

The performance of a person authentication system is a crucial issue in practical applications. Particularly the ability to reject impostors, who claim a false identity, is very important. The combination of several modalities is likely to increase the performance of unimodal systems due to the increased amount of information provided to the classifier. Such a system is also likely to be more robust to impostor accesses. Although systems which combine face recognition with speaker recognition have previously been proposed [29, 28], both modalities were treated separately and did not exploit the temporal information of the face image during articulation. The use of spatio-temporal image information provides additional information which is complementary to the static face image because it models the variability of the face and complementary to the acoustic signal because it provides additional information responsible for the speech production process. Furthermore, in practical applications, the dependence between the acoustic and the visual signal can be used to detect asynchrony between the two signals which might be caused by an impostor. In this work, a novel approach for person recognition, based on visual spatio-temporal models of talking persons is presented. The method is described for visual speech features as well as for audio-visual speech features and its further combination with a face recognition system is proposed.

1.4 Thesis overview

The next chapter, Chapter 2, introduces hidden Markov models (HMMs) which form the basis of the developed speechreading and person recognition system. Their definition and underlying assumptions are outlined and different training and recognition algorithms are described.

Chapter 3 describes physiological and psychological issues regarding speechreading by humans and discusses their implications for speechreading by machine. The visual discrimination of phonemes is analysed as well as audio-visual complementarity and integration. The evaluation of the visual contribution to intelligibility is then discussed as well as the determination of important visual features for a speechreading system.

Chapter 4 describes the two databases used for the experiments and shows some example images of different subjects and of typical speech sequences.

Chapter 5 reviews previous approaches to facial speech feature extraction and describes the method developed here. The algorithm is based on an appearance based model which represents both shape information and intensity information. The model parameters are learned from a training set and are used in image search to estimate the similarity between the image and the model. The performance of the algorithm is evaluated on two multi-speaker databases of grey-level image sequences.

Chapter 6 reviews previous speechreading approaches and describes the speechreading system developed here. The parameters, modelling the talking mouth, are recovered from the lip tracking results and represent the shape of the lip contours and the grey-level appearance of the mouth. The application of these features for speechreading is presented, by modelling their distribution by mixtures of Gaussians and their temporal dependencies by HMMs. Experimental results for an isolated and a continuous digit recognition task are described.

Chapter 7 reviews some related face recognition and speaker recognition approaches and details the proposed modality for person identification, based on spatio-temporal speaker models. The method aims to model the appearance of the articulators of talking persons as well as their temporal changes during speech production. Person identification is performed by tracking the lips of the unknown talking person and estimating the posterior probability of each model given the observed sequence of features. Person recognition experiments are presented on an isolated digit database of 12 subjects and a continuous digit database of 37 subjects.

Chapter 8 draws overall conclusions and describes possible future work.

1.5 Summary of Results

A new method for lip localisation, lip tracking and feature extraction has been developed. A robust search algorithm which considers both shape variability and intensity variability is used to fit the model to the image. A visual speech recognition system has been described based on these extracted features and hidden Markov models. Finally, a novel modality for person recognition is proposed based on spatio-temporal models of speaking persons.

The comparison of the performance of algorithms is only possible if they are tested on identical databases and for identical experiments. If the performance of algorithms is not comparable with other methods, their quality and benefit can not be assessed, which makes little overall contribution to progress. All experiments reported here were performed on two publicly available databases to allow the comparison with other methods.

An algorithm for tracking lips in grey-level image sequences has been developed, was subjectively evaluated on the Tulips1 database of 12 subjects and was found to obtain good tracking results for over 90% of the tested sequences. In addition, the algorithm was tested on the M2VTS database of 37 subjects for lip tracking on continuous speech. Shape and intensity parameters were extracted from the tracking results of both databases and were evaluated for the tasks of visual speech and speaker recognition.

The visual speech recognition system obtained a speaker independent digit recognition accuracy of 90.6% on the Tulips1 database using visual information only. This performance is slightly higher than the performance of humans with no speechreading training, who were tested on the same database. The system also outperformed all speechreading approaches which were recently evaluated for an identical recognition task by Gray et al. [81]. On the M2VTS database, the system achieved an accuracy of up to 58.5% for a speaker independent continuous digit recognition task which represents one of the largest reported speechreading tests.

The proposed method for person authentication identified up to 97.9% of persons correctly on the Tulips1 database and up to 94.6% on the M2VTS database. Person verification experiments on the M2VTS database reduced the false acceptance rate of an acoustic speaker recognition system from 2.3% to 0.5% when the visual modality was included.

Chapter 2

Hidden Markov Model Speech Recognition

The most successful approach to automatic speech recognition developed so far is the modelling of speech units by hidden Markov models (HMMs). This chapter describes the definition of HMMs, their underlying assumptions, their training, and their use in speech recognition. A more detailed description of HMMs can be found in [150].

A HMM may represent a sub-word unit like a phoneme or tri-phone, a word, or a sentence. In this chapter, for simplicity, the HMMs will be referred to as word models. Word model based recognition systems are often used in small vocabulary speech recognition tasks where one HMM is built for each word class. For large vocabulary tasks, the number of word models needed becomes very large and the training set is rarely large enough to build good models for all word classes of the vocabulary. Furthermore, the recognition process becomes computationally very expensive for a large number of word classes. Sub-word models were therefore introduced for large vocabulary recognition tasks to reduce the number of models needed for training and recognition. The sub-word models are used in a hierarchical scheme by concatenating them to construct words, and the constructed words are further concatenated to form sentences. Pronunciation dictionaries can be used to constrain the sub-word models to form possible words and language models may be used to constrain the word sequences to form plausible sentences.

2.1 Definition and Underlying Assumptions

A HMM is a finite state automaton with two concurrent stochastic processes. The state sequence models the temporal evolution of speech and the output distributions attached to the states model the speech features occurring at that state. The speech signal is assumed to be preprocessed and represented as a sequence of feature vectors, extracted at regular intervals from the sampled speech waveform. Commonly used feature representations are spectral analysis and linear prediction analysis. The observed sequence of observation vectors is denoted by

$$\mathbf{O} = \mathbf{o}(1), \mathbf{o}(2), \ldots, \mathbf{o}(T), \tag{2.1}$$

where each observation $\mathbf{o}(\tau)$ is a n-dimensional vector, extracted at time τ with

$$\mathbf{o}(\tau) = [o_1(\tau), o_2(\tau), \ldots, o_n(\tau)]^T. \tag{2.2}$$

For a given set of possible word classes, denoted by

$$\widehat{\Lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_M\}, \tag{2.3}$$

the recognition task can be expressed as

$$\Lambda^* = \arg\max_{\Lambda} P(\Lambda|\mathbf{O}), \tag{2.4}$$

where Λ represents a particular sequence of word classes. The classification criterion is referred to as *maximum a posteriori* (MAP) probability classification and the term $P(\Lambda|\mathbf{O})$ is denoted as the posterior probability. Each word class is represented by a HMM which is made up of several states according to a predefined topology. The following assumptions about speech are implicitly made by HMMs:

- The speech signal can be split into piece-wise stationary segments which are modelled by a state specific probability density function.
- The speech signal is assumed to follow a first-order Markov process, i.e. the probability of being in a certain state at time τ depends only on the state of the Markov chain at time $\tau - 1$ and not on any earlier states.
- The observation vectors are conditionally independent of previously observed vectors.
- The probability that a particular vector will be observed depends only on the current state of the process.

All of these assumptions are not true for the speech signal and several approaches have been proposed which attempt to avoid such assumptions. However, most state-of-the-art speech recognisers are still based on standard HMMs.

A HMM can be defined by the following parameter set:

- N is the number of states which are denoted by $\mathbf{s} = \{s_1, s_2, \ldots, s_N\}$. The number of states is usually determined heuristically.
- $\mathbf{A} = \{a_{ij}\}$ is the matrix of state transition probabilities from state i to state j, with $a_{ij} = P(q_j(\tau+1)|q_i(\tau))$, where $q_i(\tau)$ denotes the event of residing in state i at time τ. The state transition probabilities are assumed to be time independent.
- $\mathbf{B}$ is the vector of observation probabilities $b_j(\mathbf{o}(\tau))$ associated with each emitting state j, with $b_j(\mathbf{o}(\tau)) = P(\mathbf{o}(\tau)|q_j(\tau))$.
- π is the vector with the initial state probabilities π_i of entering the model at state i: $\pi_i = P(q_i(1))$.

A HMM can now be represented by the compact parameter set

$$\lambda = (\mathbf{A}, \mathbf{B}, \pi). \tag{2.5}$$

Since the speech signal evolves forward in time, the transition probability matrix $\mathbf{A}$ is normally constrained to only allow self-loops, by residing in the same state for several consecutive frames, or transitions from left to right. HMMs are usually classified with respect to their kind of output distribution model into Discrete HMMs (DHMM), where the output distribution is based on discrete elements and Continuous Density HMMs (CDHMM), where the output distribution is based on continuous density functions. All HMMs considered in this dissertation are based on CDHMMs. The observation probabilities are modelled as mixtures of Gaussian distributions:

$$b_i(\mathbf{o}(\tau)) = \sum_{m=1}^{M} c_{im} \mathcal{N}(\mathbf{o}(\tau), \mu_{im}, \Sigma_{im}) \tag{2.6}$$

where c_{im} is the mixture weight for state i and mixture m, and $\mathcal{N}(\mathbf{o}(\tau), \mu, \Sigma)$ a multivariate Gaussian with mean μ and covariance matrix Σ.

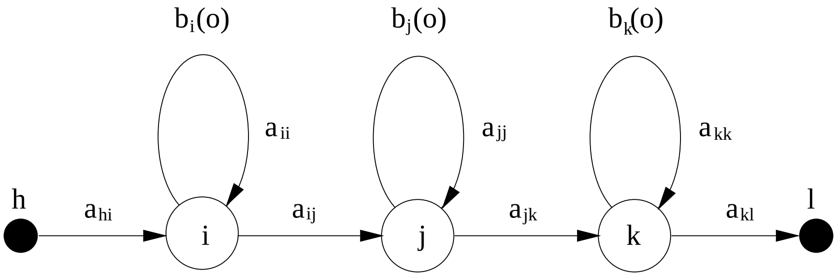


Figure 2.1: A hidden Markov Model with three emitting states and continuous output distributions.

The HMM topology used in this thesis is based on terminology of the HMM Toolkit (HTK) [187], where the first and the last state are non-emitting. The HMMs are constrained to start in the first state and to end in the last state. This topology facilitates their concatenation for continuous speech recognition tasks. An example of a HMM with five states, of which three are emitting, is shown in Figure 2.1.

2.2 Recognition

Recognition is usually performed according to the MAP probability criterion. The posterior probability $P(\Lambda|\mathbf{O})$ is required in the MAP calculation (2.4) for estimating the probability for the word sequence Λ of having produced the observed sequence $\mathbf{O}$. Usually it is not possible to calculate $P(\Lambda|\mathbf{O})$ directly but it can be obtained using Bayes' law:

$$P(\Lambda|\mathbf{O}) = \frac{P(\mathbf{O}|\Lambda)P(\Lambda)}{P(\mathbf{O})} \tag{2.7}$$

$P(\Lambda)$ represents the prior probability of the model and $P(\mathbf{O})$ the prior probability of the observation sequence. $P(\mathbf{O}|\Lambda)$ is usually referred to as the *Likelihood* of the observation, given the model Λ.

If the prior probabilities of all models $P(\Lambda)$ are assumed to be equal and the probability of the feature vector $P(O)$ to be constant, the MAP classification criterion turns into the *Maximum Likelihood* (ML) criterion:

$$\Lambda^* = \arg\max_{\Lambda} P(\Lambda|\mathbf{O}) \tag{2.8}$$

Using the ML criterion makes the following additional assumptions:

- $P(\Lambda)$ can be estimated separately and independent of acoustic data (normally from a language model).
- During recognition, $P(\mathbf{O})$ can be assumed to be constant for all models.

In this work, the ML criterion is used for training and the MAP criterion for recognition. This is the standard method used in HMM based speech recognition systems. Bourlard and Morgan [19] have however shown that ML training reduces the discriminating properties of HMMs since the training procedure maximises the probability of the training data given the model rather than maximising the probability of the model given the data with respect to rival models. They have therefore proposed the use of multilayer perceptrons for HMM training which performs parameter estimation according to the MAP criterion.

2.2.1 Baum-Welch Recognition

The total likelihood of the data given the model is obtained by summing the likelihoods of all possible paths through the model:

$$P(\mathbf{O}|\Lambda) = \sum_{\theta \in \Theta} P(\mathbf{O}, \theta|\Lambda) \tag{2.9}$$

where Θ is the set of all possible state sequences of model Λ. This method is computationally expensive since for an observation sequence of length N and a HMM with T states, N^T possible state sequences need to be considered. The number of state sequences to compute can drastically be reduced by the Baum-Welch or so-called Forward-Backward algorithm.

The Baum-Welch algorithm can be used to efficiently calculate the likelihood of an observation sequence given the model. The first-order Markov assumption states that a particular observation is only dependent on the current state and not on any previous states. Partial likelihoods for a particular state and time can therefore be obtained by summing all possible previous likelihoods, which can be computed recursively. The idea of the algorithm is that all possible sequences of the total likelihood must merge into one of N states and that the sum of the likelihoods over all states at any time gives the total likelihood.

To describe the method, two new variables are introduced, the *forward-probability* and the *backward-probability*. Only the forward-probability is necessary to compute the total likelihood for recognition, but both probabilities are required for the training algorithm. The forward probability $\alpha_i(t)$ is defined to be the joint probability of having generated the partial forward sequence up to observation t and having arrived at state i:

$$\alpha_i(t) = P(\mathbf{o}(1), \mathbf{o}(2), \ldots, \mathbf{o}(t), q_i(t)|\Lambda) \tag{2.10}$$

The backward probability $\beta_i(t)$ is defined as the joint probability of having generated the observation sequence for the rest of the path, starting at time $t+1$ and state i:

$$\beta_i(t) = P(\mathbf{o}(t+1), \mathbf{o}(t+2), \ldots, \mathbf{o}(T)|q_i(t), \Lambda) \tag{2.11}$$

As the HMMs are constrained to start in state s_1 and to exit in state s_N, the initial forward probabilities are given by

$$\alpha_i(1) = \begin{cases} 1 & \text{if } i = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

and the the final probability is given by

$$\alpha_N(T) = \sum_{i=2}^{N-1} \alpha_i(T) a_{iN}. \tag{2.13}$$

For the remaining states $2 \leq j \leq N-1$, the forward probabilities for $2 \leq t \leq T$ can be calculated recursively using

$$\alpha_j(t) = \left[\sum_{i=1}^{N} \alpha_i(t-1) a_{ij} \right] b_j(\mathbf{o}(t)). \tag{2.14}$$

For the backward probability, the initial and final conditions are defined by

$$\begin{aligned} \beta_i(T) &= \alpha_{iN} \qquad \text{for} \qquad 1 \leq i \leq N \\ \beta_N(t) &= 0. \end{aligned} \tag{2.15}$$

The backward probabilities for state $1 \leq j < N$ and $1 \leq t \leq T-1$ can then be calculated with the recursive algorithm

$$\beta_i(t) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{o}(t+1)) \beta_j(t+1). \tag{2.16}$$

and the total likelihood can be expressed by

$$P(\mathbf{O}|\Lambda) = \alpha_N(T) = \beta_1(1) = \sum_{j=1}^{N} \alpha_j(t) \beta_j(t). \tag{2.17}$$

2.2.2 Viterbi Recognition

Whereas the Baum-Welch algorithm computes the likelihood of the model for having generated the observed sequence using *all* possible sequences of states of length T, the Viterbi algorithm [179] computes the *maximum likelihood* sequence of states. The partial maximum likelihood of observing the feature sequence up to time t and being in state j is defined as

$$\phi_j(t) = \max_i[\phi_i(t-1)a_{ij}b_j(\mathbf{o}_t)] \tag{2.18}$$

Using the previous constraints for the initial state probabilities, the initial conditions of $\phi_j(1)$ are

$$\phi_j(1) = \begin{cases} 1 & \text{if } j = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.19}$$

The maximum likelihood is given by

$$\phi_n(T) = \max_i[\phi_i(T)a_{iN}]. \tag{2.20}$$

The Viterbi algorithm differs from the Baum-Welch algorithm in that the summation over all states is replaced by the maximum operation. Viterbi recognition is usually preferred over Baum-Welch recognition and requires $(N-1)T$ fewer additions per HMM in the likelihood calculation. The algorithms differ further in that the Viterbi algorithm calculates the most likely state sequence whereas the Baum-Welch algorithm calculates the total likelihood and therefore the most likely word sequence.

2.3 Baum-Welch Training

Different criteria exist for the training of HMMs. The method used here is based on the Maximum Likelihood (ML) criterion which aims to find Λ^* with

$$\Lambda^* = \arg\max_{\Lambda} P(\mathbf{O}|\Lambda) \tag{2.21}$$

The *Expectation Maximisation* (EM) technique [53] is used for the estimation of parameters, which was first studied by Baum et al. [11] for the application of HMMs and commonly referred to as the Baum-Welch re-estimation algorithm. It is designed such that

$$P(\mathbf{O}|\hat{\Lambda}) \geq P(\mathbf{O}|\Lambda) \tag{2.22}$$

where $\hat{\Lambda}$ is the new estimate of the model parameters. The EM algorithm is performed in two steps, the estimation step (E-step) and the maximisation step (M-step). The E-step augments the process with *incomplete data*, i.e. the model parameters, and the M-step maximises the likelihood $P(\mathbf{O}_T|\Lambda)$ over Λ. The following auxiliary function is introduced:

$$\mathcal{H}(\Lambda, \hat{\Lambda}) = \sum_{\theta \in \Theta} P(\mathbf{O}, \theta|\Lambda) \log P(\mathbf{O}, \theta|\hat{\Lambda}) \tag{2.23}$$

where $\hat{\Lambda}$ is the new estimate of the model. It can be shown that maximising $\mathcal{H}$ ensures that $P(\mathbf{O}_T|\Lambda)$ is not decreasing. The algorithm starts with some initial estimate of Λ and repeats the following steps until it converges at a local maximum:

$$\left.\begin{array}{ll}\textbf{E-Step:} & \text{Given } \hat{\Lambda},\ \text{compute } \mathcal{H}(\Lambda, \hat{\Lambda}) \\ & \\ \textbf{M-Step:} & \text{Compute new estimate } \hat{\Lambda} = \arg\max_{\Lambda} \mathcal{H}(\Lambda, \hat{\Lambda}) \\ & \text{Update } \Lambda \end{array}\right\} \tag{2.24}$$

The re-estimation formulas below can be derived from the EM algorithm. The transition probabilities can be re-estimated using

$$\hat{a}_{ij} = \frac{\sum_{\tau=1}^{T-1} \alpha_i(\tau) a_{ij} b_j(\mathbf{o}(\tau+1)) \beta_j(\tau+1)}{\sum_{\tau=1}^{T-1} \alpha_i(\tau) \beta_i(\tau)} \tag{2.25}$$

where $1 < i < N$ and $1 < j < N$. The transition probabilities for the non-emitting states are re-estimated by

$$\hat{a}_{1j} = \frac{1}{P(\mathbf{O}|\Lambda)} \alpha_j(1) \beta_j(1) \tag{2.26}$$

and

$$\hat{a}_{iN} = \frac{\alpha_i(T) \beta_j(T)}{\sum_{\tau=1}^{T} \alpha_i(\tau) \beta_i(\tau)} \tag{2.27}$$

For the re-estimation of the mean, variances and mixture weights, the term $L_{jm}(\tau)$ is introduced. The probability of being in state j and mixture m at time τ is

$$\begin{aligned} L_{jm}(\tau) &= P(q_{jm}(\tau)|\mathbf{O}, \Lambda) \\ &= \tfrac{1}{P(\mathbf{O}|\Lambda)} U_j(\tau) c_{jm} b_{jm}(\mathbf{o}(\tau)) \beta_j(\tau) \end{aligned} \tag{2.28}$$

where

$$U_j(\tau) = \begin{cases} a_{1j} & \text{if } \tau = 1 \\ \sum_{i=2}^{N-1} \alpha_i(\tau-1) a_{ij} & \text{otherwise} \end{cases} \tag{2.29}$$

The re-estimation formulae can now be expressed in terms of $L_{jm}(\tau)$ by

$$\hat{\mu}_{jm} = \frac{\sum_{\tau=1}^{T} L_{jm}(\tau) \mathbf{o}(\tau)}{\sum_{\tau=1}^{T} L_{jm}(\tau)} \tag{2.30}$$

$$\hat{\Sigma}_{jm} = \frac{\sum_{\tau=1}^{T} L_{jm}(\tau) (\mathbf{o}(\tau) - \hat{\mu}_{jm}) (\mathbf{o}(\tau) - \hat{\mu}_{jm})^T}{\sum_{\tau=1}^{T} L_{jm}(\tau)} \tag{2.31}$$

$$\hat{c}_{jm} = \frac{\sum_{\tau=1}^{T} L_{jm}(\tau)}{\sum_{\tau=1}^{T} L_j(\tau)} \tag{2.32}$$

The re-estimation formulae given here are for a single observation only. The extension of the algorithm for multiple observations is straight-forward and can be found in [187].

For isolated word recognition, the segmented training observations are used to independently train the models. For continuous speech recognition, the models are usually trained in two stages. The first stage is the same as for isolated word models and the second stage is performed *Embedded Training*. Embedded training is implemented by concatenating the previously trained models according to the transcription but the information about segment boundaries is ignored. Baum-Welch re-estimation is then performed on the whole sentence while models are allowed to to change their boundaries to more likely alignments. The method assumes that the segmented training data is not optimally segmented and that a better alignment can be found by embedded training.

Chapter 3

Physiology and Psychology of Speechreading

The study of speechreading by humans relies on basic research in the speech, hearing and vision sciences. Speechreading by machine is therefore a multidisciplinary subject too which requires the understanding of the basic concepts from these areas. Unfortunately, due to its multidisciplinary nature, few researchers have considered all of these fundamental disciplines in developing speechreading systems.

Researchers from the ASR community have often based automatic speechreading experiments on a single subject. Such experiments ignore the large linguistic and appearance variability between subjects which represents one of the main problems in ASR. Image analysis was often performed semi-automatically with the aid of head-mounted cameras or lip stick to highlight the lip region. These experiments have provided valuable insights into the possible benefit of speechreading systems but the imposed constraints prevent their use in real-world applications and the selected features might not provide all relevant speech information.

Researchers from the computer vision community, on the other hand, have often avoided to determine visual speech features, and have used the whole image area containing the mouth as feature vector. These approaches work well under controlled lighting and recording conditions and for a small number of subjects, but they tend to be very susceptible to differences in illumination, pose and appearance. Other researchers have reported speechreading experiments for recognition of static vowels, although it is well known in the psychology community that there is a 'many to one mapping' between the visible articulators and vowels, that vowels are articulated differently in continuous speech, and that they are highly dependent on coarticulation.

This chapter will address different issues related to speechreading. First, visual speech perception by humans will be reviewed. This is followed by a discussion of the speech production process and the visibility of speech articulators. Two models of human speech production will be described and the recovery of their model parameters from the visual speech signal will be discussed. The visual intelligibility of individual phonemes will then be analysed, as well as the complementarity of acoustic and visual features, and possible strategies for their integration. The last section aims to determine those visual features which are most important for speechreading and which should be accounted for by a speechreading system.

3.1 Speechreading by Humans

It is well known that information from the face of a talking person provides speech information and enhances the intelligibility of speech [169, 54]. Hearing-impaired[1] persons often use speechreading as the primary source of information for speech perception. Some few individuals perform speechreading to such a high degree which enables almost perfect speech perception [172]. People with normal hearing on the other hand, often speechread under difficult listening conditions to supplement or complement acoustic information. Difficult listening conditions can for example be due to interfering noise such as environmental noise, music, and multiple talkers or due to reverberation. Speechreading can improve the intelligibility of speech under such conditions [169]. Humans with normal hearing also benefit from visual information when no noise is present. This was demonstrated by Reisberg et al. [154], who have shown that normal-hearing subjects who see the talker's face perceive speech more accurately, even in noise-free conditions.

Visual speech cues play an important role in the acquisition of speech perception and speech production from early life on. Infants are aware of congruence between lip movements and speech sounds as early as three months of age [5]. Speechreading skills are acquired at early age, for example toddlers can speechread familiar words when they reach 19 months of age [55]. The visual speech modality also plays an important role in learning to speak. This was demonstrated by Mills [131], who has shown that blind children are slower in the acquisition of speech production than seeing children, for those sounds which have visible articulation.

The contribution of visual information in human speech perception can be divided into three categories: attention, redundancy and complementarity. In the first category, visual information is used to direct auditory attention to a specific speaker and to screen that speaker's speech signal from other sound sources. Redundancy takes effect mainly in noise-free environments, where the visual signal can provide information which is redundant to the acoustic signal. The third category occurs predominantly in noisy environments, in which visual perception can complement auditory perception.

Although these studies of human speechreading clearly demonstrate the benefit of visual speech information, theories of audio-visual speech perception are still in their infancy. The development of automatic audio-visual speech recognition systems would greatly benefit from a better understanding of human speech perception. The following questions are of particular interest:

- How large is the benefit of visual information in speech perception?
- Which visual features are most important in speech perception?

[1] The term hearing-impaired is used here to refer to both deaf and hard-of-hearing persons.

- Are there visual units equivalent to phones?
- How are the two information sources integrated?

The answers to these questions could help to evaluate the benefit of visual information in automatic speech recognition systems. They might lead to improved performance of ASR systems by drawing parallels between human speech perception and machine recognition. The next sections will address these questions and attempt to describe their consequences for audio-visual ASR.

3.2 Human Speech Production

This section will address the speech production process and will describe two models for speech production. Speech production models have led to several approaches to speech recognition. The recovery of these model parameters from the visual speech signal will therefore be discussed. The speech waveform is an acoustic sound pressure wave which originates from movements of the human speech production system. Figure 3.1 displays the mid-sagittal section of the human speech production system [67]. The main components which determine the speech waveform are the lungs, trachea, larynx, pharyngeal cavity (throat), oral cavity (mouth) and nasal cavity (nose).

The pharyngeal and oral cavity are commonly referred to as the vocal tract. The vocal tract begins at the vocal cord constriction at the top of the trachea and ends at the lips. The cross-sectional area of the vocal tract is nonuniform and is determined by the position of the articulators which comprise the lips, jaw, tongue and velum.

The nasal cavity, which is often referred to as the nasal tract, begins at the velum and terminates at the nostrils. The opening of the velum controls the coupling between the nasal tract and the vocal tract. For non-nasal sounds the velum is drawn up tightly towards the back of the pharyngeal cavity, thereby closing the entrance to the nasal tract.

The articulators which influence the speech production are the vocal folds, soft palate (velum), tongue, teeth, lips, jaw and air pressure. The articulators move to different positions to produce different speech sounds. During speech production, air is forced from the lungs through the trachea into the pharynx or nasal cavity. Between the trachea and the pharynx is the larynx, which is responsible for the production of voiced sounds. Speech sounds can be classified into voiced and unvoiced sounds. Voiced sounds are produced by vibratory action of the vocal cords, located in the larynx. The air passes through the glottis, the opening of the vocal cords, and the alternating opening and closing provides the quasi-periodic air pulses for voiced sounds. Unvoiced sounds are produced by turbulent flow of air created at a constriction in the vocal tract. Consonants can be *continuants* such as the *fricatives*, where the sound is sustained, or *stops* such as the /p/ sound, where pressure is build up, followed by an abrupt release of the pressure. Fricatives and stops can be both voiced or unvoiced. Constrictions of continuants can be classified as *labiodental* (upper teeth on lower lip), *interdental* (tongue behind front teeth), *alveolar* (tongue touching gum ridge), *palatal* (tongue resting on hard or soft palate), and *glottal* (vocal folds fixed and tense) .The closure of *stop* consonants can occur at various positions of the vocal tract such as the lips (bilabial), teeth (alveolar) and palate (velar) (see Tab. 3.2).

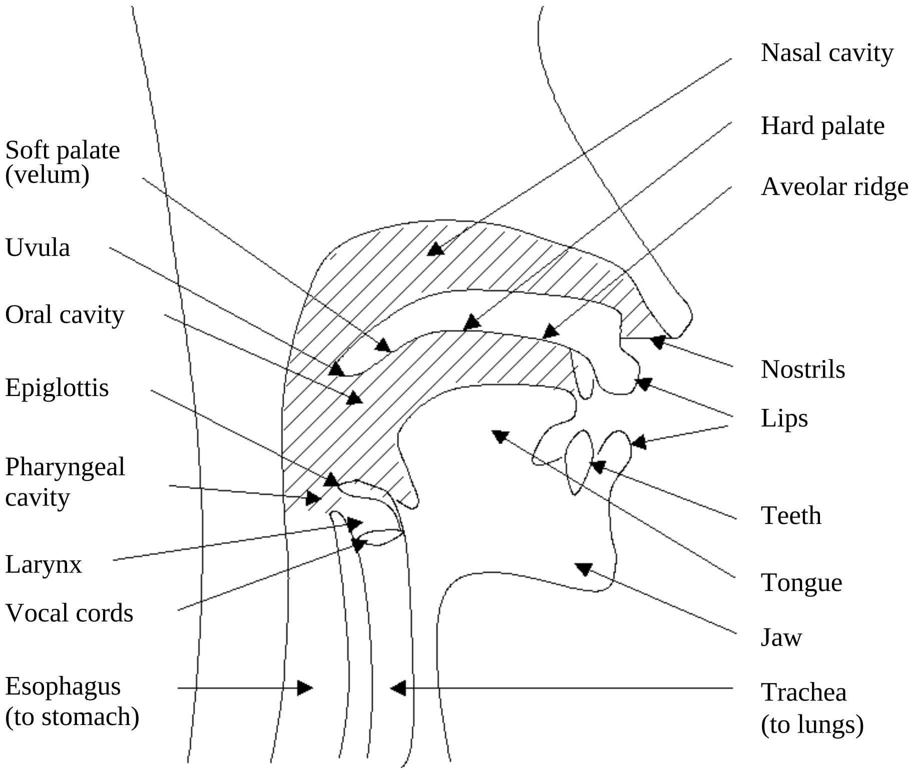


Figure 3.1: The human speech production system.

It becomes clear now, that even under the best circumstances, only the lips, the frontal part of the tongue, and the jaw are visible to the speechreader. Perfect intelligibility for an unrestricted vocabulary by speechreading alone and without the use of other knowledge sources is therefore not possible. Important parts like the vibration of vocal folds, soft palate and complete tongue shape can not be observed visually. The importance of voicing information, for example, is demonstrated by human speechreaders who perceive speech more accurately, more fluently, and with less effort when voicing information is provided by a tactile signal.

Speech production can be modelled as an acoustic filtering operation [62]. The vocal tract and nasal tract represent the acoustic filter which is excited by the air passing from the lungs through the larynx and which radiates at the lips or the nostrils. The sound source for voiced sounds is often modelled by a train of quasi-periodic puffs of air, expelled from the oscillating vocal cords. Unvoiced sound sources are usually modelled as a turbulent noise excitation, passing through a narrow constriction or by a complete closure followed by a plosive excitation. The sound propagation for an ideal vocal tract model with continuously variable cross-sectional area is complex. Multiple concatenated acoustic tubes are therefore often used to build simplified models of the vocal tract. Figure 3.2 displays an acoustic tube model for speech production using six tubes. Different cross-sections of the vocal tract are modelled by different diameters of individual tubes. The nasal tract can be modelled by adding a parallel tube which branches off at a tube representing the vocal tract. These models are simplified and do not consider effects like losses along the vocal tract, sub-glottal coupling with the vocal tract resonance, and the time-varying nature of the vocal tract. They can however serve reasonably well for the modelling of stationary speech sounds.

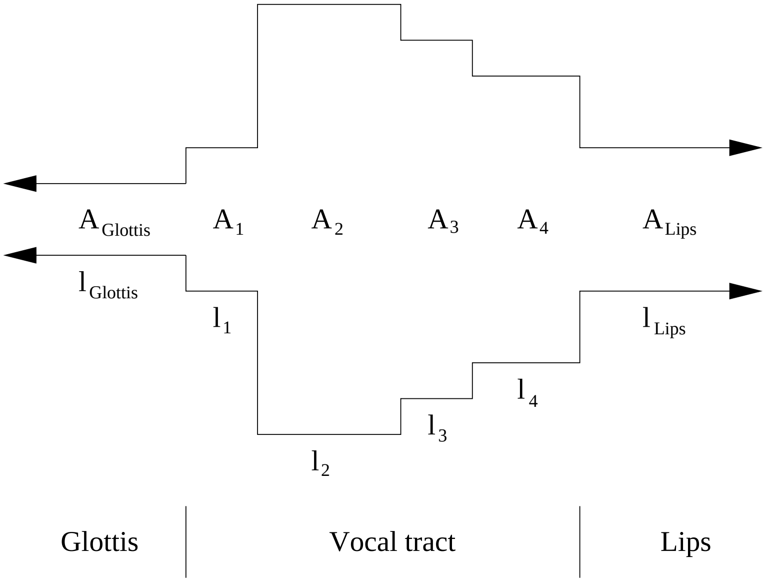


Figure 3.2: Acoustic tube model of speech production. The model is defined by the area A and length l of the individual tubes.

Considering such a model for speechreading, the visual information perceived would only provide information about the configuration of the first tube, describing the lip shape, and perhaps parts of the next few tubes, accounting for the tongue position and jaw.

Another way to model speech production is to use an electrical filter model [151] shown in Figure 3.3. This model does not try to model the physical characteristics of the human vocal tract. Instead it attempts to represent human speech production by an electric filter model, driven by an excitation source. It is therefore based on producing speech signals similar to those of human speech by using an electric circuit. Here, the excitation source for voiced speech is replaced by an impulse train generator and for unvoiced speech by a random noise generator. The vocal tract, nasal tract, and lip radiation are represented by filters models. The glottal pulse model is excited by the impulse train generator and produces a glottal pulse waveform.

Thus, voiced speech is modelled as a periodic pulse train, filtered by the glottal pulse model, vocal tract model, and the lip radiation model. Unvoiced speech is modelled by random noise, filtered by the vocal tract and the radiation model. The advantage of the electric filter model is that the parameters of the model can be derived by linear prediction analysis (LP) from the speech signal [6]. The estimated parameters describe the

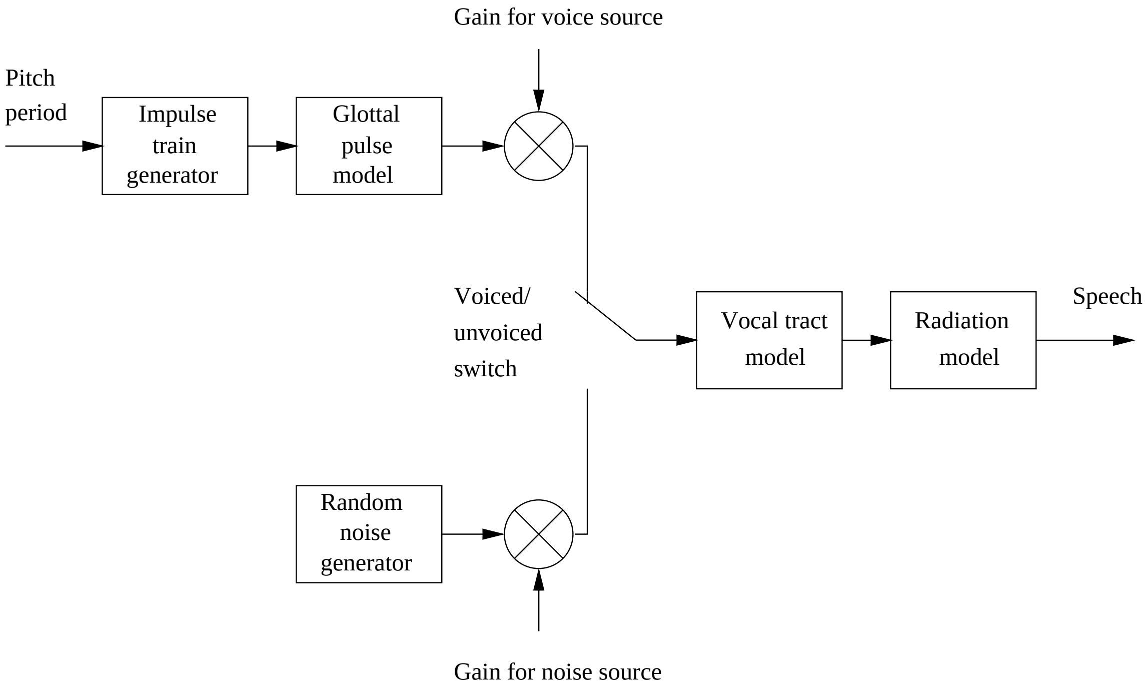


Figure 3.3: Electric filter model for speech production. After Rabiner and Shafer [151].

configuration of the model for producing the observed speech signal and have successfully been used as features for speech recognition systems. All-pole models are usually preferred over other models, since they allow simple computational techniques in linear prediction analysis to derive the model parameters from the speech signal. Many ASR systems are based on linear prediction coefficients and obtain state-of-the-art performance. The all-pole model is considered an appropriate choice for the vocal tract but not for the radiation model which has the effect of a high pass filter which usually includes a *zero* in the transfer function [52]. The model might therefore not be able to describe the configuration of the lips very well. The visual speech signal might provide information about the configuration and radiation at the lips and compensate for the sub-optimal filter radiation model.

3.3 Phonemes and Visemes

The basic linguistic unit is called *phoneme*. A phoneme is the theoretical unit for describing the linguistic meaning of speech. Phonemes have the property that if one is replaced by another one, the meaning of the utterance is changed. The English language can be classified into about $35 - 70$ phonemes, but in automatic speech recognition, usually only about 47 phoneme classes are used. These can be divided into 13 vowels, 23 consonants, 4 diphthongs, 5 semi-vowels, glottal stop, and whisper. The single-letter symbols of the ARPAbet, developed under the auspices of the United States Advanced Research Project Agency (ARPA), will be used here for the phonetic transcription. The transcriptions for

Table 3.1: Phonetic Alphabet for IPA and ARPAbet symbols.

		IPA Symbol	ARPAbet (SV)	ARPAbet (UV)	Examples
Vowels	Front	i	i	IY	beat
		I	I	IH	bit
		e	e	EY	bait
		ε	E	EH	bet
		æ	@	AE	bat
	Back	ɑ	a	AA	Bob
		ɔ	c	AO	bought
		o	o	OW	boat
		U	U	UH	book
		u	u	UW	boot
	Mid	ɝ	R	ER	bird
		ə	x	AX	ago
		ʌ	A	AH	but
Diphthongs		ɑI	Y	AY	buy
		ɑU	W	AW	down
		ɔI	O	OY	boy
		ɨ	X	IX	roses
Stop Consonants	Voiced	b	b	B	bat
		d	d	D	deep
		g	g	G	go
	Unvoiced	p	p	P	pea
		t	t	T	tea
		k	k	K	kick
Fricatives	Voiced	v	v	V	vice
		ð	D	DH	then
		z	z	Z	zebra
		ʒ	Z	ZH	measure
	Unvoiced	f	f	F	five
		θ	T	TH	thing
		s	s	S	so
		ʃ	S	SH	show
Semivowels	Liquids	l	l	L	love
		l	L	EL	cattle
		r	r	R	race
	Glides	w	w	W	want
		ʍ	H	WH	when
		j	y	Y	yard
Nasal	Non vocalic	m	m	M	mom
		n	n	N	noon
		ŋ	G	NX	sing
	Vocalic	m	M	EM	some
		n	N	EN	son
Affricates		tʃ	C	CH	church
		ʤ	J	JH	just
Others	Whisper	h	h	HH	help
	Vocalic	f	F	DX	batter
	Glottal stop	ʔ	Q	Q	

the ARPAbet and the corresponding International Phonetic Alphabet (IPA) symbols are given in Table 3.1. Both versions of the ARPAbet are shown, the single letter version (SV) and the upper case version (UV).

