Exploiting The Bimodality Of Speech In The Cocktail Party Problem

Thesis submitted to Cardiff University in candidature for the degree of Doctor of Philosophy.

Andrew James Aubrey

Centre of Digital Signal Processing
Cardiff University
2008

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ..... (candidate) Date 26/08/2008

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed ..... (candidate) Date 26/08/2008

STATEMENT 2

This thesis is the result of my own investigation, except where otherwise stated. Other sources are acknowledged by giving explicit reference.

Signed ..... (candidate) Date 26/08/2008

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organizations.

Signed ..... (candidate) Date 26/08/2008

STATEMENT 4

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, after expiry of a bar on access approved by the Graduate Development Committee.

Signed ..... (candidate) Date 26/08/2008

ABSTRACT

The cocktail party problem is one of following a conversation in a crowded room where there are many competing sound sources, such as the voices of other speakers or music. To address this problem using computers, digital signal processing solutions commonly use blind source separation (BSS) which aims to separate all the original sources (voices) from the mixture simultaneously. Traditionally, BSS methods have relied on information derived from the mixture of sources to separate the mixture into its constituent elements. However, the human auditory system is well adapted to handle the cocktail party scenario, using both auditory and visual information to follow (or hold) a conversation in a such an environment.

This thesis focuses on using visual information of the speakers in a cocktail party like scenario to aid in improving the performance of BSS. There are several useful applications of such technology, for example: a pre-processing step for a speech recognition system, teleconferencing or security surveillance.

The visual information used in this thesis is derived from the speaker's mouth region, as it is the most visible component of speech production. Initial research presented in this thesis considers a joint statistical model of audio and visual features, which is used to assist in controlling the convergence behaviour of a BSS algorithm. The results of using

the statistical models are compared to using the raw audio information alone and it is shown that the inclusion of visual information greatly improves its convergence behaviour.

Further research focuses on using the speaker's mouth region to identify periods of time when the speaker is silent through the development of a visual voice activity detector (V-VAD) (i.e. voice activity detection using visual information alone). This information can be used in many different ways to simplify the BSS process.

To this end, two novel V-VADs were developed and tested within a BSS framework, which result in significantly improved intelligibility of the separated source associated with the V-VAD output. Thus the research presented in this thesis confirms the viability of using visual information to improve solutions to the cocktail party problem.

ACKNOWLEDGEMENTS

Firstly I must thank both of my supervisors Dr Yulia Hicks and Prof Jonathon Chambers, without whom I would not have been able to undertake such a challenging and interesting subject. Their guidance and support over the last three years has been invaluable. They have my utmost respect, professionally and personally.

The funding for my PhD was courtesy of the Engineering and Physical Sciences Research Council (EPSRC) and Qinetiq, I am very grateful for their financial support. I would like to particularly thank Prof John McWhirter not only for his role with the funding from Qinetiq but also for his support and discussions over the three years.

To my colleagues at the Cardiff Centre of Digital Signal Processing research group, both past and present. I wont name you all for fear of missing someone out but you know who you are. I thank you for your advice and friendship, and helping make the PhD experience that little bit more enjoyable. Thanks to Paul Farrugia and Denley Slade for keeping the audio-visual equipment up and running. I also thank Prof Christian Jutten for allowing me the opportunity to spend several weeks in the lab at Grenoble Institute of Technology, and to Dr Bertrand Rivet for their collaboration whilst I was there.

I would also like to express my gratitude to Dr Darren Cosker for his general advice and helping me to understand some of the techniques

I had to become familiar with, and for the many hours of watching the highs and lows of the Welsh rugby team. To all my friends, both old and new, thank you for your support.

Finally thank you to my parents and sister for their continuing support. And to Kitty, for her advice, encouragement and for being there throughout.

LIST OF ACRONYMS

AAM	Active Appearance Models
ASA	Auditory Scene Analysis
ASM	Active Shape Model
AV	Audio-Visual
AV-BSS	Audio-Visual Blind Source Separation
AVSR	Audio-Visual Speech Recognition
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CDWT	Complex Discrete Wavelet Transform
CSD	Correct Silence Detection
CSSD	Cumulative Subband Squared Difference
CWT	Continuous Wavelet Transform
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DOA	Direction Of Arrival

DSM	Downhill Simplex Minimisation
DWT	Discrete Wavelet Transform
EAP	Exclusive Activity Period
EEG	Electroencephalography
EM	Expectation Maximisation
FIR	Finite Impulse Response
fps	Frames Per Second
FSD	False Silence Detection
FT	Fourier Transform
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ICA	Independent Component Analysis
IDFT	Inverse Discrete Fourier Transform
JADE	Joint Approximate Diagonalisation of Eigenmatrices
LS	Least Squares
ME	Motion Estimation
MFCC	Mel-Frequency Cepstral Coefficients
MIMO	Multiple-Input Multiple-Output
MLE	Maximum Likelihood Estimation
MPEG	Moving Picture Experts Group

NG	Natural Gradient
PCA	Principal Component Analysis
pdf	Probability Density Function
PDM	Point Distribution Model
ROC	Receiver Operating Characteristics
SD	Squared Difference
SE	Speech Enhancement
SIMO	Single-Input Multiple-Output
SIR	Signal to Interference Ratio
SNR	Signal to Noise Ratio
SOBI	Second-Order Blind Identification
SOS	Second Order Statistics
SSD	Subband Squared Difference
STFT	Short Time Fourier Transform
TDOA	Time Difference of Arrival
VAD	Voice Activity Detection
V-VAD	Visual-Voice Activity Detection
WT	Wavelet Transform

LIST OF SYMBOLS

x	Scalar quantity
$\mathbf{x}$	Vector quantity
$\mathbf{X}$	Matrix quantity
$\bar{\mathbf{x}}$	Mean vector
$\hat{\mathbf{x}}$	Estimate of original quantity $\mathbf{x}$
$(.)^T$	Transpose operator
$(.)^H$	Hermitian transpose operator
$(.)^{-1}$	Matrix inverse
$(.)^*$	Complex conjugate operator
$\|.\|$	Matrix determinant
$\|\|.\|\|_F$	Frobenius Norm
$\mathbf{1}$	Vector of Ones
$\mathbf{I}$	Identity matrix

List of Figures

CONTENTS

Chapter 1

INTRODUCTION

When someone walks into a crowded room they are surrounded by audio signals, be they speech, laughter, music, or even the ticking of a clock, but holding a conversation in such a noisy environment is something humans are able to do with great ease (providing the intruding sounds aren't overwhelming).

The study of this phenomenon began in the 1950's with the work of Colin Cherry [21, 22]. Cherry defined what is now known as *the cocktail party* problem, i.e. how to select one source of auditory input amongst the many competing sources. Indeed, the human auditory system is well adapted for such situations by utilizing both audio and visual information, but in the domain of digital signal processing this problem has yet to be resolved.

During the 1980's and 90's two different signal processing approaches emerged to solve this problem. Computational Auditory Scene Analysis (CASA), which stemmed from Albert Bregman's work on *auditory scene analysis* (ASA) [11] that described how the human auditory system organizes and processes complex mixtures of sound. The other, and focus of this thesis is Blind Source Separation (BSS). The two methods differ in several aspects. Typically CASA methods [12, 39, 123] aim to segregate a target speech signal from the background noise, whereas

BSS methods aim to separate all signals from the mixture. It is also worth noting that because BSS has received a great deal of attention during the last ten to fifteen years, it is slightly more mature than CASA with regards to dealing with convolutive (real world) mixtures of speech.

A solution to the cocktail party problem found using either approach would have many applications, including teleconferencing, security surveillance or as a pre-processing step for speech recognition.

1.1 Blind Source Separation

Blind source separation (BSS) is a process by which individual sources can be separated from measurements containing a mixture of sources. The term "blind" refers to the fact that little or no prior information about the sources is known. However, some weak assumptions regarding the nature of the sources must be made. BSS has applications in many fields of signal processing, ranging from speech processing to bio-medical and financial time series analysis [56, 64] and because of this wide variety of applications it is a well researched area of signal processing.

The origins of BSS date back to early work by Herault and Jutten [50] and since then a wide variety of BSS algorithms have been proposed. They have been developed to cover many applications but the signal mixtures, and hence the algorithms to solve them, can generally be classified into one of three categories: Instantaneous, Anechoic and Echoic.

Instantaneous algorithms rely on the presumption that there is essentially no relative delays between the sources and sensors and that

the path of the source is direct to the sensor. This is usually the case for Electroencephalography (EEG) signals. Anechoic mixtures/algorithms are the middle ground between instantaneous and echoic cases. Again only direct paths are assumed but this case allows for a delay in the direct path between source and sensor. The echoic case is a more complex problem due to reflections, which create the situation of multipath signals.

Real world signals always contain some amount of noise, either due to the sensors or from the environment in which the signals are recorded in. Accounting for noise in the mixing process does increase the complexity of the BSS problem, therefore some BSS methods filter the signals to reduce the noise before applying the BSS algorithm, whilst others include noise in the model of the mixing process and treat it as an additional source signal. An in-depth discussion of noise and methods that account for its effect on the unmixing process can be found in [24, 56]. For the work contained in this study, the level of noise is considered to be negligible, this is not uncommon [24, 56].

Early research was focused on the instantaneous case of signal mixtures, but, at least where audio source separation is concerned, the last decade has seen a shift of focus to echoic/convolutive mixtures which more realistically represents real world situations. This is the kind of signal mixture that is produced in the cocktail party environment. However, most current speech separation algorithms are uni-modal, relying solely on audio information, and there is a general consensus in the research community that by using audio information alone, BSS is reaching a limit in terms of accuracy of source separation, and that extra information (modalities) where available should be utilized [39, 52].

It is this approaching limit that is the motivation for this work. For a robust and accurate solution to the cocktail party problem, a system is required that exploits the extra information available, as humans do in a cocktail party scenario.

1.2 Why Visual Information?

Human speech is inherently bimodal, with both audio and visual components. It has been shown [118] that being able to see a speaker's face in a noisy environment greatly improves the intelligibility of that person's voice.[1] Moreover, the McGurk [78] effect also highlights the relationship between the audio and visual aspects of speech and how humans perceive speech.

Visual cues, for example are used to determine who is being addressed. Particular attention is focused upon the lips to help in deciding when the other person has started/stopped speaking and even to use the shape of the lips to help understand what is being said. Girin et al. [46] proposed one of the earliest systems for using visual information to help clean up noisy speech but few other publications have developed the idea of using visual information to aid speech separation.

The use of visual speech information is also popular in current research into speech recognition methods. Speech recognition suffers from poor performance in the presence of moderate acoustic noise, the inclusion of visual speech information has improved the results in this situation. Moreover, an audio-visual BSS method could be employed as a pre-processing step for an audio-visual speech recognition process.

[1]This is easily verified by speaking to someone in a noisy environment. First with the eyes closed, then with the eyes open and looking at the speaker's face. The speaker should seem more audible when the face can be seen.

This thesis seeks methods to answer the following questions:

- What visual features give the most useful information about an individual speaker and correlate strongest with the desired speech signal?
- How best to extract and track these features from the video?
- How can these features be modelled, for example their dynamic behaviour?
- How can visual information be integrated into a BSS algorithm in order to improve the performance?

1.3 Bi-modal speech separation

There exists very little literature in the area of bi-modal or video assisted BSS methods. However, the idea of using visual information to aid the BSS process is gaining momentum in the research community [52].

In this thesis, it is shown that by building a joint audio-visual model of a speaker and incorporating it within a convolutive BSS algorithm to control the learning rate, the convergence behaviour can be improved. In the model, the audio data are represented with Mel-Cepstral Frequency Components (MFCCs) and the visual data are described with an Active Appearance Model (AAM) [29]. The increased convergence rate is shown using firstly a Gaussian Mixture Model (GMM), and further improvement is gained by modelling the temporal dynamics of the joint audio-visual features using a Hidden Markov Model (HMM).

The high computation time and training requirements of the above method led to the investigation of a less complex visual speech model.

To this end a novel Visual-Voice Activity Detector (V-VAD) is proposed. Again, an AAM is used to model the shape and texture of the lips and the dynamics are modelled using an HMM. The difference between the proposed V-VAD and the audio-visual BSS work, is that for the proposed V-VAD, continuous speech is not modelled, nor the combined audio and visual data. Instead, only visual data related to the motion of the lips, without a speech utterance, are modelled. The motion of the lips over several frames is classified to be *speech* or *non speech* using the model of the lips motion without voice. The term "non speech" is used since a person may be silent when motion of the lips occurs, e.g. during a smile.

This research is then extended with a novel generic V-VAD which is more robust than the AAM based V-VAD. In this method, silence detection is achieved by modelling the motion flow of the area of the lips. The motion flow is obtained with a phase based motion estimation algorithm [74] and silence detection is achieved by modelling the flow field of the area of the lips using an HMM. Because of the similarities in the way people speak/smile, the models are largely person independent, as is shown by application of the model to a speaker not included in the model training data.

The two proposed V-VAD methods are evaluated in the same manner. Firstly, receiver operating characteristics (ROCs) are used to compare the classification accuracy of each method. They are then evaluated by individually using their outputs in a BSS algorithm [100] that utilizes them for solving the permutation problem inherent to BSS. The results show that the inclusion of visual information improves the alignment of permutations and thus improves the result of the BSS

algorithm compared to using only audio information.

1.4 Thesis Overview

The remainder of this thesis is structured as follows:

- In Chapter 2 a review of the relevant literature relating to BSS is given along with a review of visual feature extraction and modelling. It concludes with an overview of precedent audio-visual BSS methods.

- Chapter 3 provides the background of the shape-texture modelling technique used in this thesis, namely Active Appearance Models (AAMs).

- A novel audio-visual BSS method is presented in Chapter 4. The method uses a joint audio-visual probability model to control the learning rate of a previously published BSS algorithm.

- Chapter 5 focuses on using AAMs of the lips of a speaker for a novel visual voice activity detector (V-VAD). The method uses HMMs to model the dynamics of the AAM for speech/silence detection. The results of this method are compared to an existing V-VAD.

- A novel V-VAD is proposed in Chapter 6, based on using an HMM of a motion field of the lips of a speaker, where the motion field is obtained from a phase based motion estimation technique. The advantages of this new method over existing methods are discussed. The chapter concludes with experiments that use

the outputs from the V-VADs presented here along with another existing V-VAD in a bi-modal BSS algorithm.

- Chapter 7 provides the conclusions and suggestions for future work.

1.5 Main Contributions

The main contributions, and the publications (listed on the following page) arising from this work are:

- A novel video assisted BSS algorithm. The algorithm uses a joint audio-visual model of a speaker to control the learning rate of a penalty function based BSS algorithm [1,2].
- A novel visual voice activity detector is proposed based on capturing the dynamics of an AAM of the lips of a speaker with an HMM [3,4].
- A generic novel V-VAD is then presented. This is achieved by modelling the dynamics of a motion field of the lips of a speaker with an HMM to classify the lip motion as speech/non-speech. It is also shown that a reasonable accuracy of classification can be achieved on several subjects, some of whom are not included in the HMM training data [5].
- An audio-visual database was created during the course of this thesis, and the possibility of making it available to other research groups is being investigated.

1.6 Publications Arising From This Study

Below is a list of publications based on the novel contributions listed on the previous page

1. A. Aubrey, Y. Hicks, S. Sanei and J. Chambers, "Study of Video Assisted BSS for Convolutive Mixtures," *IEEE 12th Digital Signal Processing Workshop.* Wyoming, USA. September 2006.

2. A. Aubrey, J. Lees, Y. Hicks, and J. Chambers, "Using the Bimodality of Speech for Convolutive Frequency Domain Blind Source Separation," *in IMA 7th International Conference on Mathematics in Signal Processing.* December, 2006.

3. A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, and C. Jutten, "Two Novel Visual Voice Activity Detectors Based on Appearance Models and Retinal Filtering," *In* 15^{th} *European Signal Processing Conference (EUSIPCO).* Poland, 2007.

4. B. Rivet, A. Aubrey, L. Girin, Y. Hicks, C. Jutten, and J. Chambers, "Development and comparison of two approaches for visual speech analysis with application to voice activity detection," *Int. Conference On Auditory Visual Speech Processing (AVSP).* The Netherlands, 2007.

5. A. Aubrey, Y. Hicks, and J. Chambers, "Visual Voice Activity Detection with Optical Flow," *In preparation for submission to IET Proc. Signal Processing.*

Chapter 2

SOLVING THE COCKTAIL PARTY PROBLEM: A REVIEW

This chapter is not an all encompassing review of the literature on the cocktail party problem; the vast amount of material on the subject makes this impossible. Rather, its purpose is to place the research described in this thesis into the broader context of the study of the cocktail party problem.

The cocktail party problem was first proposed by Colin Cherry [21, 22]:

> "How do we recognise what one person is saying when others are speaking at the same time (the "cocktail party problem")?" - Colin Cherry 1954 [21]

It refers to the remarkable human ability to select and recognise one source of auditory input (e.g. speech, music) in a noisy environment. Researchers have been investigating this problem for the last few decades, yet, fifty years since Cherry's work the problem remains unsolved. Cherry did not only propose the cocktail party problem, he

also proposed the design of a machine ("filter") to solve it [21]. He suggested the following factors be considered:

- Voices originate from different directions.
- Lip reading, gestures and the like.
- Variation in the voices, male and female, mean pitch or speed, and so on.
- Different accents and other linguistic factors
- Transition probabilities based on voice dynamics, subject matter, syntax etc.

Current solutions contain one or more aspects of Cherry's proposed machine. For example, beamforming based approaches utilise a microphone array to obtain the direction of arrival (DOA) of a speaker's voice to aid in the speech separation process [105, 106]. There are also methods that exploit appropriate visual features to assist in separating or enhancing a speech signal contaminated by noise [46, 70, 100, 126]. Speech recognition relies on modelling the transition probabilities between words/letters/phonemes etc, for high recognition rates.

Attempts at solving the cocktail party problem can be broadly put into one of the following categories:

1. Blind Source Separation (BSS)
2. Computational Auditory Scene Analysis (CASA)
3. Speech Enhancement (SE)

The goal of blind source separation methods (BSS) is to separate all source signals from their mixtures, by exploiting their inherent differences. Computational auditory scene analysis (CASA) methods seek to segregate a target signal from a mixture based on principles of the human auditory system. While CASA methods perform speech enhancement, there are speech enhancement methods that do not rely on mimicking aspects of the human auditory system. Instead they rely on statistics of the signals in the mixtures, but do not separate all signals from the mixture as is the case with BSS.

During the past twenty years, the cocktail party problem has seen an increase of attention from researchers as the increase in computing power made a real time solution to the problem a possibility. The proposed solutions have traditionally used solely audio information. However, it is becoming ever more apparent, as Cherry highlighted [21], that at least in the field of BSS, extra information related to the nature of the sources is required for a robust solution to the cocktail party problem. As speech is bimodal (with both audio and visual aspects), a rational next step is to include visual features into the BSS framework to provide the additional information.

For example, beamforming based solutions [105,106] use microphone arrays to determine the position of speakers in a room, which could be obtained more accurately using visual rather than audio information, especially in non-stationary environments. Indeed, the relationship between the audio and visual aspects of speech has been noted on several occasions. Sumby and Pollack [118] reported that being able to see the speaker's face increases the intelligibility of that person's voice in a noisy environment. The McGurk effect [78] highlights the relation-

ship between the audio and visual aspects of speech and how humans perceive speech.

The focus of this thesis is the integration of visual information into the BSS framework. Therefore, CASA and SE methods will not be discussed in detail in this chapter, but will rather be mentioned in comparison to BSS methods.

This chapter is divided into the following Sections: Section 2.1 focuses on current methods of BSS, in particular the methods addressing the case of convolutive mixtures of sources. In Section 2.2 research devoted to using visual information for speech recognition, speaker identification and tracking, and methods for extracting this information from video data are discussed. In Section 2.3, methods for solving the cocktail party problem relying on both audio and visual information are discussed.

2.1 Audio Based Speech Separation

Over the past two decades there has been considerable interest in the field of blind source separation, starting with the seminal work of Herault and Jutten [50]. Initial research concentrated on instantaneous mixtures of sources [7, 8, 15, 25]. However, the instantaneous mixture model is simplistic and cannot be used to model the type of mixing that occurs in the cocktail party situation because it does not account for the multipath propagation of the speech signals as present in a real room environment. Later research switched focus to model the mixtures using convolutive models which represent the types of mixtures found in the cocktail party problem better than instantaneous mixtures.

Early solutions to the cocktail party problem using a convolutive

mixture model were based in the time domain [1, 65, 82]. However, to represent recordings taken in a real room environment where the impulse response of the room is in the order of 1000's of samples, a time domain method would be too computationally expensive to realize [109]. It is known that convolution in the time domain is equivalent to multiplication in the frequency domain. Transferring the problem into the frequency domain simplifies the convolutive mixing problem to that of instantaneous but complex valued mixing at each frequency. Thus the computational complexity is reduced, making the frequency domain the preferred choice in the current research [54, 79, 87, 90, 106, 110, 127]. Nonetheless, the reduction in computational complexity does come at a cost; the permutation problem inherent to BSS now occurs at each frequency. There has been a significant amount of research devoted to solving this problem and a selection of literature covering this topic will be discussed in Section 2.1.

A popular approach to solve the cocktail party problem is to exploit the second order statistics (SOS) of the speech signals. One of the more effective frequency domain BSS algorithms based on SOS, according to the results in [88], is the approach of Parra and Spence [87]. They exploit the non-stationarity of speech, where a least squares (LS) approach is used to minimise a cost function based on the cross power spectrum matrices of the sources. The goal being to simultaneously diagonalise these matrices at different times. The LS optimization then allows the unmixing matrix to be estimated iteratively. This method is generally regarded as one of the first frequency domain approaches to achieve a reasonable separation performance in a realistic environment, due in part to the novel solution proposed to solve the permutation

problem (discussed in Section 2.1.1).

Wang et al. [125, 127] extend the work of [87] by introducing a penalty function based approach that reduces the amplitude ambiguity and provides better shape preservation of speech signals than the original method [127]. The penalty constrains the optimization process allowing for fast convergence; provided a suitable penalty function is chosen the proposed method obtains better estimates of the original speech signals compared to the method in [87]. Robledo-Arnuncio and Juang [101] also modified Parra's algorithm and suggest a version using non-causal separation filters; however, they reported no significant improvement over the original algorithm.

Pham et al. [90] exploit the SOS of speech signals in their joint diagonalisation method. They use a variation of the cross power spectrum found in [87, 127], where the cost function is essentially a logarithmic based version of that found in [87]. It is unclear how well it separates convolutive mixtures as a comparison with another BSS method has not been provided. Mitianoudis and Davies [79] suggest a time-frequency framework that utilises a non-Gaussian model of the sources. They propose two methods for the update of the unmixing matrix, one is a modified natural gradient algorithm applied at each frequency bin, the second is a fast Newton-type ICA algorithm. The fast ICA method was shown to have a faster convergence and also achieved a better separation performance.

Beamforming is an alternative approach to BSS and is based on finding the set of spatial filters that reject/block sounds coming from the directions of interference. Saruwatari et al. [105] proposed a combined frequency domain BSS and beamforming method which decides at each

frequency bin if the direction of arrival (DOA) estimate obtained from the BSS algorithm or the beamformer is correct, and updates the unmixing matrix accordingly. The proposed method was shown to outperform a BSS algorithm for cases when the unmixing filters were learnt on short data lengths (1 second), but obtained equivalent performance for longer data lengths (5 seconds). The reason for this is that the short data lengths do not have adequate information for non-beamforming methods to learn a good estimate of the separation matrix. Parra and Alvino [86] suggest a variation of Parra's earlier work [87]. The method incorporates geometric information as a penalty constraint into the cost function described in [87] to steer the unmixing matrix so that the contributions from interfering sources are minimised. Results indicate that the combined beamforming and BSS algorithm outperforms conventional beamforming techniques.