The acoustic manifestations of phonemes are called *phones* and the manifold acoustic variations of a phone are called *allophones*. They may vary widely but should be considered as different realisations of the same phoneme.

Visually distinguishable speech units are denoted as *visemes* [65]. Whereas phones are quasi mutually exclusive, this is not the case for visemes. The mapping from phonemes to visemes is often many to one, i.e. several phonemes can correspond to the same visual configuration. Most visemes can therefore not uniquely be associated with a single phoneme but rather with a group of phonemes. The definition of viseme is therefore mis-leading since it refers to the visually *distinguishable* manifestations of phonemes rather than the visual manifestations of *all* phonemes. To avoid confusion, the term *vise* will be introduced here to denote the visual manifestation of a phoneme, equivalent to the acoustic unit *phone*. Phonemes belonging to the same viseme group are called *homophonous*. There exists no consensus about viseme categories and proposed numbers and categories vary widely.

The next two sections will address the confusability of *vises* to get an approximate idea about the expected performance of pure visual speech perception. The results of possible viseme classes are however not used in the developed speechreading system. The speechreading system treats the visual features as additional features to the acoustic features, which are associated to a particular phoneme or utterance. During the training procedure the distribution of the visual features is learned and is used for recognition as additional discriminative information. Although it might be useful for an AVSR system to cluster vises models which correspond to the same viseme class to reduce the number of parameters of the system, the approach pursued here is to let the system learn from training data which vises are confusable. The use of visemes introduces hard decisions regarding the mapping from visemes to phonemes which might not be adequate. Hard decisions should preferably be made at the latest possible stage to reduce their effect of inadequacy.

3.3.1 Speechreading Vowels

The articulatory gestures for vowels are generally stationary but their configurations required for a particular vowel are not very constrained. Although, acoustically, each vowel of a particular speaker is characterised by a unique combination of formant frequencies, in fluent speech these values are rarely attained. The production of vowels in natural speech is generally more lax and depends on coarticulation and stress. Furthermore, the speaker can vary the visible configuration of many vowels without changing their auditory characteristics.

Stevens' [166] classical work in the quantitative description of vowel articulation modelled the vocal tract as a continuous tube whose cross-sectional area varies between the glottis and the lips. The possible range of formant frequency combinations for producing vowels can then be obtained by varying as few as three parameters:

- distance between the glottis and the maximum constriction

- cross-sectional area at the constriction
- ratio of the area of lip opening to the length of lip passage

Out of these three parameters, only the third one is clearly visible which demonstrates the difficulty of speechreading vowels.

The first two formant frequencies are most prevalent and are often used to display the vowel triangle, where one axis represents $F1$ and the other axis $F2$. The three corners of the triangle are represented by /i/, /a/ and /u/. Using Stevens' model, the three vowels are subject to the following configurations:

/i/: Placing the major constriction as forward in the mouth as possible and making the cross-section as small as possible. Third parameter has little influence.

/a/: Placing the major constriction at the back of the mouth. Lip area to lip length ratio must be medium to large.

/u/: Placing the major constriction at the back of the mouth and making the cross-section small (rounding the lips). Lip area to lip length ratio must be small.

Summerfield [170] has described the necessary lip configurations for these three vowels as follows:

/i/: horizontal lip extension is advantageous but not essential.

/a/: lip rounding and protrusion must be avoided but extensive vertical lip opening is not essential.

/u/: lip rounding and protrusion are essential.

In fluent speech, the talker tends to articulate only what is essential for auditory distinctiveness but not for visual distinctiveness. The vowel /u/ will therefore retain its distinctiveness, while other vowels will become more difficult to distinguish. This is mainly due to the invisible tongue, which is the key parameter to vowel production. Further difficulties are due to coarticulation effects, variable pronunciation, and large differences in the articulation of vowels across talkers. Whereas formant frequencies provide fairly exclusive representations of vowels across male speakers [145, 166] this is not the case for vises. In a study conducted by Lesner and Kricos [107], the number of viseme categories across speakers, varied from only one up to five. The constituents of these categories varied also between the subjects.

Overall it can be concluded that speechreading vowels from natural speech is very difficult and becomes even more difficult for multiple subjects. Studies of vowel speechreading from static images of one speaker, as described in [190, 191, 183], should therefore be interpreted very carefully. Vowels are articulated differently in continuous speech, they are highly dependent on coarticulation, and they are speaker dependent. Furthermore, speechreading requires the temporal analysis of mouth movements rather than the analysis of single images.

3.3.2 Speechreading Non-Vowels

Consonants, diphthongs, and semi-vowels are usually referred to as non-vowels. In comparison to vowels, non-vowels often involve rapid articulatory gestures which attain specific target configurations. These target configurations are often more constrained than for vowels, e.g. the lips have to be closed for bilabial consonants. The number of visually distinguishable non-vowel visemes reported in the literature varies between about 9 [181] and 16 [77]. These visemes often contain phonemes requiring the same place of articulation. They are therefore often grouped with respect to their place and manner of articulation. Table 3.2 gives an overview of non-vowels[2] with respect to their place and manner of articulation.

Table 3.2: Classification of non-vowel speech sounds by place and manner of articulation.

Place of	Manner of articulation				
Articulation	Nasals	Oral stop	Fricative	Glide	Liquids
Bilabial (Lips closed)	/m/ “mom”	/p/ /b/ “pin” “bin”		/w/ “wet”	
Labiodental (Upper teeth on lower lip)			/f/ /v/ “fine” “vine”		
Interdental (Tongue behind teeth)			/T/ /D/ “thing” “then”		
Alveolar (Tongue touching gum ridge)	/n/ “noon”	/t/ /d/ “time “did”	/s/ /z/ “see” “zoo”		/l/ /r/ “like” “run”
Palatal (Tongue resting on palate)			/Z/ /S/ “azure” “she“	/y/	“yell”
Velar (Tongue closing oral cavity)	/G/ “sing”	/k/ /g/ “kick” “good”			
Glottal (Vocal folds fixed)		/Q/ glottal stop	/h/ “heat”		

The place of articulation is generally more visible than the manner of articulation. The place can often be determined from the lip shape and visibility of teeth and tongue, whereas the cues for manner are mainly governed by the larynx, velum and tongue. The larynx and the velum are not visible and the tongue might only be partially visible. The manner of articulation is therefore very hard to determine visually.

[2] The table does not include affricates, which can be considered as a combination of sounds.

The group of bilabials can normally be distinguished from other groups based on the closed lips. Within this group, the nasal /m/ and oral stops /p/ and /b/ can often be distinguished from the glide /w/, for which the lips are rounded but not completely closed. The labio-dental and interdental groups can usually be classified correctly if the upper teeth and the tongue tip is visible. The other groups are less distinct, but may be classified by finer features like the degree of protrusion or lip wrinkles.

The phonemes within the displayed groups of oral stops and fricatives differ in voicing, which is not visible. The bilabial nasals and oral stops can normally not be distinguished although some studies suggest that the degree of visual expansion of the cheeks, immediately prior to the opening of the lip opening, can be used for their discrimination [162]. Other studies indicated that the lips open more slowly for "ba" than for "pa" and still slower for "ma" [71]. The temporal frequency response of the human visual system falls off rapidly beyond about 15-20 Hz [98]. Berger [16] estimated that humans can only distinguish about 8 to 10 lip movements per second, whereas acoustic speech contains an average of about thirteen different speech sounds per second. Human speechreading might therefore be limited due to the temporal resolution of visual events. A machine speechreading system could easily double or quadruple this temporal resolution and might recognise more visual speech events than humans. The hypotheses that speechreading performance is related to temporal resolution is supported by findings of Summerfield [172] who concluded that human speechreading skills depend highly on the speed of low-level visual neural processing.

The classification of consonants is often more difficult in natural speech, where coarticulation might occur over several consecutive phonemes. For example in the syllables "klu", "kli", and "kla", the consonants /k/ and /l/ are highly dependent on the following vowel. Benguerel and Pichora-Fuller [12], for example, have shown that the intelligibility of consonants in vocalic context of /@/ at 78% can drop to 58% when spoken in a vocalic context of /u/.

Consonant recognition scores are also dependent on the talker which has been reported by Kricos [101]. Perceptual experiments have shown that both the number of viseme categories and the constituents of the viseme categories vary across talkers.

In natural speech, the number of non-vowel visemes is generally higher than the number of vowel visemes. Although several non-vowel viseme categories have been proposed in the literature, effects like natural speech, coarticulation, within/between speaker variability do influence the number of classes, which has not been thoroughly investigated. Finally, it should be noted that many non-vowel vises are also confusable with certain vowel vises, which further reduces the number of unique visemes.

3.4 Audio-Visual Complementarity

The main benefit of using visual cues is that they are often complementary to the acoustic signal, i.e. some phonemes which are difficult to understand acoustically are easier to distinguish visually, and vice versa. For example the utterances "ma" vs. "na" are highly confusable acoustically but easy to distinguish visually, based on the lip closure. On the other hand, the utterances "ba" vs. "pa" are easy to distinguish acoustically, based on the voice-onset time (VOT) [3], but highly confusable visually.

Different frequency bands convey different articulatory and acoustic cues. For example, acoustic cues to the place of articulation, mainly reside in the mid- to high frequency bands. For consonants, these cues are often of short duration and low intensity and therefore highly confusable in the presence of noise. In contrast, the visual cues reveal the place of articulation well and are quite robust when perceived by speechreading [171]. The cues for the manner of articulation tend to reside in the low-frequency bands, are of intense nature and slowly changing. These cues are therefore not very sensitive to acoustic noise. Thus, the visual signal often provides that part of speech information, mainly the place of articulation, which is acoustically most sensitive to noise. Walden et al. [181] have performed cluster analysis of the visual confusion of consonants made by trained lipreaders and derived a confusion tree for different numbers of clusters. Nine viseme groups were considered to be distinct visemes which were identified correctly at 75% of presentations. The pattern of the whole confusion tree was roughly the reverse of the confusion tree for acoustic consonant perception at different noise levels.

There exists a parallel between hearing-impairment and speech perception in noise. Most hearing impaired people suffer from losses in the sensitivity at high frequencies and reduced frequency resolution. As a result, energy in high frequencies is attenuated more and is resolved less well than energy in low frequencies. Cues for the place of articulation are therefore perceived less accurately than the place of articulation and voicing, as it is the case for speech perception in noise.

3.5 Audio-Visual Integration

The large influence of visual articulation on human perception of speech is demonstrated by the McGurk effect [130]. For this experiment, subjects are presented simultaneously with an acoustic recording of an utterance and a visual recording of a speaking face corresponding to a different utterance. The effect is that subjects perceive a sound which does neither correspond to the acoustic nor to the visual stimuli. For example, a subject hearing an audio recording of "baba" and seeing the synchronised video of a talker saying "dada" often resulted in perceiving "gaga".

How humans integrate visual and acoustic information is not well understood. Several models for human integration have been proposed in the literature. They can be divided into early integration (EI) and late integration (LI) models.

Early Integration In the EI model, integration is performed in the feature space to form a composite feature vector of acoustic and visual features. Classification is based on this composite feature vector. Figure 3.4 shows a block diagram of the scheme. The model makes the assumption of conditional dependence between the modes and is therefore more general than the LI model. It can furthermore account for temporal dependencies between the modes, such as the VOT, which are important for the discrimination of certain phonemes.

[3] The time delay between the burst sound and the movement of the vocal folds

Late Integration In the LI model, each modality is first pre-classified independently of each other. The final classification is based on the fusion of the outputs of both modalities by estimating their joint occurrence. In comparison with the early integration scheme, this method assumes that both data streams are conditionally independent. Furthermore, temporal information between the channels is lost in this approach.

Early integration models for speech recognition have for example been used in [27, 23, 2, 47, 135, 174]. A different EI model where the visual information was mapped to the frequency envelope of the acoustic signal has been described in [190, 189]. The aim of this approach was to map visual data to a common feature space which was chosen to be the frequency envelope. A similar approach but where the acoustic and the visual signals are mapped onto the motor space, represented by the coordinates of the tongue has been proposed in [158]. However, it is clear that neither humans nor machines are able to map the visual information into the motor space since the visual signal only reveals a small part of the vocal tract.

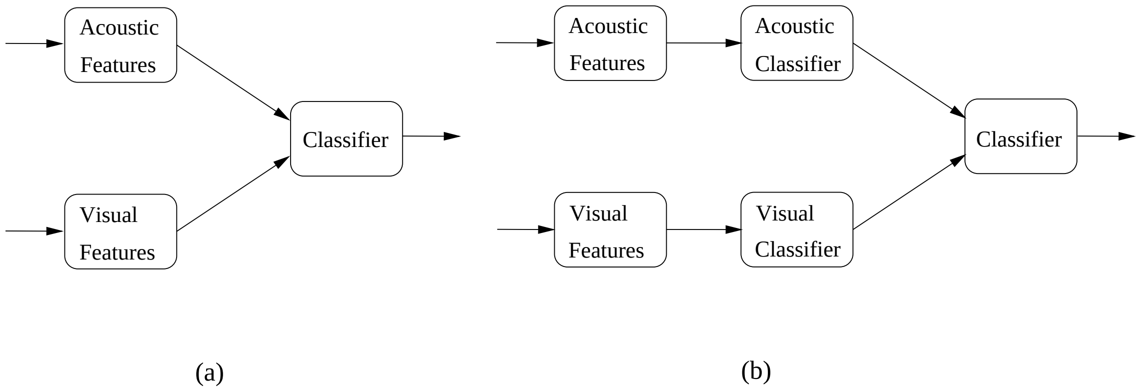


Figure 3.4: Block diagram for the *Early Integration* model (a), and the *Late Integration* model (b).

AVSR systems with late integration models have been described in [143, 168, 22, 164, 163]. Earlier findings supported the theory of the LI model and suggested that the human visual system determines the place and the auditory system the manner of articulation. The hypothesis is denoted as VPAM (Vision: Place, Acoustic: Manner) but has been disproved in [171]. It was concluded that integration of the two modalities occurs before phonetical characterisation and that the information of the two modalities is continuous rather than discrete. Erber and De Filippo [59] suggested that temporal information between mouth opening and VOT influences the perception of consonants and often enables the determination of voicing. These findings are also incompatible with the VPAM theory.

Massaro and Cohen [125] have proposed the Fuzzy Logical Model of Perception (FLMP), where integration is modelled by a fuzzy-logical process which combines the modalities by a multiplicative AND rule. This model can also be considered as an LI model. Although it is well known that temporal information between the two signals is important, only EI models can account for this information. Recently, attempts have been made to extend existing LI models by including explicit temporal information [127, 158].

It is still not well known how humans integrate different modalities, although it is generally agreed that integration occurs before speech is categorised phonetically. This was concluded by the following independent studies: Summerfield [170, 171] argues that audio-visual integration must occur before phonetic categorisation since speech can be perceived audio-visually when the acoustic signal is replaced by a signal that cannot be categorised acoustically. Such a signal can be a sequence of acoustic pulses synchronised with the times when the vocal folds close [79]. Several studies have shown that consonants which differ by the VOT such as "bi" and "pi", are distinguished based on the evidence of both modalities [59, 82]. Integration must therefore take place before phonetic categorisation. Similar evidence was found by Massaro [126], who concluded that human perception was better modelled by a EI model than by a LI model. In another study, Braida [21] has derived models of cross-modal integration from the confusion matrices of single modalities. Experimental results for consonant identification were more consistent with the prediction of the EI model than the LI model or the FLMP.

In acoustic speech perception, on the other hand, there is much evidence that humans perform partial recognition across different acoustic frequency bands. This was concluded by Fletcher's early work [68] and has led to the development of the Articulation Index (see Sec. 3.6). Fletcher performed perception experiments on CVC (consonant-vowel-consonant) syllables, where the speech signal was filtered by different low-pass and high-pass filters. He showed that phone probabilities are estimated in independent frequency bands and are optimally merged so that the estimated error of the merged estimates becomes the product of the estimation errors of each frequency band. Or, in other words, the partial recognition errors for each frequency band are independent and the combined error is obtained by the sum of the partial log-errors of all frequency bands. How the information of different streams is merged is not well understood but the processing of independent feature streams seems to explain the high robustness of human speech perception in noise. It is surprising that only recently, speech recognition research has followed the approach of processing different frequency bands independently, which has lead to highly robust recognition accuracy in noise [20].

The interesting point now is that the auditory system seems to perform partial recognition which is independent across channels, whereas audio-visual perception seems to be based on early integration of acoustic and visual features, which assumes conditional dependence between both modalities. These two hypotheses are controversial since the audio-visual theory of early integration assumes that no partial categorisation is made prior to the integration of both modalities.

3.6 Audio-Visual Intelligibility

Although the visual information provides only part of the speech information, some humans perform speechreading very efficiently which can lead to almost perfect speech perception. However, such a high performance can only be achieved by using higher level knowledge like phonology, lexica, prosody, syntax, semantics and pragmatics. Phonology describes the sequences of phonemes which are possible within a language, whereas lexical knowledge is used to constrain the sequences of phonemes to make up words. The encoding process at these two levels is aided by the prosody which considers the stress and intonation patters of speech. The encoding of word sequences is aided by the syntax which requires the word sequences to form grammatically correct sentences. Speech perception is further aided by semantic knowledge which constrains the sentences to be meaningful. Finally, the highest level knowledge used in speech perception is pragmatics, which considers the context of the conversation to form a meaningful sentence. It becomes clear now, that speechreading performance is highly dependent on the kind of knowledge sources used for speech perception. The entropy at the phoneme level is the highest and decreases with the number of knowledge sources used.

It is often desirable to evaluate the visual contribution to speech perception at a low level of phonemes or sub-words in order to avoid that other knowledge sources affect the performance. This is also the goal of speech perception tests when using nonsense words. In analogy, the performance of automatic speech recognition systems is often evaluated separately for phoneme or sub-word classes. This approach assumes a bottom-up processing, where recognition at the phoneme level is not influenced by higher level sources, and where the decision is deferred until the highest knowledge level.

An overview of lip-read vowel intelligibility spoken in a consonant context and consonant intelligibility spoken in a vocalic context can be found in [15]. The average accuracy for five or six different vowels is roughly 50% and for 22 or 23 consonants about 30%. It should be noted that since less than half of all vowels were used in the experiments, the intelligibility for all vowels should be considerably lower. Similar tests which correspond more to the actual number of visemes have been performed by Benguerel and Pichora-Fuller [12]. They have performed tests for three vowels which resulted in an average vowel intelligibility of 89%, while tests for ten consonants yielded an average intelligibility of 71%.

Auer and Bernstein [7] performed a computational analysis of the effect of the number of viseme classes on speech intelligibility. For a vocabulary of 30,000 words, they found that 10 viseme classes lead to about 19% of unique words, which rises to about 56% for 20 viseme classes and to about 77% for 30 classes. For a vocabulary of 1,000 words, the scores for the three viseme numbers were about 29%, 71% and 91%. These results can be interpreted as results achieved by a human speechreader who uses no syntactic, semantic or pragmatic knowledge and who is able to distinguish the given viseme classes. These results are encouraging, considering their potential improvement by using other knowledge sources. They also explain why speechreading without acoustic cues can lead to very high intelligibility.

Sumby and Pollack [169] investigated the visual contribution to speech perception in noise. They suggested that seeing the talker's face is equivalent to improving the signal to noise ration (SNR) by up to about 15 dB. Or, in other words, subjects who speechread can tolerate an about 15 dB lower SNR while maintaining the same performance as listening only. A 1 dB higher SNR corresponds to about a 5 % change in intelligibility. For example, seeing the speaker's face at a SNR of -6 dB improves the intelligibility rate from about 20% to about 80% [118].

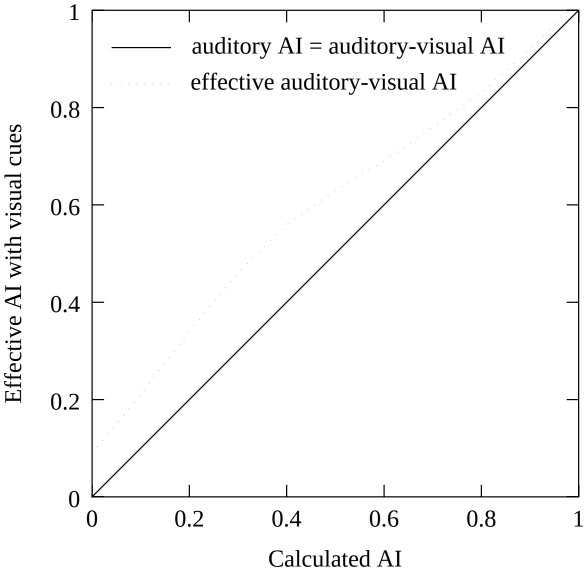


Figure 3.5: Relation between the calculated AI and the effective AI when auditory cues are combined with speechreading.

One theory for speech intelligibility is the *Articulation Theory* [69] which is based on the pioneering work of Fletcher [68, 3]. It attempts to predict speech intelligibility as a function of the intelligibility of different frequency bands. The theory has been expanded to include the effect of speechreading [4]. The method assumes that a signal's auditory articulation index (AI) determines its auditory-visual AI uniquely. The term articulation is used as the probability of correct recognition of nonsense sounds. The procedure makes the same auditory-visual prediction for different but equally intelligible frequency bands. Fig. 3.5 shows the relation of the auditory AI to the auditory-visual AI (after [4]). The contribution of speechreading at a high auditory AI is relatively small while at an auditory AI of 0.1, speechreading can roughly double the auditory AI.

The auditory-visual AI is based on experiments performed by Sumby and Pollack [169]. It does not take into account that frequency bands of roughly equal importance for auditory speech perception might be of different importance for auditory-visual speech perception. As different frequency bands convey different acoustic and articulatory cues, as shown in Sec. 3.4, it is reasonable to assume that low frequency bands would complement speechreading more than high frequency bands and lead to a higher AI, especially in noise. Grant and Braida [80] have therefore evaluated the AI for different frequency bands but found that the AI was well predicted by the ANSI standard. They concluded that the auditory-visual scores can be well predicted by the auditory score irrespective of the spectrum of the auditory signal. This implies that audio-visual speech perception can not be understood adequately in terms of feature based (place, manner) descriptions. The findings support Fletcher's theory of partial channel recognition and contradict the hypothesis of early feature integration.

3.7 Analysis of Visual Speech Features

Both the understanding of human speechreading and the development of a machine speechreading system would benefit from a better understanding of the features which contribute to visual speech information. The determination of important features could help a speechreading system to look for important features which carry speech information and to disregard those which do not contain such information or which are highly variable across subjects.

One of the earliest studies about the determination of the physical characteristics of lips during vowel production has been reported by Fromkin [70]. In this study, the horizontal extension and the vertical separation of lips were measured and it was shown that the relation of these parameters for 12 American vowels was consistent across talkers. The measurements were mainly made on photographs for static vowels.

Jackson et al. [91] used judgements of similarity of vowel lip movements between pairs of /h/-V-/g/ syllables to derive perceptual dimensions presumed to underlie vowel speechreading. The /h/-V-/g/ context was used because it produces minimal coarticulation effects on the visibility of the vowel. Similar parameters to those of Fromkin, like vertical and horizontal width of opening, and others, like the area of opening, correlated with perceptually based vowel configurations.

Geometric parameters of the lips for the French language have been performed by Abry and Boë [1] and Lallouache [103]. Benoît et al. [14] have analysed the correlation between the horizontal and vertical extension of the mouth opening and found that they are highly correlated with the area of the opening. They further found that the dependence of the two parameters was for some vowels linear and for others nonlinear.

Montgomery and Jackson [133] performed perceptual tests and analysed physical measurements of vowel lip shapes, spoken in a /h/-V-/g/ context. They reported about 50% correct perception of vowels. They found that physical measurements could only account for about half of the variability found in the perceptual data. It was concluded that other measures like the visibility of teeth and the complex characteristic of rounding are not well represented in their physical parameters. Rounding is expressed as lip protrusion, lip tension, a specific indented upper lip shape, reduced area of lip opening, lip wrinkles and reduced visibility of teeth.

Brooke and Summerfield [26] performed perception experiments using computer-controlled graphical displays of a talking face. The animation was based on facial measurements, taken from a talking person. The graphical display was represented by the outline of the head and the animation was performed by synthesising the outer and inner lip contours and the chin. Human speechreading performance for vowels on the natural face was 98% correct and on the synthetic face 69% correct. It was concluded that additional cues like the visibility of teeth and tongue tip appear to be required for accurate identification of vowels.

McGrath investigated the influence of teeth in speechreading vowels [129]. Without the visibility of teeth, the subject's response was 51% correct and with the visibility of teeth, it increased to 57%. The visibility of the teeth mainly helped to discriminate close front vowels as in ‘beeb’ from more open vowels as in ‘berb’. They also helped to distinguish rounded vowels as in ‘boub’ from unrounded vowels as in ‘barb’.

Finn [64, 63] attempted to determine geometric parameters of the lips for the recognition of consonants. She used measures like the height and width of the mouth opening, vertical spreading of the upper and lower lips and “cornering” which is related to the distance of the outer and inner lip corners. No information about teeth or lips was included. Experiments were performed on a single speaker for consonants which were preceeded and followed by the vowel /a/. Recognition was performed on the middle image of the utterance and achieved an accuracy of 87%.

Benoît et al. [13] have described experiments which confirmed that the lips play the most important role in visual speech perception but that the visibility of the chin, teeth, tongue and cheeks are important too.

3.8 Discussion

The benefit of visual information in speech perception is relatively small for clean speech but increases drastically with the noise level. The benefit is mainly due to the complementarity of visual and acoustic information. It is suggested that good speechreading performance is highly correlated with the speed of low-level visual neural processing. This might have important implications for the design of automatic speechreading systems, as the temporal resolution of such a system could be several times higher than the temporal resolution of human vision.

There exists no consensus on how many visually distinguishable units exist. The number varies across subjects and within subjects. The intelligibility of visemes is generally smaller for natural speech and for unstressed phonemes. The visibility of individual visemes depends highly on coarticulation. Only few experiments have been performed to evaluate the effect of coarticulation, therefore a complete account of its consequences is not possible. The number of visemes depends furthermore on the lighting condition, the visibility of small features in the oral cavity and on finer details like puffed cheeks, wrinkles, and so forth, which are not well understood.

There is much evidence that humans integrate the two sources of information before they categorise speech phonetically. There is however still much discussion about how the two modalities are represented and how they are integrated. Although different frequency bands convey different articulatory and acoustic cues, the auditory-visual AI was well predicted by the auditory AI, irrespective of the spectrum of the auditory signal. This implies that the two streams are conditionally independent and that some kind of precategorisation takes place. Although this is in contrast with findings about acoustic-visual speech perception, it supports Fletcher's theory of partial independent recognition across frequency bands. In summary, it seems that we still don't know well how humans process audio-visual speech.

Research in automatic speech recognition was often motivated by human perception of speech. The reason for this is obvious, it is the very high performance of human speech perception. High performance is maintained for natural speech, spontaneous speech, various dialects, different speed, difficult listening conditions, simultaneous talkers, and so forth. Current high performance ASR systems try to make use of all the knowledge sources readily used by humans such as lexica, syntax, semantics and pragmatics, in form of pronunciation dictionaries, lexicon and language models etc. One knowledge source which has only been considered recently is the use of visual information. ASR systems which do not use visual information can be considered as modelling speech perception of blind humans. Although state-of-the-art ASR systems achieve reasonably high performance under certain condition, their performance drops drastically in the presence of noise. It is therefore reasonable to expect future ASR systems to only achieve performance levels similar to the ones of humans if all sources of knowledge, including speechreading, are incorporated.

Chapter 4

Description of the Used Databases

This chapter describes the two databases considered for the experiments. A survey of existing bimodal speech databases was performed and indicated that only the Tulips1 database, collected by Movellan [137] was publicly available and which was therefore used in this work. A detailed survey on existing and planned audio-visual speech databases can be found in [35]. The second database considered is the M2VTS database which has recently been collected by Pigeon and Vandendorpe [146] in the framework of the European ACTS M2VTS project, under which part of this work was performed. Both databases are now publicly available and allow the performance comparison of algorithms from other researchers.

4.1 The Tulips1 Database

Tulips1 is an audio-visual digit database of undergraduate students from the Cognitive Science Department at the University of California, San Diego [137]. It consists of simultaneous audio and video recordings of the first four English digits, spoken twice in isolation by 12 subjects. The set of words spoken the first time will be denoted as Set 1 and the words spoken the second time as Set 2. The subjects were asked to talk into a video camera and to position themselves so that their lips be roughly centred in a feed-back display. The recordings were taken in a windowless room. Due to sub-optimal illumination, several of the image sequences contain large areas of reflections or shadows. Some of the images are out of focus and sometimes the lips extend beyond the boundary of the image frame. The database consists of speakers with different ethnic origin, and also include subjects with make-up or facial hair.

Database Description:

General

- Number of subjects: 12 (9 male, 3 female)
- Speech data: 4 isolated digits (one, two, three, four)
- Repetitions: 2 (denoted as Set 1 and Set 2)

- Total number of utterances: 96 words

Video

- Image format: 100 × 75 pixels
- Image frame rate: 30 Hz
- Pixel resolution: 8 bits grey-level
- Total number of images: 934
- Average number of frames per word: 9.7

Audio

- Sampling rate: 11,127 kHz
- Sample resolution: 8 bits

Figure 4.1 displays example images of all subject in the database and Fig. 4.2 displays example sequences of three subjects saying the word "one" (only every second image of the sequences is shown).

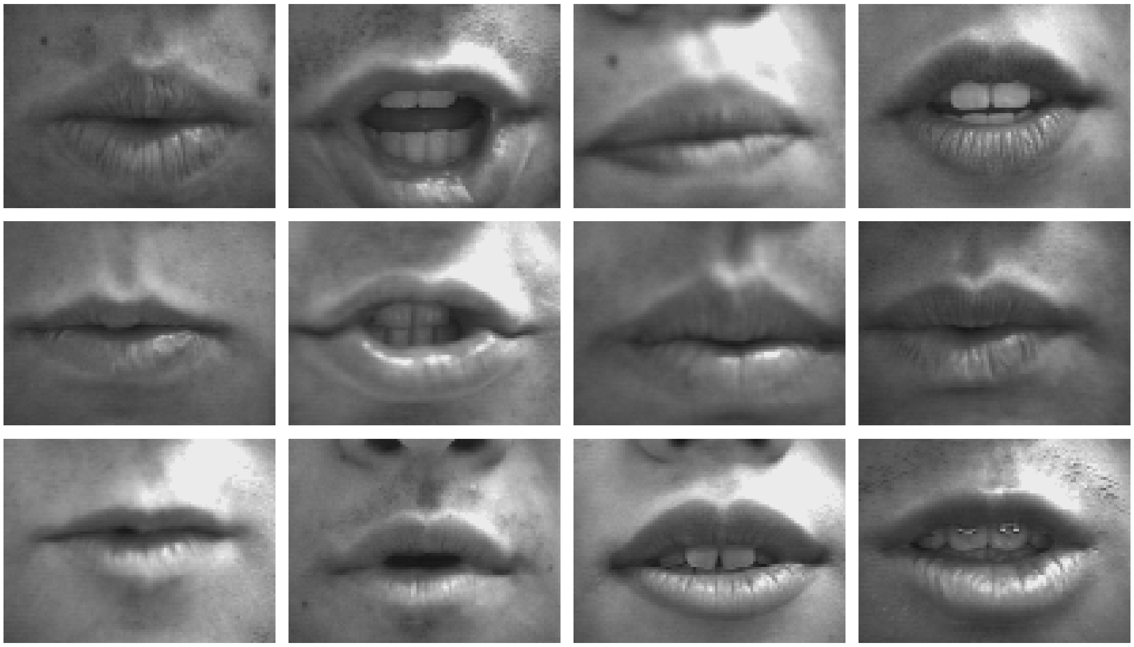
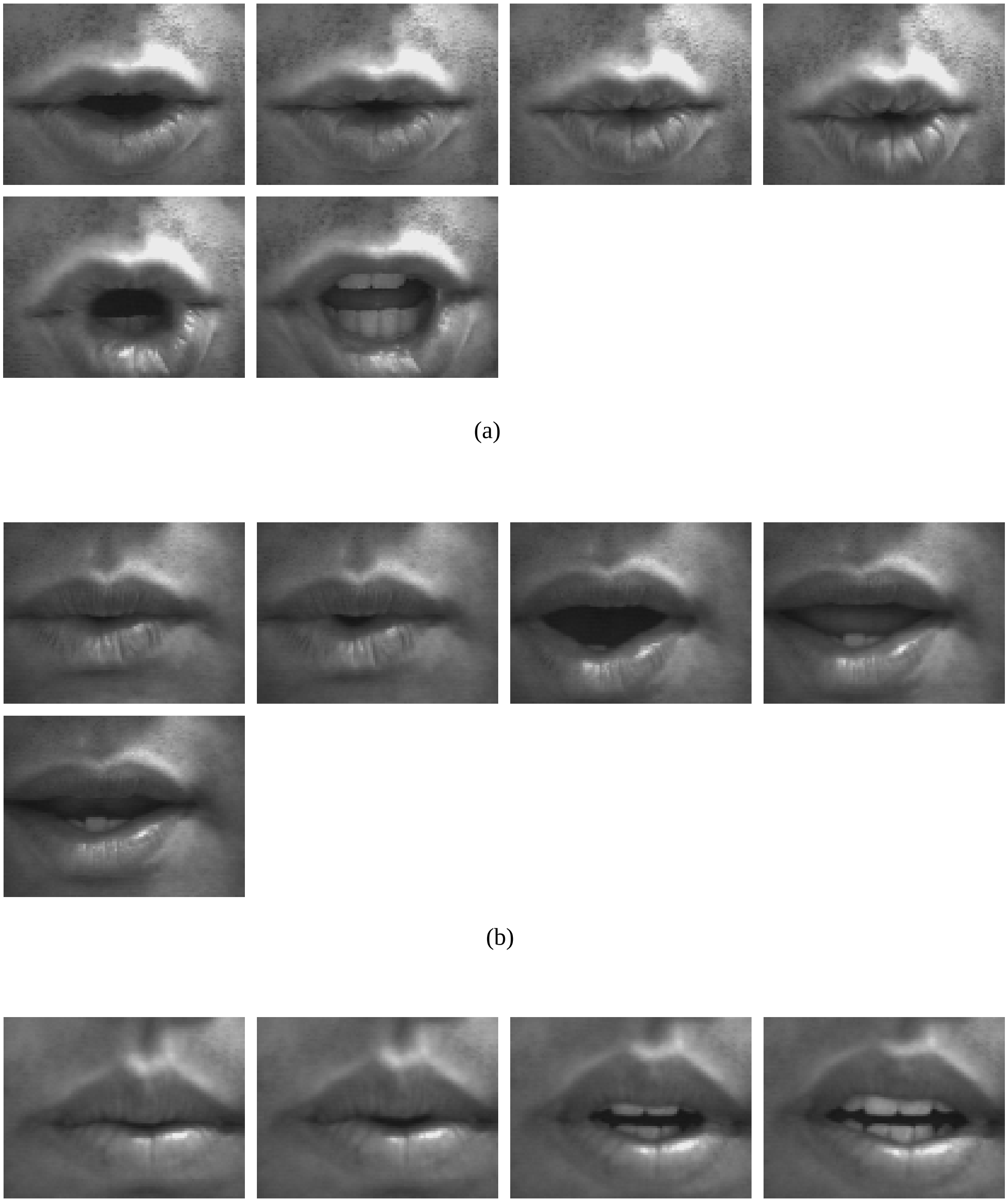


Figure 4.1: Example images of all 12 subjects of the Tulips1 database.



(a)

(b)

(c)

Figure 4.2: Example sequences of the word “three”, spoken by three subjects (a, b, c) of the Tulips1 database, displayed in raster order.

4.2 The M2VTS Database

The M2VTS audio-visual database [146] has been collected within the framework of the European ACTS-M2VTS project to address the issue of multi-modal person authentication. It contains recordings of 37 speakers pronouncing in French the digits from zero to nine. One recording consists of the audio recording and of the video recording of the spoken digit sequence. In addition, motion sequences were also recorded while the subjects were rotating their head by 90 degrees to the left and to the right. The rotation sequences were not used in the experiments reported here. Five recordings have been taken of each speaker, at one week intervals to account for minor face changes like beards and hairstyle. For each person, the most difficult shot to recognise was labelled as the 5^{th} shot. This shot differs from the others in face variation (head tilted, unshaved), voice variation (poor voice SNR), or shot imperfections (poor focus, different zoom factor). The images contain the whole head of the person in frontal view as shown in Figure 4.3. Apart from the imperfections of Shot 5, additional impairments are due to some persons who could not keep from smiling during the recordings. The original video sequences which are in colour format were converted to grey-level sequences and for all experiments reported here, only the grey-level sequences were used. Figure 4.4 displays a typical image sequence of the French digits “zero” to “neuf” (Only every 5^{th} image frame is displayed in the figure).

Database Description:

General

- Number of subjects: 37 (25 male, 12 female)
- Speech data: sentences of 10 continuously spoken digits (zero, un, deux, trois, quatre, cinq, six, sept, huit, neuf)
- Repetitions: 5, recorded at one week intervals (denoted as Shot 1 - Shot 5)
- Total number of utterances: 1850 words or 1850 sentences

Video

- Image format: 286 × 350 pixels
- Image frame rate: 25 Hz
- Pixel resolution: 8 bits grey-level (converted from colour)
- Total number of images: 27,427
- Average number of frames per word: 14.8

Audio

- Sampling rate: 48 kHz
- Sample resolution: 16 bits

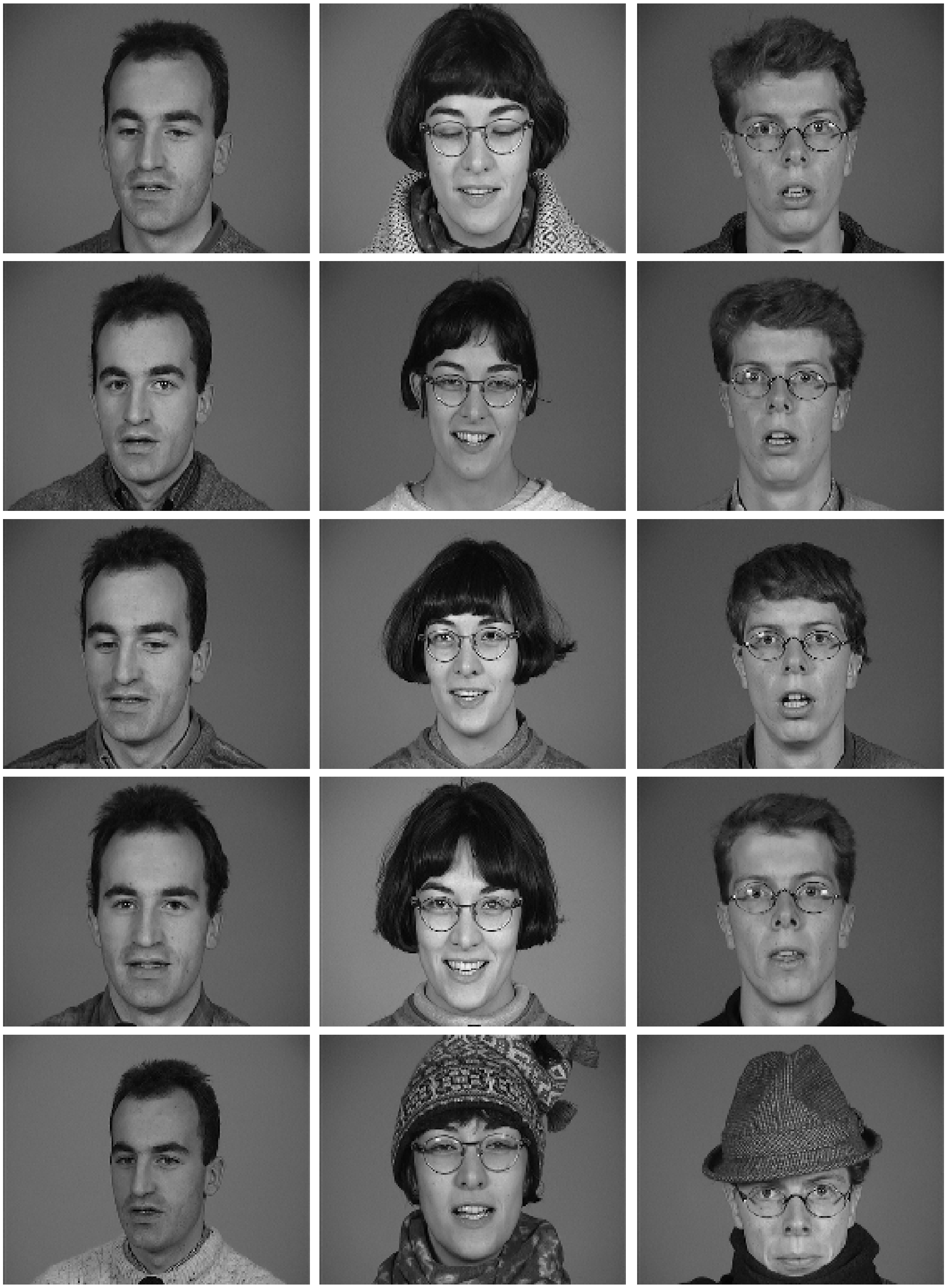


Figure 4.3: Example images from the M2VTS database. Each row represents a different session. The 5^{th} row represents the most difficult images to recognise.