More recently, several BSS algorithms applied to separation of speech signals in a cocktail party scenario have utilised available information about the activity of the speakers [23, 55, 83]. Nickel and Iyer [83] proposed a method that first detects *"exclusive activity periods"* (EAPs), i.e. periods where only one speaker is active. These are found by calculating a signal to interference measure (SIR) over short time periods, and a period is classed as an EAP if the SIR is greater than or equal to a predetermined threshold. The unmixing coefficents are found for a chosen speaker during that speaker's EAP. However, they only consider instantaneous mixtures. Huang et al. [54, 55] proposed a two-stage approach for convolutive mixtures in the frequency domain, that firstly estimates the unmixing matrix and then dereverberates the estimated speech signals. The unmixing filters for each speaker are

found during the periods when only one speaker is active and assumed to be static thereafter. The above is carried out in the same manner for each speaker. A speech dereverberation algorithm is then applied to the estimates to obtain the original speech signals. The advantage of converting a multiple-input multiple-output (MIMO) BSS into that of a single-input multiple-output (SIMO) framework is that the inherent permutation ambiguity of BSS is avoided, and the method was reported to be successful in highly reverberant environments.

Chu et al. [23] also use information about the speaker's activity and suggest a time domain method that optimizes an eigenvalue based cost function. The method is compared to that of Parra and Spence [87] and is shown to achieve better separation performance for varying levels of signal to noise ratio (SNR). However, the above methods require that the speakers remain stationary to achieve a good separation performance, due to the manner in which the unmixing filters are found.

To summarise, several methods for solving the cocktail party problem have been discussed, with focus given to frequency domain approaches, as well as recent attempts to exploit additional information from the silence periods in speech. The frequency domain is attractive to researchers due to its low computational cost compared to solutions in the time domain. However, the permutation problem inherent to BSS becomes a serious issue in the frequency domain. Thus, in the next section several recent proposals to overcome the permutation problem are discussed.

2.1.1 Solutions to the Permutation Problem

Transferring the BSS problem from the time domain to the frequency domain results in the permutation ambiguity becoming a serious problem. The permutation problem is one where the order in which the signals are recovered is unknown. For time domain solutions this is not a large problem, but in frequency domain methods the permutations must be aligned at each frequency bin so that the time domain separated signal contains frequencies from the same source. Further information on the permutation problem can be found in Chapter 4 of this thesis. There is also an amplitude (scale) ambiguity across frequencies but this is easily solved with matrix norm multiplication [56]. Several time-frequency methods have been proposed, which switch between the two domains to take advantage of each domain while avoiding their disadvantages. The permutation problem is also avoided as the independence of the estimated signals is usually evaluated in the time domain [107]. However, the time spent transforming between the two domains is significant, so for solutions to the cocktail party problem, the frequency domain is still more appealing.

Parra and Spence [87] proposed to solve the permutation problem via a smoothness constraint on the unmixing filters. The constraint essentially forces the frequency bins to align, and is achieved by limiting the length of the filter in the time domain to be much less than the size of the DFT. This has the effect of forcing the sources in the frequency domain to be continuous or smooth. However, Ikram and Morgan [57, 58] showed that in realistic environments Parra's method failed to align all the permutations, and suggested that the constraint on the filter length should be relaxed once the algorithm converges.

In [57] the authors provide an in-depth discussion of permutation inconsistency. They assume the mixing filters are known to derive ideal benchmarks of signal to interference ratio (SIR) improvements by comparing the SIR of individual sources and deciding whether or not to manually rearrange the permutations. Based on the solution to the permutation problem suggested by Parra and Spence [87], Ikram and Morgan [57] show that as the length of the unmixing filter increases to represent real room conditions, the SIR becomes worse. A solution to overcome this drawback is proposed in the form of a multistage algorithm where the separation is carried out in multiple stages. The initial mixing stage is followed by several unmixing stages, with the length of the unmixing filter increasing at each stage, where the final values of the unmixing matrix obtained at the previous stage are used as initial values of the next stage. It was found that the majority of the permutations aligned in the early stages retained their order during later stages, and there was no overall significant increase in computational complexity as the optimum number of stages was found to be two.

Sawada et al. [106] proposed to combine direction of arrival (DOA) information with interfrequency correlation of signal envelopes. They use a combination of the natural gradient and information maximization algorithms to perform the initial speech separation and then align the permutations in two stages. The first stage is to fix the permutations at those frequencies where the confidence of the DOA approach is high. The second stage is to decide the permutations for the remaining frequencies based on neighbouring correlations without changing those fixed by the DOA method. The reasons for using a combination of the

two approaches are that the DOA method is robust, as misalignment at one frequency does not affect other frequencies, whereas correlation is not robust since misalignment can affect neighbouring frequencies. However, the DOA approach is not as precise as the interfrequency correlation approach. In the same paper they also propose a method utilizing the harmonic structure of the signals that aligns permutations at low frequencies, where DOA estimation is difficult.

As mentioned in the introduction chapter, there is a general consensus in the research community that BSS using audio information alone is reaching a limit in terms of accuracy of source separation. As human speech is bimodal in nature, the natural way to proceed as intimated by Cherry [21] is to incorporate visual information of the speakers into BSS techniques.

In fact, there have already been attempts to use properties of conversational speech, such as silence periods, in finding the unmixing filters or solving the permutation problem. In a noisy environment these silence periods might be found more accurately using visual information [67].

Furthermore, beamforming based methods rely on being able to accurately determine the DOA, both of which can be achieved to a degree of success using only audio information, but both require the SNR to be high. Visual information could provide both of these; in particular, the DOA could be found more accurately in high noise environments as the visual data are immune to audio noise. Therefore, in the next section, tracking in video and visual feature extraction are discussed, as such visual information can be combined with corresponding audio information to form a multi-modal BSS system.

2.2 Visual Feature Extraction, Modelling and Tracking

A BSS system for speech separation typically uses purely audio information. To create a multi-modal BSS system, visual information is sought that is highly correlated with the audio data. There is a wide variety of visual feature extraction techniques available depending on what information is required. The ideas presented in this thesis exploit visual features which are related to speech, consequently the extraction of facial feature information is the prime goal. To locate faces in a video sequence, face detection and tracking methods are used [66], [81]. It may be the case that the participants are moving around a room, in this situation the participants themselves are tracked and then the face region is found and extracted. This section will highlight recent research on modelling visual features, feature extraction and tracking techniques.

2.2.1 Speaker localization

For tracking moving people in a room, particle filtering is a popular technique [19,20,66,80]. Checka et al. [19] developed a system to track the movements of several people around a room using two cameras and a microphone array consisting of 16 microphones. The cameras are in adjacent corners of a square room and the microphones are placed in groups of four, one group under each camera with the remaining two groups evenly spaced about the center of the wall containing the cameras. A particle filter is used to determine the number of speakers and their speech activity based on audio and video information. The results show that they were able to track up to three people at a time and determine who was speaking at that instant. The downside of this

method is that the data are not processed in real time but offline. In addition the system is only able to operate with at most one person speaking at a time; there was no investigation into the effects of the background noise on the performance of the system.

Liu et al. [66] use particle filtering to estimate the current position of the face, using prior probabilities of object motion and likelihood models of colour and edge are to estimate the location of the face. The data are processed in real time and the system is able to track partially occluded faces. Darrell et al. [36] used a single camera and two microphones to identify what portions of the video signal correspond to a particular audio signal by maximising the mutual information from the video and audio data.

Another interesting development in tracking techniques is the design of *smart rooms* that enable this [13, 19, 43, 80, 108]. In [13], Busso et al. have designed and developed a smart room capable of participant tracking and identification using multiple camera views in real time. Speaker identification is achieved using MFCCs (mel frequency cepstral coefficients) of each individual's speech and modelling the MFCCs using a GMM (Gaussian mixture model). For tracking, each participant (active and non-active speaker) is identified using multiple camera views. A Gaussian background model is used to segment moving areas in the scene and for areas where large variations are detected the foreground pixels are extracted and turned into regions. The multiple regions (one from each camera view) are combined to form a 3-D silhouette which is then converted to a visual hull. The active speaker is identified using time difference of arrival (TDOA) for speech location, which is then compared to each participant's visual location and a decision is made.

Conversation overlaps of the four participants were allowed, however, there was no detailed analysis of their effect on the performance of the system.

Siracusa et al. [108] developed a similar system that combines audio and visual cues. The possible participants are found with a head tracker that relies on view based appearance models of a head, and the views from multiple cameras are employed to find the orientation of a person's head. The audio cue is derived from the TDOA between microphones in an array. The audio-visual cues are then combined, and the audio-visual cues with the highest synchrony are assumed to be from the same speaker. The system is also able to determine the orientation of the head and uses this as the basis for inferring to whom someone is speaking.

2.2.2 Facial Feature Extraction

The extraction of facial features typically requires the face to be found in an image in the first instance. Yang et al. [129] produced a survey of current approaches to face detection and classified them into four categories, knowledge based, feature based, template matching and appearance based methods.

Knowledge based methods are developed based on rules derived from the developer's own knowledge of faces. The rules are simple and describe the facial features and their relationship, e.g. faces usually have eyes that are usually symmetric to each other along with a nose and mouth [128].

Feature based methods assume that there are features on a human face that do not change in different poses and lighting conditions. The

aim of these methods is to extract facial features such as eyes, nose, mouth, eyebrows and hairline using edge detectors [119]. Then build a statistical model that describes their relationship and decide if a face exists or not. Skin colour and skin texture have also been used to separate out faces from other objects in images [62].

Template matching uses several standard face patterns to describe the face as a whole or its individual features. The existence of a face is determined based on the correlation values of a given input image and the standard pattern. Active Shape Models (ASMs) proposed by Cootes et al. [26] can be considered as deformable templates.

Appearance based methods also use templates, but are learnt from a set of training images, e.g. Eigenfaces [121] and Active Appearance Models (AAMs) [29] which are an extension of ASMs. Statistical analysis and machine learning aid in finding the relevant characteristics of faces, which can be modelled using distribution models or discriminant functions. The dimensionality of the data is usually high so it is reduced to decrease the computational complexity.

A popular face detection method is the real-time detector of Viola and Jones [124]. It is based on the AdaBoost algorithm [44], which creates a strong classifier by combining several weak classifiers. Viola and Jones also defined a new type of image representation called *integral image* that represents the image in a reduced form, allowing for a considerable reduction in computation time. The method provides accurate detection of several faces in an image while still running in real time. This method has been extended by others such as Cristinacce and Cootes [35] for detecting individual facial features such as the eyes.

Cootes and Taylor [29] introduced AAMs as a way of modelling

selected features. An AAM is a joint statistical model of shape and texture, where a single appearance parameter defines a corresponding texture and shape vector. The parameters of the statistical model are learnt from a set of training images, as explained in detail in Chapter 3 of this thesis. AAMs are widely used for feature modelling and have been used in face recognition [41] and facial animation [32,33].

The disadvantage of using AAMs is that the tracking process does not tolerate large head movements and evaluating model parameter values requires significant time. An extension to this is proposed by Cootes and Taylor [27] that allows for more robust tracking over several frames. An alternative approach is view based appearance models [30, 108] that allow for movement of the head, however they require models of the face from different angles and are more complex models than the standard AAM.

Once the face is found (if it needed to be), the next step is to extract those visual speech features that have a high correlation with the audio speech. One naturally assumes the lips are the most important visual aspect, however it has been shown that there is also useful information contained in the cheeks [5,60,130]. Much of the research on extracting visual speech information has been for the purpose of improving Audio-Visual Speech Recognition (AVSR) [76, 91, 92], but has been used for speaker identification as well [69]. More recently, a new research topic has emerged, Visual Voice Activity Detection (V-VAD) [67,113], where the aim is to determine if a speaker is active (speaking) or not using solely visual information.

Matthews at al. [76], compared Active Shape Models (ASM), AAMs and a novel cascade filter (*sieves*) for extracting information about the

speaker's lips for the purpose of AVSR. They find that AAMs have a slight advantage over sieves but this is not significant. Their results show that ASMs have the worst performance. Similar to the Eigenface technique of Turk and Pentland [121], Bregler and Konig proposed Eigenlips to be used as the visual cue for AVSR [10]. Examples of other visual feature extraction techniques used within AVSR are optical flow [120], taking the *discrete cosine transform* (DCT) of the mouth region and also *lip geometric features*, where useful information such as height, width and area of the lips are extracted from the video. A comparison of these features for AVSR can be found in the work of Potamianos et al. [92].