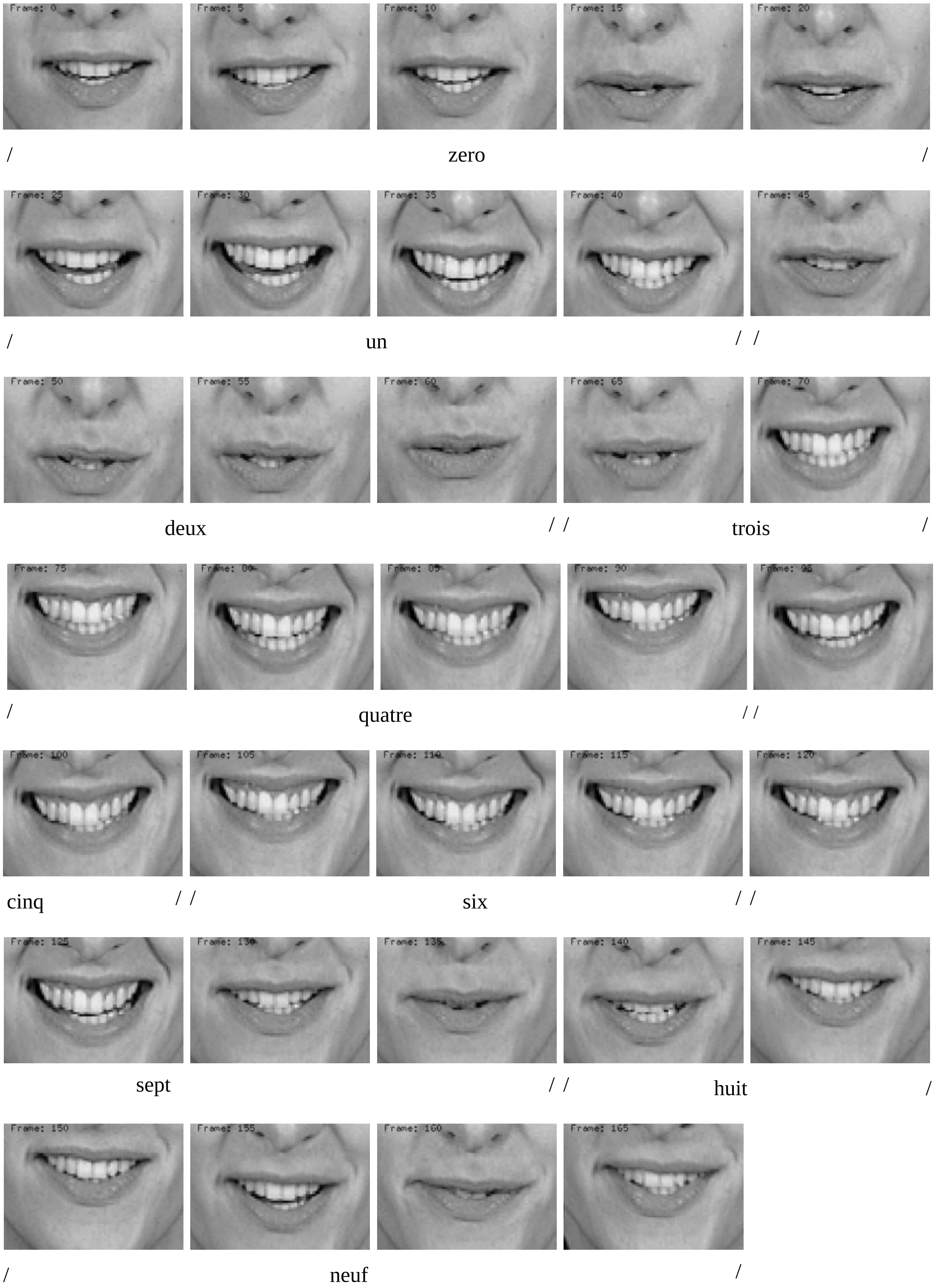


Figure 4.4: Example sequence from the M2VTS database.

Chapter 5

Feature Extraction

This chapter describes a method for visual feature extraction to be used for speech recognition and speaker recognition applications. An appearance based model is used to represent both the shape of the lip contours and the grey-level intensity around the mouth area. The model is used to localise and track the lips in image sequences and to extract important features. Experimental results are presented for two different databases.

Facial speech feature extraction and modelling has become an important issue in both, automatic speech processing and automatic image processing. Potential applications of these methods include

- acoustic-visual speech recognition [143]
- acoustic-visual person authentication [110]
- lip synchronisation [34]
- speech-driven talking heads [140]
- speech-based image coding [134].

Robust and accurate facial feature analysis is a difficult object recognition problem, because of large appearance differences between and within subjects and differences due to varying environmental conditions. Inter-subject variability is mainly due to different appearance of lips, skin, facial hair, and make up, whereas intra-subject variability is largely caused by speech production and facial expressions, but it can also be due to wet lips or teeth, causing visual reflection. Environmental differences depend on the illumination and include the changes in the pose of the face.

Motivated by psychological studies, several researchers have developed speechreading systems which have demonstrated the potential use of visual information to improve the robustness of acoustic speech recognition systems in noise. While these systems have validated the benefit of visual speech information, there is still much discussion about how to determine which visual features are important for speechreading, how to represent them, and how to extract them automatically in a robust manner [117]. Most approaches for speechreading have constrained or circumvented the feature extraction problem by marking the subjects' lips with colour or reflective markers, by recording the lip movements with a head mounted camera, by using one subject only for the experiments, by hand segmenting the lip region, or by using very controlled lighting conditions. This chapter attempts to solve the feature extraction problems without the use of such constraints.

The task of feature extraction can be divided into feature localisation and feature tracking. The first step involves the localisation of the features in an image, while the second step is concerned with tracking the features over an image sequence and extracting important information useful for a certain application. Feature localisation is usually more difficult than feature tracking, since the coarse position, scale, illumination and subject identity are not known. These variables will however remain fairly constant within an image sequence which facilitates lip tracking. Feature localisation is often performed under certain constraints, e.g., by using images where the face is already segmented from the background, or by using images which contain the mouth area only.

5.1 Previous Approaches

The problem of facial feature extraction has been studied extensively, particularly for the application of face recognition. Feature extraction normally consists in the localisation of salient features like the eyes and the mouth and sometimes also the face outline. For face recognition applications, it is usually assumed that the person has a neutral facial expression with the mouth closed, which simplifies the localisation of the mouth.

5.1.1 Feature Localisation

Deformable Templates

Yuille et al. [192, 193] developed the deformable template method which was used for feature extraction and person identification. The method is based on simple geometric shapes which can deform and move under the influence of image evidence, in an attempt to minimise an energy function. Although certain constraints for shape deformation are used, it is difficult to ensure that the model only deforms into legal shapes and at the same time is flexible enough to resolve fine contour details. Several other researchers have described similar methods [90, 39].

Craw et al. [50] presented a model based approach for feature detection where the model is derived from a training set. Simulated annealing is used to find a maximum of an objective function, which considers image gradients and shape deformation.

A unified approach for feature localisation was described by Lanitis et al. [104, 106, 105]. The approach is based on statistical models which were derived from a training set of face images. Shape deformation is modelled by a point distribution model (PDM) [46] and active shape model (ASM) search [42] is used to locate the face together with the facial features in the image. Grey-level appearance is modelled using flexible grey-level models which are attached to points on the shape model and which are used in image search. This method has also been applied to faces with varying facial expressions. The approach for facial feature extraction describe here is closely related to the PDM method.

Template Matching

Brunelli and Poggio [30] have used a correlation method to locate facial features. The grey-levels of the images are usually first normalised before templates of different scales are used to find a maximum correlation. The normalisation process is however critical and sensitive to different illumination, pose, rotation, and scale. Also, the method does not deal well with deformations. Moghaddam and Pentland [132, 141] described an approach based on the Karhunen-Loéve (K-L) expansion to construct *eigen-features* which are used to find facial features. An exhaustive search is performed over the entire image and at different scales, and a density measure is used to compute a local measure of target saliency. This method can be considered an extension of the correlation method but where image variability is taken into account. The drawback of the scheme is that it does not model image deformation explicitly and that the spatial relationship between the different features is not used in image search.

Knowledge based Approaches

Yang and Huang [186] described an attempt to detect faces and facial features using a knowledge based method consisting of three levels. Face detection was based on pyramid images which were used to search for possible face candidates. Edge detection and multi-binarisation methods were then employed on these face candidates to find facial features. These were evaluated by a number of hand crafted rules to make a final decision. The method is highly dependent on the rules defined by the user and it might be difficult to define such rules for a large number of conditions and faces.

A more elaborate knowledge based approach has been described by Yow and Cipolla [188], who presented a system which uses a family of Gaussian derivative filters to search and extract human facial features from the image. Steerable and scalable filters are applied to find line features of the eyes, eye brows, nose, and mouth. The features are grouped into partial face groups and a belief network is used for verifying possible face candidates. The simple representation of facial features might however not hold for a large number of subjects and for different facial expressions.

Reisfeld et al. [156, 155] developed a generalised symmetry measure as a computer vision analogy to attention and fixation. The symmetry transform does not require knowledge about the shape of the object and assigns a continuous symmetry measure to each point in the object. Knowledge about the geometry of the face is then applied to the symmetry measure to locate eyes and mouth.

Duchnowski et al. [56] described a neural network based method for mouth and mouth corner localisation. The mouth corners are used to place and scale a window over the face image which serves as feature vector for a speechreading system. The first network is trained on edges obtained by the Sobel operator and gives a coarse position of the mouth. The second network is trained on horizontal gradients and is used to find the mouth corners. This approach learns the appearance of the lip corners from a training set but due to the use of gradient information and the limited information contained in the lip corners, it might not be robust for a large number of persons and recording conditions.

Visual Motion Analysis

Wolf et al. [182] described a method for mouth detection for the purpose of speechreading. The technique is based on edge detection, thresholding, and triangulation of the eyes with the mouth. However, it was not always found to produce stable results, especially if the eyes were closed. The method was therefore augmented by motion information obtained from the difference of consecutive images after filtering and thresholding.

Mak and Allen [121] have described an approach for the localisation of the mouth, as a preprocessing step for lip tracking. The algorithm is based on morphological filtering, subtraction of subsequent images, thresholding, and cluster analysis. This method assumes that the intensity of the mouth is considerably lower than that of the skin. The largest cluster is assumed to represent the mouth and its centre is used as the centre of the mouth. These assumptions might be invalid for persons with beard or persons with dark skin.

5.1.2 Feature Tracking and Extraction

The main approaches for extracting visual speech information from image sequences can be grouped into *image based*, *geometric feature based*, *visual motion based*, and *model based approaches*.

In the *image based* approach, the grey-level image containing the mouth is either used directly or after some pre-processing as feature vector. The advantage of this method is that no data is disregarded. The disadvantage is that it is left to the classifier to learn the nontrivial task of finding the generalisation for image variability (translation, scaling, 3D rotation, illumination) and linguistic variability (inter/intra speaker variability). Another disadvantage is the high dimension and high redundancy of the feature vector. It might appear that such approaches will work if a large enough collection of training examples is available, which cover all possible image variations. However, just the linguistic variability, present in low-dimensional acoustic speech features, causes major problems in acoustic speech modelling, despite the availability of very large databases. This indicates that high-dimensional visual features, which contain additional image variability of several different origins, will lead to even larger difficulties in visual speech modelling.

The *visual motion based* approach can be considered a sub group of the image based approach. It is assumed that the visual motion during speech production contains relevant speech information. Visual motion information is likely to be robust to different speakers and to different skin reflectance, however, the algorithms usually do not calculate the actual flow field but the visual flow field. A further problem consists in the extraction of relevant features from the flow field.

The *geometric feature based* approach assumes that certain measures such as the height or width of the mouth opening are important features. The automatic extraction of these features is however not trivial and most of these approaches have used semi-automatic methods or have painted the lips of the talker to facilitate feature extraction. The advantage of the method is that the features are represented in a compact form. The disadvantage consists in the difficult automatic extraction of these features and the subjective choice of the features to consider.

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

Image based approaches

Yuhas et al. [190, 189] presented a speechreading system where the whole grey-level image containing the mouth area was used as feature vector for the ASR system. The image was sub-sampled to provide an image of 20 pixels. Similar feature exaction methods were described in [183, 22]. A method which only uses the horizontal and vertical grey-level vectors, centred at the mouth, as feature vectors has been proposed in [182]. Silsbee [164, 163] has described another image based method based on vector quantisation, where 17 codevectors describing different mouth configurations were selected by hand. The approach was however described as being extremely sensitive to differences in width and height of the lip opening and very sensitive to the presence/absence of teeth.

Wu et al. [183] have compared the grey-level image representation with binary representations of the mouth, similar to those of Petajan [143], and geometric features. The binary image represented the oral cavity, either including or not including the teeth. The geometric features were represented by the height, width and area of the cavity which were obtained from the binary image. Recognition results for static images of vowels were slightly better for the binary representations with teeth and for the geometric features, than for grey level images. The experiments however seem too small to draw major conclusions about feature representations. Furthermore, the recognition of speech from static images simplifies the problem substantially.

Techniques for the reduction of the feature space based on a K-L expansion of the images containing the mouth area have been proposed in [23, 27, 57, 138]. A multi-scale nonlinear decomposition method based on *sieves* has been described in [128]. These methods reduce the features space considerably and are generally less sensitive to noise. The theoretical foundation for these approaches is however not very clear. The extracted features account for both shape variability and intensity variability. The features might therefore be less discriminative for a large population of speakers with different appearances and mouth shapes.

Movellan [137] described a feature extraction method where the image was first normalised and symmetrised using the vertical mid-line of the image as the axis of symmetry.

Only one half of the symmetrised image was kept for further processing. The other half of the image was replaced by the pixel by pixel differences of consecutive images. The resulting images were then filtered, sub-sampled, and scaled by a function which approximates histogram normalisation over the whole image sequence. This resulted in a 300-dimensional feature vector per image. The temporal features have been shown to improve the recognition accuracy considerably. Gray et al. [81] have compared this feature representation with optical flow and PCA and found it to outperform both methods. A short-coming of the method is that the area of interest (AOI) was used directly without knowing the exact position and scale of the lips. Gray et al. have therefore used the lip tracking results obtained by Luettin et al. [111], to position and scale the AOI around the lips. This resulted in substantially higher recognition performance, which suggests that all image based methods might obtain higher performances by the use of a lip tracker to normalise the AOI prior to further processing.

Visual Motion Analysis

Mase et al. [123] described a speechreading system based on optical flow, which was calculated on four windows near the mouth. Parameters were extracted from average flow vectors in each window. The underlying assumption of this approach was to estimate the major muscle activities involved in speech production. The 2D optical flow, however, only corresponds to the 2D motion flow under certain conditions: the radiance has to remain constant during motion, the object has to be rigid, and the motion should be smooth. All these assumptions are incorrect for lip motion analysis.

A method for feature extraction based on block matching has been presented in [121]. The image is divided into squares of equal size and exhaustive search is used to find the best estimated placement. Problems were reported for regions which were uniform in shading and texture. If the oral cavity area changes and teeth and tongue become visible, the block matching results might no longer correspond to actual motion.

Geometric Features

Petajan [143, 144] was probably the first researcher to develop a speechreading system. The system was based on geometric features of the mouth opening, like height, width and area. A simple thresholding technique was used to find the mouth opening and linear time-warping was applied to match test sequences to training templates. A similar feature extraction method was later used by Goldschen [77], who developed a continuous speechreading system. While these systems have obtained very good speechreading results, a major problem was the extraction of the geometric features which could only be performed with human assistance.

Finn et al. [63] described a speechreading system where features were extracted by measurements on highly reflective dots, placed around the speaker's mouth. A similar feature extraction method was used in [168]. Other researchers have highlighted the lips with colour to facilitate visual processing and extraction of geometric features [2, 51, 97]. A system where 3D information of the face is recovered and used for speech recognition has been described in [119, 47]. Small reflective markers were placed on the subject's face which was illuminated by a infrared stroboscope. The perspective coordinates of the markers were recovered by two cameras and 3D coordinates were obtained by means of a stereo-photogrammetric procedure. Approaches using markers or lipstick mainly serve to determine the usefulness of visual features but they can not be applied to practical applications.

Deformable Models

The application of *deformable templates* [193] to lip-tracking has been described in [85, 153]. The outline of the lips is modelled by a set of hand coded polynomials, which are matched onto the outline of the lips, represented by the image gradient. Since the deformation of deformable templates is constrained by the initial choice of polynomials, they are often not able to resolve fine contour details. Image search is performed by fitting the template to image gradients, assuming strong edges at the lip contours. This assumption is however often violated as the gradient along the contour is dependent on the speaker, illumination, reflection, facial hair, visibility of teeth, and mouth opening. A similar approach but based on colour information has been presented in [40].

Kass et al. [96] have described active contour models, so-called *snakes*, for lip-tracking. These are able to resolve fine contour details but shape constraints are difficult to incorporate and one has to compromise between the degree of elasticity and the ability to resolve fine contour details. Bregler et al. [24, 25] described a method based on *snakes* for tracking the outer lip contour, but where the contour was constrained to lie in a sub-space learned from a training set. The energy function for image search considers the distance of the contour to the sub-space, an internal energy which is minimal when the contour follows a straight line, and the sum of gradients along the contour. The weights for each contribution were determined empirically. Similar to deformable templates, the approach assumes that image gradients are well suited to represent the lip contours. After performing speech recognition tests with the extracted shape parameters, Bregler et al. found that the information provided by their tracking results were not distinctive enough to give reasonable recognition performance. They therefore used the components of the grey-level matrix, centred and scaled around the lips, as speech features. Another modified version of *snakes* together with the use of colour information has been proposed by Chiou [37, 36] and a version which was used in combination with deformable templates has been reported in [88].

An approach based on *splines* [8] and Kalman filters [75] has been described in [51, 97]. Shape constraints were imposed on the deforming template by limiting the number of degrees of freedom. This was performed by fitting the spline to two extreme mouth shapes. Image search was done by searching for high contrast edges. However, tracking was only stable when a lip-stick was worn to enhance the contrast around the lips. Ramos Sanchez et al. [161] combined a spline based lip model with a colour based image search. Lip boundary search was formulated as a two-class ('lip' or 'skin') classification problem. They found however the lip chromaticity not to be adequately represented by a unimodal distribution. They therefore constructed individual chromaticity models for each subject, which prevents the use of the model for multi-speaker or speaker independent feature extraction. A similar colour based approach but where shape modelling was based on Bezier curves has been described in [180]. The energy function considered internal and external modal forces, dynamic constraints, and colour information from the image. Their relative contribution to the energy function was determined empirically.

5.1.3 The Need for Better Feature Extraction Methods

It is often argued that visual information in a speech recognition system is most beneficial in real world applications where noise is present. If the aim of a visual-acoustic speech recognition system is therefore to be used in such an environment, the robustness of the visual subsystem in such an environment is of crucial importance. The requirement of the feature extraction method can be divided as follows:

- Robustness to speakers: ethnic background, male/female, facial hair, make-up, age.
- Robustness to pose: translation, 3D rotation, scale.
- Robustness to illumination: natural/artificial, direction.
- Extraction of important speech related features, preferably invariant to the previous three factors.

In order to be robust to all these sources of variability encountered in real world applications, the feature extraction method should use as much knowledge about the scene as possible. One way of incorporating such knowledge is to build a model of the object. Image based methods and motion analysis methods which do not use any model knowledge are unlikely to achieve the desired performance.

Most model-based systems developed so far have relied on heuristics about shape deformation which can lead to shape deformation which is either too limited or too flexible. Such limited specificity of shape deformation can decrease the robustness of the image search which maps the model to the image.

Previous model based methods have also used heuristics about the appearance of the lip contours as information for the image search. The assumptions about strong gradients at the lip contours are often incorrect and can lead to ambiguities with other gradients in the image. This can lead to poor search performance.

Important visual speech features are thought to be contained in the contour of the lips and in finer image details such as the teeth, tongue, wrinkles, and protrusion. Several contour models are not able to represent the principal deformations of the lip contours well and often only consider the outer lip contour. Image based feature extraction methods on the other hand often mainly represent intensity changes due to mouth opening and are not able to resolve fine mouth details.

The approach described here is to learn typical shape deformation from a training set. Typical intensity variation around the lip contours is learned from training data too. These patterns of variability are used in image search to map the model to the image and to extract speech dependent shape and intensity information from this mapping.

5.2 Appearance based Models

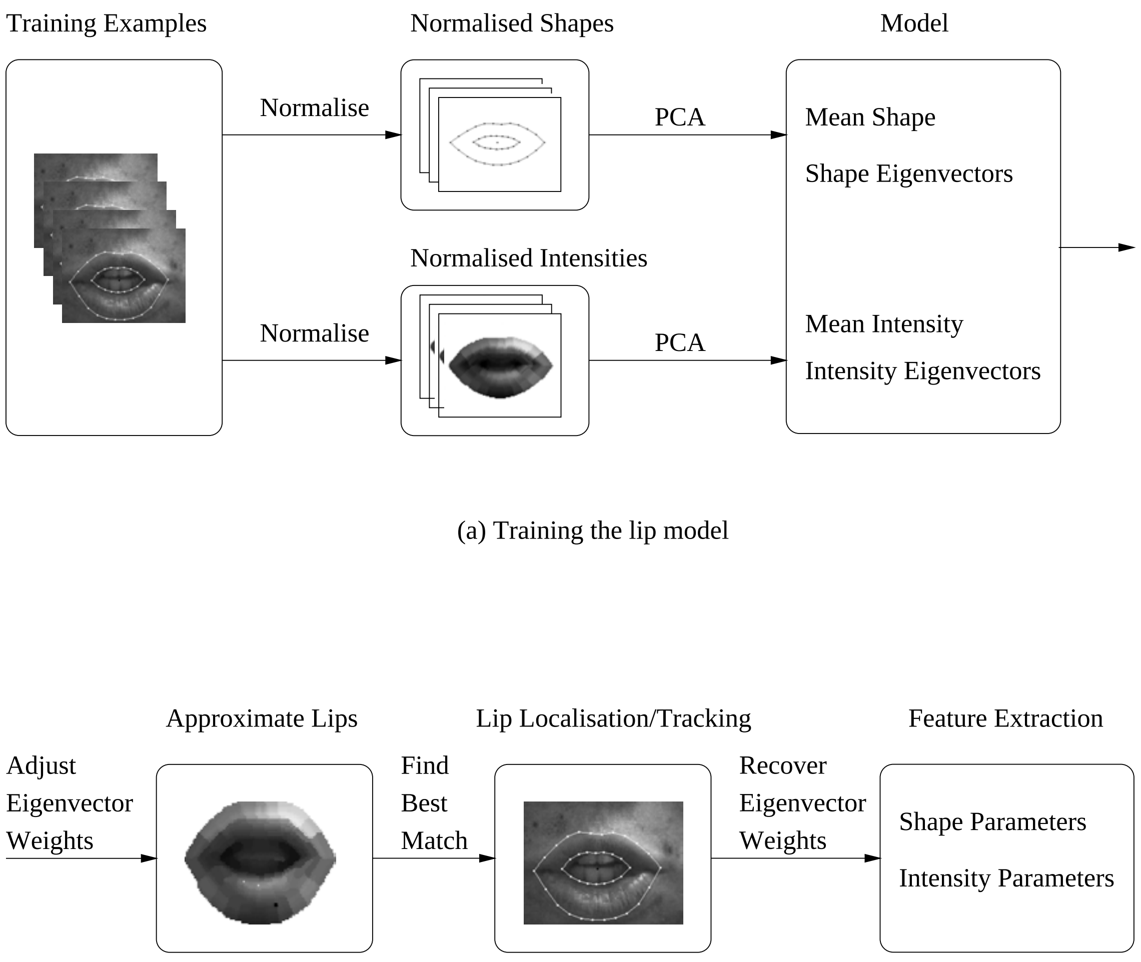
The remainder of this chapter describes the development of an appearance based model (ABM) for lip localisation, lip tracking, and feature extraction. One of the issues to address in a model based approach is to choose an appropriate description of the visible articulators. We are modelling a physical process, so we could describe this process in terms of physical movements and positions of the articulators that determine the vocal tract. Specifically, for visual analysis, we could attempt to estimate muscle action from the image such as in [124, 60]. However, the musculature of the face is complex, 3D information is not present, muscle motion is not directly observable, and there are at least thirteen groups of muscles involved in the lip movements alone [83]. Furthermore, using optical flow computation to estimate such action might not be appropriate due to violation of its underlying assumptions.

An appearance-based model was therefore chosen to represent the visual speech articulators. In order to construct such a model, we need to determine which features the model will incorporate and how they are represented. For the application of speech recognition, the model should represent those features important for speech recognition and disregard those which account for speaker identity, pose, and illumination. For speaker recognition applications, on the other hand, the method should describe features which are specific to a persons identity and disregard all others. Here, it will be assumed that visual speech and identity information can not be separated easily. The same features which are thought to be important for speech recognition will therefore also be used for speaker recognition. This assumption is usually also made in acoustic speaker recognition.

It is generally agreed that most visual information is contained in the lip contours, especially the inner lip contour, but it has been shown in Chap. 3 that the visibility of teeth and tongue provides important speech cues too. We therefore would like to have a model which describes both the shape of the inner and outer lips and the intensity around the mouth area. Models based on point distribution models (PDM), also called active shape models (ASM) when used in image search [46, 44, 42, 43] are used for modelling shape deformation. PDMs are flexible models which represent an object by a set of labelled points. The points describe the boundary or other significant locations of an object. The average shape and the principal modes of variation are captured from a labelled training set.

5.3 Overview of their Training and Usage

Figure 5.1 shows the block diagram of the different stages for training and using the models. A training set of images must be labelled by hand by placing points around the contour of the object to model. The shapes are decomposed into a weighted sum of basis shapes using a Karhunen-Loéve expansion. The mean and the covariance matrix of the coordinates of the shape points are computed and the eigenvectors which represent the basis shapes are obtained by Principal Component Analysis (PCA). Similarly, the grey-levels in an area around the model points are decomposed into a weighted sum of basis intensity images which are obtained by PCA. The model structure is described by the



(a) Training the lip model

(b) Using the lip model

Figure 5.1: Block diagram of the training and use of the models. Principal component analysis (PCA) is performed on a labelled training set of object examples which reveal the principal modes of shape and intensity variation. The means as well as the eigenvectors are used during localisation and tracking to fit the model to the image by using a cost-minimisation function. Shape and intensity features can be recovered from these tracking results and can serve as features for speech or person recognition applications.

mean shape, mean intensity, basis shapes, and basis intensities.

Any shape and intensity can be synthesised using the mean shape, mean intensity, and a linear combination of their basis vectors. The weights of the shape and intensity eigenvectors are denoted as the *model parameters*. The model is used to locate and track the lips in image sequences. This is realised by a cost function which matches the fit between the model and the image, and a minimisation function which finds the best match. Once the best match is found, the model parameters describing the configuration of the model for this match are recovered and can be used as features for further processing.

5.4 Shape Modelling

PDMs are used for shape modelling in order to avoid the use of heuristic assumptions about legal shape deformation. Instead, knowledge about legal shape deformation is obtained by examining a representative training set. This leads to a description of local and global deformations with a small set of parameters and constrains the shape model to only deform to shapes similar to the ones seen in the training set.

5.4.1 Labelling Training Shapes

Labelling training shapes consists of placing model points around the boundary of the training images to allow the statistical analysis of these coordinates. The training examples therefore need to be labelled in a consistent manner in order to be able to compare equivalent points from different shapes. Here, the two outer corner points of the lips are used as reference points. Their distance is defined as scale, their orientation to the horizontal as the angle, and their centre as the origin. The other points are placed at equal horizontal distance along the lip contours. Two different models of the lips were built: *Model SC* describes the outer contour of the upper and lower lips and *Model DC* describes the outer and inner contour of the upper and lower lips. Figure 5.2 displays *Model DC* with translation t_x and t_y, scale s and angle θ.

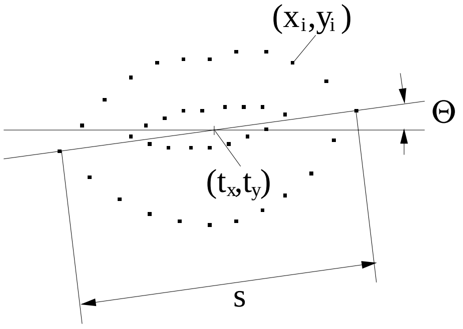


Figure 5.2: *Model DC* representing the outer and inner lip contour with 38 control points. To generate a certain lip shape, the translation (t_x, t_y), rotation (Θ), and scale (s) are required in addition to the coordinates of the model points.

It should be noted that the labelling of the training set can be considered an ill-posed problem. The boundary of the outer lip contour, especially of the lower lip, can not always be determined uniquely. In some cases there is no visible boundary at the lower contour while in other cases two boundaries appear, or the boundary is occluded by facial hair.

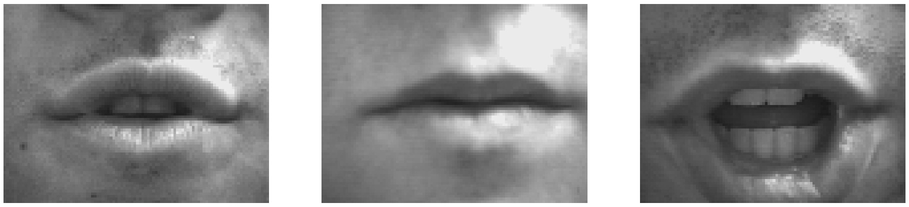


Figure 5.3: Examples form the Tulips1 database for which the labelling of lip contours was difficult. The outer lower lip contour is often not very distinct from the skin and the border of the inner contour can be confused with the texture inside the mouth.

Figure 5.3 shows three examples for which the determination of the lip contour is ambiguous. In the first two images, the lower lip is not distinct while in the third image, the determination of the inner lip contour is difficult due to the absence of a clear boundary between the lips and the mouth. The inner lip contour is normally determined as the visible border between the skin texture and the mouth opening. Where this contour lies on the skin depends on the mouth opening and in particular on the protrusion, which generally reveals more of the skin inside the mouth. This border is therefore neither the actual inner lip contour nor does it always correspond to the same part of the lips. To minimise these ambiguities it was attempted to label all examples in the same way, i.e. in cases of ambiguity at the outer lip contour, the contour which corresponds to the texture difference rather than the skin wrinkle was used as border. Similarly, the inner lip contour was determined as the visible border between the texture and the mouth opening.

The i^{th} shape in the training set ($i = 1 \dots N_e$) is described by a vector $\mathbf{v}_i$ with

$$\mathbf{v}_i = (x_{i0}, y_{i0}, x_{i1}, y_{i1}, \dots, x_{iN_s-1}, y_{iN_s-1})^T \tag{5.1}$$

where (x_{ij}, y_{ij}) are the coordinates of the j^{th} point ($j = 0 \dots N_s - 1$) of the i^{th} shape. The training shapes are then normalised by scaling to unit width, zero translation, and zero rotation using

$$\mathbf{x}_i = M\left(\frac{1}{s}, -\theta\right)[\mathbf{v}_i - \mathbf{t}_c] \tag{5.2}$$

with translation

$$\mathbf{t}_c = (t_x, t_y, t_x, t_y, \dots, t_x, t_y)^T \tag{5.3}$$

and $M(s, \theta)[\mathbf{x}]$ performing a rotation by θ and a scaling by s of $\mathbf{x}$ using

$$M(s,\theta)\begin{bmatrix} x_{ij} \\ y_{ij} \end{bmatrix} = \begin{pmatrix} (s\cos\theta)x_{ij} - (s\sin\theta)y_{ij} \\ (s\sin\theta)x_{ij} - (s\cos\theta)y_{ij} \end{pmatrix}. \tag{5.4}$$

The normalisation process ensures that model points of different examples can be related without distortion from scale, translation or rotation.

Automatic labelling of training examples is still a problem in the PDM framework. A solution for moving objects has been presented by Baumberg and Hogg [10], where the object boundary was obtained by subtracting the background from the moving object. Hill and Taylor [87] have proposed an algorithm for the automatic generation of landmark points, however, the algorithm still requires that the contour of the object is segmented, prior to the landmark generation. Both methods can therefore not easily be applied to the labelling of lip contours.

5.4.2 Modelling Shape Variability

A shape can be approximated by a weighted sum of basis shapes which are obtained by a K-L expansion. Given a set of N_e normalised labelled shapes, we can calculate the mean shape $\overline{\mathbf{x}}$ and the covariance matrix $\mathbf{S}_s$. The eigenvectors and eigenvalues of the covariance matrix are obtained by principal component analysis (PCA). The eigenvectors with the largest eigenvalues describe the most significant modes of variation, i.e. the variance described by an eigenvector is equal to its corresponding eigenvalue.

A normalised shape can now be approximated by

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{P_s b_s} \tag{5.5}$$

where $\mathbf{P_s} = (\mathbf{p}_{s1}, \mathbf{p}_{s2}, \ldots, \mathbf{p}_{sT_s})$ is the matrix of the first T_s ($T_s < N_s$) column eigenvectors corresponding to the largest eigenvalues and $\mathbf{b_s} = (b_{s1}, b_{s2}, \ldots, b_{sT_s})$ a vector containing the weights for the eigenvectors. The number of eigenvectors used to describe the main modes of variation is normally much smaller than the number of variables of a shape instance.

Figure 5.4 and Figure 5.5 show the mean shape of the models and the first four modes of variation by ±2 standard deviation (*s.d.*), computed over training examples from the Tulips1 database. In order to project a lip model into an image, the scale, rotation, and translation parameters are needed in addition to the shape parameters.

For *Model SC*, the first mode of variation mainly changes the position of the lower lip contour. The second mode primarily describes the position of the upper lip contour and the third mode accounts for asymmetry. Subsequent modes describe finer contour details. For *Model DC*, the first mode mainly changes the lower lip, the second mode changes the upper lip and the third mode describes the mouth opening. Subsequent modes account for asymmetry and finer contour details.

An earlier version of these models which only considered vertical deformation of the model points has been described by Luettin et al. [109]. The method was sufficient for the single contour model but could not fully describe the large variability of the inner contour of the double contour model.

The approach assumes that the principal modes are linearly independent, although there might be non-linear dependencies present. For objects with non-linear behaviour, linear models reduce the specificity of the model and can generate implausible shapes, which lead to less robust image search. They also require more modes of variation than

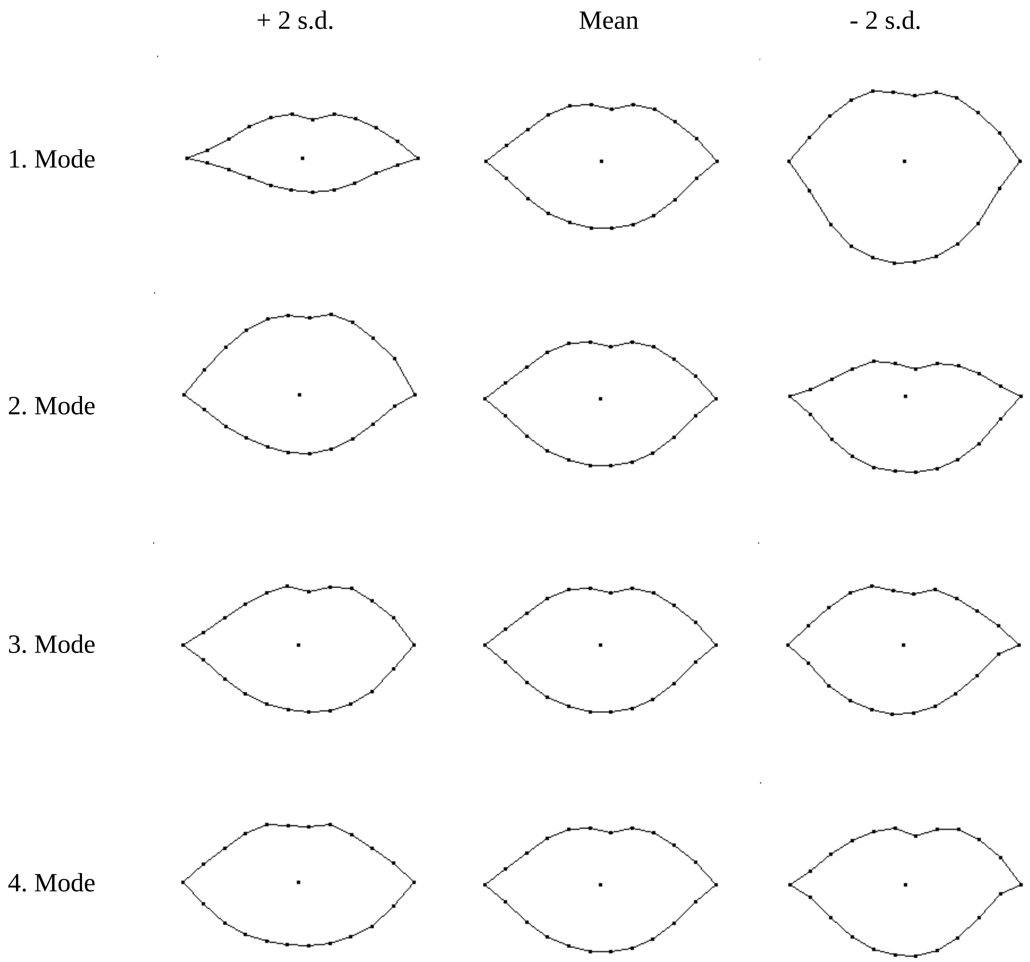


Figure 5.4: The middle column displays the mean shape for *Model SC* and the other columns show the four most significant modes of shape variation as the sum of the mean shape and different basis shapes.

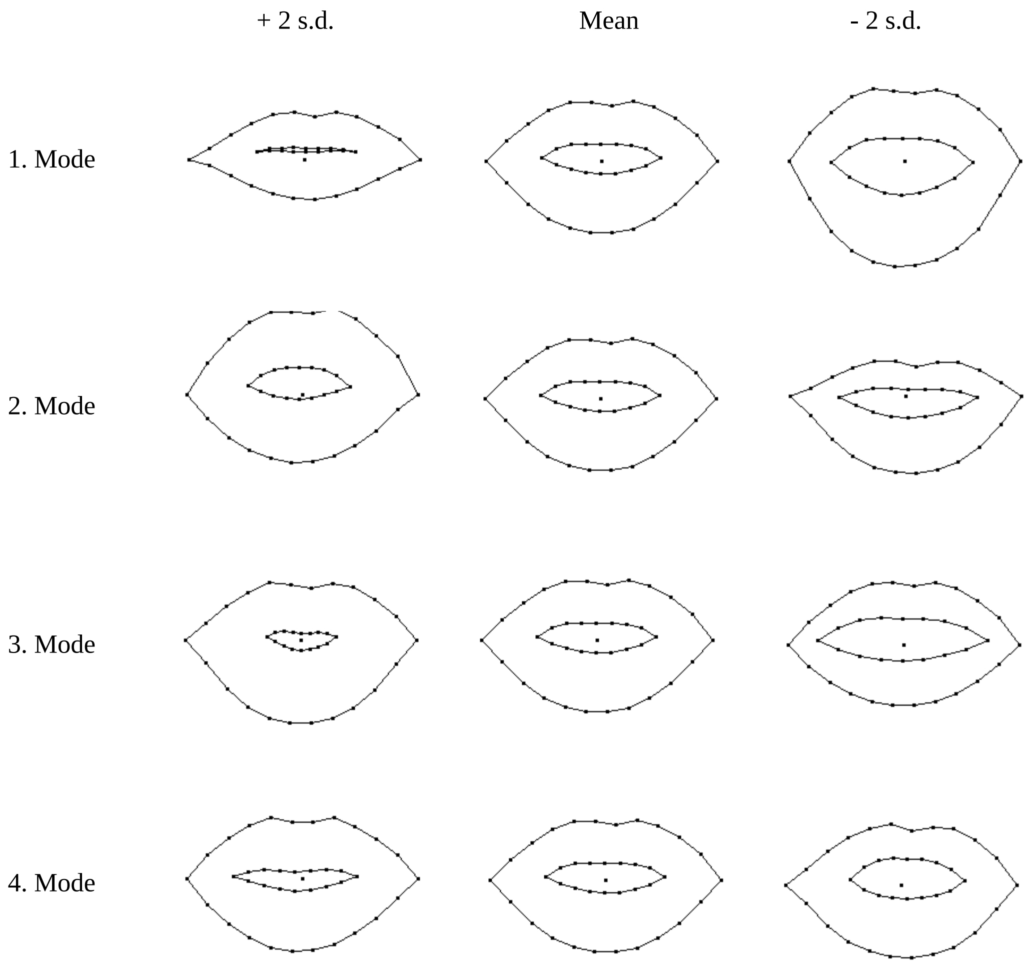


Figure 5.5: The middle column displays the mean shape for *Model DC* and the other columns show the four most significant modes of shape variation as the sum of the mean shape and different basis shapes.

the true number of degrees of freedom of the object. The specificity of a model can be improved by a nonlinear process, e.g. by nonlinear PCA as described by Souza et al. [165].

5.5 Intensity Modelling

Intensity modelling is described here as a mean for a robust image representation to be used for image search in locating and tracking lips. In Chapter 6 and Chapter 7 the intensity model will be used to provide important features for speech and speaker recognition, which are obtained from the results of the image search. In order to use PDMs for image search, we would like to have a cost function, which measures the fit between the model and the image. This fit should provide information about how likely it is that the model, positioned at a certain location and with a certain shape, matches the actual lips in the image. A minimisation function will then be used to search the image for possible candidates. The candidate with the best fit will be assumed to represent the actual position and shape of the lips in the image. We therefore need to find a way of representing dominant image features of the lip contours which we try to match with a certain representation of our model.

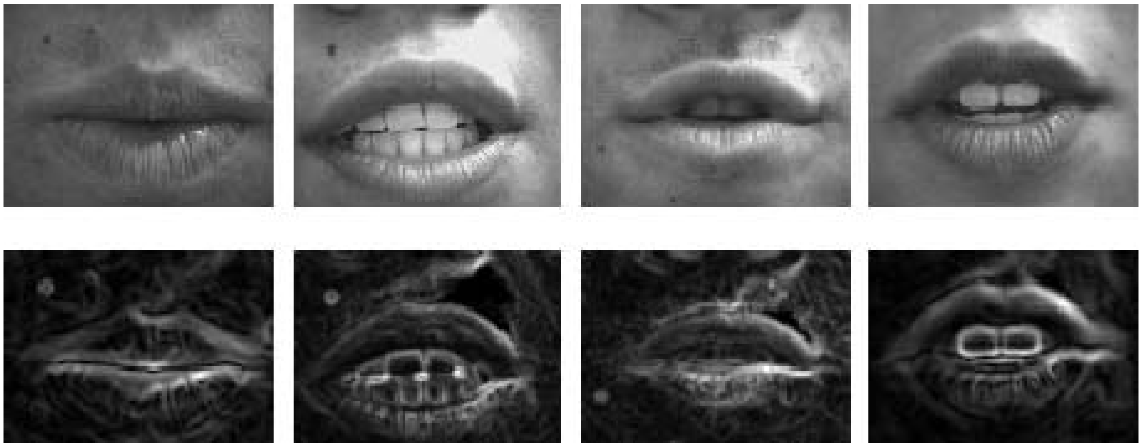


Figure 5.6: Example images of the Tulips1 database (first row) and their gradient magnitude images (second row).

The most common approach for representing contours is to use edges or gradients. It is however well known that the representation of object boundaries by edges are ill-posed and therefore do not have unique solutions [147]. Particularly for lips, the appearance of contours is highly variable – even for the same person. The gradient values at the outer lip contour are often strong at the upper lip and week at the lower lip. The gradients at the inner contour are highly dependent on mouth opening and the visibility of teeth and tongue. Gradients are also dependent on the speaker (make up, facial hair, ethnic origin) and illumination. Furthermore, edges of the lip contours are often confused with other gradients, which can originate from specularity, visibility of teeth and tongue, shadows, or facial hair.

Figure 5.6 shows some example images of the Tulips1 database used for the experiments and their gradient magnitude images after Gaussian smoothing. The examples show clearly the difficulties gradient-based search methods are faced with and that gradient information is not an appropriate way to represent dominant features of lip contours.

In analogy to the statistical description of the lip deformation we want to avoid the use of heuristics for image search and rather learn the grey-level appearance at the contour from a training set. Assuming that grey-level changes are not only important at each contour point but also in regions around each point, the statistics of the actual grey-level appearance are captured around each model point and their main modes of variation are estimated from a training set.

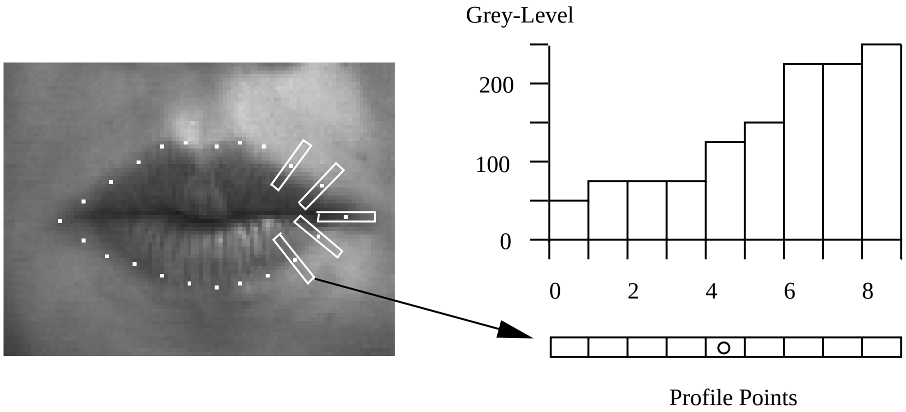


Figure 5.7: Grey-level profile extraction. The grey-level vectors are sampled perpendicular to the lip contour and centred at the model points.

Following [44], one dimensional profiles $\mathbf{g}_{ij}$ of length N_p are sampled perpendicular to the contour and centred at point j for each training image i as shown in Fig. 5.7. But instead of calculating individual mean profiles and covariance matrices for each model point, the profiles of all model points are concatenated to construct a global profile vector $\mathbf{h}_i$ for each training image i as described in [84]:

$$\mathbf{h}_i = (\mathbf{g}_{i0}, \mathbf{g}_{i1}, \ldots, \mathbf{g}_{iN_s-1})^T \tag{5.6}$$

The global mean profile $\overline{\mathbf{h}}$ of dimension $N_i = N_s N_p$ and the global covariance matrix $\mathbf{S}_i$ is then calculated, and the eigenvectors and eigenvalues of the covariance matrix $\mathbf{S}_i$ are obtained by PCA. The eigenvectors with the largest corresponding eigenvalues describe the main modes of grey-level variation seen in the training set. Any profile can be approximated using

$$\mathbf{h} = \overline{\mathbf{h}} + \mathbf{P}_i \mathbf{b}_i \tag{5.7}$$

where $\mathbf{P}_i = (\mathbf{p}_{i1}, \mathbf{p}_{i2}, \ldots, \mathbf{p}_{iT_i})$ is the $N_i \times T_i$ matrix of the first T_i $(T_i < N_i)$ column eigenvectors corresponding to the largest eigenvalues and $\mathbf{b}_i$ a vector containing the weights for each eigenvector. A lip model Ω can now be represented by the following notation:

$$\Omega = \left(\overline{\mathbf{h}}, \overline{\mathbf{x}}, \mathbf{b_s}, \mathbf{b}_i, t_x, t_y, s, \theta\right) \tag{5.8}$$

This notation contains all parameters required to approximate the shape, intensity, and pose of a certain lip image.

The intensity weight vector $\mathbf{b}_i$ represents the intensity around the lip contour and accounts for features like the intensity of the oral cavity, visibility of teeth and tongue, and finer details like protrusion. These parameters therefore represent important features with respect to speech recognition and person identification.

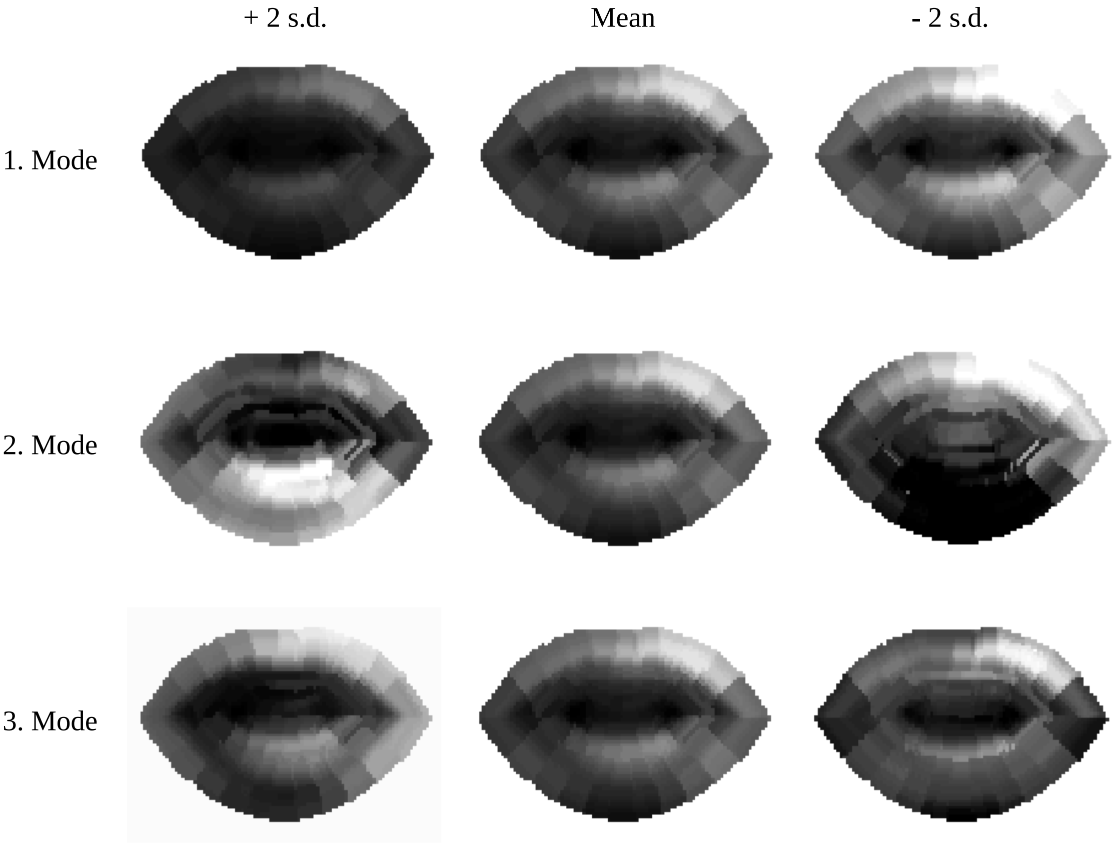


Figure 5.8: The principal modes of intensity variation of *Model DC*. The middle column displays the mean intensity and the other columns represent the three most significant modes of intensity variation for ± 2 s.d.

This feature extraction approach is similar to the local grey-level models described by Lanitis et al. [104, 106, 105], who performed an individual PCA on each model point. Here, a global grey-level model is used, assuming that variances across model points are correlated. It is also related to the *eigen-lips* reported by Bregler et al. [23]. Bregler et al. placed a window around the mouth area on which PCA was performed. Since the window does not deform with the lips, the eigenvectors of the PCA mainly account for intensity variation due to different lip shape and mouth opening. In comparison, the PCA space described here deforms with the lip contours and therefore describes intensity variation which is independent of the lip shape. The obtained intensity information is therefore complementary to the contour information provided by the shape model.

In order to visualise the intensity models, the grey-levels between the profile vectors are interpolated and smoothed with a Median filter. The mean intensity and the first three modes of intensity variation using this method are shown in Figure 5.8. All grey-level models are displayed using the mean shape. Although this might generate unrealistic appearances, since the grey-levels are correlated with the mouth opening, it was considered to be acceptable for visualisation purposes.

The grey-level models show the area covered by the profile vectors: the mouth opening, the lip area, and the skin around the lips. The first mode mainly accounts for global intensity variation, the second mode describes the contrast between the upper and the lower lip, and the third mode mainly represents the contrast between lips and skin. Subsequent modes describe finer variations, such as lighting direction, specularity, and visibility of teeth and tongue.

This method for intensity modelling assumes that the variances of intensity profiles at different model points are correlated with each other as they are expected to be due to illumination effects and different skin and lip intensity. The profile model captures the global variation across speakers and the variation within speakers. Particularly during speech production, the intensity variation inside the mouth is subject to the largest intensity variation. In analogy to shape modelling, the intensity modes are assumed to be linearly independent.

5.6 Image Search

The task of image search is to localise the lips in the image. This is normally performed by matching the model to the image at different locations and with different shapes and by choosing the hypothesis with the highest likelihood. In this work two different cost functions are investigated to describe the fit between the model and the image: an image-gradient based function and a function based on the intensity model. Both methods use the Downhill Simplex Method (DSM) for finding the position and shape with the minimum cost.

5.6.1 Cost Function based on Gradients

Several researchers [85, 24, 23, 153, 32] have described lip tracking experiments using gradient based cost functions. An approach based on gradient information was therefore implemented and tested for both shape models, *Model SC* and *Model DC*. The images were first Gaussian filtered before the gradient magnitude was calculated. Examples are shown in Fig. 5.6. The cost E_g was defined as the negative sum of the gradient magnitudes over all model points given by

$$E_g = -\sum_{i=1}^{N_s} \sqrt{dx_i^2 + dy_i^2} \tag{5.9}$$

where dx_i and dy_i represent the gradients at point i in horizontal and vertical direction, respectively. The search function tries to minimise the cost function f with respect to the following parameter set:

$$f(t_x, t_y, s, \theta, \mathbf{b}_s) \tag{5.10}$$

First experiments showed that the algorithm performed very poorly and that the results were highly dependent on the initial estimate being close to the actual lip location. This was thought to be due to the sparse occurrence of gradients over the image. The gradient magnitude images were therefore filtered with an exponential function to blur the gradients over a larger area and to make the search more robust. Still, the performance of the algorithm was very low.

Hennecke [85] has used a slightly different method by calculating horizontal edges and has used a cost function which differentiates between positive and negative edges. This approach makes the following assumptions about the intensity I: $I_{skin} > I_{lips} > I_{mouthopening}$. This method was also implemented but was not found to provide satisfactory results. The assumptions made were for example invalid in the case of shadows below the lower lip, visibility of teeth, subjects with dark skin, and subjects with beards.

5.6.2 Cost Function based on Intensity Models

This cost function uses the intensity model to describe the fit between the image and the model. It can be considered an extension to correlation based matching with the difference that grey-level variability is taken into account. It can also be viewed as an extension to the *eigen-face* approach [175] for detection, which takes grey-level variability into account but which does not model shape variability explicitly. The search method described here considers both shape variability and grey-level variability in the calculation of the cost E_p^2, which is calculated by the following two steps:

$$\left.\begin{array}{l} \text{1. Given } t_x, t_y, s, \theta, \mathbf{b}_s, \text{ compute } \mathbf{b_i} \text{ using (5.12) to align model to image.} \\ \text{2. Given } \mathbf{b_i}, \text{ calculate cost } E_i^2 \text{ using (5.13)} \end{array}\right\} \tag{5.11}$$

To account for grey level variation captured in the training set, the model profile is first aligned to the image profile $\mathbf{h}$ as closely as possible, using the mean intensity and a weighted sum of the first few basis intensities. As the eigenvectors in $\mathbf{P}_i$ are orthogonal, the parameter vector $\mathbf{b}_i$ describing the weights for the basis images can be found using

$$\mathbf{b}_i = \mathbf{P_i}^T(\mathbf{h} - \overline{\mathbf{h}}). \tag{5.12}$$

To measure how well a model fits the image, the sum of square errors E_i^2 between the image profile and the aligned model profile is calculated using

$$E_i^2 = (\mathbf{h} - \overline{\mathbf{h}})^T(\mathbf{h} - \overline{\mathbf{h}}) - \mathbf{b}_i^T\mathbf{b}_i. \tag{5.13}$$

Note, that the cost function is optimised with respect to the same parameters as for gradient based search using (5.10). The intensity weight parameter vector $\mathbf{b}_i$ can be obtained from the cost function and can be used for further processing.

Cootes et al. [42] have used the following measure to estimate how well the model fits the profile:

$$E_c^2 = \sum_{j=1}^{T_i} \frac{b_{ij}^2}{\lambda_j} + \frac{E_i^2}{0.5\lambda_{T_i}} \tag{5.14}$$

where λ_i is the eigenvalue corresponding to the ith eigenvector with $\lambda_i \leq \lambda_{i+1}$. This measure considers both the distances between the considered modes from the mean $E_{1,T}$ (first term) and the distance not explained by the considered modes $E_{T+1,N}$ (second term). The notation $E_{n,m}$ refers to the residual error for the modes n to m. The relative weighting of both terms assumes that the sum of the squares of residuals are Gaussian distributed and have a variance of $0.5\lambda_{T_i}$. Moghaddam and Pentland [132] have shown that the optimal value for $0.5\lambda_{T_i}$ is the arithmetic mean of the eigenvalues $(\lambda_{T_i+1}, \ldots, \lambda_{N_i})$.

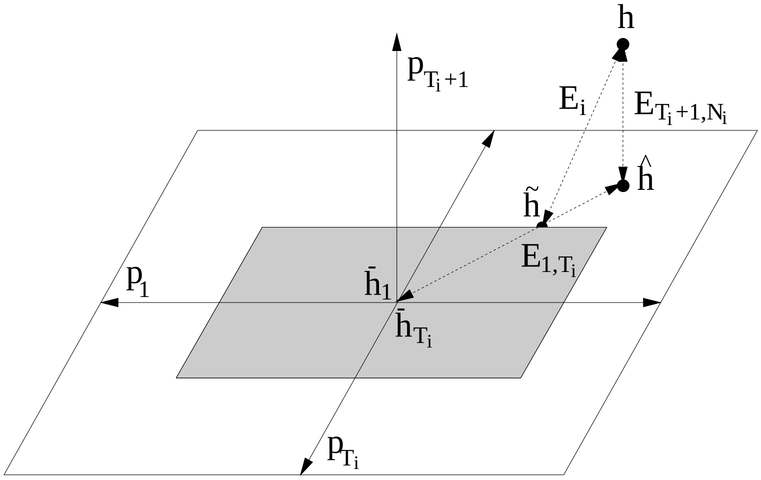


Figure 5.9: A simple model for $T_i = 2$ and $N_i = 3$ with mean $\bar{\mathbf{h}} = (\bar{h}_1, \bar{h}_T)^T$ and two eigenvectors $\mathbf{p}_1, \mathbf{p}_{T_i}$. The *best fit* $\hat{\mathbf{h}}$ is the projection of $\mathbf{h}$ onto the surface spanned by $\mathbf{p}_1$ and $\mathbf{p}_{T_i}$ which results in the residual error E_{T_i+1,N_i}. Constraining all parameters b_i to stay within a certain limit (shaded area) results in the *limited fit* $\tilde{\mathbf{h}}$ and the residual error E_i.

Since this measure penalises values far from the mean, it is unlikely to be appropriate for the application of lip localisation and tracking, where the intensities vary considerably for different subjects and mouth opening. In this case it is more desirable to assign equal prior probabilities to instances within a certain limit and to constrain the model parameters to stay within these limits. This strategy was implemented here, by using the sum of residual square errors E_i^2 as distance measure but forcing all intensity modes to stay within $\pm 3 s.d.$, which accounts for 99% of intensity variation.

Figure 5.9 illustrates the different error measures for a simple model with two modes of variation. Along the directions $\mathbf{p}_i$ for which $i \leq T_i$, the weights b_i are not considered in the error function E_i, but they are constrained to lie within the limits of ± 3 *s.d.* For the point $\mathbf{h}$ the *best fit* $\hat{\mathbf{h}}$ is the projection of $\mathbf{h}$ onto the surface spanned by $\mathbf{p}_1$ and $\mathbf{p}_{T_i}$, resulting in the residual error E_{T_i+1,N_i}. The *limited fit* $\tilde{h}$ is obtained by limiting the weight vectors which results in the residual error E_i.

Several lip tracking approaches use an additional cost component for shape deformation [85, 24, 180], which penalises shapes far from the mean shape and gives preference to shapes close to the mean shape. This is clearly not desirable and can result in search errors for shapes far from the mean shape. Here, the use of a component for shape deformation is avoided, but each shape mode is constrained to stay within ±3 s.d. during image search, in analogy to the limits for the intensity modes, which accounts for 99% of shape variation. Equal prior probabilities are assigned to all model shapes within these limits.

5.6.3 Minimisation Function

The minimisation function is to used to find the optimum choice of parameters, i.e. those which give the minimum cost. Although we are interested in finding the global minima, the cost function usually has several local minima. The degree of difficulty for finding the global minimum depends largely on the constraints on the images, under which the algorithm should work. Constraints can be used for illumination, pose, within/between subject variability, and the AOI containing the lips.

Here, for lip localisation it is assumed that the coarse position, scale and angle of the mouth is known. Most lip tracking applications require the face to be located first followed by a rough localisation of facial features like eyes and mouth. No other constraints were applied and the algorithm was tested for the illumination, pose, and subject variabilities covered by the used databases.

Cootes et al. [42] use a minimisation method which evaluates the best fit for each model point separately and then updates the global shape parameters. This two-step process is iterated until convergence. For our application this method did not to perform satisfactory. One reason for this was the often very weak image evidence at the lower outer lip contour which often suggested false movements. The second reason was due to the large variability of profile vectors which are clearly correlated. This correlation across profile vectors is not accounted for by isolated profile evaluation. Grey-level profile correlation is mainly due to differences across subjects, different degrees of mouth opening, and varying visibility of teeth.

The Downhill Simplex Method (DSM) was used instead [139, 149] to find a minimum of the cost function. The DSM has the important property of not requiring derivatives of function evaluations and that it can get out of local minima. A simplex is a geometrical figure consisting, in n dimensions, of $n+1$ vertices. For two dimensions it is a triangle and for 3 dimensions it is a tetrahedron. One of the vertices corresponds to the initial guess and the other points define vector directions that span the n-dimensional vector space. The distance of each of these points to the initial guess represent the initial perturbation in that direction. The method iteratively evaluates the cost in each of these perturbation directions and performs one of the following steps:

Reflection: reflect the vertex with the highest cost in the plane defined by the remaining vertices.

Expansion: expand the reflected vertex.

Contraction: contract the reflected vertex.

Multiple contraction: contract all vertices around the lowest point.

The DSM will always converge to a minimum of the function, although to find the global minimum is not guaranteed. The model is initialised with the mean shape and placed in a random location in the AOI which is the starting point for the search algorithm. The algorithm then minimises the cost function (5.10). The search process begins by evaluating initial perturbations of each parameter, which were chosen to be $2s.d.$ An evaluation is performed in two steps, at first the optimal profile weights are calculated for a given instance using (5.12), then the cost is computed using (5.13). It gradually moves and deforms the model to shapes which give a lower cost and changes the perturbations dynamically until a certain number of iterations is reached or until the difference in cost falls below a certain threshold. The maximal number of iterations was set to 500 and the threshold value was set to 10^{-6}. The algorithm has proven to be robust to various initial parameters and copes well with local minima. Figure 5.10 shows an example of image search and the results after a few iterations.

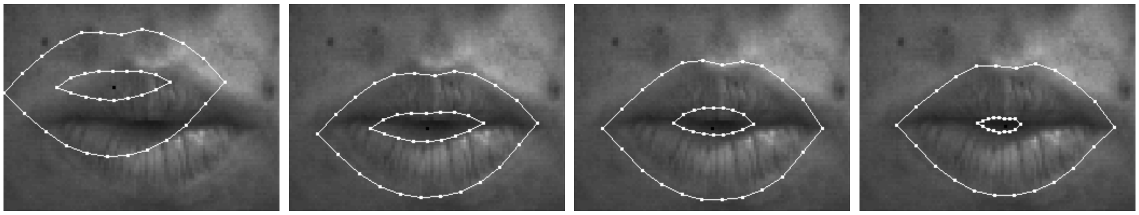


Figure 5.10: Off-centre initialisation of *Model DC* and image search results after 5, 10, and 20 iterations.

Lip-tracking is initiated by locating the lips in the first image as described above. For consecutive frames, the previous frame is used as the initial estimate of the lip position and the search is performed as for lip localisation. Although constraints could be introduced to limit the search to stay within certain limits during tracking, for simplicity, the same constraints were used as for locating the lips.

5.7 Experiments on the Tulips1 Database

The computer vision algorithms were implemented under the computer vision research environment TINA, which was developed at Sheffield University [177]. The shape and profile models were built using 190 training examples for *Model SC* and 250 examples for *Model DC* drawn from all 12 speakers and covering a representative set of mouth shapes.

All examples were taken from Set 1 and contained hand labelled points. *Model SC* was represented by 22 points, *Model DC* by 38 points and the dimension of a grey-level profiles was 19. This resulted in a global grey-level vector of dimension 722. *Model SC* consisted of 8 shape modes, which covers 87.1% of shape variation and *Model DC* consisted of 10 shape modes, which covers 84% of shape variation, estimated from the corresponding covariance matrices. For image search, 12 intensity modes were used for *Model 1* and 20 modes for *Model DC*.

5.7.1 Performance Evaluation

There exist two different strategies for the evaluation of feature extraction algorithms. One strategy is to omit direct evaluation of the algorithm at the feature extraction level, since only the final result of the system is of importance and since the evaluation of feature extraction methods is often subjective. Here, the final result of the system could be visual speech recognition performance or person identification performance. The other strategy is to evaluate the performance of the feature extraction algorithm separately, to be able to relate miss-classification of the final system either to the feature extraction process or to the classification method. Furthermore, the evaluation is helpful in comparing different feature extraction algorithms independent of further processes.

For the Tulips1 database, it was attempted to evaluate the performance of the lip tracking algorithm directly. Ideally, this would require all images to be labelled with the correct outline of the lips, which would need to be done by hand. Instead we choose to judge the performance by visual inspection into three categories. A search result was classified as *Good* if the lip contour was found within about one quarter of the lip thickness deviation. It was classified as *Adequate* if the outline of the contour was found between one quarter and half the lip thickness deviation and it was classified as a *Miss* otherwise. For tracking results, all of the images had to be classified as *Good* in order to classify the whole sequence as *Good*. For *Adequate* tracking classification, all of the images had to be classified as either *Adequate* or *Good*.

5.7.2 Lip Localisation

All tests for lip localisation were performed on the first image frames of *Set 2*. For localisation tests, the position of the model was initialised at a position off the centre. This makes the search process more realistic, assuming that only a rough estimate of the lip position is known. The shape parameters were initialised with the mean shape parameters. Figure 5.10 shows the initial placement of the model in the image and results of the image search after a few iterations.

Examples of lip localisation results for all subjects are shown in Figure 5.12. The two localisation results classified as *Miss* are depicted in Figure 5.11. Localisation results for both models can be found in Table 5.1. For both models, there was only 1 image out of 48 which was classified as *Miss*, all other images were classified as *Good*.

To compare the intensity based search function with gradient based search, a second experiment was conducted using the gradient image as a criteria for the cost function.

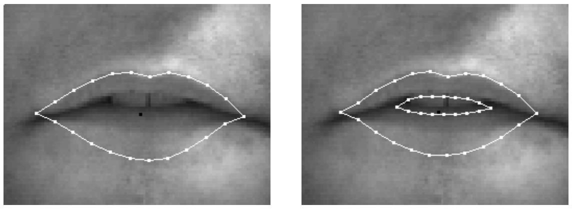


Figure 5.11: Examples of lip localisation results using *Model DC* and intensity model based search, which were classified as *Miss*.

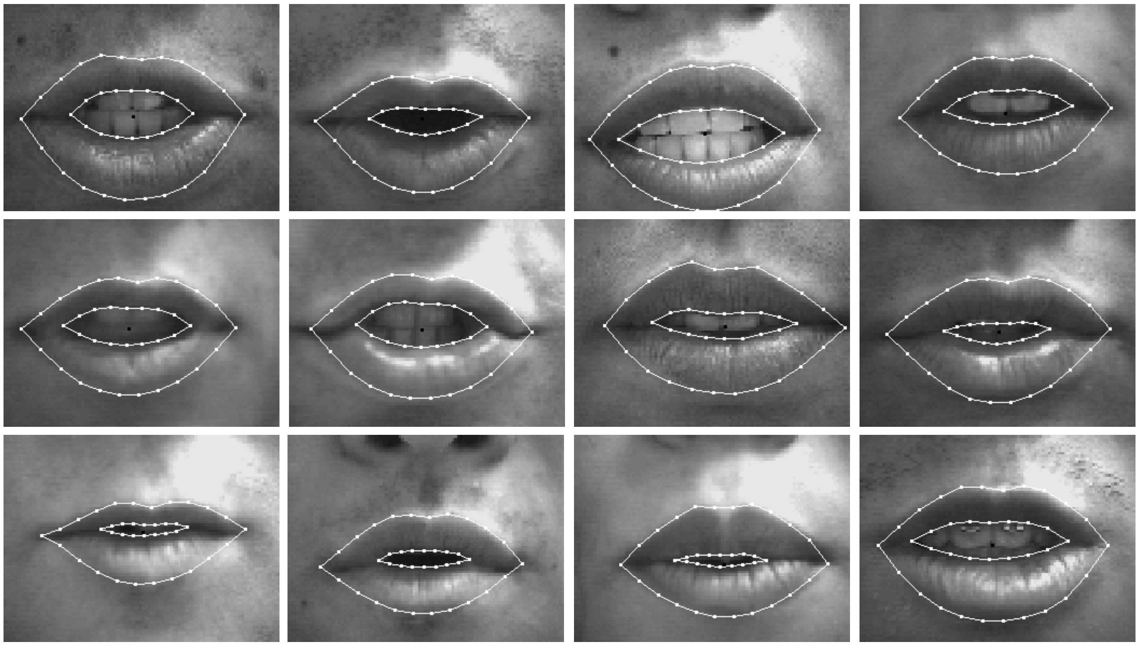


Figure 5.12: Examples of lip localisation results for all subjects, using *Model DC* and intensity model based search, which were classified as *Good*. Image search copes well with different subjects and different illumination and finds the inner lip contour despite the large intensity differences due to the visibility of the teeth and mouth opening.

The images were first Gaussian filtered before calculating the gradient magnitude. The resulting images were smoothed with an exponential function to blur the gradients over a large image area and to make the search more robust. The cost was defined as the negative sum of the gradient image at all model points according to (5.9). This method is similar to the ones described in [85, 24]. Because the results for lip localisation were so poor, another test was performed where the model was initialised at the centre of the image.

Table 5.1: Results for lip localisation using intensity based image search or gradient based image search. The model was either initialised at the centre or off the centre.

Cost	Initialisation	Model	Good (%)	Adequate (%)	Miss (%)
Intensity Model	off-centre	SC	97.9	0	2.1
		DC	97.9	0	2.1
Gradient Magnitude	off-centre	SC	6.3	6.3	87.5
		DC	4.2	12.5	83.3
	centre	SC	18.8	22.9	58.3
		DC	20.8	8.3	68.8

Table 5.1 shows the results for lip localisation using gradient information. Miss-localisation was mainly caused by the small gradient of the outer lower lips, reflections on the lips, and gradients originating from the teeth. Initialising the model at the centre of the image improved the results only slightly. These results clearly demonstrate the advantage of grey-level models over gradients in image search.

5.7.3 Lip Tracking

Lip tracking tests were performed on all image sequences of *Set 2* and were initialised by placing the model at the centre of the image. Tracking examples of *Good* performance are shown in Figure 5.13 and for *Adequate* performance in Figure 5.14. Table 5.2 summarises tracking results for both models. For *Model SC*, results for lip-tracking are similar to the ones for localisation. For *Model DC*, the performance for lip-tracking is lower than for localisation. This was mainly due to tracking errors where the model was aligning to the teeth instead of the inner lip contour.

With this configuration, the algorithm usually converged after 200 to 400 iterations. On a Sun Sparc 4 workstation, one function evaluation took about 20 msec, resulting in a total of about 7.6 sec processing time per image. Computation time was of no concern for the experiments, thus no attempts for reducing the computation time were done.

The assumption of a linear model seems to be appropriate for shape modelling but might be inadequate for profile modelling. Most of the errors were due to the inner lip contour latching on to the teeth. The profiles covering the mouth opening basically change between three different intensity levels: (1) mouth closed, (2) mouth open and teeth not visible, (3) mouth open and teeth visible. It is therefore unlikely that this variability can

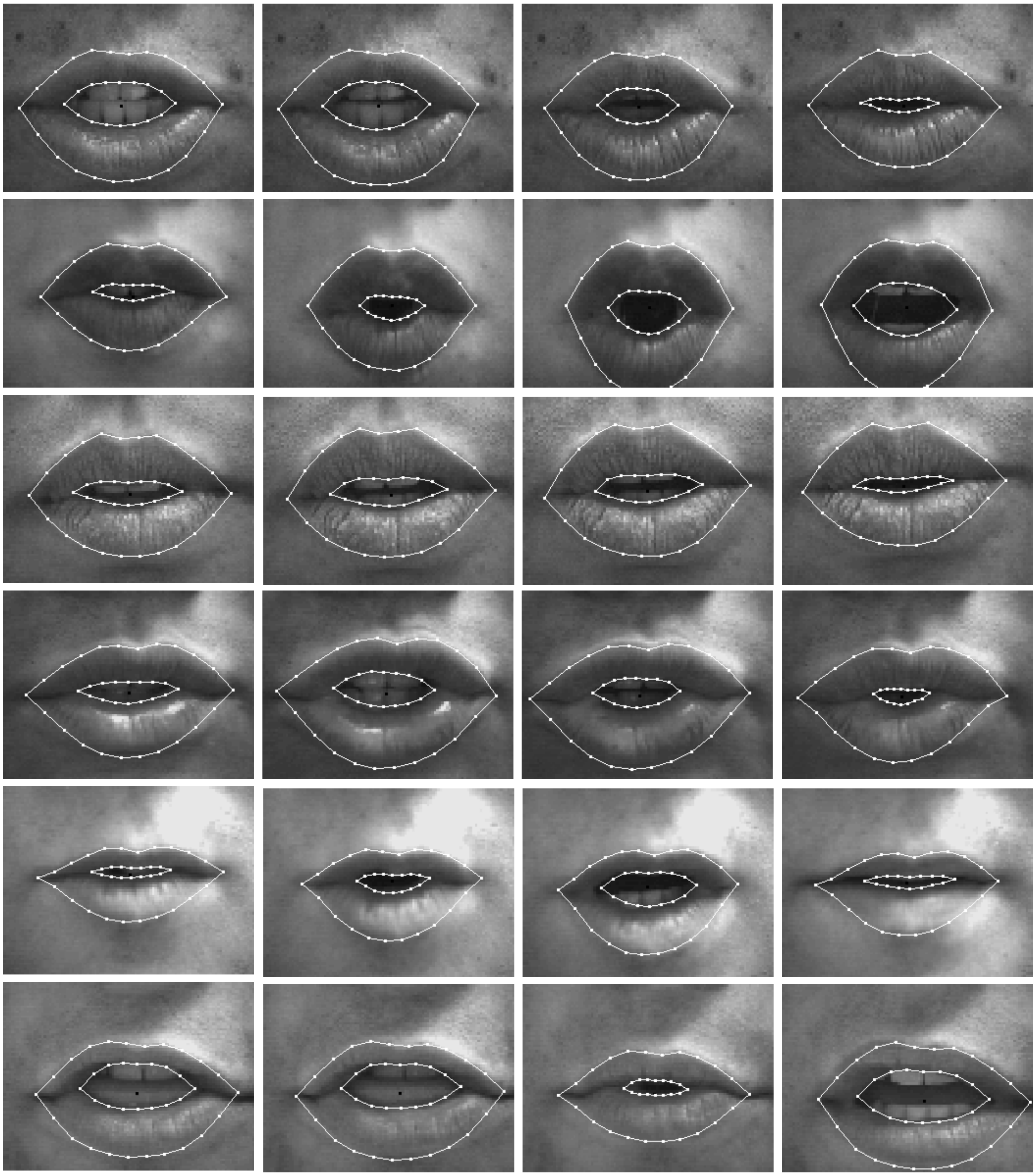


Figure 5.13: Examples of lip tracking results using *Model DC* and intensity model based search, which were classified as *Good*. The tracking results are very accurate across subjects despite large appearance differences and varying visibility of teeth and tongue.

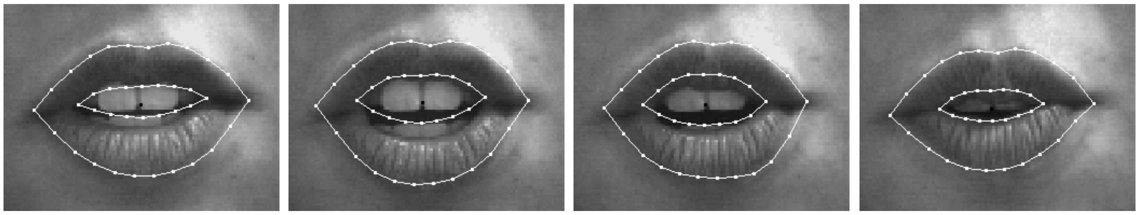


Figure 5.14: Example of a lip tracking sequence using *Model DC* and intensity model based search, which was classified as *Miss*. The inner model contour first latched onto the teeth and later missed the upper lip contour.