Active contours (Snakes) [61, 115] are dynamic elastic curves that deform due to an energy minimization criterion to fit the shape boundary of an object. ASMs can be thought of as 'smart snakes' [115], as they have some prior knowledge of the shape boundary that is being sought. Delmas and Lievin [38] and Eveno et al. [42] have both used active contours for lip tracking. In [38] the corners of the mouth are first identified and used as starting points for the active contour and they achieve a good degree of tracking accuracy, but the authors point out that their tracker does not perform well when the tongue and gums are visible. Eveno [42] proposed a quasi automatic tracker based on a new active contour called *jumping snakes*. The snake grows from a single landmark (keypoint) to fit the boundary of the lips. Additional landmarks are then placed on the lip boundary to define its shape. Results indicate that the proposed tracker is comparable to landmarks placed manually.

Finally, comparisons of different visual feature extraction techniques

are provided by Çetingül et al. [17, 18] for the purpose of speech reading and speaker identification, and Sargin et al. [104] who measure the audio-visual correlation of several visual features. In [104], the audio-visual correlation is measured by combining audio information (MFCCs) with either the 2D-DCT of intensity of the lips region, the 2D-DCT of optical flow vectors of the lips region or the lip shape. They found that the 2D-DCT of optical flow vectors provided the highest correlation with the audio feature. Çetingül et al. [17] experimented with two similar features, they used the 2D-DCT of motion vectors obtained from the region of the lips, and also pure motion vectors that were obtained from the lip boundary. They used temporal correlations (encoded by an HMM) of the individual features for the purpose of speaker identification, noting that adding shape information to the pure motion vectors from the lip boundary improves the identification process. The authors expand upon this work in [18] by not only considering additional features for speaker ID but also experimenting with which feature or combination of features would be best suited for speech reading. The features they considered were the 2D-DCT of motion vectors found from the lips region, 2D-DCT of the motion vectors from the lips boundary, and a combination of the latter with the lips shape. For finding the lips shape they use an active contour method based on that proposed by Eveno et al. [42]. They also experimented with combining lip intensity information, found by computing the 2D-DCT of the lip region intensity values, with the 2D-DCT motion vectors. Results showed that lip motion was more useful for speech recognition than intensity information, as combining motion and intensity information improved speaker identification, but the speech recognition rate decreased.

This concludes the overview of extracting visual information from video of speakers in a cocktail party like scenario. In the next section, speech enhancement or speech separation methods are discussed that use visual information to aid in solving the cocktail party problem.

2.3 Audio-Visual Speech Separation

In Section 2.1, solutions to the cocktail party problem were discussed that relied solely on audio information. Early on into research of solving convolutive mixtures it was suggested that for solving realistic mixtures of speech, using audio information alone might not be sufficient [45]. Solutions for convolutive mixtures using just audio information have been partially successful, however, a recent survey [88] compared approximately two dozen BSS algorithms and found there to be little difference between the results of the several best methods. Furthermore Haykin and Chen [52] have also noted that more than just audio information would be required for a robust solution to the cocktail party problem. This section focuses on previous attempts to combine audio and visual information to improve speech separation performance. The type of visual information used is typically either the location of the speakers, visual facial features (lip shape, texture) or a combination of the two. Methods of obtaining this information were highlighted in Section 2.2.

One of the earliest works on the subject of using visual information for a source separation problem is by Darrell et al. [36]. Their system was able to identify where in a region of video an audio source is located and enables the user to enhance the voice of one of the speakers. Girin et al. [46] proposed a method that enhances noisy speech using a

filtering approach, where the filter coefficients are estimated with the aid of lip shape information. Girin et al. then extended this work to propose one of the first audio-visual BSS methods [45]. Based on instantaneous mixtures of speech, the method finds the unmixing matrix by maximising the audio-visual coherence. Simultaneously, Okuno et al. [85] proposed to use video information to provide the locations of speakers for a beamforming based BSS method. They showed that the use of visual information led to a significant improvement over the performance of a purely audio based method in a convolutive mixture of speech signals. Furthermore, a similar method to that proposed by Girin et al. [46] was the basis of a preprocessing stage in a speech recognition framework proposed by Goecke et al. [47]. The noisy speech signals were first subjected to an audio-visual speech enhancement stage before being processed by a speech recognition system. Their results show that enhancing the speech before processing by an audio-only speech recognizer led to a decrease in the recognition error rate, however better results were obtained using an AVSR system with no speech enhancement. This is possibly because the speech enhancement stage assumes a fairly simple mixture of noise and speech, and the conditions in the experiments were more complicated than instantaneous mixing. The coherence of the audio and visual data was better captured by the AVSR system [47]. Further investigation into speech enhancement using audio-visual information was performed by McCowan et al. [77] and Maganti et al. [70]. In [77], an audio-visual tracker was utilised to provide the locations of people in a room using a microphone array and several video cameras. The estimated locations were used in a beamformer to enhance a particular person's speech. Their proposed

beamforming method was shown to enhance the speech to a similar quality of a headset or lapel microphone, even for situations when the speaker was not facing the microphone array. This work was expanded upon by Maganti et al. [70] and further experiments conducted to evaluate the quality of the speech enhancement by using it as the input speech to an audio only speech recognition system. They reported that the recognition error rates for the audio-visual beamformer were better than those of an audio-only beamformer, and comparable to those of a lapel worn microphone in a real meeting room environment with two speakers.

For solving the cocktail party problem using BSS methods, visual information can be employed to either find the unmixing filters, solve the permutation problem or both. Initial research considered instantaneous mixtures of speech. Sodoyer et al. [114] extended the work of Girin et al. [45, 46] by proposing a novel audio-visual blind source separation (AV-BSS) algorithm, where the visual information is used to estimate the separation matrix for an instantaneous mixture of speech signals. Further experiments were conducted by Sodoyer et al. in [111, 112] and initial investigations into the use of audio-visual information to help in solving the permutation problem were performed. The performance of their previous AV-BSS algorithm [114] is compared to the JADE [15] algorithm and it is shown to outperform both the standard JADE algorithm and a modified JADE algorithm where visual information is used as a post processing step for solving the permutation problem. Although simple mixtures of speech were used, mixtures of several people were considered.

As the field of audio-only BSS methods for convolutive mixtures

matured, researchers realised more information was required in order to improve the separation performance, so AV-BSS methods began to be developed. Rivet et al. [99] proposed to solve the permutation problem using audio-visual (AV) coherence of the speech signals. A joint AV probability model containing visual and audio data was formed, where the visual data were provided in the form of geometric lip shape parameters (height and width) and audio data in the form of spectral characteristics. First the mixtures were separated using the audio BSS method by Pham et al. [90], then the joint model is used to estimate the permutation matrix to further separate the mixtures. Rivet et al. [96,98] extended this, offering a solution to the scaling ambiguity problem of BSS.

An alternative method of using the audio-visual information was suggested by Wang et al. [126]. They also build a joint audio-visual probability model, but the visual information is encoded this time using an AAM of lip characteristics supplying the visual information and MFCCs supply the audio data. The AAM provides better shape description than the visual feature in [96] and also provides texture information of the inside of the mouth, which was not available in [96]. The joint AV model is then used as the penalty function in the BSS algorithm from [127] to find the separation matrix. They show that the use of visual information improves the quality of the separated signals, compared to using no visual information. Sanei et al. [103] incorporate the location of the speakers into a frequency domain penalty function based approach. The location of the speakers relative to the microphones is utilised to constrain the estimation of the separation filters. Also due to the manner in which the update at each iteration is per-

formed, the permutation problem is claimed to be completely solved.

More recently, visual voice activity detectors (V-VADs) have been used to provide additional information for BSS. Rivet et al. [100] used the output of a V-VAD to solve the permutation problem by post processing the output of a BSS method. In [97] Rivet et al. proposed a geometric BSS method that uses the inactivity of a source to estimate the separation filter matrix. The main advantage of this technique is the low computational cost.

2.4 Summary

The work of Cherry [21,22] was identified as the foundation of the cocktail party problem. Audio only approaches to speech separation were summarised and the frequency domain methods highlighted as most suitable for separation of sources measured in a room environment. The permutation problem encountered in frequency domain methods and possible solutions were described. Next, the visual information that can be exploited in multi-modal audio separation was reviewed and facial feature extraction identified to have a major importance. Finally, previous research in the field of audio-visual blind speech separation was reviewed as the platform for the research activity in this thesis.

Chapter 3

MODELLING VISUAL FEATURES

This chapter provides an introduction to the visual feature extraction techniques used in Chapters 4 and 5 and also the modelling techniques used throughout this thesis. It also covers the process by which the video and speech data used in this thesis are acquired.

As previously mentioned, a key challenge in audio-visual speech separation (AVSS) is to ensure that the chosen features have a high correlation with the audio information. As speech is the focus of the work, the most natural features to use would be those involved with the production of speech and the most visible components of speech production are the lips. Chapter 2 discussed several methods for modelling visual features, and in particular ones that have been used previously for modelling the lips (or the lips and surrounding area) in audio-visual ASR (automatic speech recognition). Audio-visual ASR research has been ongoing since the mid 1980's [89], and many approaches to extracting visual features have been proposed. As in this work, mouth features are primarily used in audio-visual ASR, and the successful techniques developed therein are useful for this work. One of the more popular methods of feature extraction is the AAM. When extracting

facial features, there is normally the problem of firstly detecting the face and mouth, and then tracking the desired feature. Detecting the face and mouth region is not considered as this can be achieved using a technique such as the Viola-Jones face detector [124]. It is taken as given that this can be performed efficiently, and so the problem of mouth feature extraction is the focus here.

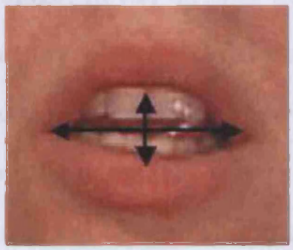
The following section provides a brief overview of mouth feature extraction methods. It is by no means exhaustive, only a few commonly used methods are considered. The remainder of the chapter is as follows: Sections 3.2 and 3.3 focus on the Active Appearance Model (AAM) developed by Cootes et al. [28, 29] as it is used as the visual descriptor for the work contained in Chapters 4 and 5. Section 3.4 details how the audio and visual data are obtained. Sections 3.5 and 3.6 contain an overview of the speech features used and statistical data modelling techniques respectively, and Section 3.7 concludes the chapter with a summary.

3.1 Feature Extraction

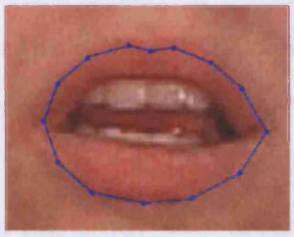
As mentioned previously, research into audio-visual ASR has been ongoing since the mid 1980's, and many of the methods developed are pertinent to work in this thesis. They can be roughly grouped into three categories [81]: lip contour based features, low level video pixel based features and features that are a combination of both. Lip contour features consist of the inner and/or outer lip contour shape, which are then modelled in a statistical model such as a point distribution model (PDM), or alternatively geometric parameters such as lip height and width (See Figure 3.1). Video pixel features consist of applying appro-

priate transforms on a region of interest (ROI) [81] such as the speaker's mouth area, and using the transformed pixel values as the features. Examples of such transforms are: the discrete cosine transform (DCT), the discrete wavelet transform (DWT) or a PCA projection. There are also examples of methods where the low-level (pixel based) and high-level (contour) features are combined to provide both shape and appearance features, such as the AAM. Comparisons using the above techniques in an audio-visual ASR context are given in [76, 81, 92].

In this work it was decided to use the AAM appearance parameters, as they are widely used to model the lips due to their ability to incorporate both high level data (shape) and low level data (texture) into a single statistical model. This has an advantage over other techniques in that it provides more information than low or high level methods alone.



(a) Lip Height and Width



(b) Lip Shape

Figure 3.1. Example of visual features extracted from the lips.

3.2 Overview of Active Appearance Models

AAMs were first proposed by Cootes et al. [28, 29] as an extension to the active shape model (ASM) [26]. In an AAM, a single *appearance parameter* vector describes both shape and texture and is frequently used to model the face [27, 41, 48, 49]. An AAM is built in three main stages [29, 30], and where the result of the final stage is a set of appearance parameters. The appearance parameters can also be used to recreate the data from the dataset, or alternatively new data not found in the original dataset can be created. This possibility has been exploited in areas such as speech driven facial animation [34]. The process of building the AAM is discussed next.

3.2.1 Landmarking Images

To build an AAM, the images in the dataset must first be landmarked. The landmarks are typically placed to define a feature in the video frame. The placement and number of landmarks on the feature should be consistent throughout the dataset.

For example, landmarks are used in this study to define the contour of the lips, which is achieved by placing a landmark on each corner of the lips and also one in the centre of the top and bottom lips. The remaining landmarks are then placed at equal distance between these as shown in Figure 3.2.

The datasets used in this study have several thousand frames and manually landmarking these would be highly time consuming. To overcome this a semi-automatic process is employed, whereby several frames are landmarked and used as training examples for an automatic landmarking algorithm. The automatic landmarking algorithm used in this

thesis is a modified version of Luettin's [68] downhill simplex minimisation (DSM) tracking algorithm proposed by Cosker [31]. The result of the automatic landmarking is checked for accuracy and any mistakes are manually corrected [1].

3.2.2 Point Distribution Models

Point distribution models (PDMs) are linear models of shape variation and are required for constructing the AAM. In the application described in this thesis, the PDM describes the lip shape variation of a speaker in a set of video frames. The PDM is calculated from a set of landmarked images, and for the purpose of the work in this thesis the landmarks are placed on the lips contour, as illustrated by Figure 3.2.

Landmarks placed on the lips contour define its shape. So for N landmark points in d dimensions, the shape of the lips can be represented by an Nd vector. This vector is formed by concatenating the elements of all the landmarks. For the 2-D images used in this work, each landmark is represented as coordinates (x_i, y_i); so for a single image, the $2N$ element vector $\mathbf{x}$ is:

$$\mathbf{x} = (x_1, \ldots, x_N, y_1, \ldots, y_N)^T \tag{3.1}$$

For a given set of j images there are $\mathbf{x}_1, \ldots, \mathbf{x}_j$ such vectors, and in each vector in the set, the coordinates are in the same order.

Once the images have been landmarked they need to be aligned into a common coordinate frame. Cootes and Taylor [26] use a method called Procrustes analysis that aligns each shape so that the sum of

[1] The Matlab code for the semi automatic landmarking process, and much of the code to build the AAMs was kindly provided by Dr Darren Cosker.

Figure 3.2. Image of lip region landmarks, with a zoomed in view of landmarks below, where the landmarks are connected by a solid line.

distances D of each shape to the mean shape is minimized.

$$D = \sum_{i=1}^{j} |\mathbf{x}_i - \bar{\mathbf{x}}|^2 \tag{3.2}$$

where $\mathbf{x}_i$ is a shape vector and $\bar{\mathbf{x}}$ is the mean shape. To align all images in a set the following iterative approach can be used [26, 115]:

1. Translate each shape vector so its center of gravity is at the origin.

2. Choose one example as an initial estimate of the mean shape $\bar{\mathbf{x}}$, and scale to have unit length i.e. $||\bar{\mathbf{x}}|| = 1$.

3. Record this estimate as the vector $\bar{\mathbf{x}}_0$.

4. Align each shape vector with the current mean shape $\bar{\mathbf{x}}_0$ using the Procrustes method outlined in [26, 115].

5. Calculate a new mean from the aligned shapes, align the new mean with $\bar{\mathbf{x}}_0$.

6. If not converged, return to step 4. The process is deemed to have converged if the current estimate of the mean does not change significantly when compared to the previous mean:

$$||\bar{\mathbf{x}}_{new} - \bar{\mathbf{x}}_{old}|| < lim_x$$

where lim_x is a predefined value.

Now that the set of points $\mathbf{x}_j$ have been aligned, these vectors form a distribution in the Nd dimensional space in which they reside. To make the data more manageable it is necessary to reduce the dimensionality. For example, in the application considered in this thesis, each appearance vector has several hundred dimensions and without reducing the number of dimensions the computation time would be impractical. One well known approach to reduce the dimensionality is to use PCA, which allows any of the original points to be approximated using a model with fewer than Nd parameters. PCA also has other properties that are discussed in Chapter 4 of this thesis. The steps for calculating the principal components are as follows:

1. Compute the sample mean vector of the data.

$$\bar{\mathbf{x}} = \frac{1}{j}\sum_{i=1}^{j}\mathbf{x_i} \tag{3.3}$$

2. Compute the sample covariance matrix of the data

$$\mathbf{S} = \frac{1}{j-1}\sum_{i=1}^{j}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^T \tag{3.4}$$

and $(.)^T$ denotes vector transpose.

3. Compute the eigenvectors, ϕ_i and eigenvalues, λ_i of $\mathbf{S}$ using $\mathbf{S}\phi_i = \lambda_i\phi_i$, and sort them so that $\lambda_i \geq \lambda_{i+1}$, i.e. in descending order of energy.

Large eigenvalues correspond to large variations in the underlying data set, and also provide the **modes of variation** (see [29] for more discussion on this). The training set can be approximated using:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_s\mathbf{b}_s \tag{3.5}$$

where $\mathbf{P}_s = (\phi_1, \phi_2, \ldots, \phi_t)$ and contains the t eigenvectors corresponding to the largest eigenvalues and $\mathbf{b}_s$ is a t dimensional column vector that indicates how much variation is exhibited with respect to each of the eigenvectors. Similarly, the shape parameter $\mathbf{b}_s$ can be represented as:

$$\mathbf{b}_s = \mathbf{P}_s^T(\mathbf{x} - \bar{\mathbf{x}}) \tag{3.6}$$

Variation of $\mathbf{b}_s$ allows new shapes to be defined. Furthermore, restricting the variation of the elements b_i of $\mathbf{b}_s$ to within $\pm 3\sqrt{\lambda_i}$ (3 standard deviations) ensures the new shape is similar to those present in the original data. It is also necessary to decide on a value for the number of modes to keep (t). The easiest is to choose a t that explains a chosen percentage of the variance exhibited in the training set. This

can be done by equating the total variance of the training data to the sum of all eigenvalues, $\lambda_{total} = \sum \lambda_i$, and then choose the t largest eigenvalues such that:

$$\sum_{i=1}^{t} \lambda_i \geq \alpha \lambda_{total} \qquad 0 \leq \alpha \leq 1 \tag{3.7}$$

where α defines the proportion of the total variation to retain, e.g. 0.98 for 98%.

3.3 Texture Models

To model accurately a complete set of images, models of both shape and texture from each image are needed. The grey level values across the selected region of the image is referred to as texture.[2] Once the shape model is built, each image is warped (using triangulated piece-wise affine warping [28,31]) from its landmark points to the mean shape to obtain a shape free patch. This patch is then used to build a statistical model of the texture variation within the region. Effectively, the shape free patch becomes the region of interest (ROI) of the required texture features.

3.3.1 Statistical Models of Texture

Using a triangulation algorithm [28, 31], the selected feature in each image is warped to the mean shape to obtain a shape free patch. The intensity information from the shape free patch is formed into a texture vector $\mathbf{g}_{im}$. The effect of lighting variation is minimized by normalizing the images by applying a scaling α, and offset β. The following

[2]Hereafter, texture refers to the grey level values unless otherwise stated.

procedure is described in [28,29].

1. Choose a texture as an initial estimate of the mean $\bar{\mathbf{g}}$

2. For the texture vector $\mathbf{g}_{im}$, find a scaling value $\alpha = \mathbf{g}_{im}.\bar{g}$ and offset $\beta = (\mathbf{g}_{im}.1)/n$ where n is the number of elements in $\mathbf{g}_{im}$ and 1 is a vector of ones.

3. Obtain a normalised texture vector $\mathbf{g}$ by applying the scaling and offset to $\mathbf{g}_{im}$ so that $\mathbf{g} = (\mathbf{g}_{im} - \beta.1)/\alpha$

4. Calculate a new estimate for the mean $\bar{\mathbf{g}}$, repeat from step 2 until $\bar{\mathbf{g}}$ converges, (i.e. until there is little change in the mean value):

$$||\bar{\mathbf{g}}_{new} - \bar{\mathbf{g}}_{old}|| < lim_g$$

where lim_g is a predefined value.

Once the texture data have been normalised, PCA is applied to the normalised vector set to obtain a texture model:

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_\mathbf{g}\mathbf{b}_\mathbf{g} \tag{3.8}$$

where $\bar{\mathbf{g}}$ is the mean normalised grey level vector, $\mathbf{P}_\mathbf{g}$ contains a set of orthogonal modes of intensity variation and $\mathbf{b}_\mathbf{g}$ contains a set of grey level parameters. As with the PDM, varying the grey level parameters $\mathbf{b}_\mathbf{g}$ allows new textures to be created, under the same constraint of three standard deviations. As texture vectors can be very large (especially in comparison with the shape vectors) the application of PCA can provide a significant reduction in the size of these vectors whilst retaining the majority of the texture information.

3.3.2 Combined Shape and Appearance Models

The shape and texture of any example in the set can be described by the shape and texture parameters, $\mathbf{b_s}$ and $\mathbf{b_g}$. By concatenating these vectors a joint shape and texture vector $\mathbf{b}$ is defined as [29, 30]:

$$\mathbf{b} = \begin{pmatrix} \mathbf{W_s b_s} \\ \mathbf{b_g} \end{pmatrix} = \begin{pmatrix} \mathbf{W_s P_s^T (x - \bar{x})} \\ \mathbf{P_g^T (g - \bar{g})} \end{pmatrix} \tag{3.9}$$

where $\mathbf{W_s}$ is a diagonal matrix of weights for each shape parameter to scale the parameters of $\mathbf{b_s}$ to lie in the range of values for $\mathbf{b_g}$. $\mathbf{W_s}$ may be found by calculating the energy ratio of total shape variation to the total intensity variation:

$$\mathbf{W}_s = r\mathbf{I} \tag{3.10}$$

where

$$r^2 = \frac{\sum_{i=1}^{t_s} \lambda_{is}}{\sum_{i=1}^{t_g} \lambda_{ig}} \tag{3.11}$$

where λ_{is} and λ_{ig} are the shape and texture eigenvalues respectively and t_s and t_g are the total number of shape and texture eigenvalues retained from building the shape and texture models. There will also be correlations between the shape and texture variations as they were obtained from the same dataset. Applying PCA to the set of concatenated shape and texture vectors gives:

$$\mathbf{b} = \mathbf{P}_c \mathbf{c} \tag{3.12}$$

where $\mathbf{P_c}$ contains the eigenvectors and $\mathbf{c}$ is a vector of *appearance parameters* controlling both shape and texture variation.

Due to the linearity of the model it is possible to express the shape

and grey-levels directly as a function of $\mathbf{c}$ [29,30]:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P_s}\mathbf{W_s^{-1}}\mathbf{P}_{cs}\mathbf{c} \quad , \quad \mathbf{g} = \bar{\mathbf{g}} + \mathbf{P_g}\mathbf{P}_{cg}\mathbf{c} \tag{3.13}$$

and,

$$\mathbf{P}_c = \begin{pmatrix} \mathbf{P}_{cs} \\ \mathbf{P}_{cg} \end{pmatrix} \tag{3.14}$$

Equations (3.13) can be summarized further:

$$\begin{aligned} \mathbf{x} &= \bar{\mathbf{x}} + \mathbf{Q}_s\mathbf{c} \\ \mathbf{g} &= \bar{\mathbf{g}} + \mathbf{Q}_g\mathbf{c} \end{aligned} \tag{3.15}$$

where

$$\begin{aligned} \mathbf{Q}_s &= \mathbf{P}_s\mathbf{W}_s^{-1}\mathbf{P}_{cs}\mathbf{c} \\ \mathbf{Q}_g &= \mathbf{P}_g\mathbf{P}_{cg}\mathbf{c} \end{aligned} \tag{3.16}$$

This completes the definition of the AAM and the acquisition of the audio-visual data is considered next.

3.4 Audio-Visual Data Collection

Figures (3.3) and (3.4) below show the set up of the equipment in the Intelligent office situated in the CDSP Lab (Center of Digital Signal Processing).