Table 5.2: Results for lip tracking using intensity based image search.

Model	Good (%)	Adequate (%)	Miss (%)
Model SC	95.8	2.1	2.1
Model DC	91.7	6.3	2.1

accurately be described by a uni-modal distribution.

5.8 Experiments on the M2VTS Database

For the M2VTS database, only *Model DC* was built with consisted of 38 points, the same number as for the Tulips1 database. The lip model was built from training samples of the first three shots. The dimension of the grey-level profiles was reduced to 7 but the distance of the profile points was increased from one pixel to two pixels. This resulted in a global grey-level vector of dimension 266. The 14 most significant shape modes and the 10 most significant intensity modes were used for lip tracking and for further processing.

5.8.1 Lip Tracking

The images contain the whole head of the person in frontal view as shown in Figure 4.3. Since a coarse location of the mouth area was not known, only lip tracking and no lip localisation experiments were performed. Tracking was initialised by placing the lip model with the mean shape and estimated scale and angle, at the lip centre of the first image of an image sequence.

Due to the much smaller dimension of the intensity model, image search was faster than for the Tulips1 models. However, the smaller dimension of the intensity model covers a smaller sample area for image search which might sacrifice robustness of the search. It might also reduce the information contained in the extracted intensity features for further processing.

The average number of function evaluations per image was about 370. A function evaluation took about 4.6 msec, resulting in about 1.7 sec for processing one image. The

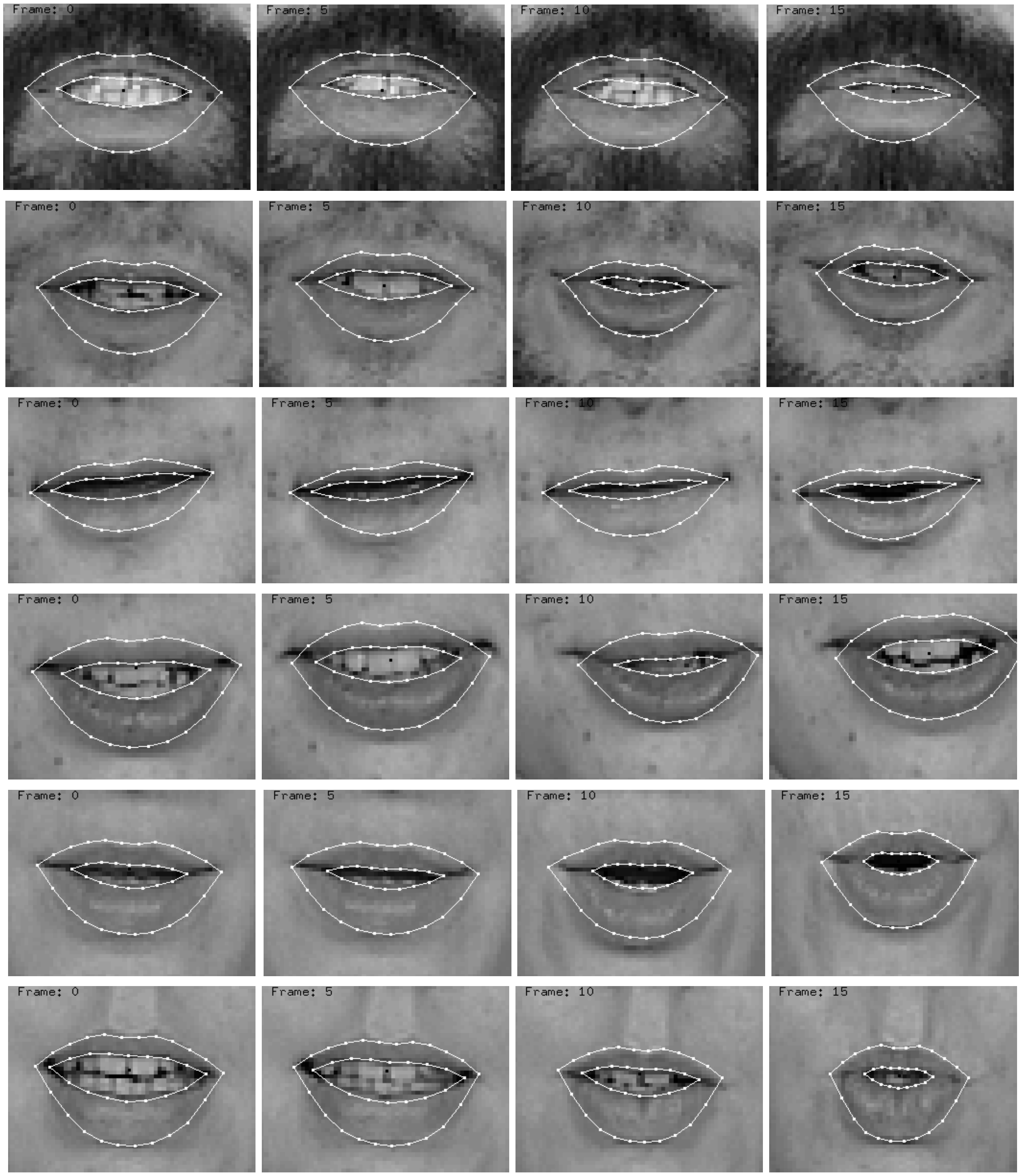


Figure 5.15: Examples of lip tracking results on the M2VTS database. The first row demonstrates the robustness of the algorithm for subjects with beard despite the reverse contrast between lips and skin. It can also be seen that the search method copes well with the tracking of the inner lip contour which is subject to the same contrast changes.

original colour images of the database were converted to grey-level images for the experiments. Tracking experiments were performed on all 5 shots which consists of 27,427 images. This represents the most extensive lip tracking test found in the literature.

Whereas for the Tulips1 database, subjective evaluation of the tracking performance was still feasible, it would be very laborious to perform this task for the M2VTS database. Direct performance evaluation of the tracking algorithm was therefore omitted. Instead, it was tried to evaluate the combined performance of the feature extraction process and the classification process which uses these features. The features are used in Chapter 6 for speech recognition and in Chapter 7 for person recognition. Classification errors, for both applications, might therefore be due to inaccurate tracking results or due to the classification algorithm. Examples of lip tracking results are shown in Fig 5.15.

5.9 Discussion

An appearance based lip model has been described for lip localisation, lip tracking and feature extraction. Image search is very robust since the model can only deform into shapes similar to the ones found in the training set. The model is able to track the lip contours of different person and different skin and lip intensities by using a statistical model for the intensity distribution around the lips.

Other lip tracking approaches are often dependent on heuristics like weights for individual shape parts, thresholds for edges, or penalties for deformation, to make the algorithm work. The method described here almost completely avoids the use of such constraints, the only constraints being the statistically based limits for the shape and intensity modes. Both the shape and the intensity variability are learned from a training set which seems to be the only realistic way of model acquisition.

The cost function assumes that the intensity distributions are well represented by a unimodal distribution. However, the intensities inside the mouth mainly dependent on the mouth opening and the visibility of teeth and tongue, and are therefore more likely to follow a multi-modal distribution. The lip tracking performance might be improved by constructing a more specific intensity model of the lips.

The lip tracking experiments consist of the largest experiments found in the literature with respect to the number of subjects and the size of the database. The system has achieved state-of-the-art performance and was shown to perform considerably better than image gradient based search techniques. The model can be represented by a small number of shape and intensity modes to any desired degree of accuracy. The model parameters can easily be recovered from the tracking results and be used for subsequent applications such as speech recognition or speaker recognition.

Chapter 6

Visual Speech Recognition

This chapter describes the developed speechreading system [115, 113, 116] based on the features extracted from the lip model during lip tracking. The features represent the shape of the lips and the intensity around the mouth and contain therefore relevant speech information. The features are modelled by Gaussian distributions and their temporal dependencies by hidden Markov models. Word models based are trained on these features based on Maximum Likelihood (ML) estimation and speech recognition is performed using the maximum posterior probability criterion.

One of the main difficulties in speechreading is to cope with the large variability across speakers, due to individual appearance and individual lip movements (see for example Fig. 5.13 and Fig. 5.15). Speaker independent tests, using different speakers for training and testing, were therefore performed to see how well the system generalises to new speakers.

Recognition experiments are described for an isolated digit recognition task using visual features only. Results are also presented for audio-visual speech recognition tests using composite acoustic-visual feature vectors. Finally, extensive tests for a speaker independent recognition task of continuously spoken digits are described using visual features only.

6.1 Previous Approaches

While different approaches for feature extraction have been discussed in Chapter 5, here, different methods for speech modelling using those features are reviewed. Most approaches are based on modelling techniques which are widely used in acoustic speech recognition. Besides speech recognition, the use of visual features has also been proposed for improved speech segmentation in noise [120]. However, current speech recognition technologies try to avoid to make hard decisions like phoneme segmentation at an early stage. Instead, the segmentation is deferred until higher level knowledge sources have been considered.

Template Matching of Static Images

A very simple method for visual speech recognition is the comparison of features extracted from static images with stored templates [64, 190, 183]. Such approaches ignore the temporal evolution of features over time and assume that an image, representing the important features of a given utterance, can be captured from the image sequence. It was however shown in Chapter 3 that the visual properties of the speech articulators are very different for continuous speech and that coarticulation has a strong influence on them. The main challenge in both acoustic and visual speech recognition lies in the recognition of continuous and natural speech. Recognition of static images simplifies the recognition procedure extremely and makes only little contribution to the state-of-the-art in ASR.

Template Matching of Image Sequences

Petajan's early system [143, 144] compared the extracted sequence of geometric parameters with stored templates of feature sequences. No time warping was used to align the sequences. Differences in the word length were simply handled by filling the end of the shorter feature sequence with zeros until it had the same length as the longer sequence. This early approach already performed recognition on image sequences, although different speaking rates are not taken into account.

The system was later extended by vector quantisation and dynamic time warping (DTW) [142]. A mouth image codebook was created by vector quantisation which was based on the dark region area of the oral cavity. The quantised images were used in a DTW algorithm to find the best path, between the unknown sequence and the stored template. Other DTW based systems include the work of Rao et al. [153, 51, 97].

Mase and Pentland [122, 123] used linear time-warping to match the observed feature vectors with stored templates. The original features consisted of eight components, made up of a horizontal and a vertical velocity component for all four windows, placed around the mouth. These were reduced by principal component analysis to only two features which accounted for about 75% of the total variance. The two components were normalised for the distance measure with respect to their variance.

DTW represents an important milestone in ASR but most current ASR systems are based on HMMs which allow more accurate modelling of speech, thus enabling higher ASR performance. DTW is however still used in the decoding stage of ASR systems because the Viterbi algorithm (Chap. 2) can be interpreted as a DTW algorithm. Some of the disadvantages of DTW based speech modelling techniques are the distance measures between the features which doesn't consider their feature distribution and the temporal modelling which is not very specific and often based on heuristic penalties.

Discrete HMMs

Petajan's system [143, 144] was extended by Goldschen [77, 76] to a continuous speechreading system using discrete HMMs. A codebook of 64 mouth images was created with a clustering algorithm based on 13 oral-cavity features. Distinct viseme groups were then determined using a HMM similarity metric [152] and a clustering algorithm. Context dependent sub-word models were trained based on these viseme classes, similarly to acoustic sub-word models. This study suggests that acoustic sub-word modelling techniques are also applicable to the visual speech signal. Silsbee and Bovik [164, 163] also employed discrete observation HMMs using a set of 17 visual code-vectors. The code-vectors are based on preprocessed grey-level images and were determined by hand.

Continuous Density HMMs

Brooke et al. [27] have used continuous density HMMs (CDHMM) to train acoustic-visual tri-phone models on a digit database. Composite feature vectors were constructed by concatenating the acoustic filter bank outputs with visual PCA coder outputs. The method is similar to common acoustic sub-word modelling techniques with the difference that the acoustic feature vector is augmented with visual features. Other speechreading approaches based on CDHMMs are described in [137, 37, 36, 81, 86]. The speechreading system described here also uses CDHMMs for visual speech modelling.

Connectionist Approaches

The speechreading system reported by Bregler et al. [22] used a modular time-delay neural network (TDNN) which consists of an input layer, a hidden layer, and a phone state layer. The network was trained by back-propagation. At the phone state layer, DTW is used to find the optimal path of phone-hypotheses for the word model. Variations of these networks were described in [57]. A similar system, based on a modified TDNN has been described by Wolff et al. [182].

Bregler et al. [23, 25] have described another connectionist approach for combining acoustic and visual information into a hybrid MLP/HMM speech recognition system [19]. The acoustic and visual feature vectors were concatenated to form a composite feature vector. The MLP is trained to estimate the posterior probabilities of the phonemes given the audio-visual data. The likelihoods are obtained from the posterior probabilities and used as the emission probabilities for the HMM.

6.2 Overview of the Speechreading System

An overview of the speechreading approach proposed here is shown in Fig. 6.1. The system utilises the lip tracking algorithm described in Chap. 5. Lip tracking and feature extraction is first performed on the training set. The extracted features and their corresponding labels of the spoken utterance are then used to train one HMM for each utterance class. The training process is initialised by a prototype HMM defining the HMM structure and initial parameter values. Training is performed according to ML estimation. During the recognition stage, lip tracking and feature extraction is performed and the probability of each model for having generated the observed features is estimated. The HMM with the highest probability is chosen as the recognised utterance.

The same procedure is used for isolated and for continuous speech recognition. The difference in the training process of continuous speech is that an additional training cycle is performed which does not use time alignment information. For audio-visual speech recognition, the feature extraction process includes the extraction of acoustic features as

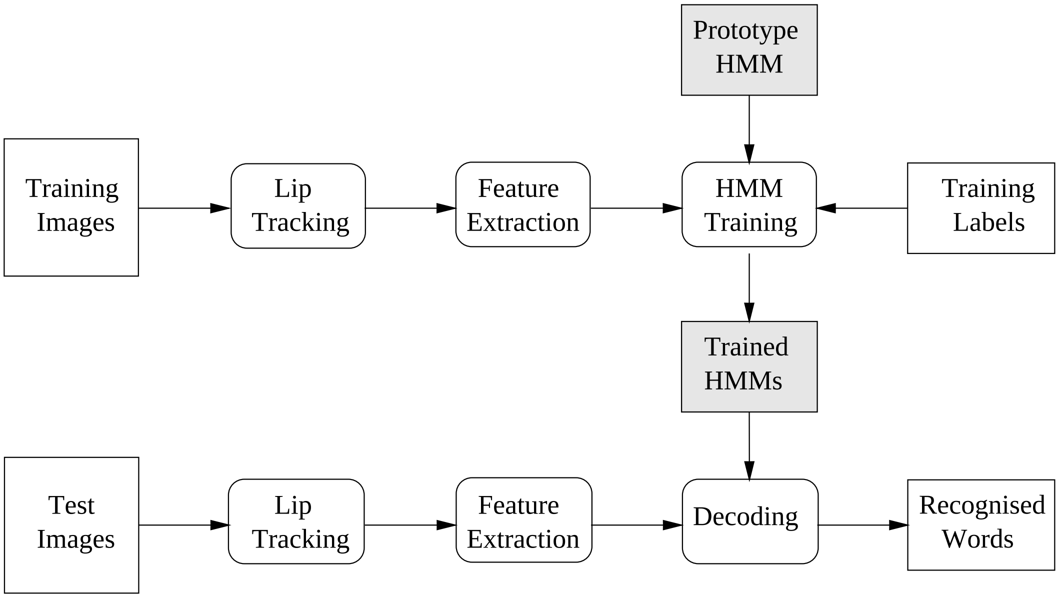


Figure 6.1: Overview of the speechreading system.

well. The experiments described here are speaker independent, i.e. the subjects in the test set do not appear in the training set.

6.3 Visual Speech Features

6.3.1 The Correspondence Problem

The method described here is related to techniques for face decomposition [99] and face recognition [175], where the face is decomposed into a weighted sum of eigenvectors computed by a Karhunen-Loéve (K-L) expansion. Its application for speechreading has been described in [23, 27]. One of the disadvantages using the K-L expansion method is the so-called correspondence problem. The algorithm estimates the intensity difference between images using a set of basis images where pixel $I_1(ij)$ of image I_1 is compared with pixel $I_2(ij)$ of image I_2. Particularly in the case of deformable objects, this leads to a comparison of pixel intensities which do not correspond to the same object part. In lip motion analysis this might be due to non-rigid deformation of the lips, mouth opening, visibility of teeth and tongue, or due to individual differences. Although this variability might be implicitly modelled by a K-L decomposition, the manifold of possible lip instances projected onto the basis images is not linear since the weights still account for the intensity rather than the deformation.

For speech and speaker recognition, a representation is desirable which separates shape variability from intensity variability. Detailed shape information should be independent of intensity and intensity information should account for the same part of the object and be independent of shape. The method described here is related to the so-called shape-free features [106, 105]. Individual profiles move with the contour model and are placed at the centre of the contours points and perpendicular to the contour. The intensity features are therefore independent of shape since they always describe the intensity around the same lip contour point. The intensity profiles inside the mouth are dependent on the mouth opening and on the visibility of teeth and tongue but they alway account for the grey-levels next to the lip contours.

6.3.2 Shape Features

In Section 3.7 it was shown that the inner and outer lip contours are the most important visual speech features. The shape parameters obtained from the tracking results are therefore used as features for the speech recognition system. Translation, rotation and scale parameters are not used as features since they are unlikely to provide speech information. Furthermore, translation and rotation might be due to head movement and absolute scale is difficult to estimate and is speaker dependent. Assuming accurate lip tracking, all shape features are invariant to translation, rotation (2D), scale, and illumination. All shape parameters b_{si} are normalised with respect to their corresponding eigenvalue λ_{si}:

$$\tilde{b_{si}} = \frac{b_{si}}{\lambda_{si}} \qquad \text{for all } 1 \leq i \leq N_s \tag{6.1}$$

The parameters account for both variability across speakers and variability due to speech production. The task of the speech recognition system is therefore to learn models which account for speech information and which are invariant to speaker variability.

6.3.3 Intensity Features

Lip shape information provides only part of the visual speech information. Other information is contained in the visibility of teeth and tongue, protrusion, and finer details. The intensity parameters of the lip model are therefore used as features to provide information complementary to the shape features. The intensity of a 2D image reflects its actual 3D shape and provides information not covered by the shape model. In general, the intensity image depends on the shape of the object, its reflectance properties, the light source and the viewing angle. For a Lambertian surface[1], the image radiance or brightness $I_r(x, y)$ at a particular point in the image is proportional to the irradiance or illumination $I_i(x, y)$ at that point. The radiance depends on the irradiance of the illumination source $I_i(x, y)$ and the angle θ between the surface normal and the direction toward the illumination source:

$$I_r(x, y) = \frac{1}{\pi} I_i(x, y) \cos\theta \qquad \text{for } \theta \geq 0 \tag{6.2}$$

This equation directly relates the 3D shape to intensity and is fundamental to the methods for recovering shape from shading [89]. The recovery of 3D shape from shading is possible under certain constraints, i.e. it is normally assumed that the angle of the illumination

[1] An ideal Lambertian surface is one that appears equally bright from all viewing directions and reflects all incident light, absorbing none.

source is known and that the surface is smooth. For our application of the face image, the shape from shading problem becomes very difficult. The illumination source is generally not known, the surface is disrupted at the oral opening, the oral opening itself is not smooth, and the reflectance properties of facial parts are not homogeneous and generally not known. Here, the motivation behind the use of intensity information is therefore not to reconstruct the 3D shape but to use it to implicitly represent 3D information and to represent information about the position and configuration of facial parts based on their individual brightnesses. In analogy to the shape parameters, the intensity parameters b_{ij} are normalised with respect to their corresponding eigenvalues λ_{ij}:

$$\tilde{b_{ij}} = \frac{b_{ij}}{\lambda_{ij}} \qquad \text{for } 1 \leq j \leq N_i \tag{6.3}$$

The determined intensity modes describes variation due to different subjects, different illumination, and speech production. In analogy to the shape features, the speech recognition system has to learn models which account for speech information and which are invariant to other intensity variability. Given correct tracking results, the features are invariant to translation, scale and rotation, but not to illumination as is the case for shape features.

6.3.4 Dynamic Features

Both the shape and intensity parameters are extracted for each image to form a frame base feature vector. Although consecutive feature vectors describe the dynamic information about the features implicitly they do not provide explicit information about temporal characteristics. Much visual speech information is contained in the dynamics of lip movements rather than the actual shape or intensity. Furthermore, dynamic information might be more robust to linguistic variability, i.e. intensity values of the lips and skin will remain fairly constant during speech, while intensity values of the mouth opening will vary during speech. On the other hand, intensity values of the lips and skin will vary between speakers, but temporal intensity changes might be similar for different speakers and robust to illumination. Similar comparisons can be made with shape parameters. Dynamic parameters (delta parameters) of the shape and intensity vectors were therefore used as additional features.

Absolute scale, which was defined as the distance between the lip corners, is difficult to estimate. If the whole face is available, it could be estimated as a relative measure with respect to eye distance for example. But this might be unreliable due to differences across subjects. Scale changes might however convey more important speech information than absolute scale and is easier to estimate. The change in scale was therefore also included in the delta parameters of the feature vector.

The dynamic features $\Delta\mathbf{o}(\tau)$ were given by

$$\Delta\mathbf{o}(\tau) = \frac{\mathbf{o}(\tau+1) - \mathbf{o}(\tau-1)}{2} \tag{6.4}$$

where the observation $\mathbf{o}(t)$ might represent the shape, intensity or scale.

Although delta features are widely used in ASR systems and can improve their performance considerably, their interpretation is not very clear. It should also be noted that their use violates the HMM assumption, that speech can be split into quasi-stationary segments, and the assumption that the observed vectors are conditionally independent of previously observed vectors.

6.4 Visual Speech Modelling

Hidden Markov Models (HMMs) are currently the most successful stochastic approach to acoustic speech recognition and are used as the basic stochastic framework in this thesis for visual speech modelling. A visual observation $\mathbf{O}$ of an utterance is represented by a sequence of visual feature vectors which are obtained from the lip tracker, with

$$\mathbf{O} = \mathbf{o}(1), \mathbf{o}(2), \ldots, \mathbf{o}(T), \tag{6.5}$$

where $\mathbf{o}(t)$ is the feature vector extracted at time t. The feature vector can consist of a combination of shape parameters $\tilde{\mathbf{b}}_s$, intensity parameters $\tilde{\mathbf{b}}_i$, and additional delta parameters $\Delta\mathbf{o}(t)$. In analogy to the acoustic speech signal, it is assumed that the feature vectors follow continuous probability distributions which is modelled by a mixture of Gaussian distributions. It is also assumed that temporal changes during speech are piece-wise stationary and follow a first-order Markov process. Thus, each HMM state represents several consecutive visual feature vectors. These assumptions are not strictly true but are also often made in the acoustic domain. They can be improved by increasing the number of states, which decreases the duration of the stationary states, but this increases the complexity of the models and the number of parameters to estimate.

The same assumption as for acoustic speech features is made, where the observation probabilities are uncorrelated and only dependent on the current state of the Markov chain and not on any previous states or observations. Much research effort has been devoted to develop methods which avoid these assumption. Despite these efforts, standard HMMs are still used in many state-of-the-art speech recognition systems. A more detailed study and discussion about the underlying assumptions of HMMs can be found in [19].

6.4.1 Training

As usually done in the case of small vocabulary ASR, whole word models were used for all experiments. In this case, each word class was thus represented by one HMM. The small size of the database and vocabulary prevented the use of sub-word models but the training of visual sub-word models is straight forward if an initial phonetic transcription is available. The HMMs only allowed self-loops and sequential transitions from the current to the next state as represented in Fig. 2.1.

The initial state probabilities are set to zero for all states but the first. The remaining parameters are estimated from the extracted model parameters of the training set. Each HMM is initialised by linear segmentation of the training vectors onto the HMM states, followed by iterative Viterbi alignment and computation of the means and variances for each state. In the case of multiple mixtures, the vectors for each state are clustered by a modified K-Means algorithm. The models are further re-estimated based on ML estimation using the Baum-Welch procedure (Sec. 2.3). Furthermore, the prior probabilities of the models $P(\lambda_i)$ are assumed to be uniform for all models. Also, no discriminant training was applied, consequently, $P(\mathbf{O})$ was assumed to be constant during training.

For continuous word recognition tasks, the models were trained using two training stages. In the first stage, the models are trained using the segmented training data, which is similar to the procedure of isolated word training. This procedure was performed to obtain some initial estimates of the model parameters. The second re-estimation procedure was performed on the whole training sentences using *embedded training*. This consists in concatenating the word models according to the transcription, but without segment information, and performing a Baum-Welch re-estimation of all model parameters simultaneously for each training sentence. This procedure is usually used in acoustic speech recognition and optimises the training of the models by ignoring segment information, thereby letting the models find the optimal model boundaries. One motivation for this procedure is that the provided transcription might not be optimal or might not be available at all.

6.4.2 Recognition

Speech recognition is performed based on the maximum posterior probability (MAP) criterion, given by

$$\arg\max_i P(\Lambda|\mathbf{O}) \tag{6.6}$$

where Λ represents a particular word string and $\mathbf{O}$ the observation sequence. In the case of isolated speech recognition, the term Λ will represent only one class, whereas in continuous digit experiments, the number of digits is not known a priori. The posterior probability can be estimated from the likelihood and the prior probabilities using Bayes rule:

$$P(\Lambda|\mathbf{O}) = \frac{P(\mathbf{O}|\Lambda)P(\Lambda)}{P(\mathbf{O})}. \tag{6.7}$$

The prior probabilities of all classes $P(\lambda_i)$ are assumed to be equal, consequently, also the prior probabilities of all possible word strings Λ are equal. Furthermore, the prior probabilities $P(\mathbf{O})$ are equal for all word classes during recognition. The recognition criterion (6.6) therefore reduces to the maximum likelihood criterion. The Viterbi algorithm is used to calculate the most likely state sequence.

6.4.3 Visualisation of Learned HMM States

The mean vectors of a HMM state represents the mean shape and intensity weights for that state and the sequence of mean vectors therefore describes the sequence of quasi-stationary mouth instances for a given utterance. The means of the states can be used to synthesise mouth images which represent these sequences.

Figure 6.2 shows the HMM states learned from the training sequences of 11 speakers, for the words “one” and “three”, using one mixture component per state. The models are

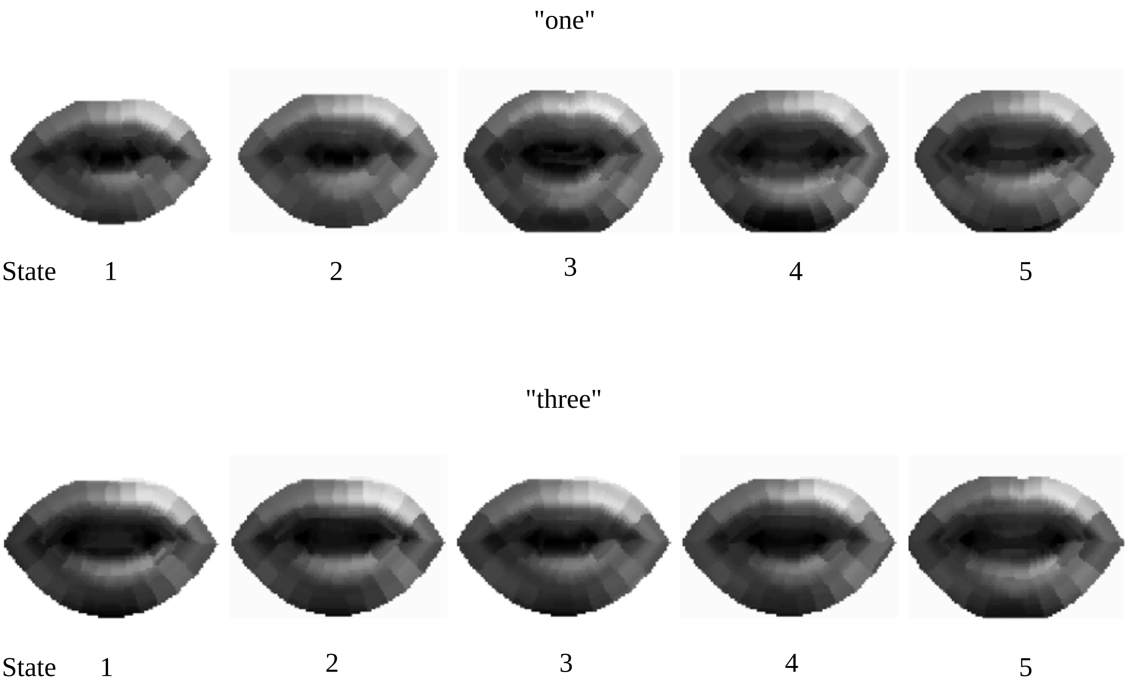


Figure 6.2: Learned sequence of HMM states for the words *one* and *three* using *Model DC*.

therefore average word models of 11 subjects. More detailed and subject specific articulation models might be obtained by using training data of only one person. The feature vector consisted of 10 shape and 20 profile parameters. The images show the synthesised lip model using the means and the weighted sum of the first 10 basis shapes and the first 20 basis profiles.

The differences of lip shape and intensity at the mouth opening, as well as their temporal differences can be clearly seen on these examples. The mouth opening in the first two states is smaller for model 'one' than for model 'three' but at state three it is smaller for model 'three'. The intensity inside the mouth is higher at the end for both models, representing the visibility of teeth. For model 'three' it is also high at the first state, indicating the visibility of the tongue.

6.5 Acoustic-Visual Speech Modelling

Although the focus of this work is on visual speech processing, a simple audio-visual speech recognition architecture is described and experiments for AVSR in acoustic noise conditions are described. The method does not aim to present an optimal integration strategy of the acoustic and visual modalities, but to illustrate the potential benefits of visual information in an acoustic speech recognition system.

6.5.1 Acoustic Speech Features

Acoustic speech features were processed in the form of *mel* frequency cepstrum coefficients (MFCC) [150]. This is a common acoustic feature extraction approach which often outperforms other parameter representations. The signal was first pre-emphasised with the transfer function $(1 - 0.96z^{-1})$ to increase the relative energy of the high frequency spectrum. 24 filter bank coefficients were then calculated from overlapping Hamming windows of 50 msec duration at 30 Hz. The frame period of 30 Hz was chosen to provide acoustic features at the same rate as the visual features, which facilitates integration of the two modalities. The filter bank outputs are warped to the mel-scale, using triangular filters, equally spaced along the mel-scale. The mel-scale is approximately linear below 1 kHz and logarithmic above, it is based on psycho-acoustical experiments and corresponds to units of perceived frequency. 12 MFCC parameters were then computed from the resulting log mel-frequency filter bank coefficients using the discrete cosine transform. Each acoustic feature vector consisted of 26 parameters: 12 MFCC coefficients, 12 delta MFCC coefficients, normalised log-energy and delta normalised log-energy.

6.5.2 Acoustic-Visual Feature Integration

This section briefly discusses issues regarding the integration of the visual and acoustic modalities and a simple integration technique is described. Issues like optimal integration are however out of the scope of this section and are not discussed. More elaborate schemes for audio-visual integration can be found in [167, 174, 94, 136].

Although there has been much work in modelling sensory integration by humans, the underlying process of human sensor fusion is not well understood and still subject to much discussion. Research in audio-visual speech perception has been sparked by the McGurk effect [130] and much research in human sensory integration has aimed at developing human sensory integration models which describe this effect. The illusion of perceived sounds for conflicting audible and visible stimuli confirm the combined audio-visual speech perception of humans. However, the perceived illusion might equally well occur in the case of conflicting acoustic stimuli, for example where one frequency band corresponds to phone A and the second frequency band to phone B. A similar illusion is likely to occur in this case and the integration of different frequency bands might be performed the same way as the integration of different modalities.

Here, the problem of sensory integration is addressed from an information theory point of view. For machine recognition, the bimodal speech signal can be considered as an observation vector consisting of acoustic and visual features. According to Bayesian decision theory, a maximum posterior probability classifier (MAP) is denoted by

$$\Lambda^* = \arg\max_{\lambda} P(\Lambda|\mathbf{O}^a, \mathbf{O}^v) \tag{6.8}$$

where Λ represents a particular word string, $\mathbf{O}^a$ represents the sequence of acoustic feature vectors $\mathbf{O}^a = \mathbf{o}^a(1), \mathbf{o}^a(2), \ldots, \mathbf{o}^a(T)$ and $\mathbf{O}^v$ the sequence of visual feature vectors $\mathbf{O}^v = \mathbf{o}^v(1), \mathbf{o}^v(2), \ldots, \mathbf{o}^v(T)$. The a posteriori probability can be obtained according to Bayes rule:

$$P(\Lambda|\mathbf{O}^a,\mathbf{O}^v) = \frac{P(\mathbf{O}^a,\mathbf{O}^v|\Lambda)P(\Lambda)}{P(\mathbf{O}^a,\mathbf{O}^v)} \tag{6.9}$$

If the two modalities are independent, the likelihood $P(\mathbf{O}^a,\mathbf{O}^v|\lambda_i)$ becomes

$$P(\mathbf{O}^a,\mathbf{O}^v|\lambda_i) = P(\mathbf{O}^a|\lambda_i)P(\mathbf{O}^v|\lambda_i) \tag{6.10}$$

which assumes that the likelihoods of the visual and acoustic modalities can be estimated independently.

In the work described here, the two modalities are assumed to be *conditionally dependent* which was implemented by constructing composite feature vectors of synchronised acoustic and visual parameters, sampled at equal frame rate. As discussed in Chapter 3, there are several arguments which support the hypothesis that the two modalities are conditionally dependent. Assuming conditional dependence is the more general case which includes the case of conditional independence as a model where the parameters of the acoustic and visual modalities are uncorrelated. A disadvantage of this method is that it normally requires more model parameters to be estimated. For the experiments described here, the dimension of the visual modality was similar to the dimension of the acoustic modality and caused no problems in parameter estimation.

Composite feature vector sequences $\mathbf{O}^{av}$ were constructed by concatenating the acoustic and visual feature vectors for each observation using

$$\mathbf{o}^{av}(\tau) = [o_1^a(\tau), o_2^a(\tau), \ldots, o_{N_a}^a(\tau), o_1^v(\tau), o_2^v(\tau), \ldots, o_{N_v}^v(\tau)]^T \tag{6.11}$$

The training and recognition processes are performed the same way as for visual features.

The described integration method which assumes *conditional dependence* has the following potential drawbacks:

- Each acoustic and visual feature component is assigned equal weight during training and classification, but the reliability of information conveyed by the features might vary.

- The quasi-stationary events of the acoustic and visual modalities might be different which is not well modelled by composite feature vectors.

- The model requires the availability of synchronised features at equal frame rate. This might be achieved by a non-optimal compromise between the acoustic and the visual frame rate, or by some kind of frame interpolation.

- The method is unlikely to be robust to high noise levels which occur in only one of the modalities.

In a scenario where either the acoustic channel or the visual channel is contaminated by noise, the test conditions will be different to the training conditions and consequently the performance of the system will be sub-optimal. For such a case of noise contamination which is limited to one modality or even a sub-band of features of that modality, integration schemes assuming conditional independence seem more likely to lead to optimal recognition performance [20].

6.6 Experiments

The speech recognition tests are based on the features extracted from the lip tracking procedure described in Chapter 5. An overview of the training and recognition procedure is illustrated in Fig. 6.1. All speech recognition experiments were performed for speaker independent tasks, using different speakers for training and testing. Speaker independent tests show how well a method generalises for new speakers and are therefore more difficult than multi-speaker or speaker dependent tests. All experiments were performed using the leave-one-out method [72]: Given a total of N_p subjects, one subject is excluded, the models are trained on the remaining $N_p - 1$ subjects, and the excluded subject is used for testing. The procedure is repeated N_p times, each time leaving a different subject out for testing, to test all subjects. Finally, the results are averaged over all subjects. All experiments were performed using the HMM toolkit HTK version 1.5 [187].

The used HMMs were constrained to only allow self-loops and sequential transitions between the current and the next state. In acoustic speech modelling, the number of states is often chosen so that a certain number of states is used per phoneme (usually around 3). In visual speech modelling, the number of quasi-stationary segments is likely to be different to the number of segments in acoustic speech. Experiments were therefore performed for HMMs with different numbers of states to investigate the number of quasi-stationary segments present in the visual speech signal.

The number of mixture components used to represent the distribution of a HMM state is an important but difficult problem. If the number is too small, the distribution will not be able to adequately model the discriminative characteristics of a feature distribution. If the number is too large, the training data might not be large enough to estimate the large number of parameters of the model. No method is known for estimating the optimal number of mixture components a priori. Experiments were therefore performed for different numbers of components.

6.6.1 Isolated Visual Digit Recognition on Tulips1

The Tulips1 database was used for these experiments which consists of 12 subjects. This resulted in 11 subjects to train each word model for each leave-one-out cycle. The number of training samples per model was 22. Recognition results were averaged over all speakers, resulting in a total of 96 test digits. All tests were performed for *Model SC* and *Model DC*. The words can be phonetically transcribed as follows:

one:	/W/	/AH/	/N/
two:	/T/	/UW/	
three:	/TH/	/R/	/IY/
four:	/F/	/AO/	/R/

The visual confusability of these four words can be roughly estimated by comparing the phone by phone confusability using viseme tables from other researchers (for example [171, 77]). The following observations can then be made: The first phoneme of each word is distinct. The second phonemes are confusable for the words two, three and four. For the third phoneme[2], one and three are confusable as well as four and two. The four words are therefore mainly distinct in the first phoneme.

Table 6.1: Word recognition rate (%) using all shape and intensity features and additional delta features for HMMs with six states and one mixture component.

Model	Shape + Δ	Intensity + Δ	Shape + Intensity + Δ
Model SC	81.3 %	78.1 %	82.3 %
Model DC	77.1 %	83.3 %	88.5 %

A large number of experiments was performed by varying the number of HMM states from one to six and the number of mixtures from one to three. HMMs with full covariance matrices were also tested but the results were generally worse than for diagonal variances. Six states was the maximal number which could be used, since the minimum number of frames of some utterances was also six and since the HMM structure did not allow to skip states. The maximal number of mixture components which could be trained was around three and depended on the number of states. For example, given 22 training examples with an average frame number of ten, and a HMM with 6 states, results in an average of only about 37 training frames per states. The feature vector either contained shape or intensity parameters or both. Additional tests were performed by including delta features.

Best results were obtained by HMMs with six states and one diagonal variance vector. This indicates that the number of quasi-stationary states of the speech signal might be at least as large as the frame rate. The results also suggest that not more that one mixture component could be estimated reliably due to the small training size. For experiments using all shape modes and all intensity modes in the feature vectors as well as all delta features, recognition rates of 82.3% were obtained for *Model SC* and 88.5% for *Model DC*. More details are given in Table 6.1. For *Model SC*, the shape parameters led to higher performance than the intensity parameters, while for *Model DC* the opposite was the case.

The shape features and intensity features used in the experiments describe principal shape and intensity variability. This variability might however not be directly related to speech information. For example, shape variability might account for lip shapes of different persons and intensity variability might be due to illumination or different skin and lip intensities across subjects. Although the training algorithm ideally learns the features which are important for word discrimination, in practice the training set is rarely large and balanced enough to ensure this. Furthermore, no discriminative training was used to train the models.

Another set of tests was therefore performed in order to analyse the contribution of each mode towards the recognition performance. This was done by performing separate experiments for each individual shape and intensity mode [114, 113]. Results for individual shape modes for *Model SC* are depicted in Figure 6.3 and for *Model DC* in Figure 6.5.

[2] Since the second word only consists of two phonemes, the second phoneme is used here.

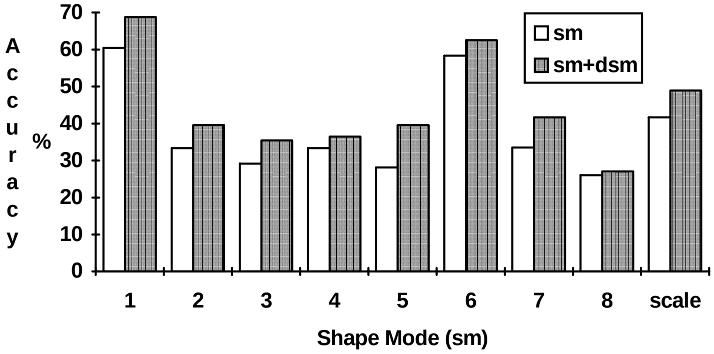


Figure 6.3: Recognition accuracy for each individual shape mode (sm) and optional delta shape mode (dsm) using *Model SC*.

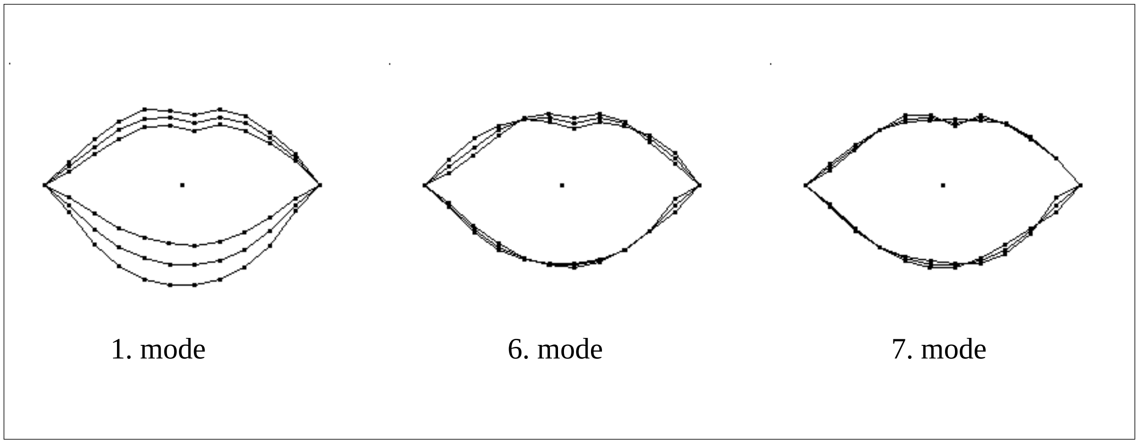


Figure 6.4: The first three shape modes (1, 6, 7) of *Model SC* with the highest recognition performance (not counting scale). The modes of variation are laid over the mean shape.

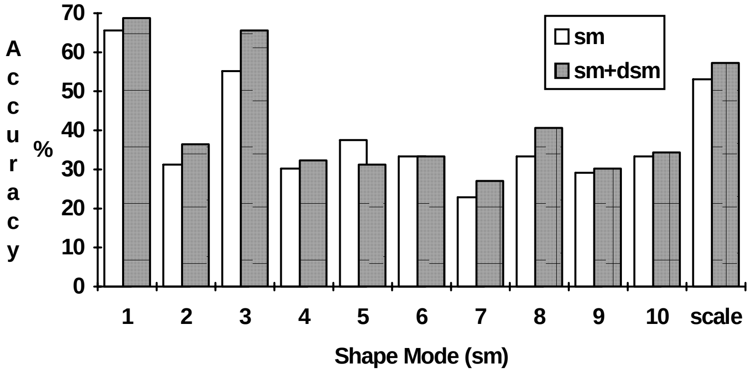


Figure 6.5: Recognition accuracy for each individual shape mode (sm) and optional delta shape mode (dsm) using *Model DC*.

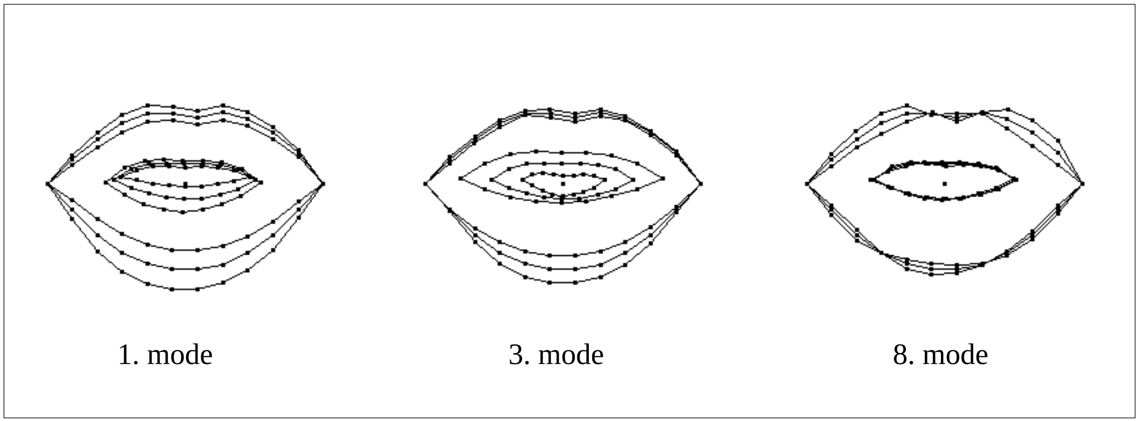


Figure 6.6: The first three shape modes (1, 3, 8) of *Model DC* with the highest recognition performance (not counting scale). The modes of variation are laid over the mean shape.

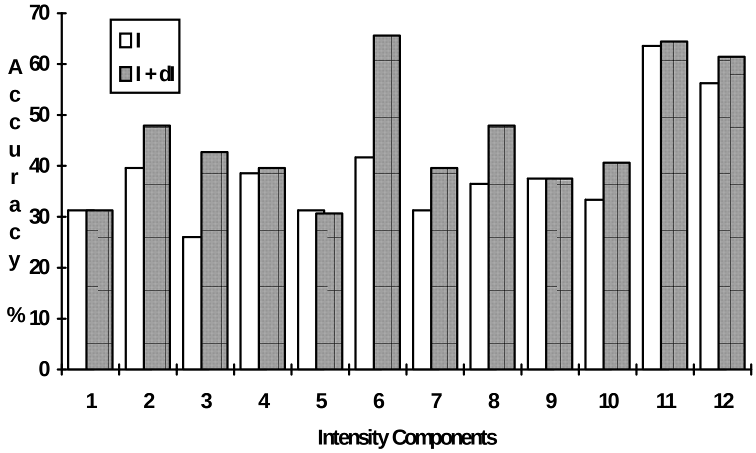


Figure 6.7: Recognition accuracy for each individual intensity mode (I) and optional delta intensity mode (dI) using *Model SC*.

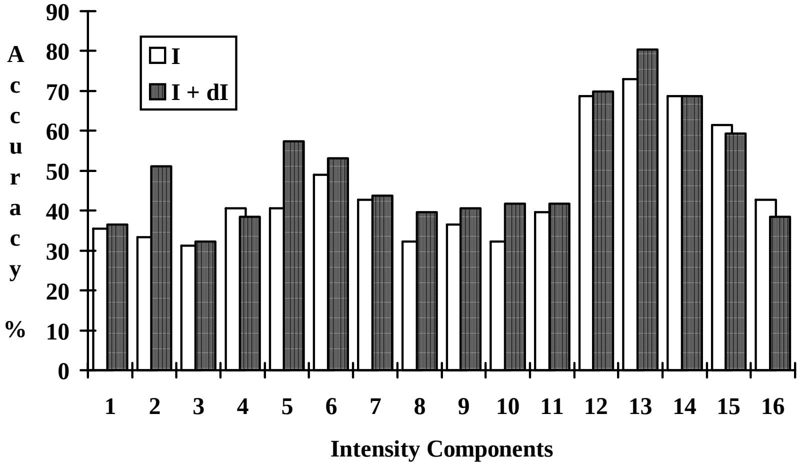


Figure 6.8: Recognition accuracy for each individual intensity mode (I) and optional delta intensity mode (dI) using *Model DC*.

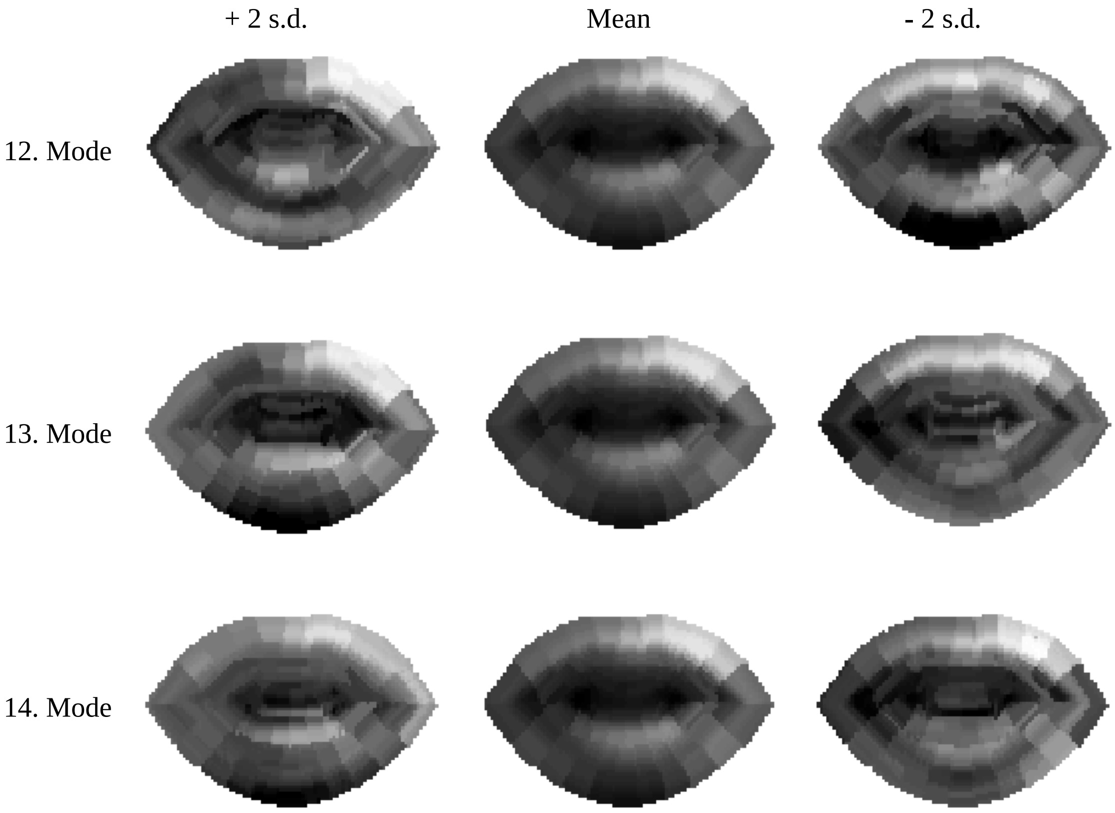


Figure 6.9: Intensity modes 12-14 for the mean shape of *Model DC*, which contributed most to recognition performance.

It can be seen that most of the parameters led to an accuracy between 30% and 40%, which is close to random classification of 25%. A few parameters, however, led to high performance and the shape modes with the highest individual recognition accuracy for *Model SC* are displayed in Fig. 6.4 and for *Model DC* in Fig. 6.6. The results indicate that for both models, the first mode, which accounts for the lower lip contour, is the most discriminant feature for the given task. Furthermore, the inclusion of delta information improved the recognition accuracy in almost all cases.

Results for recognition test using individual intensity modes can be found in Fig. 6.7 and Fig. 6.8. The most discriminant intensity modes of *Model DC* are displayed in Fig. 6.9. For both models, the most discriminant intensity modes are not the modes which account for the largest variability but rather modes which describe finer intensity variability. The most discriminant modes for *Model DC* as shown in Fig. 6.9 are hard to interpret but they mainly seem to account for different intensities at the mouth opening.

Table 6.2 displays the recognition rate for the five most discriminant parameters (shape modes 1, 3 and intensity modes 12-14). Additional tests were performed by including delta

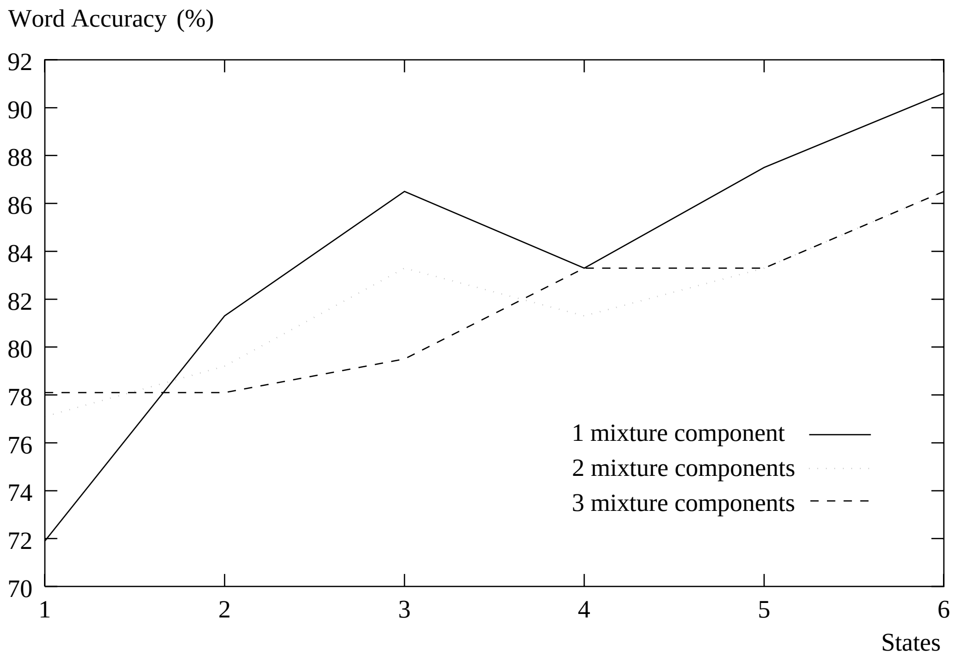


Figure 6.10: Recognition rate using *Model DC* as a function of HMM states and mixture components.

parameters and delta scale. Using the reduced feature vector, all results are higher than for the original feature vector. The results for shape features are similar for *Model SC* and *Model DC* but for intensity parameters, significantly higher accuracies are obtained with the double contour model *Model DC*. This seems obvious since the double contour model provides important information about the mouth opening. The inclusion of delta features improved the performance of the system in almost all cases, indicating the importance of dynamic information and their robustness to speaker variability and illumination. The confusion matrix for the words is shown in Table 6.3 and the word accuracy for each individual test subject in Table 6.4.

Table 6.2: Word recognition rate using the five most discriminant features.

Features	Model SC	Model DC
Shape	72.9 %	75.0 %
Intensity	65.6 %	89.6 %
Shape + Intensity	84.4 %	87.5 %
Shape + Δ	83.3 %	81.3 %
Intensity + Δ	82.3 %	89.6 %
Shape + Intensity + Δ	86.5 %	90.6 %

Table 6.3: Confusion matrix for the system with 90.6 % recognition rate (rows represent the actual digits, columns the recognised digits).

	one	two	three	four
one	23	0	1	0
two	0	24	0	0
three	2	0	21	1
four	2	0	3	19

Figure 6.10 displays the performance of the system as a function of the number of states per HMM and for different numbers of mixture components. The performance generally increases with the number of states. Best performance is obtained by using only one mixture component which indicates that the training set was too small to estimate more than one mixture component reliably. It can also be seen, that for HMMs with only one state, more training data per state is available, which allows more accurate modelling of several mixture components and which leads to higher performance for these models.

Table 6.4: Word accuracy for each individual test person using the system with 90.6% recognition rate.

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Accuracy (%)	100	87.5	87.5	75	100	100	75	100	100	75	100	87.5

It is interesting to know how the recognition system would perform for a much larger number of subjects. We can estimate the confidence interval for such a population in which a certain proportion of samples will fall. For the system with 90.6% accuracy, the confidence intervals for 95% of the sample population is

$$CONF\{84.8\% \leq Accuracy \leq 96.5\%\}.$$

This result suggests that for a large population of speakers, the speechreading system will obtain accuracies within the estimated confidence interval for 95% of the subjects, of which the 12 subjects of the Tulips1 database are a representative set.

Overall best results of 90.6% were obtained by *Model DC*, using shape and intensity features with additional delta features. This is approximately equivalent to the performance achieved by humans with no lip-reading knowledge who were asked to lip-read on the same database [137]. Those subjects achieved an average of 89.93% while the average performance of hearing impaired subjects with lip-reading knowledge was 95.49%. The performance of the described system is slightly higher than the performance of untrained humans but lower than the performance of trained lipreaders.

6.6.2 Isolated Acoustic-Visual Digit Recognition on Tulips1

This section describes digit recognition experiments using either acoustic features, visual features, or both combined. Conditional dependence was assumed for acoustic-visual experiments which was implemented by the use of composite acoustic-visual feature vectors. All experiments were performed on the Tulips1 database using the leave-one-out method. The acoustic signal was contaminated by additive noise to simulate an application scenario where background noise is present.

Speech models using only acoustic features were tested independently on a large number of HMM architectures. The number of states was varied between one and six and the number of diagonal mixture components between one and four. Full covariance matrices were also tested. The feature vector consisted of 12 MFCC parameters, 12 delta MFCC parameters, normalised log-energy and delta normalised log-energy. The acoustic recognition system obtained the highest performance of 94.8% using HMMs with two states and three mixture components.

Best results for audio-visual recognition tests in noise-free conditions were obtained using HMMs with four states and three mixtures. This HMM topology was used for further tests where the acoustic signal contained different levels of additive noise. Approximately white Gaussian noise was added to the audio signal resulting in SNRs from 20 dB to -20 dB. The noise was only added to the tests set, the training set always consisted of noise-free data.

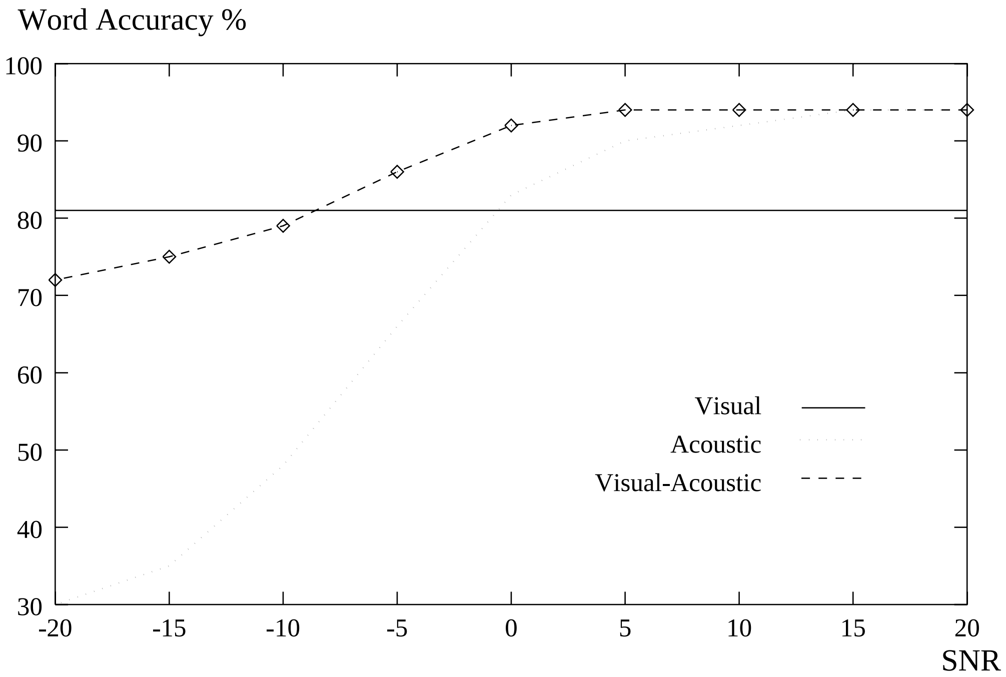


Figure 6.11: Recognition rate as a function of acoustic SNR for acoustic only, visual only, and acoustic-visual features.

Figure 6.11 displays the recognition results for acoustic, visual, and acoustic-visual feature vectors as a function of the SNR. The accuracy of the visual system for this HMM architecture is 81.3% and is constant over the whole ranges of acoustic SNRs. At a SNR of -20 dB, the performance of the acoustic system drops down to 30.2% whereas for the acoustic-visual system it only falls to 72.9%. The performance of the AVSR system drops below the performance of the visual only system at very low SNRs. This is due to the miss match between the training set and the noise contaminated test set.

The experiments show that the performance of an acoustic only ASR system can be considerably improved in the presence of noise by using visual information and by integrating both modalities at the feature level.

6.6.3 Continuous Visual Digit Recognition on M2VTS

The M2VTS database was used to perform speaker independent digit recognition experiments. Since no transcription of the speech data was available, the word boundaries were found by a HMM based speech recognition system which was used to segment and label the sentences [92]. The recogniser used the known sequence of digit word models, which were trained on the *Polyphone* database of IDIAP [38], and performed "forced alignment". In addition, the recogniser also performed silence detection, which were cut out of the digit segments. The training data of the visual features are therefore based on acoustic rather than visual segmentation. It is likely that acoustic and visual segmentation is not identical, e.g. the visual signal is likely to change more smoothly than the acoustic signal and might not contain quasi-stationary periods. Furthermore, visual anticipation might precede the acoustic signal. However, the acoustic signal provides normally more information than the visual signal with respect to phoneme boundaries, which favours the use of acoustic information to segment the visual speech signal.

Two recognition tasks were performed: one task was continuous word recognition of the spoken sentences and the second task was defined as word recognition of the segmented sentences. Although the first task is continuous speech recognition, the sequence of the digits was always the same from "zero" to "neuf". The words were therefore always spoken in the same context, which usually simplifies continuous speech recognition. The second task can be considered as isolated word recognition but where the words were spoken continuously. It was mainly performed to obtain the recognition accuracy, given the acoustic segmentation. All experiments were performed for speaker independent tests on the first four shots of the database using the leave-one-out procedure. One experiment therefore consisted of 37 leave-one-out tests, each made up of 1440 training words and 40 test words. This resulted in a total of 1480 test words spoken by 37 subjects.

Visual features were obtained from lip tracking results as described in Chap. 5. The visual feature vector consisted of 14 shape parameters, 10 intensity parameters, scale, and their temporal difference parameters. This resulted in a 50-dimensional feature vector. Different HMM architectures were investigated by varying the number of states (1 - 10) and the number of mixture components (1, 2, 4, 8). For HMMs with only sequential transition probabilities, the number of frames of a training or test example must be at least as large as the number of HMM states. Those segments, where the number of frames was below

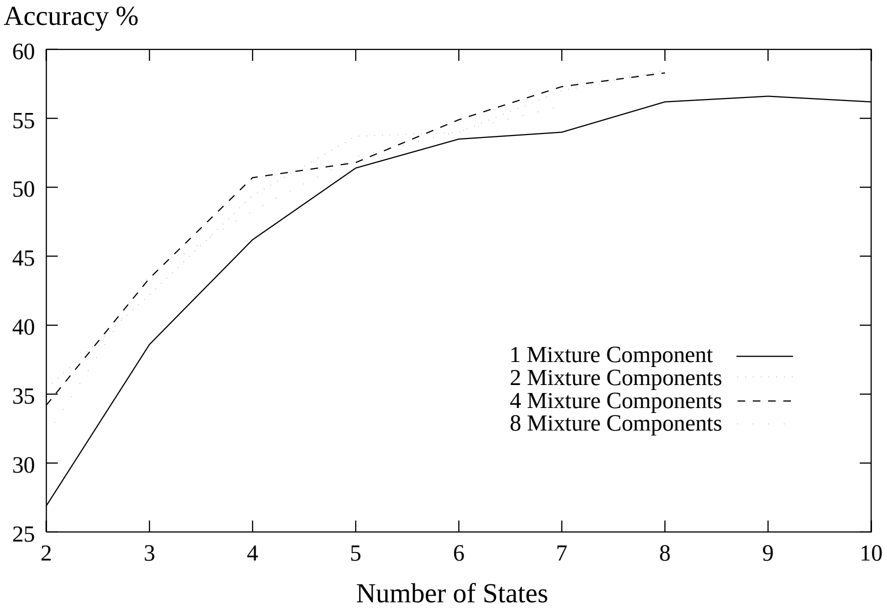


Figure 6.12: Continuous speaker independent digit recognition rate rate for different number of mixtures as a function of the number of HMM states using visual features only.

Table 6.5: Speaker independent recognition accuracy using HMMs with 8 states and 2 mixture components per state for different training and test procedures.

	Segmented Recognition	Continuous Recognition
Segmented Training	60.2%	51.3%
Embedded Training	58.7%	58.5%

the number of HMM states, were therefore excluded in the training and recognition of segmented digits.

The models for the segmented digit recognition task were trained on the segmented training samples. The models were initialised by linear segmentation, followed by Viterbi-alignment. This was followed by re-estimation using the Baum-Welch algorithm. Multiple mixture components were initialised by splitting the distributions of trained states using a modified k-means clustering algorithm. The models for continuous word recognition were trained with the same procedure but a second re-estimation procedure was performed on the whole training sentences using *embedded training*. The results are given in the percentage of words correctly recognised using

$$\%Correct = \frac{H}{N} \times 100\%, \tag{6.12}$$

and by the accuracy which is computed by

$$\%Accuracy = \frac{H - I}{N} \times 100\%, \tag{6.13}$$

where H is the number of correct words, I the number of insertions and N the total number of words in the test set. The alignment between the actual word labels and the recognised word labels was computed by a DTW algorithm.

Highest continuous recognition results at 58.5% accuracy were obtained using HMMs with eight states and two mixture components per state. This performance is very high, considering the high visual confusability between several words and the difficult task of speaker independent continuous speech recognition. The results for this HMM architecture are shown in Table 6.5, using different training and test procedures. Although continuous speech recognition is usually much more difficult than isolated speech recognition, the error rates for the procedure of segmented training and testing are not much smaller than for embedded training and continuous recognition. This might however be due to the inadequate acoustic segmentation of the visual training words. This assumption is supported by the fact that the segmented recognition results for embedded training were lower than for segmented training. Continuous recognition results on the other hand improved considerably after embedded training. The relatively high performance for continuous word recognition might however be due to the fact that the words were always spoken in the same context.

Results for continuous digit recognition for different numbers of states and mixtures are summarised in Fig. 6.12. The accuracy increased substantially with the number of HMM states. This suggests that the visual speech signal does not contain quasi-stationary segments like the acoustic speech signal or that the quasi-stationary segments are of smaller duration than the sampling period. The performance also generally increased with the number of mixture components. For HMMs with a large number of states, only a few mixture components could be trained due to the reduced number of long enough training examples.

The confusion matrix for segmented word recognition using HMMs with five states and one mixture component is shown in Tab. 6.6 (rows represent the actual digits, columns the recognised digits). In addition to the previous notation, D is the number of deletions and S the number of substitutions. Five segments had fewer than five frames and were excluded from the experiments. The performance of the system for different words varies considerably. Visually more distinct words like “zero”, “deux” and “trois” are easier to recognise than visually less distinct words like “quatre”, “cinq” and “sept”. It can also be seen from the table, that the latter three words were mainly confused with another. The confusion matrix for continuous word recognition is shown in Tab. 6.7. Similar to the segmented recognition tests, words which are visually more distinct like “zero”, “trois”, and “neuf” obtain high recognition rates, whereas visually less distinct words like “quatre”,“cinq”, “six”, and “sept” are harder to distinguish. These less distinct digits are subject to very little facial movements which therefore often result in “deletion” errors, where no digit was recognised at all. Deletion errors accounted to about half of the total errors. It can be seen from Tab. 6.7 that the number of substitution errors is much smaller than for segmented

Table 6.6: Confusion matrix for *segmented visual digit recognition* using HMMs with 5 states and 1 mixture component per state.

```
------------------------ Overall Results -------------------------
PHONE:  %Corr=57.36, Acc=57.29 [H=846, D=0, S=629, I=1, N=1475]
------------------------------------------------------------------
          z    u    d    t    q    c    s    s    h    n
          e    n    e    r    u    i    i    e    u    e
          r         u    o    a    n    x    p    i    u
          o         x    i    t    q         t    t    f
                         s    r                           Del [ %c / %e ]
 zero   121    5    3    0    5    1    1    7    2    3    0  [81.76/ 1.83]
   un     3   97    6    1    9    6    6    5    4   11    0  [65.54/ 3.46]
 deux     2    2  108    1    6    5    4    2    6   12    0  [72.97/ 2.71]
trois     0    0    2  113    5    0    1    3   21    2    0  [76.87/ 2.31]
quatr     1   17    3    0   48   21   32    6    7   11    0  [32.88/ 6.64]
 cinq     1    4    6    1   27   37   50   16    1    4    0  [25.17/ 7.46]
  six     0    4    6    0   12   22   94    6    1    3    0  [63.51/ 3.66]
 sept     3    5    3    2   13   18   37   54    3   10    0  [36.49/ 6.37]
 huit     0    3    8   19    6    3    1    5  100    3    0  [67.57/ 3.25]
 neuf     5   15   13    1    4    1    1   24    9   74    0  [50.34/ 4.95]
  Ins     0    0    0    0    0    0    0    1    0    0    1 / 629 / 0
```

Table 6.7: Confusion matrix for *continuous visual digit recognition* using HMMs with 8 states and 2 mixture component per state.

```
------------------------ Overall Results -------------------------
PHONE:  %Corr=62.03, Acc=58.45 [H=918, D=332, S=230, I=53, N=1480]
------------------------------------------------------------------
          z    u    d    t    q    c    s    s    h    n
          e    n    e    r    u    i    i    e    u    e
          r         u    o    a    n    x    p    i    u
          o         x    i    t    q         t    t    f
                         s    r                           Del [ %c / %e ]
 zero   132    0    2    1    0    0    0    0    2    0   11  [96.35/ 0.34]
   un     0   95    1    2    6    3    3    6    2    2   28  [79.17/ 1.69]
 deux     3    0   92    1    0    1    3    2    8    5   33  [80.00/ 1.55]
trois     0    2    0  108    2    0    2    0    6    4   24  [87.10/ 1.08]
quatr     0    1    3    0   72    2    4    4    7    0   55  [77.42/ 1.42]
 cinq     1    4    2    0    3   59   10    6    2    3   58  [65.56/ 2.09]
  six     0    2    1    1    9    3   61    1    6    1   63  [71.76/ 1.62]
 sept     1    3    3    0    4    5    4   80    3   10   35  [70.80/ 2.23]
 huit     5    2    3    9    1    3    1    1  102    5   16  [77.27/ 2.03]
 neuf     3    1    8    1    1    3    0    3    2  117    9  [84.17/ 1.49]
  Ins     1    1    2    1    5    6   11   10   10    6   46 / 230 / 304
```

speech recognition.

In conclusion, the system achieved very high recognition rates for both the segmented and particularly the continuous speechreading task considering the high confusability of several words and the difficulty of the task. Most of the recognition errors seem to be due to the identical appearance of some words which prevents perfect visual recognition for this task.

6.7 Comparison With Other Approaches

The performance of the speechreading system described here can be compared with the speechreading system described by Movellan [137] for isolated visual digit recognition. Movellan performed recognition tests using the same training and test protocol as the one described in Sec. 6.6.1. He reported a digit recognition accuracy of 89.58% but stated later that he could not replicate the reported performance and that his replicable results were 87%. These results are slightly lower than the performance of the best system reported here which has a generalisation rate of 90.6%.

Gray et al. [81] have performed a comparison of different image based feature extraction methods and have also used the same training and test protocol. They investigated a low pass filtering approach together with delta information, principal component analysis, and optical flow. The different methods were tested on two different datasets. One dataset consisted of the raw images of the database. For the second dataset, they used the lip tracking results described in Chap. 5 for normalising the images with respect to scale, translation and planar rotation. Their results for both datasets can be found in Table 6.8.

Table 6.8: Speaker independent digit recognition accuracy on the Tulips1 database for the appearance based model (ABM) described here and the systems described by Gray et al. [81]. Gray et al. obtained the results for normalised images by using the lip-tracking results of the ABM reported here.

Feature Extraction	Accuracy		Feature Dimension
ABM	90.6%		11
	Raw Images	Normalised images using ABM	
Low-pass + delta	85.4%	90.6%	300
PCA	67.7%	85.4%	300
Low-pass + optical flow	61.5%	68.8%	290
Optical flow	63.5%	66.7%	140

The performance obtained by the appearance based lip model is significantly better than all of the tested image based methods. When the lip tracking results were used to normalise the images, the performance of the *Low-pass + Delta* representation achieved equivalent results to the lip model. However, these results could only be achieved by using the lip model to track the lips over the image sequences. A further drawback of that image based approach is that the dimension of the feature vector is over 27 times larger than that of the lip model.

Image based approaches for speechreading require the location of the mouth to be known, which is normally not addressed in those approaches and often circumvented by using images which contain the pre-segmented mouth region only. The experiments by Gray et al. have demonstrated that the performance using coarsely segmented images can be enhanced considerably if the exact position of the lips is known. This indicates that all image based approaches may require a lip tracker to normalise the images in order to obtain a high performance.

6.8 Discussion

The described visual feature vector consisting of shape and intensity information presents a novel approach to the representation of visual speech information. The feature vector is of low dimension and has important properties like the invariance to translation, scale, angle, and in the case of the shape parameters also to illumination. The intensity parameters are related to other eigen-decomposition approaches but differ in that the intensity profiles deform with the shape model and therefore represent information about the same object part.

Speaker independent speechreading tests were performed for an isolated digit recognition task on a database of 12 subjects. Most previous speechreading systems have been tested on a smaller number of subjects and were not speaker independent. However, visual features vary considerably across speakers due to individual appearances and mouth movements, which makes speaker independent speechreading very difficult. This is one of the first studies to perform speaker independent speechreading. The system achieved a recognition rate approximately equivalent to the performance achieved by untrained human lipreaders. Audio-visual recognition experiments were performed for different acoustic SNRs which clearly showed that visual information can improve the performance of acoustic speech recognition systems.

The continuous digit recognition experiment represents one of the first speaker independent continuous speechreading tests and consists of the largest number of speakers. For this task, the speechreading system obtained a word accuracy up to 58.5%. About half of the recognition errors were due to deletions which suggests that some HMMs were very similar. The accuracy varied considerably between digits, as a result of the high visual confusability of several French digits. The recognition rate generally increased with the number of states which indicates that the period of quasi-stationary state is similar to the visual sample period and might be even shorter.

The speechreading system was compared on an identical task with several image based approaches and obtained considerably better results than all other methods. Furthermore, the dimension of the feature vector of the best image based method was over 27 times higher than that described here.

The emphasis of this chapter was to describe a speechreading system based on a robust feature extraction method rather than to describe a large vocabulary speechreading system. I believe that the bottleneck of most existing systems developed so far is their feature extraction method. Several approaches have already shown that once visual features are available, their modelling for continuous speech recognition tasks is straightforward.

Although HMMs make certain assumptions about the speech signal that are neither true for the acoustic nor for the visual speech signal, state-of-the-art speech recognition systems are still based on standard HMMs, and the visual speech recognition experiments reported here obtained good results using the same kind of models. It seems however likely that methods that account for the apparently non-stationary nature of visual speech features will lead to higher performances.

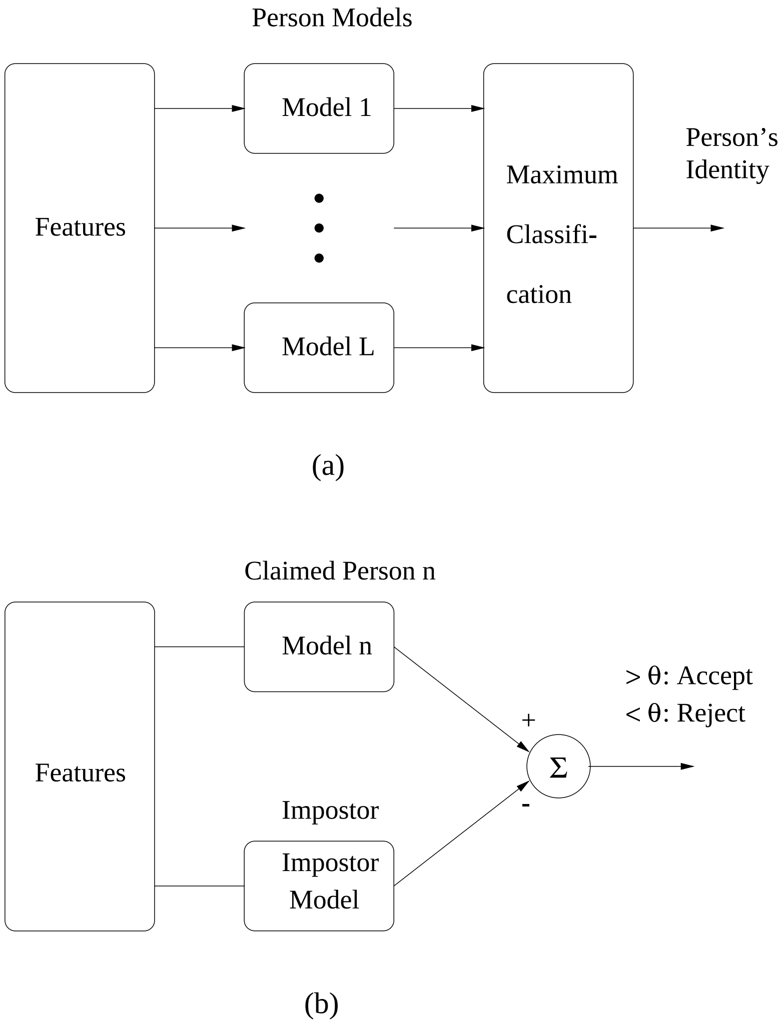
Chapter 7

Visual Speaker Recognition

This chapter describes the proposed person recognition system based on visual spatio-temporal analysis of the talking face. The lip model described in Chapter 5 is used to locate and track the lips of a talking person and to extract speaker dependent information from the motion sequence. Models of speaking persons are built from a training set of these extracted features. Person recognition is performed by extracting the same kind of spatio-temporal features from the test person and by choosing that person, whose model has the highest posterior probability, as the identified person. The method is related to common acoustic speaker recognition systems with the difference that the features are extracted from the visual rather than the acoustic signal. It is also related to common person recognition systems based on facial image analysis. The novelty of this system is that visual spatio-temporal models of talking persons are built and used for person authentication.

The recognition of a person's identity by machine is a difficult problem which has received considerable attention over the last decade. Person recognition can be divided into two tasks: identification and verification. Identification is concerned with determining that person from a closed set, whose features best match the features of the person to identify. Person identification therefore assumes that only enrolled persons will access the system. Verification is concerned with validating the claimed identity of a person from an open set. Therefore, verification also has to be able to reject subjects, referred to as *impostors*, that were not enrolled in the system. Figure 7.1 displays a block diagram of a person identification and a person verification system. Verification is normally performed by comparing the claimed client model with an impostor model. In acoustic speaker recognition, the impostor model is usually a world model, which represents the average model of a large number of speakers, or it is a cohort model, which represents the model of that speaker, closest to the claimed speaker. A decision is made by either accepting or rejecting the person, based on the difference between the likelihoods of the claimed speaker model and the impostor model.

Acoustic speaker recognition can be further divided into text-dependent and text-independent tasks. For the text-dependent task the speech used to train and test the system is constrained to be the same, while for the text-independent task it is unconstrained.