The two spot lamps were used in conjunction with the overhead room lighting to eliminate shadows from the lip area. Audio-visual data were recorded simultaneously on separate systems but were manually checked to ensure they were synchronized. All audio-visual data were recorded in the same manner, however the early data recordings used in Chapter 4 could not be used for the experiments in Chapters 5 and 6

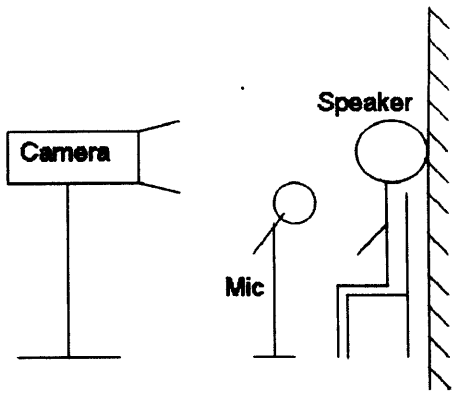


Figure 3.3. Side view of video capture setup.

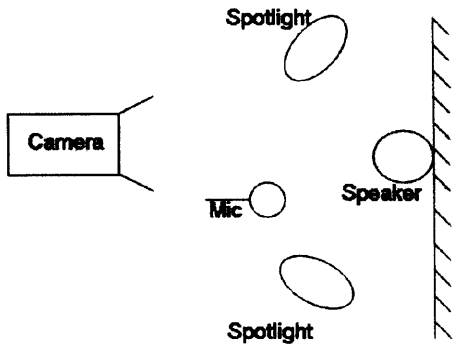


Figure 3.4. Top down view of video capture setup.

as they were mainly continuous speech and so the data were inadequate to test the visual voice activity detectors. Therefore, the data used in each chapter is described in that chapter.

For the work based on AAMs, 17 landmarks were used to define the lip shape for all speakers. Once all of the frames in the dataset had been landmarked and any errors corrected, the appearance model can be constructed. Figure (3.5) contains examples of several landmarked frames from one of the datasets. To check how accurately the appearance model approximates to the original, the modes of variation of the AAM can be considered. Figure (3.6) shows the first two modes of variation between ± 3 s.d. (the center image in each row is the mean).

This figure confirms the key information in the lip shapes is retained after the dimension reduction.

Figure 3.5. Consecutive landmarked frames, reading left to right, top to bottom.

Figure 3.6. First two modes of variation, varying $\pm$ 3 s.d. in steps of 1 s.d.

3.5 Speech Features

The experiments in Chapter 4 use a joint model of audio-visual speech. The method of modelling the visual data has been discussed, the method for robustly coding the audio information is described next.

The speech features used are Mel-Frequency Cepstral Coefficients (MFCCs), which are popular in speech recognition research [76]. MFCCs are found by warping the frequencies of audio data onto the *mel scale* [116]. The mel scale is an approximation of the nonlinear frequency response of the human ear. The mel coefficients are distributed approximately linearly up to 1kHz, and nonlinearly (logarithmicly) above 1kHz and may be calculated in the following manner [14, 31, 37]:

1. Divide the audio signal into windows. Window size can range between 15-30ms.
2. Compute the DFT of the signal in each window.
3. Compute the magnitude of the signal spectrum.
4. Scale the result of the previous stage using a Mel filter bank and take the log.
5. Obtain the MFCCs by calculating the DCT (discrete cosine transform) of step 4.

The Mel filter bank is a collection of triangular filters and is evenly spaced along the Mel Frequency Scale with 50% overlap as depicted in Figure 3.7.

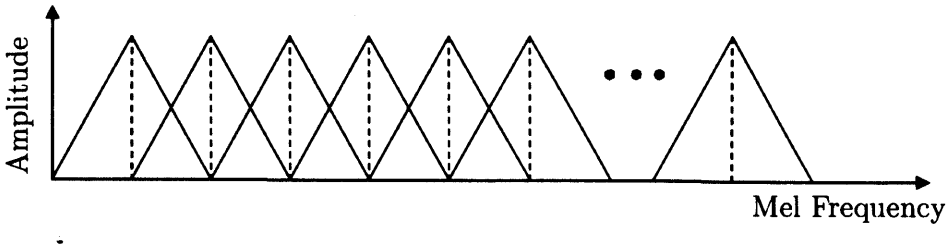


Figure 3.7. Mel filter bank on Mel frequency scale.

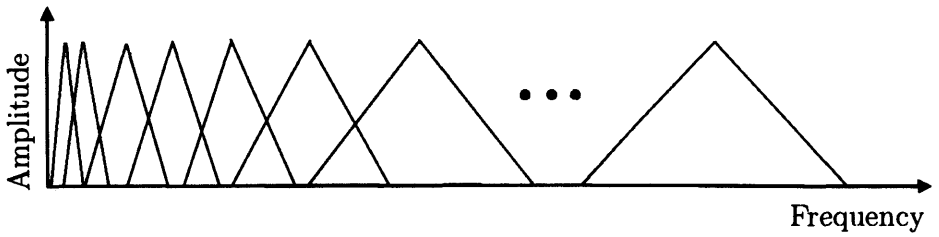


Figure 3.8. Mel filter bank converted to the original frequency scale.

Figure 3.8 is the result of converting the Mel frequency filter bank of Figure 3.7 into the original frequency domain, where the relationship between the mel-frequency and the physical frequency is defined as [37]:

$$Mel(f) = 1127.01048 \times log_e(1 + f/700). \tag{3.17}$$

The resulting amplitudes of step 5 above are the MFCCs. In this work a 20ms Hamming window is used which results in 12 MFCCs which are used as the audio feature. As a final step Cepstral Mean Normalisation (CMN) is used to normalise the MFCCs, as this reduces the distortion caused by the transmission channel (e.g. the microphone) [14,31]. Essentially, CMN is the subtraction of the mean MFCC vector from the set of MFCC vectors [31].

3.6 Modelling selected features

Both the audio and visual feature distributions are not guaranteed to be linear throughout the multi-dimensional distribution. Therefore the data would be more accurately modelled using a non-linear modelling technique such as a Gaussian Mixture Model (GMM). A k component GMM of a distribution of a variable vector $\mathbf{x}$ may be defined as:

$$p(\mathbf{x}) = \sum_{j=1}^{k} \alpha_j p(\mathbf{x}|j) \tag{3.18}$$

where α_j are the prior probabilities of the j^{th} Gaussian mixture and $p(\mathbf{x}|j)$ is a d dimensional multivariate Gaussian distribution defined as:

$$p(\mathbf{x}|j) = \frac{\exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^T \mathbf{C}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j))}{\sqrt{(2\pi)^d |\mathbf{C}_j|^{1/2}}} \tag{3.19}$$

where $\boldsymbol{\mu}_i$ and $\mathbf{C}_j$ are the centres (means) and covariance matrices of the Gaussians respectively and $|.|$ denotes the matrix determinant. The GMM parameters α_j, $\boldsymbol{\mu}_i$ and $\mathbf{C}_j$ are typically estimated using the Expectation Maximisation (EM) algorithm [40, 56]. Depending on the data, maximum likelihood estimation (MLE) can also be used to find these parameters, but because the data used to train the models in this thesis were incomplete, EM provides a more accurate estimation [40]. There is no standard method to determine the correct number of components k to use, and as such the value of k will be found empirically.

However, as the data vary with time, and taking the change in the shape of lips over time as an example, a GMM will not contain information on the temporal relationship of the shape of the lips. The shape of the lips at each time instance may be valid but the overall sequence may be invalid (with regard to the training data). For this reason the data is also modelled using a Hidden Markov Model (HMM). HMMs have been used since the mid 1970s in speech recognition research [37]. Originally, they were used to model audio features but more recently have been used to model audio-visual speech features. An HMM allows the temporal dependencies of data to be modelled using a probability transition matrix, where each element of the matrix represents the

conditional probability of transitioning from one state to another. The states in the HMMs employed are represented as single Gaussians obtained from a GMM.

An HMM is defined by the following:

- The number of states k in the model. In this work each state is a single Gaussian and the state at discrete time t is defined as $q_k(t)$
- The state transition probability matrix $\mathbf{A} = \{a_{ij}\}$, where

$$a_{ij} = P(q_j(t+1)|q_i(t)), 1 \leq i, j \leq k \tag{3.20}$$

  that is, a_{ij} is the probability of moving from state i at the current time interval to state j at the next time interval.

- The observation probability distribution $\mathbf{B} = \{b_j(O)\}$

$$b_j(O) = P(O_t|q_j(t)) \tag{3.21}$$

  which is the probability of observation O_t belonging to state j.

- The initial probabilities of being in state i at $t = 1$

$$\pi = \pi_i \tag{3.22}$$

It is common to define the model parameters of an HMM as:

$$\lambda = (\mathbf{A}, \mathbf{B}, \pi) \tag{3.23}$$

There are three basic issues with HMMs [40, 93]

1. Given an observation sequence $\mathbf{O} = O_1 O_2 \dots O_N$ and the model parameters $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$, determine the probability of the observation sequence $P(\mathbf{O}|\lambda)$.

2. Given an observation sequence $\mathbf{O} = O_1 O_2 \dots O_N$ and the model $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$, determine the hidden state sequence $\mathbf{Q} = q_1 q_2 \dots q_N$ that best explains the observations.

3. Given the number of states, find the model parameters $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$ that maximise $P(\mathbf{O}|\lambda)$.

Only problems 1 and 3 are of importance in the context of the work in this thesis. For the training of the HMM (problem 3) the widely used Baum-Welch re-estimation procedure is used. Secondly, given an observation from previously unseen data the probability the model generated this new data must be calculated, which is achieved by calculating the forward probabilities [93]. Both the Baum-Welch and the method of calculating the forward probabilities are described in excellent detail by Rabiner [93].

3.7 Summary

In this chapter the methods for obtaining both audio and visual speech features have been described, as have methods to build statistical models of these features. Detailed information has been given on statistical shape modelling along with grey level models and how to combine the two to produce the widely used active appearance model. The work presented in Chapter 4 utilises both the audio and visual speech features to form a combined audio-visual model of speech data for use in a novel audio-visual BSS algorithm, while in Chapter 5 a novel Vi-

sual Voice Activity Detector (V-VAD) is described which requires an AAM of a speaker's lips to detect silence phases in speech. The GMM and HMM discussed in Section 3.6 are used throughout the thesis for modelling speech features.

Chapter 4

USING AUDIO-VISUAL SPEECH IN A PENALTY FUNCTION BASED BLIND SOURCE SEPARATION FRAMEWORK

4.1 Introduction

The motivation of this thesis is to propose efficient methods for solving the *cocktail party problem* [21] which take advantage of all available information, including visual data. A popular technique for solving this problem is Blind Source Separation (BSS), therefore in this chapter an existing BSS approach is adapted to include audio-visual information in order to improve its convergence behaviour. In Chapter 3 methods for encoding audio and visual speech information were discussed, namely Mel-Frequency Cepstral Coefficients (MFCCs) and Active Appearance Models (AAMs). The novel method presented in this chapter uses these techniques to exploit the audio-visual coherence of speech, which

is shown in this chapter to help control the convergence behaviour of a frequency domain blind source separation algorithm. The method is based on maximising the coherence by employing an audio-visual joint probability model which is built using the statistical modelling techniques discussed in Chapter 3, namely the Gaussian Mixture Model (GMM) and the Hidden Markov Model (HMM). The statistical model (GMM or HMM) is then incorporated into the frequency domain algorithm of Wang et al. [125, 127] to control the convergence behaviour. Experiments are conducted that compare the convergence rate when modelling the audio-visual data with a GMM to modelling the data with an HMM. Simulations show that including a statistical model provides an increased rate of convergence when compared to using raw audio data (i.e. no model). Furthermore, the difference between using HMMs and GMMs as statistical models is the presence of the transition probabilities that capture the temporal dynamics of the signal, which is in line with the requirements for solving the cocktail party problem listed by Cherry [21, 22].

The next Section (4.2) provides an introduction to blind source separation, Sections 4.3 and 4.4 discuss the case of instantaneous and convolutive mixtures of signals respectively. A general overview of frequency domain BSS, together with a detailed discussion of the algorithm proposed by Wang et al. [127] is provided in Section 4.5. A novel audio-visual BSS algorithm is presented in Section 4.6 with the results of simulation in Section 4.7. The chapter is concluded in Section 4.8.