(a)

(b)

Figure 7.1: Block diagram of person identification (a) and person verification (b). In person identification, the person model which best matches the feature sequence is selected as the identified person. In verification, the probability of a claimed person model is compared with that of an impostor model and the person is either rejected or accepted, based on the difference of the probabilities.

The text-independent task can be constrained to consist only of speech from a limited vocabulary which is then called vocabulary-dependent task. The performance of speaker recognition systems is highly dependent on the following factors:

- channel difference between training and test set
- background and channel noise
- speaker variability due to speaking rate, speaking level, voice disguises or cold
- time interval between training and test
- training and test duration

Face recognition performance on the other hand is mainly dependent on the following conditions:

- illumination (type, intensity, direction)
- pose (scale, 3D rotation, translation)
- short time changes (speaking, facial expression)
- long time changes (hairstyle, beard, glasses, makeup, natural changes)

Face recognition tasks can be divided into static and dynamic tasks. Static recognition tasks are based on static images such as photographs on which visual analysis is performed. Most person recognition applications, however, require the system to perform image analysis from an image sequence. This can be performed by selecting from an image sequence that frame which is most suitable for face recognition. A recent survey [33] has shown that very little research has been done in face recognition from image sequences. Kittler et al. [100] have recently proposed to track the lips of a talking person and to use the image with the maximal vertical lip distance for the recognition system. Selecting the optimal image from an image sequences might however require more sophisticated algorithms. Information like the 3D pose, facial expression, illumination might be analysed in addition to more detailed mouth shape information. The method described here aims at using temporal mouth information explicitly for person recognition, therefore avoiding the selection of an appropriate image frame for image analysis.

Whereas the multi-modal nature of speech has received much attention from the speech recognition community, person identification research has mainly been performed by one modality only, with face recognition [160, 178, 33] and speaker recognition [74] being two of these modalities. A difficult problem encountered in visual speech recognition is the modelling by visual speech features, which usually convey speech, speaker, illumination, and pose information [113]. Speaker recognition research has largely ignored the visual modality of speech and has concentrated on the acoustic speech signal to represent a speaker.

A problem encountered in face recognition is that the appearance of a face can change considerably during speech production and due to facial expressions. In particular the mouth is subject to considerable changes but at the same time it is also one of the most distinctive parts of a face [30]. Face recognition research has largely ignored this variability and mainly focused on experiments of static face images with neutral expressions.

The problem of speaker variability in speechreading and the problem of appearance variability in face recognition, have led to the new approach of person recognition described here [112]. The method is based on spatio-temporal analysis of image sequences of the talking face. A deformable shape model is used to track and parameterise the lip boundaries during speech production. A grey-level model based on principal component analysis deforms with the shape model and provides intensity information about the mouth area. Both shape and intensity features serve as features as well as their temporal dependencies, which are used to build models of speakers. Two different modelling techniques are proposed. One is based on HMMs with mixtures of Gaussian distributions and the other is based on Gaussian mixture models (GMMs) of speakers. The HMM can be viewed as the general model which includes the GMM as a specific HMM with only one state. Person recognition for both methods is performed by estimating the posterior probability of each model for having generated the observed sequence of features. This modality for person recognition is proposed to enhance current unimodal systems like face and speaker recognition and two approaches for its combination with acoustic speaker recognition systems are described.

7.1 Previous Approaches

This section reviews some previous approaches for static face recognition, speaker recognition, and audio-visual person recognition.

7.1.1 Face Recognition

Template Matching

Kirby and Sirovich [99] have used a Karhunen-Loéve (K-L) decomposition for representing faces economically, which was extended by Turk and Pentland [175, 176] to the application of face recognition. A face is decomposed into a weighted sum of basis images (eigen-faces) which are obtained by principal component analysis. The weights of the basis images of an unknown image are compared to the weights of known subjects and used for recognition. This method was thereafter used by several other researchers, either directly or after preprocessing for face recognition. The method usually assumes that a face is normalised, for example with respect to eye location, and that it can be compared in vector representation with stored templates. The eigen-face approach has shown to obtain high recognition performance, although the theoretical foundation for this approach is not very clear. One problem of the method is that it is very sensitive to the hairstyle and the background of the face. The background is therefore often de-emphasised by multiplying the face with a two-dimensional Gaussian window, centred on the face. A major objection to the method is that it leads to global, non-topographic representations of the face which is purely based on intensities and which ignores deformation and changes in feature localisation.

The eigen-face approach has been extended in several ways. Eigen-features of the eyes, mouth and nose were included, to make the approach more robust to occlusion and to emphasise the classification process on these features, which are thought to be most discriminant for face recognition [141]. In the same paper, view-based eigen-spaces were proposed, where separate eigen-spaces are constructed for each combination of scale and orientation. In the recognition process, the orientation and scale is first determined by selecting the eigenspace which best describes the input image. Once the proper view and scale is determined, recognition is performed using the eigenvectors of that view-space.

Cotrell et al. [48] used a multilayer perceptron for facial feature extraction and classification. During recognition, face images are projected onto a subspace in the hidden layers of the network, which turned out to be similar to the eigen-face space. Since the network uses the raw grey-level images and no higher level knowledge about the face, the method might be very sensitive to illumination, pose, and facial expressions.

Geometric Features

One of the earliest works in face recognition has been described by Kanade [95] and was based on geometric features. Brunelli and Poggio [30] compared geometric features with templates for face recognition. 35 geometric features were extracted from the face which consisted of measures of the eyebrows, nose, mouth, chin, and face width. A sophisticated version of template matching was used for the eyes, nose and mouth. The template matching method achieved slightly higher results than the geometric feature method, but it was concluded that these results are difficult to generalise. Interestingly, the mouth template had the highest inter-ocular distance of the single templates and also of the whole face template. This suggests that the mouth is the most variable facial part across individuals. But it is obvious that the mouth template is also most sensitive to small changes like facial expressions.

Template and Shape Matching

Craw et al. [49] represented a face by 59 key points and used the coordinates to normalise the shapes before performing eigen-analysis on the grey level images. A similar approach was followed by Lanitis et al. [104, 106, 105]. They used point distribution models (PDMs) to model the contour and salient features of the face. The contour of the face was used to warp the intensities of the test face onto a normal face shape. The normalised intensities were then represented by a weighted sum of basis shapes using the K-L decomposition. In addition to the intensity weights, they also used the weights of the basis shapes and the weights of local basis images centred at each model point for recognition. The method described here is closely related to this approach.

Lades et al. [102] followed an approach where faces are represented by the responses of a set of 2D Gabor filters of different orientation and scale. The filters are placed at each node of a grid of equal distances, which is laid over the face. The distance measure is based on the responses of the Gabor filters and on the deformation of the grid.

Audio-Visual Person Recognition

Person recognition combining face recognition and speaker recognition have been proposed in [29, 28] where face recognition was performed on the static face image with a neutral facial expression and speaker recognition on the acoustic speech signal. The method did not make use of the visual motion of the face during speech production. Classification was based on two classifiers, one for the visual subsystem and one for the acoustic subsystem. The performance of the integrated system outperformed the performance of both individual subsystems.

Facial Motion Analysis

Several approaches towards facial motion analysis have been proposed for expression recognition. Methods based on optical flow computation over the facial expression sequence to estimate human emotion have been described in [124, 185, 184]. Physically based models for facial expression recognition have been proposed in [173, 61, 60]. These methods employ complex models of the head including skin and musculature and attempt to estimate the movements of facial parts from the image sequence. Bartlett et al. [9] used a combination of optical flow, principal component analysis and feature measurements for classifying facial actions. Black and Yacoob [18] described a method based on parametric flow models. Local models of the eyes, eye brows, and mouth which describe affine transformations and curvature are used to track facial movements which are classified by a rule based method of facial actions. All of these facial expression recognition methods were used to estimate human emotion. No attempts have been made to use them for the application of person recognition.

7.1.2 Speaker Recognition

Statistical Measures

Furui [73] described a method for speaker recognition based on long-term statistical measures of the feature vectors, such as the mean and variance. The long term statistical measures were however found to condensate much of the individual spectral characteristics and to lack in discriminating power.

Bimbot et al. [17] proposed an approach based on second-order statistical measures computed from the speech vectors. The mean vector and covariance matrix is computed for each speaker. Speaker recognition is performed based on different distance measures between these speaker statistics.

Vector Quantisation

Li and Wrench [108] have described an approach for speaker recognition based on vector quantisation. The training vectors of each speaker are clustered using vector quantisation to generate a speaker specific codebook. In the recognition stage, the test vectors are compared to the codebooks and the distances are accumulated over the entire utterance. This method allows text dependent and text independent recognition.

Hidden Markov Models

Hidden Markov Speaker models have been used by Rosenberg et al. [159] for text dependent speaker verification. Speaker dependent HMMs are trained for each speaker and are used during the test stage to find the models of that speaker which result in the smallest distance to the sequence of test features. The same models were also used for text independent recognition, were the utterances were first recognised by a speaker independent speech recognition system. Speaker recognition tests were then performed using concatenated HMMs according to the recognised utterances.

The Gaussian Mixture Model (GMM) described in [157] can be viewed as a HMM with only one state with a mixture of Gaussian distributions [148]. The underlying assumption of the method is to represent different acoustic classes by different mixtures and to ignore the temporal information of acoustic events. The method can therefore be used for text independent speaker recognition. In the work described here, both HMMs and GMMs are investigated for visual speaker modelling.

7.2 Overview of the Visual Speaker Recognition System

A block diagram of the two methods for visual speaker recognition is displayed in Fig. 7.2. Similar to the speechreading system described in Chap. 6, the speaker recognition system uses the lip tracking algorithm described in Chap. 5 for lip tracking and visual feature extraction. Person dependent models are trained on the features extracted from a training set. While for the HMMs, the word sequence or transcription of the training sequences needs to be known to train speaker dependent word models, for the GMMs, the transcription is not required. Recognition is performed by lip tracking and feature extraction on the test sequences, followed by maximum posterior probability classification. Similar to the training stage, the sequence of utterances needs to be known for the HMM based method but it is not required for the GMM based system.

7.3 Feature Extraction

Face recognition and facial expression recognition require detailed analysis of the whole face but most facial motion during speech production occurs around the mouth area. The technique for person authentication described here aims to evaluate the discriminative information contained in facial movements during speech production. Only the image region enclosing the mouth is therefore considered for feature extraction. The proposed modality may later be combined with a static face recognition module or with an acoustic speaker recognition module. But for the moment, we are only interested in evaluating the discriminative information contained in the image sequence of the mouth area.

The approach is related to appearance base face recognition methods based on intensity information and shape information [104, 106, 105]. It is assumed that the grey-level distribution around the mouth area and the shape of the inner and outer lip contour contain discriminative information. The novelty of the method described here is in the temporal modelling of the mouth deformation for person recognition. During speech production the

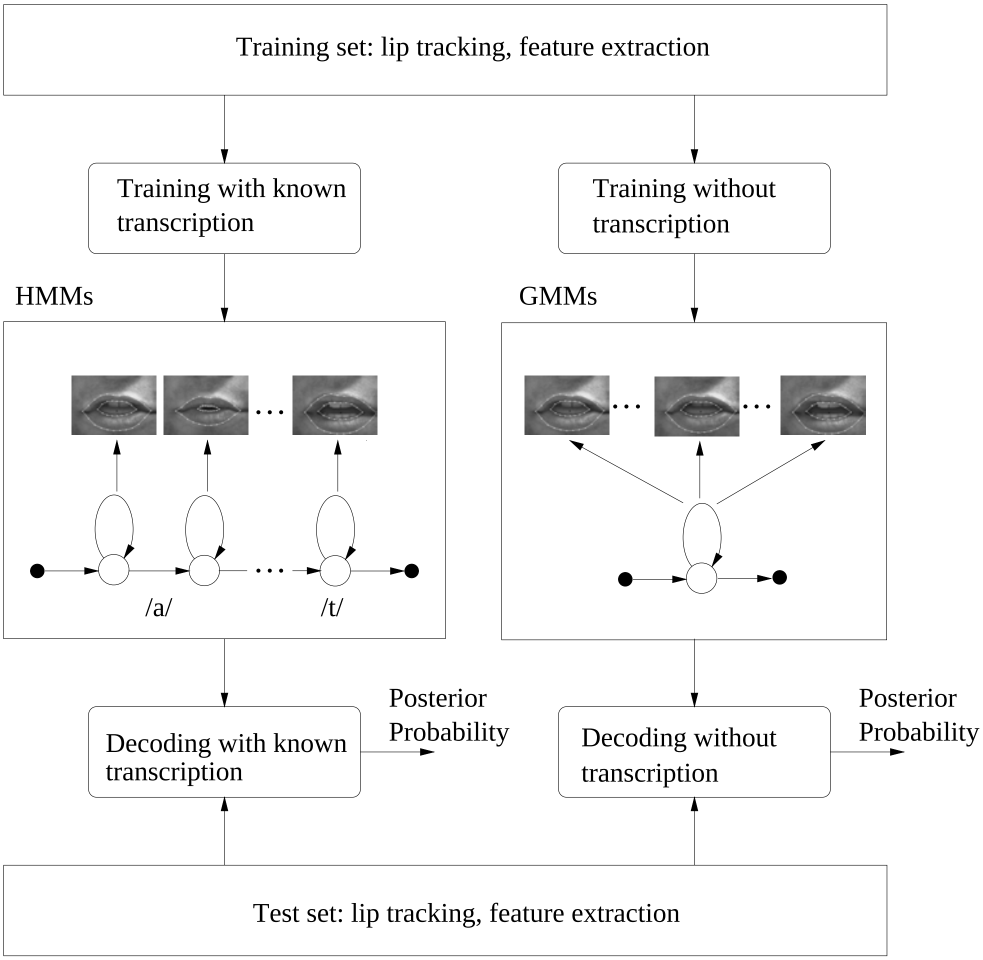

Training set: lip tracking, feature extraction
Training with known transcription
Training without transcription
HMMs
GMMs
/a/
/t/
Posterior Probability
Decoding with known transcription
Posterior Probability
Decoding without transcription
Test set: lip tracking, feature extraction

Figure 7.2: Block diagram of the two proposed methods for visual speaker recognition based on HMMs and GMMs.

lip contours deform and the intensities at the mouth area change due to lip deformation, mouth opening and visibility of teeth and tongue. Besides speech information, the features contain spatial information specific to the speech articulators of a person and temporal information specific to the way a person speaks.

The use of spatio-temporal image information provides additional information which is complementary to the static face image because it models the variability of the face, and complementary to the acoustic signal because it provides additional information responsible for the speech production process. This modality closes the gap between acoustic speaker recognition and visual face recognition by allowing multi-modal person authentication, based on a common signal source, the speech production process. Furthermore, the dependence between the acoustic and visual signal can be used to detect asynchrony between the two signals which might be due to an impostor.

The lip model described in Sec. 5 is used for tracking the lips of a speaking person and for the extraction of visual speech features. The shape and intensity parameters are normalised with respect to their corresponding variances using (6.1) and (6.3), respectively. The normalised shape parameters $\tilde{b}_s$ and intensity parameters $\tilde{b}_i$ serve as features for the speaker recognition system. Assuming correct tracking performance, the features are invariant to scale, translation and rotation. The shape features are also invariant to illumination. Two different models are proposed for representing talking persons: the Gaussian Mixture Speaker Model and the Hidden Markov Speaker Model.

7.4 Hidden Markov Speaker Models

This approach is based on HMMs and is related to modelling methods used in acoustic speaker recognition. A speaker is represented by a set of HMMs modelling the set of speech classes spoken by that person. The speech classes may, for example, be phonemes or words. One model is constructed for each subject and each speech class. The method is based on the following assumptions:

- Word or sub-word HMMs trained on visual features contain speaker dependent information, specific to a person's articulators.
- The sequence of features contains temporal speaker dependent information specific to the way a person talks.
- Any distribution attached to a state can be modelled by a mixture of Gaussian distributions given a large enough number of components.

The method is text-dependent, thus, it requires the test sequence to be known and to only contain utterances which are part of the training set. The scheme may also be applied for text-independent tasks by transcribing the test sequence with a speech recognition system and by using the transcriptions to run the speaker recognition system.

The speaker models are trained according to the maximum likelihood criterion using the Baum-Welch algorithm. Each person i is represented by the set of word models

$$\hat{W}_i = \{w_{i1}, w_{i2}, \ldots, w_{iN}\} \tag{7.1}$$

Identification is based on maximum a posteriori classification using

$$i^* = \arg\max_i P(W_i|\mathbf{O}), \tag{7.2}$$

where W_i represents the string of word models of person i and $\mathbf{O}$ the observation sequence. The prior probabilities of all speakers are assumed to be equal and the prior probabilities of the observation sequences are constant for all speakers which leads to the ML decision rule

$$i^* = \arg\max_i P(\mathbf{O}|W_i). \tag{7.3}$$

The optimal state sequence is obtained using the Viterbi algorithm. For continuous speech, the known sequence of utterance models is simply concatenated and the most likely state sequence is calculated over all models.

7.5 Gaussian Mixture Speaker Models

This approach is based on Gaussian Mixture Models (GMMs) [157] which can be regarded as a HMM with only one state. The motivation behind the use of GMMs is as follows:

- Any distribution can be modelled by a Gaussian mixture model given a large enough number of mixture components.
- Individual Gaussian components represent broad lip shapes and intensity images.
- The model ignores the dependence between observations and words, it is therefore text-independent.
- The temporal information of the feature sequence might contain only little speaker-dependent information.

In comparison to HMMs, this method ignores the temporal information contained in the sequence of observations. It assumes that the mixture components represent an underlying set of possible mouth instances which are characteristic for a certain person. Only one model is built per speaker which represents all utterances of that speaker. This method is therefore text independent and ignores the possible dependence between observations and words. The distribution of a person's feature vectors is modelled by a Gaussian mixture distribution:

$$P(\mathbf{o(t)}|\lambda_i) = \sum_{m=1}^{M} c_{im}\mathcal{N}(\mathbf{o}, \mu_{im}, \Sigma_{im}) \tag{7.4}$$

where c_{im} is the mixture weight for mixture m of speaker i, and $\mathcal{N}(\mathbf{o}, \mu, \Sigma)$ a multivariate Gaussian with mean μ and covariance matrix Σ. The mixture weights satisfy the constraint

$$\sum_{m=1}^{M} c_{im} = 1. \tag{7.5}$$

A speaker i is represented by the parameter set of the GMM by

$$\lambda_i = \{c_{im}, \mu_{im}, \Sigma_{im}\} \qquad \text{with } i = 1, \ldots, M. \tag{7.6}$$

Training is based on the ML criterion and was implemented by the Baum-Welch algorithm. Identification is performed by MAP estimation:

$$\hat{\imath} = \arg\max_i P(\lambda_i|\mathbf{O}) = \arg\max_i \frac{P(\mathbf{O}|\lambda_i)P(\lambda_i)}{P(\mathbf{O})} \tag{7.7}$$

where $\hat{\imath}$ represents the identified person. In analogy to the HMM approach, the prior probabilities $P(\lambda_i)$ are assumed to be equal and $P(\mathbf{O})$ is constant for all speakers which leads to the ML criterion

$$\hat{\imath} = \arg\max_i P(\mathbf{O}|\lambda_i) = \arg\max_i \prod_{t=1}^{T} P(\mathbf{o}(t)|\lambda_i) \tag{7.8}$$

where $P(\mathbf{o}(\mathbf{t})|\lambda_i)$ is given in (7.4).

7.6 Acoustic-Visual Speaker Models

This section addresses the problem of integrated audio-visual speaker models and describes two different techniques. Issues in multi-modal fusion for speaker recognition are closely related to multi-modal fusion for speech recognition, as described in Sec. 6.5.2. Multi-modal systems can be divided into classifiers where the different features are fused at the feature level by concatenating both feature vectors, or by the combination of both modalities at a later stage. The choice of the fusion method is mainly dependent on the assumption of conditional independence of the two modalities and of the availability of synchronised features. If conditional independence is assumed two separate models might be constructed for both modalities. This might also be necessary if different frame rates would complicate the feature fusion at the frame level. Fusion at the feature level is the more general case which avoids making assumptions about conditional independence.

7.6.1 Composite Feature Models

This method assumes that both modalities are conditionally dependent. The motivation and drawbacks for this integration method are similar to those for acoustic-visual speech recognition described in Sec. 6.5.2. For this method, composite feature vectors $\mathbf{O}^{av}$ are constructed by concatenating both feature vectors at the frame level using (6.11).

The training and recognition procedures are performed the same way as for visual features. Acoustic features are represented as MFCC coefficients as described in Sec. 6.5.1, which were extracted at the same frame rate as the visual features to facilitate their combination. The composite feature vectors are used for both the HMM and the GMM speaker models.

7.6.2 Parallel Feature Models

The work described in this section was performed jointly with Pierre Jourlin, Dominic Genoud and Hubert Wassner [92, 93]. A different method for audio-visual speaker modelling is proposed based on separate models for visual and acoustic features. The motivation behind this approach are the following:

- The reliability of the information of both modalities is different and should therefore be weighted accordingly.
- The quasi-stationary events of the acoustic and visual modalities are different and should therefore be modelled by individual HMMs with different topologies.
- Parallel models facilitate the use of different sampling rates for different modalities.

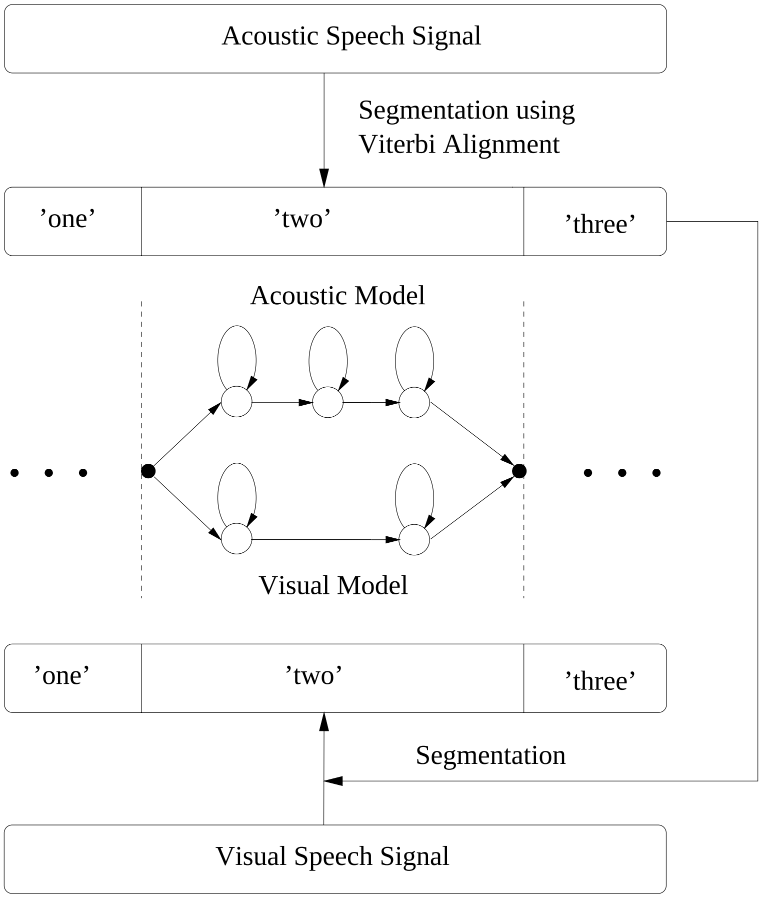


Figure 7.3: Block diagram of the Parallel Feature Model.

A block diagram of the method is shown in Fig. 7.3. A detailed description of the system can be found in [92, 93]. Here, only a brief description of the technique is given.

The acoustic training and test sequences are segmented by a speech recognition system based on the known sequence of words using forced alignment. The acoustic segmentation is then used to segment the visual speech sequences. Separate HMMs are trained on the acoustic and the visual feature vectors. The models can therefore have different HMM structures. For example, since the frame rate of the visual signal is often smaller than for the acoustic signal, a smaller number of states might be chosen for the visual models. The acoustic and visual streams of an acoustic-visual model are constrained to be time synchronous at the beginning and end of the model but can be asynchronous within the model. The likelihood of each modality is estimated for the whole test phrase and acoustic-visual classification is performed on the weighted sum of these likelihoods. The weights for the modalities were estimated on a separate validation set.

7.7 Experiments on the Tulips1 Database

The Tulips1 database was used for acoustic, visual, and audio-visual speaker identification experiments. Both speaker modelling methods were tested, GMMs and HMMs. Audio-visual experiments were performed for composite feature models only.

The utterance set spoken the first time, Set 1, was used as training set and the repetitions, Set 2, as test set. Both sets therefore consisted of only one example per word and speaker. Due to the small training size, experiments for the HMMs were performed using two different methods. In the traditional way, one HMM was trained for each word and each speaker which will be denoted as *Multiple HMMs*. In the second method, a single HMM was trained for all four words, resulting in four training examples per model. This method will be referred to as *Single HMM*. Besides the HMM approach, experiments were also performed for Gaussian mixture models.

The HMMs were constrained to only allow self-loops and sequential transitions between the current and the next state. Issues regarding the number of states and the number of mixture components of the models are similar to those in speech recognition, described in Sec. 6.6. Experiments were therefore performed for HMMs with different numbers of states and different numbers of mixtures.

Initial experiments have shown that the variances in individual mixture components can become very small, particularly in cases with small training data or with a large number of mixture components. Small variances can produce a singularity in the likelihood function of the model which can reduce the recognition performance. A variance limiting scheme was therefore introduced by constraining the variances to be equal or larger than a certain minimum variance value σ^2_{min}. Minimum variance values of 0.01, 0.001 and 0.0001 were evaluated, best performance was obtained with a value of 0.001.

7.7.1 Evaluated Features

The models were evaluated for different kinds of visual and acoustic features and combinations of both. Visual features were obtained from lip tracking experiments as described in Sec. 5 using *Model DC*. This model was chosen in favour of *Model SC* because it describes a larger region of the mouth and is therefore more likely to contain more person dependent information. Visual features consisted of either 10 shape features or 20 intensity features or a combination of both. The features represent the normalised weights obtained from the tracking results.

Acoustic features were represented by 12 MFCC parameters as described in Sec. 6.5.1. The features were extracted at 30 frames per second (*MFCC_30*) to coincide with the frame rate of the visual features and to facilitate their integration. Since this frame rate is considerably smaller than the usual frame rate of 100 frames per second, a second set of MFCC parameters was tested to evaluate the effect of the reduced frame rate. For the second set a frame rate of 100 frames per second was used and a window length of 30 msec, which are denoted as *MFCC_100*. Audio-visual experiments were only performed with the acoustic features *MFCC_30*.

Experiments were also performed by including delta features in either of the feature vector. Delta features might contain speaker dependent information and in the case of intensity features, they might be robust to different illumination.

7.7.2 Multiple HMMs

For the Multiple HMM method, one HMM was constructed per subject and word class, resulting in a total of 48 models. Since only one training example per model was available from the database, a sequential training procedure was used including several parameter tying methods. Re-estimation of the model parameters is based on the models trained in the previous step and is performed with the Baum-Welch algorithm as follows:

- Estimating variances for one global model.
- Reestimating means, mixture weights, and transition probabilities for subject independent word models.
- Reestimating the mean and mixture weights for subject dependent word models.

All HMMs therefore share the same variances and the transition probabilities of any word class are tied for all subjects. Only the means and mixture weights are estimated individually for each class and each subject. Identification is performed with the models corresponding to the spoken word. The likelihood for each speaker is estimated and the speaker with the highest likelihood is chosen as the identified person.

Table 7.1: Person identification rate for multiple HMMs (M-HMM), single HMM (S-HMM) and Gaussian mixture model (GMM), using different kinds of visual and acoustic features.

Task	Shape	Int.	Shape + Int.	MFCC_30	MFCC_100	Shape + Int. + MFCC_30
M-HMM	72.9%	89.6%	91.7%	70.8%	81.3%	93.8%
S-HMM	83.3%	95.8%	97.9%	70.8%	85.4%	97.9%
GMM	81.3%	97.9%	95.8%	60.4%	75.0%	95.8%

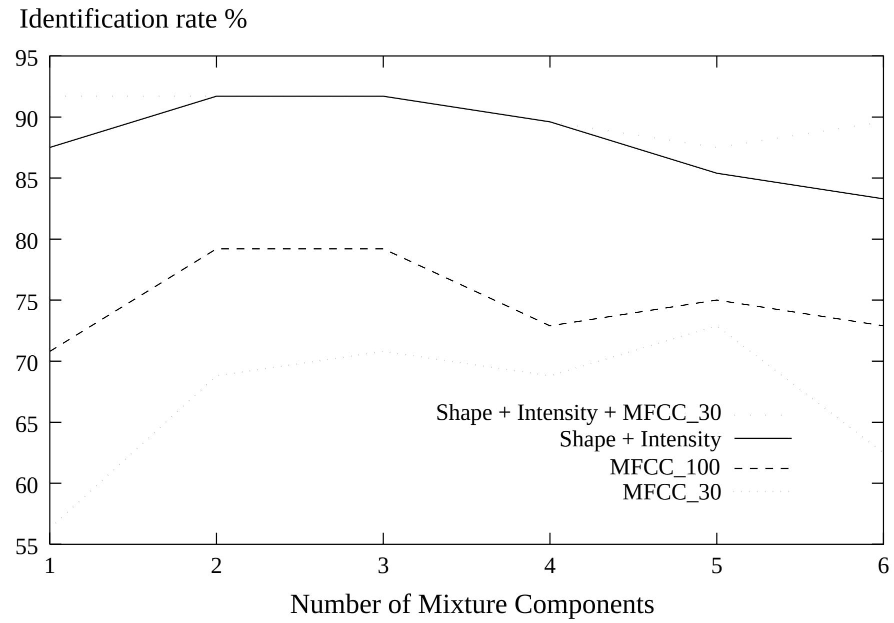


Figure 7.4: Speaker identification performance using multiple HMMs, as a function of the number of mixture components.

Results of the tests are summarised in Table 7.1. Figure 7.4 plots the performance of the different features as a function of the number of mixture components. Best results were obtained by using about 2-5 components, depending on the used feature vectors. The performance for intensity parameters was higher than for shape parameters and the combination of both parameter sets outperformed both single parameter sets. Combining the visual features with the acoustic features led to slightly higher performance than for both single modalities.

Surprisingly, the performance with visual features was substantially higher than with acoustic features. Even the acoustic features *MFCC_100*, with a more than three time higher frame rate performed significantly worse than the visual features. One reason for the poor performance using the acoustic features might be the very small training and test durations used. Table 7.2 displays the average training and test durations for the different models. Typical durations for acoustic speaker recognition systems are about

Table 7.2: Average training and test duration for the different speaker models.

Task	Training time in sec	Test time in sec
M-HMM	0.32	0.32
S-HMM	1.29	0.32
GMM	1.29	0.32

30-90 seconds for training and about 1-10 seconds for testing [157]. Speaker recognition performance is often evaluated as a function of training duration and test duration and typically the performance increases if either of those durations is increased. Compared to acoustic features, the visual features achieved considerably higher performance for short training and test durations.

7.7.3 Single HMM

The Single HMM method was introduced to circumvent the problem of parameter estimation given a very small amount of training data. It can be viewed as a mixture between the HMM and the GMM, in that one multi-state model is used to model all words. One HMM was built per subject, representing all utterances. The motivation behind this approach is to construct one model which represents different word classes by different mixture components. Parameter estimation was not as critical as in text dependent mode and was performed as follows:

- Estimating variances, means, and mixture weights for one global model.
- Reestimating mean, mixture weights, and transition probabilities for a speaker independent model.

Only the variances are therefore tied across all models. The identification experiments are text-independent since only one model is trained per speaker.

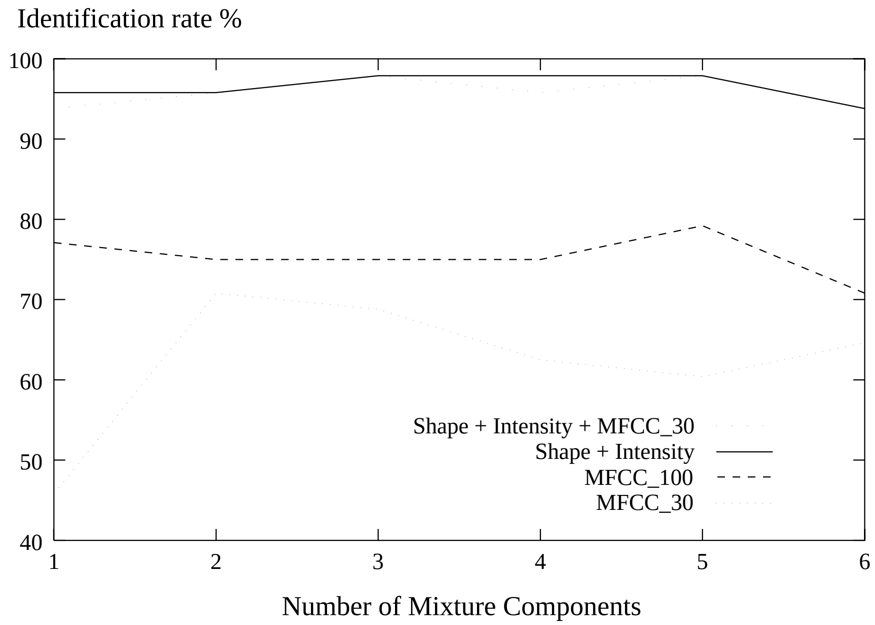


Figure 7.5: Speaker identification performance using a single HMM, as a function of the number of mixture components.

Table 7.1 shows the results for this modelling method and Fig. 7.5 displays the performance for different numbers of mixture components. Best performance was obtained by using both shape and intensity parameters. The performance for this model is higher than for the Multiple HMM method, which might be due to the larger training date available per model. For both HMM approaches, the performance was higher for intensity parameters than for shape parameters. Similar to the multiple HMM method, both acoustic feature representations performed worse than the intensity features and worse than the combination of shape and intensity features.

7.7.4 Gaussian Mixture Speaker Model

For speaker recognition tests by GMMs, a single GMM was trained per speaker. Experiments were performed for different feature representations and for different numbers of mixture components. As only one model was used to represent a certain speaker, the identification tests were text independent.

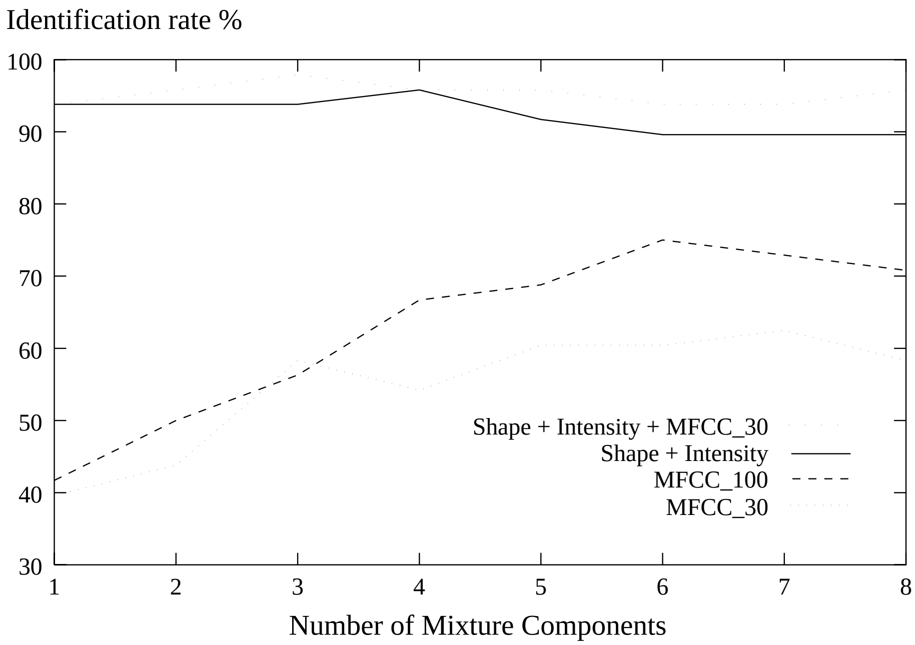


Figure 7.6: Speaker identification performance using GMMs as a function of the number of mixture components.

Identification results for different feature representations can be found in Table 7.1. The identification rate as a function of the number of mixture component is displayed in Fig. 7.6. The GMM method obtained comparable results to both HMM models across all visual feature representations. The performance for acoustic features is slightly lower than for the HMM modelling techniques. Highest results are obtained by the visual intensity features. Similar to the HMM methods, the performance for the visual features is considerably higher then for both acoustic feature sets. Using visual features or acoustic-visual features resulted in very high performance even for very few mixture components.

7.8 Experiments on the M2VTS Database

This section describes person recognition experiments on the M2VTS database using either Gaussian mixture models (GMMs) or hidden Markov models (HMMs) for speaker modelling. Visual features were extracted from lip tracking results as described in Chap. 5. The first three shots of the database were used for training and the 4^{th} and 5^{th} shot for the tests. One shot of a person consists of the sequence of digits from "zero" to "neuf". This resulted in an average training time of 17.8 seconds and an average test time of 5.9 seconds. Visual feature vectors were composed of 14 shape parameters and 10 intensity parameters.

7.8.1 Gaussian Mixture Speaker Models

This model was tested for visual features only. The feature vector was composed of either shape features or intensity features or both combined. Additional experiments were performed by including first order temporal difference parameters (Δ) in the feature vector. One GMM was build per person from the training data and the whole test sequence was used for identification.

Table 7.3: Results for person identification tests on the M2VTS database using GMMs with four mixture components. Results are shown for different feature vectors using either Shot 4 or the more difficult sequences of Shot 5 as test set.

Parameters	Shot 4	Shot 5
shape	56.8%	48.7%
shape + Δ	62.2%	56.8%
intensity	86.5%	73.0%
intensity + Δ	86.5%	73.0%
shape + intensity	78.4%	67.6%
shape + intensity + Δ	89.2%	70.3%

Results for visual speaker identification using GMMs with four mixture components are shown in Tab. 7.3. The intensity features seem to convey more person dependent information than shape features and obtain person identification performance up to 86.5% for Shot 4. Delta information seems to improve the performance for shape features but not for intensity features. The performance using Shot 5 as test shot, which is the most difficult shot to recognise, is considerably lower than the performance for Shot 4.

Figure 7.7 displays the identification rate for different feature vectors as a function

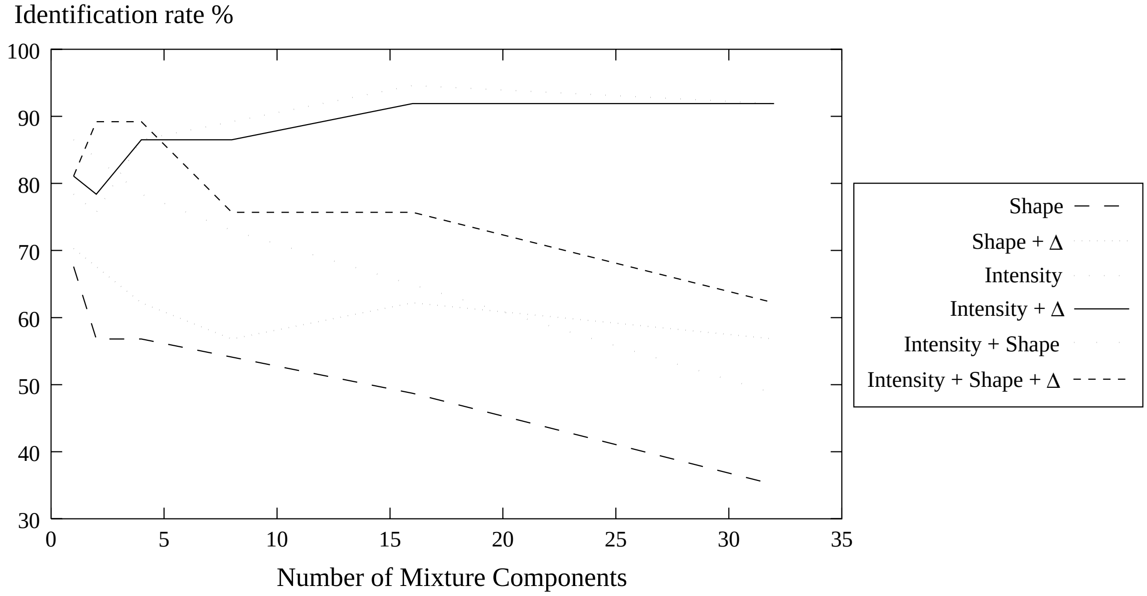


Figure 7.7: Speaker identification performance on the M2VTS database using Shot 4 as test sequence. Results are shown for different feature representations as a function of the number of mixture components.

of the number of mixture components. Results are shown using Shot 4 as test shot. For intensity parameters the performance increases with the number of mixture components and reaches a saturation at about 16 mixtures. Intensity difference parameters do not improve the identification rate. The performance of shape parameters is considerably lower than for intensity parameters, but shape difference parameters improve the identification rate. Using combined shape and intensity parameters outperforms the intensity parameters only for models with a small number of mixture components. The inclusion of delta parameters generally leads to higher identification rates. Overall best results are obtained using intensity features and 16 mixture components, which led to an identification rate of 94.6%.