4.2 Overview of Blind Source Separation

The goal of BSS is to recover independent source signals (sources) from a mixture of signals. Solutions to the problem date back to the early work of Herault and Jutten [50] and in the last two decades many solutions have been proposed. Also the subject has been regarded as a hot topic since the mid 1990's as it has applications in several signal processing areas, such as biomedical, communications and speech signal processing. The mixing of the source signals can be modelled in three distinct ways, as Instantaneous, Anechoic or Convolutive (Echoic) mixtures. Instantaneous mixtures assume that there are similar delays between the sources and sensors, and that there are no reflections/echos. Anechoic mixing simply represents the source to sensor transmission delay, while convolutive mixtures represent the different source to sensor delays and also reverberations (echoes) of the sources. Within the BSS community the cases of instantaneous or convolutive mixtures are mainly researched, and the work in this chapter focuses on the area of convolutive BSS. The related mathematics for these models is provided in Sections 4.3 and 4.4.

The most widely used method for performing BSS is Independent Component Analysis (ICA) which is discussed next.

4.2.1 Independent Component Analysis

ICA originated at the same time as Herault and Jutten [50] proposed their framework for BSS using a neural network, although the technique wasn't formally defined until several years later [56]. ICA is probably the most widely used technique for performing BSS and as such it is not uncommon for the two terms to be used interchangeably by researchers

in the area.

Independent component analysis is a technique for decomposing (unmixing) multidimensional data into components that are as statistically mutually independent from each other as possible. The underlying principle of ICA is that the original source signals must be statistically independent from each other: such that given two sources, s_1 and s_2, s_1 must not convey any information about s_2 and vice-versa. Mathematically, the sources are independent if and only if their joint probability density function can be factorised as:

$$p(s_1, s_2) = \prod_{i=1}^{2} p_i(s_i)$$

i.e. the product of the marginal probability density functions of the original sources $p_1(s_1)$ and $p_2(s_2)$. It is worth mentioning that the technique of principal component analysis (PCA) which was discussed in Chapter 3, is superficially related to ICA. PCA can also be used to obtain underlying information of the dataset so that the dimensionality may be reduced, however PCA only decorrelates the data. While independence implies uncorrelatedness the opposite is not true (except in the Gaussian case), therefore PCA cannot be used to separate independent sources. PCA can also be used to filter out noise in the signal before the application on ICA. Further information, including examples of the difference between PCA and ICA can be found in [24, 56, 64].

4.2.2 Ambiguities of ICA/BSS

Ideally, ICA would find the original sources that were mixed together, however due to certain indeterminacies this is not possible. While ICA

is a very useful technique for revealing hidden factors/components of multivariate data, its solution is subject to two indeterminacies regarding the estimated sources, namely the scaling and (more importantly) the permutation ambiguities:

- Scaling ambiguity:

  It is not possible to determine the variances (energy) of the original independent components. However some researchers, such as in [56], force the estimated sources to have unit variance.

- Permutation problem:

  The order in which the independent components are found cannot be determined since multiplication is commutative. In frequency domain BSS methods, the permutation ambiguity is more serious as the permutation problem now exists at individual frequency bins.

Therefore perfect separation cannot be achieved as the energy and order of the original sources cannot be determined exactly without additional information or assumptions.

4.3 Instantaneous BSS

Traditionally, the BSS problem was solved by modelling the sources as having been mixed instantaneously. N real and zero mean source signals that are mixed in a transmission channel, and then detected by M sensors, can be represented at discrete time index (t) by:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{4.1}$$

where $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]^T$ represents the M observed mixtures, $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^T$ are the original source signals and $[.]^T$ denotes the vector transpose. $\mathbf{A}$ is the scalar matrix channel through which the sources were transmitted. $\mathbf{A}$ is better known as the mixing matrix and is of size M by N. $\mathbf{n}(t)$ denotes the M dimensional noise vector that is independent of the source signals. When the number of sensors is greater than the number sources $M > N$ the problem is said to be over determined, when $M < N$ it is said to be under determined and when $M = N$ it is even determined. To simplify the problem, for the remainder of this chapter it is assumed that there is an equal number of sources and sensors so that $\mathbf{A}$ is now an N by N matrix (i.e. a square matrix) and also that there is no noise present.

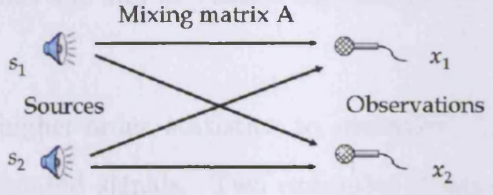


Figure 4.1. Diagram of instantaneous mixing of two signals s_1 and s_2.

The objective of BSS is to recover the original sources $\mathbf{s}(t)$ from the mixtures $\mathbf{x}(t)$ based on the assumption that the source signals are independent. For this purpose, $\mathbf{A}$ is required to be an invertible square matrix such that:

$$\mathbf{s}(t) = \mathbf{A}^{-1}\mathbf{x}(t) \tag{4.2}$$

However, the mixing matrix $\mathbf{A}$ is unknown and so its exact inverse cannot be found. Instead a square matrix $\mathbf{W}$ is found, such that:

$$\hat{\mathbf{s}} = \mathbf{W}\mathbf{x} \tag{4.3}$$

where $\hat{\mathbf{s}}$ is the best estimate of the original sources. Ideally $\mathbf{W} = \mathbf{A}^{-1}$, but in practice we have:

$$\mathbf{WA} = \mathbf{PD} \tag{4.4}$$

where $\mathbf{P}$ is a permutation matrix and $\mathbf{D}$ is a diagonal scaling matrix. The matrices $\mathbf{P}$ and $\mathbf{D}$ refer to the permutation and scale indeterminacies mentioned in section 4.2.2. In the over determined case (non square matrix), $\mathbf{W}$ can be represented as the pseudo inverse of $\mathbf{A}$: $\mathbf{W} = \mathbf{A}^{\dagger}$.

4.3.1 Instantaneous BSS Approaches

Various BSS algorithms have been proposed for solving instantaneous mixtures of signals and can be classed into several groups [24]. Examples are:

- Utilising higher-order statistics to maximise the independence of the estimated signals. Two examples of algorithms that use fourth-order statistics are: the JADE (Joint Approximate Diagonalisation of Eigenmatrices) algorithm proposed by Cardoso [15], and the algorithm proposed by Comon [25].

- Methods exploiting the second order statistics of the signals have also been proposed. The SOBI (Second-Order Blind Identification) algorithm [8] is a popular second order method that exploits the temporal correlation of the signals. Algorithms exploiting the second order statistics are generally less computationally demanding than those methods based on higher-order statistics.

- Information theoretic based methods use mutual information or the entropies of the sources as a measure of independence. A well

known method from this group is the Infomax algorithm proposed by Bell and Sejnowski [7].

The nature of the sources has a significant effect on what assumptions can be made about them. The environment in which they were recorded also influences the choice of the best mixture model. The work in this chapter uses speech signals which are non-stationary. In addition, a cocktail party environment creates a mixing channel which is itself non-stationary due to the speakers moving around and doors opening.

The instantaneous mixture model cannot be used when solving the cocktail party problem due to the fact this it does not account for the cross-talk and echoes that occur between the sources and sensors. In a real world environment the sensors will record convolved mixtures of the original sources due to reflections and reverberations the sources will undergo. The signals recorded in the real world are better modelled using a convolutive model. In the next section the convolutive mixture model is discussed as well as time and frequency domain methods which use this mixture model. Particular attention is given to the frequency domain method of Wang et al. [127] as it forms the basis of the novel BSS algorithm presented in Section 4.6.

4.4 Convolutive BSS

The instantaneous mixture model is insufficient to model the multipath channel encountered in a cocktail party environment; the convolutive model is on the other-hand generally sufficient. Although the instantaneous and convolutive mixture models are different, the independence assumption outlined in Section 4.2.1 remains. Figure 4.2 is an example

of convolutive mixing.

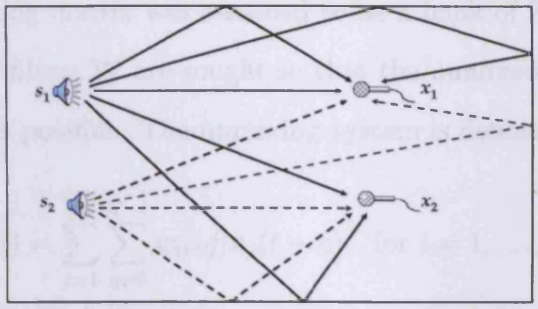


Figure 4.2. Diagram of convolutive mixing of two sources in a two dimensional room-like environment.

For N independent sources $\mathbf{s}$ and N observed signals $\mathbf{x}$, the convolutive model can be written in matrix form as:

$$\mathbf{x}(t) = \mathbf{A}(p) * \mathbf{s}(t-p) \tag{4.5}$$

where $*$ denotes convolution. For the instantaneous case, the elements of the mixing matrix $\mathbf{A}$ were scaling elements, but in the convolutive mixing case the elements represent filter coefficients, often in polynomial form. The convolutive model given in (4.5) can be rewritten as:

$$x_i(t) = \sum_{j=1}^{N} \sum_{p=0}^{P-1} a_{ij}(p) s_j(t-p) \quad \text{for } i = 1, \ldots, N \tag{4.6}$$

where the mixing matrix elements $a_{ij}(p)$ are typically considered to be FIR filter coefficients [63], and P is the filter length.

As stated earlier, the goal of any BSS method is to separate the observed signals $x_i(t)$ back into the original signals $s_i(t)$. Because the system is considered *blind*, neither the original signals nor the mixing matrix are available, one can only make assumptions of their statistics. As with the instantaneous case, an unmixing matrix is sought such that

the best estimate $y_i(t)$ of the original signals $s_i(t)$ can be obtained using it. As the mixing matrix was assumed to be a bank of FIR filters, the unmixing FIR filters **W** are sought so that the unmixed signals are as independent as possible. The unmixing system is defined as:

$$y_i(t) = \sum_{j=1}^{N} \sum_{q=0}^{Q-1} w_{ij}(q) x_j(t-q) \quad \text{for } i = 1, \ldots, N \tag{4.7}$$

where $w_{ij}(q)$ is the i, j^{th} coefficient of the q^{th} unmixing FIR filter of tap length Q, and $y_i(t)$ are the estimated source signals.

Algorithms for convolutive BSS have been suggested in both the time domain and in the frequency domain. However, the lengths of the filters in convolutive mixtures can be in the order of 1000's of samples, dependent upon the size of the room, the reverberation and sampling frequency of the signals. For such long filter lengths, time domain solutions will be computationally expensive, frequency domain solutions offer a more efficient solution. A detailed comparison of time and frequency domain algorithms can be found in [84].

4.5 Frequency Domain Convolutive BSS

Time domain solutions to the blind source separation problem can have a high computational complexity for a real room because of the need for large filter lengths, due to the complex mixing environment. This fact contributed to the popularity of frequency domain solutions, which have the advantage of greater computational efficiency. The convolution in the time domain is reduced to complex multiplication at each frequency bin[1] in the frequency domain, i.e. the convolutive mixture

[1]Frequency bin refers to a range of frequencies over which the signal energy is measured

in the time domain becomes an instantaneous but complex mixture at each frequency bin in the frequency domain.

However, frequency domain solutions are not without their inherent problems. The main problem stems from the ambiguities of the estimated sources discussed in Section 4.2.2. Because each frequency bin is treated as a separate problem, the permutation problem now occurs at each bin and must be solved in such a way that the permutations are consistent across all frequency bins. If the permutations are not consistent then once converted back into the time domain the signals will contain contributions from the other sources, thus negating the advantage of solving the BSS problem in the frequency domain.

To transform from the time domain into the frequency domain, the discrete Fourier transform (DFT) of the signal is taken, and it has also been reported [2] that there exists a trade off between the frame size T of the DFT and the length of the room impulse response P whereby too large or too small a value of T can cause the independence assumption between sources to no longer be valid. Further discussion on the selection of T can be found in [2, 57].

4.5.1 Frequency domain Permutation Problem

Many methods have been suggested to solve the permutation problem and can be grouped into two main categories [88], where the categories can be further subdivided into several groups (including but not limited to):

1. Consistency of the filter coefficients.

    - Exploitation of separation matrix spectrum continuity.

- Using beamforming techniques.

2. Consistency of the spectrum of the recovered signals.

   - Exploiting similarities in the signal envelope.
   - Psychoacoustic filtering.