7.8.2 Hidden Markov Speaker Models

The results for experiments using HMMs are based on the work described in [92, 93]. The visual feature vector was composed of 14 shape parameters, 10 intensity parameters and the scale. One HMM was constructed for each person and each digit and the number of states was chosen to be one or two, depending on the digit length. All HMM state distributions were modelled by a single Gaussian distribution with a diagonal covariance matrix. Identification was based on MAP classification using the whole test sequence. The first three shots were used for training and the 4^{th} and 5^{th} shot for tests.

The performance of the visual modality is substantially lower than for the acoustic modality. However, the combination of both sub-systems improves the identification rate of the acoustic system for Shot 5 from 97.2% to 100%. Jourlin et al. have also performed

Table 7.4: Person identification tests on the M2VTS database using Hidden Markov Speaker Models. Results are shown using either Shot 4 or Shot 5 as test shot (after [92, 93]).

Modality	Shot 4	Shot 5
Visual	82.3%	72.2%
Acoustic	100.0%	97.2%
Acoustic-Visual	100.0%	100.0%

speaker verification experiments using parallel feature models and have shown that the false acceptance rate of the acoustic subsystem could be reduced from 2.3% to 0.5% by the acoustic-visual subsystem. The false rejection score of both systems was 2.8%.

7.9 Discussion

This chapter has described a novel approach for person recognition based on visual speaker models. Speaker models are trained on visual speech features extracted from lip tracking results which represent both shape and intensity information. Gaussian mixture models or hidden Markov models are used to build spatio-temporal models of talking persons. The proposed method bridges the gap between face recognition systems and speaker recognition systems by modelling the temporal visual changes during speech production. It is therefore well suited for combination with these methods to build multi-modal authentication systems.

Visual speaker identification experiments on the Tulips1 database resulted in high performance using either GMMs or HMMs. Surprisingly, the performance of the system using visual features was higher than the performance for the acoustic features. These results suggest that the visual features provide more discriminant information than acoustic features for the given task with very short training and test periods.

Experiments on the M2VTS database demonstrated high performance for both the visual GMM system and the visual HMM system. In contrast to the experiments on the Tulips1 database, the performance of the visual system was substantially lower than for the acoustic system. This could be due to the following factors:

- The visual frame rate, which is about three times lower than the acoustic frame rate.
- Incorrect tracking performance, which was not evaluated on the M2VTS database.
- The reduced visual intensity features used for the M2VTS database.
- The used feature representation might be inappropriate.
- The modelling technique itself, which might be inappropriate.

- The discriminative information of the visual modality might be smaller than for the acoustic modality.

How much each of these factors affects the identification rate of the system is not known. Future research should therefore investigate the influence of these factors on the performance of visual speaker recognition systems.

Acoustic speaker recognition performance typically increases if either the training duration or the test duration is increased above the usual durations. The performance of the visual system, however, turned out to be very high even for very small training and test periods, when tested on the Tulips1 database. This is likely to be due to the shape invariant intensity features, which appear to be less variable during speech production than shape features. Intensity features therefore provide features which are robust to mouth deformation which enables to train and test models with very short feature sequences.

For intensity features, results were generally worse by the inclusion of delta features. This might be due to the small size of the training set or due to the non-discriminative nature of these parameters for speaker recognition. Whereas in speech recognition, delta features have been shown to convey important speech information which is robust across speakers, in speaker recognition, the absolute features seem to contain most discriminative information.

Initial experiments for combining the proposed modality with an acoustic speaker verification system have shown that the false acceptance error could be reduced by a factor of 4.6. These results indicate that both modalities contain complementary information and that the proposed method can improve the performance of acoustic speaker recognition systems.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

The thesis has described methods for the extraction of visual speech features and their modelling for visual speech recognition and visual speaker recognition.

Lip localisation, lip tracking, and feature extraction is based on a deformable model, which is learned by examining the statistics of a representative training set. This ensures that the model can only deform in ways consistent with the training set. The shape is decomposed into a weighted sum of basis shapes using a Karhunen-Loéve (K-L) expansion. Similarly, the intensity around the mouth area is decomposed into a weighted sum of basis shapes using a K-L expansion. The considered intensity areas deform with the shape model to represent shape independent intensity information. This enables the modelling of both shape and intensity with a small number of parameters.

Image search is performed by computing the cost between the model and the image using the Downhill Simplex Method to find a minimum. The cost is defined as the sum of the square errors between the intensity model and the image. The prior probabilities of different shapes and different intensities are assumed to be equal within certain statistical limits. This formulation enables robust image search for various subjects and for uncontrolled lighting conditions. The method was tested on two publicly available databases and was found to achieve state-of-the-art performance. It has been shown to outperform gradient based search methods, and to model both shape deformation and object appearance more specifically than previously reported methods.

Speech features can be recovered from lip tracking results and describe shape and intensity appearance of the mouth. The features are invariant to scale, translation and rotation, the shape features are also invariant to illumination. The sample space of the intensity model deforms with the shape model and therefore represents intensity features which are independent of the lip shape. The intensity features mainly account for speech information like the visibility of teeth, tongue, and protrusion which is complementary to the shape information provided by the shape parameters.

A speechreading system has been presented which models visual speech by hidden Markov models using the extracted visual parameters as features. Using visual information only, the system obtained performance levels on a digit recognition task similar to the performance of humans with no lipreading knowledge. The proposed method outperformed all image based approaches which were compared by Gray et al. [81] for the same recognition task. Gray et al. have also shown that the performance of image based approaches can be considerably improved by the use of the lip model described here.

Visual features vary considerably across speakers due to individual appearances and different mouth movements, which makes speaker independent speechreading very difficult. The experiment on the M2VTS database is one of the largest studies about speaker independent continuous speechreading. Results have shown that the proposed method generalises well to new speakers. The system obtained an accuracy of up to 58.5% for a speaker independent continuous digit recognition task of French digits, using visual features only.

The visual speech recognition experiments suggest that the period of quasi-stationary segments of visual speech features is similar to the used sample periods (33.3 msec and 40 msec) or even shorter. The assumption of pice-wise stationary segments made by HMMs might therefore not be appropriate for visual speech. These results further indicate that higher frame rates for the image acquisition might lead to higher recognition performance. This is also supported by psychological studies which suggest that high speechreading performance by humans is highly correlated with the speed of low-level visual neural processing.

A new modality for person authentication has been introduced based on spatio-temporal modelling of visual speech features. The extracted features provide detailed information about shape, intensity, and their temporal dependencies during speech production. They therefore do not only contain speech information but also specific information about a person's articulators and about the way a person speaks. Two speaker modelling techniques have been proposed, Gaussian mixture speaker models which is a text independent method and hidden Markov speaker models which performs text dependent recognition.

Both methods have been shown to obtained good identification performance on the Tulips1 database using visual speech features only. The identification system achieved 97.9% correct identification on the Tulips1 database of 12 subjects using visual features only, which was considerably higher than the performance using acoustic features. This indicates that for short training and test periods, visual features lead to higher performance than acoustic features. The visual speaker identification system obtained up to 94.6% correct identification on the M2VTS database of 37 subjects.

The authentication method can be considered as an approach between face recognition and speaker recognition and it is proposed for combination with these modalities to build multi-modal verification systems. Two multi-modal architectures were described, the composite feature model and the parallel feature model. Both models were shown to improve the performance of unimodal systems. The benefit of the visual modality has been demonstrated for verification experiments on the M2VTS database, where the integrated system reduced the false acceptance rate of the acoustic subsystem from 2.3% to 0.5%.

8.2 Future Work

There are several possibilities to improve the the performance of lip model with regards to robustness, specificity, and computation time during image search:

- Image search could be made more specific by separating variability due to subject identity and variability due to speech production, for example such as described in [58]. Both sources of variability should be used for locating the mouth in the first images since both the identity and the configuration of the mouth is not known. But during tracking, the parameters accounting for identity could be frozen and only the parameters which account for speech production could be used during image search.

- Shape variability and intensity variability have so far been treated separately although they appear to be clearly related. Particularly the intensity values at the mouth opening are clearly correlated with the shape of the mouth opening. The combined processing of shape and intensity information as described in [58] is therefore likely to represent the model more compactly and to reduce the possibility of illegal shape - intensity parameter combinations.

- The method currently requires an initial estimate of the lip position relatively close to the actual position. This requirement could be relaxed by the use of multi-resolution image search [45]. This consists on a coarse-to-fine search strategy which also uses models of different resolution. The search is initialised at the coarsest resolution and progressively changes to smaller resolutions. Image search is therefore much faster and more robust.

- To increase the robustness and speed of the tracking algorithm it might be advantageous to use a tracking algorithm for predicting future observations. The Kalman filter consists of both, a system model and a measurement model, which makes it appropriate for this application.

- Some applications allow to acquire the image of the mouth of the talking person in a relatively constrained manner. These include the use of head sets with integrated cameras and masks of an air-plane pilot. In most applications, however, the pose of the face is not known and the illumination can be highly variable. To apply the system for a real-world application, where only the coarse position of the person is known, an algorithm with locates and tracks head and mouth are highly desirable. The mouth model can then be combined with a face or head model to make image search more robust and to constrain the deformation and pose of the lips to conform with the other facial parts.

- In order to use the proposed feature extraction and modelling methods in real world applications it needs to be able to deal with 3D rotations of the head. The used images did not include images with large rotations in 3D, but is has been shown in [106] that a similar model of the face is able to recover vertical and horizontal rotation of at least ± 20 degrees. It should therefore also be possible for the lip model to recover such rotations and to provide features which are robust to such variability.

With respect to ASR, it seems inevitable that future system will have to make use of all knowledge sources used by humans, which includes speechreading, to achieve performance levels close to those of humans. Similarly, person authentication systems should make use of several knowledge sources, to improve the reliability of the systems. Although the presented speechreading system and the visual speaker recognition system have been shown to obtained high performances, there are still several issues that require further research particularly with respect to their influence on the performance of the two systems. Among these are:

- The optimal frame rate of the visual features needs to be investigated and their influence on the performance of the system.
- The performance of the lip tracker plays a crucial role in the quality of the visual features. The possibility to evaluate the performance separately would therefore be advantageous.
- So far only simple feature representations have been evaluated which either describe the lip contours or the intensity around the mouth area. Further research is necessary to investigate alternative features, particularly representations which consider temporal properties.
- The appropriateness of HMM as modelling technique needs to be further evaluated, particularly with respect to the assumptions made by them.
- The main benefit of the described visual speech and speaker modelling techniques is their ability to improve the performance of unimodal systems. A central issue is therefore the combination of different modalities, which requires further research.

Bibliography

[1] C. Abry and L. J. Boe. Laws for lips. *Speech Communication*, 5:97–104, 1986.

[2] A. Adjoudani and C. Benoît. Audio-visual speech recognition compared across two architectures. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1563–1566, 1995.

[3] J. B. Allen. How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577, 1994.

[4] ANSI. Methods for calculating the articulation index (ANSI S3.5-1969). American National Standard Institute, New York, 1969.

[5] E. Aronson and S. Rosenblum. Space perception in early infancy: perception within a common auditory-visual space. *Science*, 172:1161–1163, 1971.

[6] B. S. Atal and L. S. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50:637–655, 1971.

[7] E. T. Auer and L. E. Bernstein. Homophenity in speechreading: Effects of phonemic equivalence classes on the structure of the lexicon. In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 169–178. Springer Verlag, Berlin, 1996.

[8] R. Bartels, J. Beatty, and B. Barsky. *An Introduction to Splines for Use in Computer Graphics and Geometry Modelling.* Morgan Kaufmann, 1987.

[9] M. S. Bartlett, P. A. Viola, and T. J. Sejnowski. Classifying facial action. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 823–829. MIT Press, Cambridge, MA, 1996.

[10] A. Baumberg and D. Hogg. Learning flexible models from image sequences. In *Third European Conference on Computer Vision*, volume 1, pages 299–308. Springer Verlag, 1994.

[11] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.

[12] A. P. Benguerei and M. K. Pichora-Fuller. Coarticulation effects in lipreading. *Journal of Speech and Hearing Research*, 25:600–607, 1982.

[13] C. Benoît, T. Guiard-Martigny, B. Le Goff, and A. Adjoudani. Which components of the face do humans and machines best speechread? In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 315–328. Springer Verlag, Berlin, 1996.

[14] C. Benoît, T. Mohammadi, and C. Abry. A set of French visemes for visual speech synthesis. In G. Bailly and Christian Benoît, editors, *Talking Machines: Theories, Models and Designs*, pages 485–504. Elsevier Science Publishers, North-Holland, 1992.

[15] C. Benoît, T. Mohammadi, and S. Kandel. Audio-visual intelligibility of French speech in noise. *Journal of Speech and Hearing*, 37:1191–1203, 1994.

[16] K. W. Berger. *Speechreading: Principles and Methods.* National Education Press, Baltimore, 1972.

[17] F. Bimbot, I. Magrin-Chagnolleau, and L. Mathan. Second-order statistical measure for text-independent speaker identification. *Speech Communication*, 17(1-2):177–192, 1995.

[18] M. J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Fifth International Conference on Computer Vision*, pages 374–381. IEEE Computer Society Press, July 1995.

[19] H. A. Bourlard and N. Morgan. *Connectionist Speech Recognition, A Hybrid Approach.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.

[20] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. In *Proceedings of the International Conference on Spoken Language Processing*, pages 422–425, 1996.

[21] L. Braida. Crossmodal integration in the identification of consonants. *Quarterly Journal of Experimental Psychology*, 43A(3):647–677, 1991.

[22] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving connected letter recognition by lipreading. In *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pages 557–560, Minneapolis, 1993. IEEE.

[23] C. Bregler and Y. Konig. Eigenlips for robust speech recognition. In *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pages 669–672, Adelaide, 1994.

[24] C. Bregler and S. M. Omohundro. Surface learning with applications to lipreading. In J.D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6. Morgan Kaufmann, 1994.

[25] C. Bregler and S. M. Omohundro. Nonlinear manifold learning for visual speech recognition. In *IEEE International Conference on Computer Vision*, pages 494–499. IEEE, Piscataway, NJ, USA, 1995.

[26] N. M. Brooke and A. Q. Summerfield. Analysis, synthesis and perception of visible articulatory movements. *Journal of Phonetics*, 11:63–76, 1983.

[27] N. M. Brooke, M. J. Tomlinson, and R. K. Moore. Automatic speech recognition that includes visual speech cues. In *Proceesings of the Institute of Acoustics*, volume 16, pages 15–22, 1994.

[28] R. Brunelli, D. Falavigna, L. Stringa, and T. Poggio. Automatic person recognition by acoustic and geometric features. *Machine Vision Applications*, 8(5):317–325, 1995.

[29] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):955–966, October 1995.

[30] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, Oct 1993.

[31] U. Bub, M. Hunke, and A. Waibel. Knowing who to listen to in speech recognition: visually guided beamforming. In *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, 1995.

[32] D. Chandramohan and P. L. Silsbee. A multiple deformable template approach for visual speech recognition. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, volume 1, pages 50–53, 1996.

[33] R. Chellappa, C.L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–740, May 1995.

[34] T. Chen, H. P. Graf, and K. Wang. Lip synchronization using speech-assisted video processing. *IEEE Signal Processing Letters*, 2(4):57–59, 1995.

[35] C. C. Chibelushi, F. Deravi, and J. S. D. Mason. Survey of audio-visual speech databases. Technical report, University of Wales Swansea, Swansea, UK, 1996.

[36] G. I. Chiou and J. N. Hwang. Lipreading from color motion video. In *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pages 2158–2161, 1996.

[37] G. I. Chiou. *Active Contour Models for Distinct Feature Tracking and Lipreading*. PhD thesis, University of Washington, 1995.

[38] G. Chollet, J. L. Cochard, A. Constantinescu, and P. Langlais. Swiss French Polyphone and Polyvar : Telephone speech databases to study intra and inter speaker variability. Technical report, IDIAP, Martigny, 1995.

[39] G. Chow and X. Li. Towards a system for automatic facial feature detection. *Pattern Recognition*, 26(12):1739–1755, 1993.

[40] T. Coianiz, L. Torresani, and B. Capril. 2D deformable models for visual speech analysis. In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 391–398. Springer Verlag, Berlin, 1996.

[41] R. Cole, L. Hirschmann, L Atlas, and et al. The challenge of spoken language processing: research directions for the nineties. *IEEE Trans. on Speech and Audio Processing*, 3(1):1–20, 1995.

[42] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam. Use of active shape models for locating structures in medical images. *Image and Vision Computing*, 12:355–365, Jul-Aug 1994.

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

[44] T. F. Cootes, C. J. Taylor, A. Lanitis, D. H. Cooper, and J. Graham. Building and using flexible models incorporating grey-level information. In *Proceedings of the International Conference on Computer Vision*, pages 242–246, 1993.

[45] T. F. Cootes, C. J. Taylor, and A. Lanitis. Active shape models : Evaluation of a multi-resolution method for improving image search. In *Proceedings of the British Machine Vision Conference*, pages 327–336, 1994.

[46] T. F. Cootes and C. J. Taylor. Active shape models - smart snakes. In *Proceedings of the British Machine Vision Conference*, pages 266–275. Springer Verlag, 1992.

[47] P. Cosi, M. Dugatto, F. Ferrero, E. Magno Caldognetto, and K. Vagges. Phonetic recognition by recurrent neural networks working on audio and visual information. *Speech Communication*, 19:245–252, 1996.

[48] G. W. Cottrell and M. K. Fleming. Categorisation of faces using unsupervised feature extraction. In *Proceedings of the International Conference on Neural Networks*, volume 2, pages 65–70, 1990.

[49] I. Craw and P. Cameron. Face recognition by computer. In *Proceedings of the British Machine Vision Conference*, pages 498–507, 1992.

[50] I. Craw, D. Tock, and A. Bennett. Finding face features. In *Proc. European Conference on Computer Vision*, pages 92–96, 1992.

[51] B. Dalton, R. Kaucic, and A. Blake. Automatic speechreading using dynamic contours. In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 373–382. Springer Verlag, Berlin, 1996.

[52] J. R. Deller Jr., J. G. Proakis, and J. H. L. Hansen. *Discrete-time processing of speech signals.* Macmillan Publishing Company, New York, NY, 1993.

[53] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(B):1–38, 1977.

[54] B. Dodd and R. Campbell, editors. *Hearing by Eye: The Psychology of Lip-reading.* Lawrence Erlbaum Associates Ltd., London, 1987.

[55] B. Dodd. The acquisition of lip-reading skills by normally hearing children. In Dodd and Campbell [54], pages 163–175.

[56] P. Duchnowski, M. Hunke, D. Buesching, U. Meier, and A. Waibel. Toward movement-invariant automatic lip-reading and speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 1, pages 109–112. IEEE, Piscataway, NJ, USA, 1995.

[57] P. Duchnowski, U. Meier, and Alexander Waibel. See me, hear me: Integrating automatic speech recognition and lip–reading. In *International Conference on Spoken Language Processing*, 1994.

[58] G. J. Edwards, A. Lanitis, C. J. Taylor, and T. F. Cootes. Statistical models of face images — improving specificity. In *Proceedings of the British Machine Vision Conference*, 1996.

[59] N. P. Erber and C. L. De Filippo. Voice-mouth synthesis of tactual/visual perception of /pa, ba, ma/. *Journal of the Acoustical Society of America*, 64:1015–1019, 1978.

[60] I. A. Essa and A. P. Pentland. Facial expression recognition using a dynamic model and motion energy. In *Proc. 5th Int. Conf. on Computer Vision*, pages 360–367. IEEE Computer Society Press, July 1995.

[61] I. A. Essa and A. Pentland. A vision system for observing and extracting facial action parameters. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 76–83, 1994.

[62] G. Fant. *Acoustic Theory of Speech Production.* Mouton, The Hague, 1960.

[63] E. Kathleen Finn and Alan A. Montgomery. Automatic optically based recognition of speech. *Pattern Recognition Letters*, 8(3):159–164, 1988.

[64] K. Finn. *An Investigation of Visible Lip Information to be used in Automatic Speech Recognition.* PhD thesis, Georgetown University, Washington, D. C., 1986.

[65] C. G. Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11:796–804, 1968.

[66] J. L. Flanagan, A. C. Surendran, and E. E. Jan. Spatially selective sound capture for speech and audio processing. *Speech Communication*, 13:207–222, 1993.

[67] J. L. Flanagan. *Speech Analysis, Synthesis, and Perception*, volume 3 of *Kommunikation und Kybernetik in Einzeldarstellungen.* Springer Verlag, Berlin, 2nd edition, 1972.

[68] H. Fletcher. *Speech and Hearing in Communication.* Krieger, New York, 1953.

[69] N. R. French and J. C. Steinberger. Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America*, 19:90–119, 1947.

[70] V. Fromkin. Lip position in American English vowels. *Language and Speech*, 7:215–225, 1964.

[71] O. Fujimura. Bilabial stops and nasal consonants: A motion-picture study and its acoustical implications. *Journal of Speech and Hearing Research*, 4:233–247, 1961.

[72] K. Fukunaga. *Introduction to statistical pattern recognition.* Academic Press Ltd., London, 2 edition, 1990.

[73] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-29(2):254–272, 1981.

[74] S. Furui. An overview of speaker recognition technology. In *Proceedings of the ESCA Workshop on Speaker Recognition, Identification, and Verification*, pages 1–9, 1994.

[75] A. Gelb, editor. *Applied Optimal Estimation.* MIT Press, Cambridge, MA, 1974.

[76] A. J. Goldschen, O. N. Garcia, and E. Petajan. Continuous optical automatic speech recognition by lipreading. In *28th Annual Asilomar Conference on Signals, Systems, and Computers*, 1994.

[77] A. J. Goldschen. *Continuous Automatic Speech Recognition by Lipreading.* PhD thesis, George Washington University, Washington, D. C., 1993.

[78] Y. Gong. Speech recognition in noisy environments: A survey. *Speech Communication*, 16:261–291, 1995.

[79] K. W. Grant, L. D. Braida, and R. J. Renn. Single-band amplitude envelope cues as an aid to speechreading. *Quarterly Journal of Experimental Psychology*, 43A:621–645, 1991.

[80] K. W. Grant and L. D. Braida. Evaluating the articulation index for auditory-visual input. *Journal of the Acoustical Society of America*, 89(6):2952–2960, 1991.

[81] M. S. Gray, J. R. Movellan, and T. J. Sejnowski. Dynamic features for visual speechreading: A systematic comparison. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, Cambridge, MA, 1997.

[82] K. P. Green and J. L. Miller. On the role of visual rate information in phonetic perception. *Perception & Psychophysics*, 38(3):269–276, 1985.

[83] W. J. Hardcastle. *Physiology of Speech Production.* Academic Press, New York, NY, 1976.

[84] J. Haslam, C. J. Taylor, and T. F. Cootes. A probabilistic fitness measure for deformable template models. In *Proceedings of the British Machine Vision Conference*, pages 33–42, 94.

[85] M. E. Hennecke, K. V. Prasad, and D. G. Stork. Using deformable templates to infer visual speech dynamics. In *28th Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, November 1994.

[86] M. E. Hennecke, D. G. Stork, and K. Venkatesh Prasad. Visionary speech: Looking ahead to practical speechreading systems. In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 331–350. Springer Verlag, Berlin, 1996.

[87] A. Hill and C. J. Taylor. Automatic landmark generation for point distribution models. In *Proceedings of the British Machine Vision Conference*, pages 429–438, 1994.

[88] S. Horbelt and J. L. Dugelay. Active contours for lipreading - combining snakes with templates. In *GRETSI Symposium on Signal and Image Processing*, pages 717–720, 1995.

[89] B. K. P. Horn. *Robot Vision.* McGraw-Hill, New York, 1986.

[90] C. L. Huang and C. W. Chen. Human facial feature extraction for face interpretation and recognition. *Pattern Recognition*, 25(12):1435–1444, Dec 1992.

[91] P. L. Jackson, A. A. Montgomery, and C. A. Binnie. Perceptual dimensions underlying lipreading performance. *Journal of Speech and Hearing Research*, 19:796–812, 1976.

[92] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication*, Lecture Notes in Computer Science, pages 319–326. Springer Verlag, 1997.

[93] P Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. *Pattern Recognition Letters*, 1997. to appear.

[94] P. Jourlin. Handling disynchronization phenomenon with HMM in connected speech. In *Proceedings of the 8th European Signal Processing Conference (Eusipco'96)*, volume 1, pages 133–136, 1996.

[95] T. Kanade. *Computer Recognition of Human Faces.* Birkhaeuser, Basel and Stuttgart, 1977.

[96] M. Kass, A. Witkin, and Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, pages 321–331, 1988.

[97] R. Kaucic, B. Dalton, and A. Blake. Real-time lip tracking for audio-visual speech recognition applications. In *Proceedings of the European Conference on Computer Vision*, volume 2, pages 376–386. Cambridge, 1996.

[98] D. H. Kelly. Motion and vision II. stabilized spatio-temporal threshold surface. *Journal of the Optical Society of America*, 69(10):1340–1349, 1979.

[99] M. Kirby and L. Sirovich. Application of the Karhunen-Loéve procedure for the characterisation of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:103–108, 1990.

[100] J. Kittler, Y. P. Li, J. Matas, and M. U. Ramos Sanchez. Lip-shape dependent face verification. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication*, Lecture Notes in Computer Science, pages 61–68. Springer Verlag, 1997.

[101] P. B. Kricos. Differences in visual intelligibility across talkers. In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 43–53. Springer Verlag, Berlin, 1996.

[102] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. v.d. Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, Mar 1993.

[103] M. T. Lallouache. *Un Poste Visage-Parole Couleur. Acquisition et Traitement Automatique des Contours des lèvres.* PhD thesis, Institut National Polytechnique de Grenoble, 1991.

[104] A. Lanitis, C. J. Taylor, and T. F. Cootes. An automatic face identification system using flexible apperance models. In *Proceedings of the British Machine Vision Conference*, pages 65–74, 1994.

[105] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic face identification system using flexible appearance models. *Image and Vision Computing*, 13:393–401, Jun 1995.

[106] A. Lanitis, C. J. Taylor, and T. F. Cootes. A unified approach to coding and interpreting face images. In *Proc. 5th Int. Conf. on Computer Vision*, pages 368–373. IEEE Computer Society Press, July 1995.

[107] S. A. Lesner and P. B. Kricos. Visual vowel and diphthong perception across speakers. *Journal of the Academy of Rehabilitative Audiology*, 14:252–258, 1981.

[108] K. P. Li and E. H. Wrench Jr. An approach to text-independent speaker recognition with short utterances. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 555–558, 1983.

[109] J. Luettin, N. A. Thacker, and S. W. Beet. Active shape models for visual speech feature extraction. In D. G. Storck and M. E. Hennecke (editors), editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 383–390. Springer Verlag, Berlin, 1996.

[110] J. Luettin, N. A. Thacker, and S. W. Beet. Learning to recognise talking faces. In *Proceedings of the International Conference on Pattern Recognition (ICPR'96)*, volume IV, pages 55–59. IAPR, 1996.

[111] J. Luettin, N. A. Thacker, and S. W. Beet. Locating and tracking facial speech features. In *Proceedings of the International Conference on Pattern Recognition (ICPR'96)*, volume I, pages 652–656. IAPR, 1996.

[112] J. Luettin, N. A. Thacker, and S. W. Beet. Speaker identification by lipreading. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, volume 1, pages 62–65, 1996.

[113] J. Luettin, N. A. Thacker, and S. W. Beet. Speechreading using shape and intensity information. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, volume 1, pages 58–61, 1996.

[114] J. Luettin, N. A. Thacker, and S. W. Beet. Statistical lip modelling for visual speech recognition. In *Proceedings of the 8th European Signal Processing Conference (Eusipco'96)*, volume I, pages 137–140, 1996.

[115] J. Luettin, N. A. Thacker, and S. W. Beet. Visual speech recognition using active shape models and hidden Markov models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, volume 2, pages 817–820, 1996.

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

[117] J. Luettin, M. Vogt, and C. Bregler. Machine recognition and applications. In D. G. Storck and M. E. Hennecke (editors), editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 549–555. Springer Verlag, Berlin, 1996.

[118] A. MacLeod and A. Q. Summerfield. Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, 21:131–141, 1987.

[119] E. Magno Caldognetto, K. Vagges, N. A. Borghese, and G. Ferrigno. Automatic analysis of lips and jaw kinematics in VCV sequences. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 2, pages 453–456, 1989.

[120] M. W. Mak and W. G. Allen. Lip-motion analysis for speech segmentation in noise. *Speech Communication*, 14(3):279–296, 1994.

[121] M. W. Mak and W. G. Allen. Lip-tracking system based on morphological processing and block matching techniques. *Signal Processing: Image Communication*, 6(4):335–348, Aug 1994.

[122] K. Mase and A. Pentland. Lipreading: Automatic visual recognition of spoken words. In *Optical Society of America Topical Meeting on Machine Vision*, pages 1565–1570, 1989.

[123] K. Mase and A. Pentland. Automatic lipreading by optical flow analysis. *Systems and Computers in Japan*, 22(6), 1991.

[124] K. Mase and A. Pentland. Recognition of facial expression from optical flow. *IEICE Transactions*, E 74(10):3474–3483, 1991.

[125] D. W. Massaro and M. M. Cohen. Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 9:753–771, 1983.

[126] D. W. Massaro. Speech perception by ear and eye. In Dodd and Campbell [54], pages 53–83.

[127] D. W. Massaro. Bimodal speech perception: A progress report. In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 79–103. Springer Verlag, Berlin, 1996.

[128] I. Matthews, J. A. Bangham, and S. Cox. Audiovisual speech recognition using multiscale nonlinear image decomposition. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, volume 1, pages 38–41, 1996.

[129] M. McGrath, A. Q. Summerfield, and N. M. Brook. Roles of lips and teeth in lipreading vowels. In *Proceedings of the Institute of Acoustics*, volume 6, pages 401–408, 1984.

[130] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[131] A. E. Mills. The development of phonology in the blind child. In Dodd and Campbell [54], pages 145–161.

[132] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *IEEE International Conference on Computer Vision*, pages 786–793. IEEE, Piscataway, NJ, USA, 1995.

[133] A. A. Montgomery and P. L. Jackson. Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America*, 73(6):2134–2144, 1983.

[134] S. Morishima, K. Aizawa, and H. Harashima. An intelligent facial image coding driven by speech and phoneme. In *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pages 1795–1798, 1989.

[135] J. R. Movellan and G. Chadderdon. Channel separability in the audio-visual integration of speech: A Bayesian approach. In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 473–488. Springer Verlag, Berlin, 1996.

[136] J. R. Movellan and P. Mineiro. Modularity and catastrophic fusion: A Bayesian approach with applications to audio visual speech recognition. Technical Report 97.01, Department of Cognitive Science, UCSD, San Diego, CA, 1997.

[137] J. R. Movellan. Visual speech recognition with stochastic networks. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7. MIT press, Cambridge, 1995.

[138] H. Murase and R. Sakai. Moving object recognition in eigenspace representation: Gait analysis and lip reading. *Pattern Recognition Letters*, 17(2):155–162, February 1996.

[139] J. A. Nelder and R. Mead. A simplex method for function optimization. *Computing Journal*, 7(4):308–313, 65.

[140] F. I. Parke. A model for human faces that allows speech synchronized animation. *Journal of Computers and Graphics*, 1(1):1–4, 1975.

[141] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 84–91. IEEE, Los Alamitos, CA, USA, 1994.

[142] E. D. Petajan, N. M. Brooke, B. J. Bischoff, and D. A. Bodoff. An improved automatic lipreading system to enhance speech recognition. In E. Soloway, D. Frye, and S.B. Sheppard, editors, *Proc. Human Factors in Computing Systems*, pages 19–25. ACM, 1988.

[143] E. D. Petajan. *Automatic Lipreading to Enhance Speech Recognition.* PhD thesis, University of Illinois at Urbana-Champain, 1984.

[144] E. D. Petajan. Automatic lipreading to enhance speech recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 40–47, 1985.

[145] G. Peterson and H. Barney. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24:175–184, 1952.

[146] S. Pigeon and L. Vandendorpe. The M2VTS multimodal face database. In *Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication*, Lecture Notes in Computer Science. Springer Verlag, 1997.

[147] T. Poggio and V. Torre. Ill-posed problems and regularization analysis in early vision. In *Proceedings of the DARPA Image Understanding Workshop*, pages 257–263, 1984.

[148] A. B. Poritz. Linear predictive hidden Markov models. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1291–1294, 1982.

[149] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C.* Cambridge University Press, second edition, 1992.

[150] L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition.* Prentice-Hall, Englewood Cliffs, NJ, 1993.

[151] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals.* Prentice Hall, Englewood Cliffs, New York, 1987.

[152] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[153] R. R. Rao and R. M. Mersereau. Lip modelling for visual speech recognition. In *28th Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, 1994.

[154] D. Reisberg, J. McLean, and A. Goldfield. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In Dodd and Campbell [54], pages 97–113.

[155] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Context free attentional operators: the generalized symmetry transform. *International Journal of Computer Vision*, 14(2):119–130, 1995.

[156] D. Reisfeld and Y. Yeshurun. Robust detection of facial features by generalized symmetry. In *Proceedings of the 11th IAPR International Conference on Pattern Recognition*, pages 117–120, 1992.

[157] A. Reynolds and C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, 3:72–83, 1995.

[158] J. Robert-Ribes, M. Piquemal, J.-L. Schwartz, and P. Escudier. Exploiting the sensor fusion architectures and stimuli complementarity in AV speech recognition.

In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 193–210. Springer Verlag, Berlin, 1996.

[159] A. E. Rosenberg, C. H. Lee, and S. Gokoen. Connected word talker verification using whole word hidden Markov model. In *ICASSP-91*, pages 381–384, 1991.

[160] A. Samal and P. Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition*, 25(1):65–77, 1992.

[161] M. U. Ramos Sanchez, J. Matas, and J. Kittler. Statistical chromaticity models for lip tracking with B-splines. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication*, Lecture Notes in Computer Science, pages 69–76. Springer Verlag, 1997.

[162] J. S. Scheinberg. Analysis of speechreading cues using an interleaved technique. *Journal of Communication Disorders*, 13:489–492, 1980.

[163] P. L. Silsbee and A. C. Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):337–351, 1996.

[164] P. L. Silsbee. *Computer Lipreading for Improved Accuracy in Automatic Speech Recognition*. PhD thesis, University of Texas, 1993.

[165] P. D. Souza, T. F. Cootes, C. J. Taylor, and E. C. Di-Mauro. Non-linear point distribution modelling using a multi-layer perceptron. In *Proceedings of the British Machine Vision Conference*, volume 1, pages 107–116, 1995.

[166] K. N. Stevens and A. S. House. Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America*, 27:484–493, 1955.

[167] D. G. Stork and M. E. Hennecke, editors. *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*. Springer Verlag, Berlin, 1996.

[168] D. G. Stork, G. J. Wolff, and E. P. Levine. Neural network lipreading system for improved speech recognition. In *Proceedings International Joint Conference on Neural Networks*, volume 2, pages 289–295, 1992.

[169] W.H. Sumby and I. Pollak. Visual contributions to speech intelligibility in noise. *J. Acoustical Society of America*, 26:212–215, 1954.

[170] A. Q. Summerfield. Audio-visual speech perception, lipreading and artificial simulation. In Lutman M. E. and M. P. Haggard, editors, *Hearing Science and Hearing Disorders*, pages 131–182. Academic Press, New York, 1983.

[171] A. Q. Summerfield. Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd and Campbell [54], pages 3–51.

[172] A. Q. Summerfield. Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London, Series B*, 335:71–78, 1992.

[173] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6), 1993.

[174] M. J. Tomlinson, M. J. Russell, and N. M. Brooke. Integrating audio and visual information to provide highly robust speech recognition. In *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, volume 2, pages 821–824, 1996.

[175] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[176] M. Turk and A. Pentland. Face recognition using eigenfaces. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, June 1991.

[177] University of Sheffield. *TINA Algorims, Programmer, and User Guide*, 1996.

[178] D. Valentin, H. Abdi, and G. W. Cottrell. Connectionist models of face processing: A survey. *Pattern recognition*, 27(9):1209–1230, 1994.

[179] A. J. Viterbi. Error bounds for convolutional codes and an asymtotically optimum decoding algorithm. *IEEE Trans. on Information Theory*, 13(2):260–269, 1967.

[180] M. Vogt. Fast matching of a dynamic lip model to color video sequences under regular illumination conditions. In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 399–408. Springer Verlag, Berlin, 1996.

[181] B. E. Walden, R. A. Prosek, and A. Montgomery. Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, 20:130–145, 1977.

[182] G. J. Wolff, K. V. Prasad, D. G. Stork, and M. E. Hennecke. Lipreading by neural networks: Visual preprocessing, learning and sensory integration. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6. MIT Press, 1994.

[183] J. Wu, S. Tamura, H. Mitsumoto, H. Kawai, K. Kurosu, and K. Okazaki. Neural network vowel recognition jointly using voice features and mouth shape image. *Pattern Recognition*, 24(10):921–927, 1991.

[184] Y. Yacoob and L. S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642, 1996.

[185] Y. Yacoob and L. Davis. Computing spatio-temporal representations of human faces. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 70–75. IEEE, Los Alamitos, CA, USA, 1994.

[186] G. Yang and T. S. Huang. Human face detection in a complex background. *Pattern Recognition*, 27(1):53–63, 1994.

[187] S. J. Young, P. C. Woodland, and P. C. Byrne. *HTK Version 1.5: User, Reference and Programmer Manual.* Entropic Research Laboratories, Washington, DC, 1993.

[188] K. C Yow and R. Cipolla. Towards and automatic human face localization system. In *Proceedings of the British Machine Vision Conference*, pages 701–710, 1995.

[189] B. P. Yuhas, M. H. Goldstein, T. J. Sejnowski, and R. E. Jenkins. Neural network models of sensory integration for improved vowel recognition. *Proc. IEEE,* 78(10):1658–1668, October 1990.

[190] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski. Integration of acoustic and visual speech signals using neural nets. *IEEE Commun. Mag.*, pages 65–71, November 1989.

[191] B. P. Yuhas and M. H. Goldstein. Comparing human and neural network lip readers. *J. Acoustical Society of America*, 90(1):598–600, 1991.

[192] A. L. Yuille, D. S. Cohen, and P. W. Hallinan. Feature extraction from faces using deformable templates. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, 1989.

[193] A. L. Yuille, P. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8:99–112, 1992.

Index