For consistency of the filter coefficients, prior knowledge about the mixing filters must be known or assumed. Further details of which are given in [88].

4.5.2 Frequency Domain Algorithms

Convolutive BSS solutions can be transformed into the frequency domain using a T point discrete Fourier transform (DFT) of the signal x_i:

$$X_i(\omega, t) = \sum_{\tau=0}^{T-1} x_i(t+\tau) w(\tau) e^{-j2\pi\omega\tau} \tag{4.8}$$

where $w(\tau)$ is a window function (Hamming or Hanning are typical). The frequency domain representation of the time domain mixing model (4.5) can be written as [2, 58, 87, 88, 127]:

$$\mathbf{X}(\omega, t) = \mathbf{A}(\omega)\mathbf{S}(\omega, t) \tag{4.9}$$

where $\mathbf{X}(\omega, t) = [X_1(\omega, t) \ldots X_N(\omega, t)]^T$ and $\mathbf{S}(\omega, t) = [S_1(\omega, t) \ldots S_N(\omega, t)]^T$ are the time-frequency representations of the observed mixtures and source signals respectively at each frequency bin ω. $\mathbf{A}(\omega)$ is again assumed to be an invertible square matrix where the elements $a_{ij}(\omega)$ are the frequency domain representation of the time domain room impulse response $a_{ij}(p)$ of (4.6). Equation (4.9) shows that the convolutive mixture of (4.6) has now been simplified

to be separate complex multiplications at each frequency bin ω. As was noted in the previous section, a backward model can be defined to separate the mixtures into estimates of the original sources:

$$\mathbf{Y}(\omega, t) = \mathbf{W}(\omega)\mathbf{X}(\omega, t) \tag{4.10}$$

where $\mathbf{Y}(\omega, t) = [Y_1(\omega, t) \dots Y_N(\omega, t)]^T$ are the estimates of the original sources $\mathbf{S}(\omega, t)$. The unmixing matrix $\mathbf{W}(\omega)$ is a square matrix containing the elements $w_{ij}(\omega)$ that are the frequency domain representation of the unmixing filter coefficients $w_{ij}(q)$. The aim of frequency domain algorithms is to estimate the parameters of $\mathbf{W}(\omega)$ such that the signals $\mathbf{Y}(\omega, t)$ are statistically mutually independent. The N estimated signals $Y_N(\omega, t)$ are then transformed back into the time domain by applying an inverse DFT. To summarise, frequency domain blind source separation generally consists of:

- Transforming the time domain signals into the frequency domain.
- Finding the optimum unmixing matrix at each frequency bin.
- Separating the mixture of signals.
- Transforming the unmixed signals back to the time domain.

Many algorithms for frequency domain BSS have been proposed based on different criteria for obtaining an optimum unmixing matrix, some of which are discussed by Pedersen et al. [88] in their survey of convolutive BSS algorithms.

4.5.3 Penalty function based Convolutive Blind Source Separation

The basis of the novel audio-visual algorithm described in this chapter is the frequency domain penalty function based algorithm of Wang et al. [125,127], and it is introduced in detail here. The proposed method of Wang et al. jointly diagonalizes the autocorrelation matrices at different times for each frequency bin. The algorithm incorporates a penalty function into the cross-power spectrum based cost function of the algorithm proposed by Parra and Spence [87]. The inclusion of the penalty function not only results in a faster convergence to the optimum solution, but also a better performance in terms of a reduction in the amplitude ambiguity and an SIR improvement at least as good as that of [87]. The autocorrelation matrix of the observed signals at a single frequency bin, at multiple times can be expressed as [87,127]:

$$\bar{R}_X(\omega,t) = \frac{1}{D}\sum_{d=0}^{D-1}\mathbf{X}(\omega,t+dT)\mathbf{X}^H(\omega,t+dT) \qquad (4.11)$$

where T is the size of the DFT, D is the number of intervals used to estimate each autocorrelation matrix and $(.)^H$ denotes Hermitian matrix. This can be rewritten as [87]:

$$\bar{R}_X(\omega,t) = \mathbf{A}(\omega)\Lambda_s(\omega,t)\mathbf{A}^H(\omega) \qquad (4.12)$$

where $\Lambda_s(\omega,t)$ is the diagonal covariance matrix of the source signals. Provided N is large enough, $\Lambda_s(\omega,t)$ can be modelled as a diagonal matrix because of the independence assumption [87]. The independence assumption is based on the principle that as long the signals are non-stationary, $\Lambda_s(\omega,t)$ will change over time and (4.12) will be a function

of time t. Inversely, the backward model can be defined as:

$$\hat{\Lambda}_s(\omega,t) = \mathbf{W}(\omega)[\bar{R}_X(\omega,t)]\mathbf{W}^H(\omega) \tag{4.13}$$

The aim is to find a $\mathbf{W}(\omega)$ that diagonalizes $\bar{R}_X(\omega,t)$ across all frequencies for K time blocks simultaneously for $k = 1 \ldots K$, or similarly $\mathbf{W}(\omega)$ that forces all of the off diagonal elements to be zero.

Parra [87] defined the cost function as:

$$\mathcal{J}(\mathbf{W}(\omega)) = \arg\min_{\mathbf{W}} \sum_{\omega=1}^{T}\sum_{k=1}^{K} \|E(\omega,k)\|_F^2 \tag{4.14}$$

where $\|.\|_F^2$ is the squared Frobenius norm. This is a least squares (LS) estimation problem, where $E(\omega,k)$ is defined as:

$$E(\omega,k) = \mathbf{W}(\omega)[\bar{R}_X(\omega,k)]\mathbf{W}^H(\omega) - \hat{\Lambda}_s(\omega,k) \tag{4.15}$$

Wang [127] incorporates a penalty term into the cost function, defining it as:

$$\mathcal{J}(\mathbf{W}(\omega)) = \arg\min_{\mathbf{W}} \sum_{\omega=1}^{T}\sum_{k=1}^{K} \{\mathbf{J}_M(\mathbf{W}(\omega,k)) + \lambda \mathbf{J}_C(\mathbf{W}(\omega,k))\} \tag{4.16}$$

where λ is a weighting factor and:

$$\mathbf{J}_M(\mathbf{W}(\omega,k)) = \|E(\omega,k)\|_F^2 \tag{4.17}$$

The penalty term $\mathbf{J}_C(\mathbf{W}(\omega,k))$ is a matrix constraint defined as:

$$\mathbf{J}_C(\mathbf{W}(\omega,k)) = \|diag[\mathbf{W}(\omega) - \mathbf{I}]\|_F^2 \tag{4.18}$$

where $diag(.)$ is an operator to zero the off diagonal elements of a matrix. The problem in (4.16) is also a LS estimation problem and the gradients of the cost function in (4.16) can be derived as [127]:

$$\frac{\partial \mathcal{J}}{\partial \mathbf{W}^*(\omega)} = 2\sum_{k=1}^{K}\Big\{2[E(\omega,k)]\mathbf{W}(\omega)[R_X(\omega,k)] + \lambda diag[\mathbf{W}(\omega) - I]\Big\} \tag{4.19}$$

The optimal unmixing matrix $\mathbf{W}(\omega)$ can be obtained using the well known stochastic gradient algorithm, using the gradient formulated in (4.19). The implementation requires a parameter to control the step size of the adaptation of $\mathbf{J}_M(\mathbf{W}(\omega,k))$ and $\mathbf{J}_C(\mathbf{W}(\omega,k))$, the respective forms are [127]:

$$\mu_{J_M}(\omega) = \frac{\alpha}{\sum_{k=1}^{K}\|\mathbf{R}_X(\omega,k)\|_F^2} \tag{4.20}$$

$$\mu_{J_C}(\omega) = \frac{\xi}{\varsigma + \sum_{k=1}^{K}\|\frac{\partial \mathcal{J}_C}{\partial \mathbf{W}^*(\omega)}\|_F} \tag{4.21}$$

where α, ξ and ς are scalar values, adjustable for adaptation.

This penalty-function based frequency domain BSS algorithm represents the state-of-the-art in audio only separation [88]. However, its performance is still limited by the permutation problem and the non stationarity of the environment. The natural next step as suggested by Cherry [21,22] is to incorporate additional information, specifically, visual information of the speaker's face.

4.6 Audio-Visual Speech Separation

Combining audio and visual speech information has previously been discussed in the speech recognition literature [89, 92], while the area of audio-visual BSS is relatively new, and as such there is very little

literature on the subject. An overview of audio-visual BSS is provided in Chapter 2 of this thesis. Outside the area of speech recognition, research on combining audio-visual information has developed in several areas listed below:

- Audio-visual tracking whereby cameras are placed in a controlled environment such as a room, with a number of microphones. A speaker can then be tracked using either audio or visual information, or a combination of both [19, 20]. *Smart rooms* are an extension of this in which not only are the speakers tracked but also a chosen speaker is identified using audio-visual information [13, 43, 80].

- Speech enhancement techniques that use audio-visual information to improve the intelligibility of speech in a noisy environment [36, 46, 47, 53, 70].

- Audio-visual blind source separation algorithms that incorporate visual information into a BSS framework to aid finding the optimum unmixing matrix [3, 112, 114, 126] or to process the output of a BSS algorithm to solve the permutation problem [96, 97, 100].

Initial research focused on incorporating visual information into instantaneous BSS frameworks [94, 111, 114], while more recently convolutive mixtures of speech have been considered [96–99, 126]. Rivet et al. [96] proposed an audio-visual method that exploits the coherence of audio-visual speech to solve the permutation problem. A joint probability model of audio-visual features is formed and used to post process the estimated signals obtained from the BSS algorithm of Pham et al. [90]. The visual features were the lip height and width, and the

audio features were spectral characteristics. Simultaneously, Wang et al. [126] proposed a method that also exploits the coherence of audio-visual speech, again using a joint audio-visual probability model, except that their visual feature was an AAM of the lips and the audio data were MFCCs. The joint audio-visual model was used as the penalty function in their BSS algorithm [127] discussed in Section 4.5.2. More recently, Rivet et al. [97] have used the output of a visual voice activity detector to post process the output of their BSS algorithm to solve the permutation problem. In the following section a novel audio-visual BSS algorithm is outlined and its performance is assessed with a comparison to an audio only BSS algorithm.

4.6.1 Video Assisted Blind Source Separation

The frequency domain penalty function BSS algorithm described in the previous section is expanded upon here to incorporate a joint audio-visual speech model. The aim of the proposed approach is to maximise the coherence between a set of visual features $\mathbf{v}$ and a set of audio features $\mathbf{a}$ to provide a criterion for controlling the learning rate of the second order frequency domain BSS algorithm of Wang et al. [127] discussed in Section 4.5.3.

Exploiting the coherence of audio-visual speech has been discussed before [96,126]. Wang et al. [126] proposed a method that incorporates a joint audio-visual model into the penalty function framework of [127] (see Section 4.5.3) by substituting the penalty function $\mathbf{J}_C(\mathbf{W}(\omega, k))$ (Equation (4.16)) with the output of the joint audio-visual probability model. The audio component is generated at each iteration and is therefore a function of $\mathbf{W}(\omega, k)$. The reported effect was an increase

in separation performance due to a reduction of the permutation effect [126].

To extract the speaker of interest in this work the audio-visual coherence is also exploited. A model of the visual features of the speaker's lips over a set of video frames is obtained using the AAM method proposed by Cootes et al. [29]. The corresponding audio features are extracted from the audio signals using MFCCs. An overview of both AAMs and MFCCs can be found in Chapter 3 of this thesis. It has been previously mentioned that there are several alternative techniques to extract the visual features, AAMs were chosen as they can model visual features in high detail using a reduced dimension dataset, while MFCCs were chosen for their ability to mimic the non-linear frequency resolution of the human ear. The set of N_a audio feature vectors $\mathbf{a}_s = [a_{s1} \dots a_{sN_a}]^T$ and N_v visual feature vectors $\mathbf{v}_s = [v_{s1} \dots v_{sN_v}]^T$, where N_a and N_v are respectively the number of audio and visual features, are concatenated to provide joint audio-visual feature vectors:

$$\mathbf{u}_s = [\mathbf{v}_s^T, \mathbf{a}_s^T] \tag{4.22}$$

where subscript s denotes speaker. It should be noted that not all of the appearance parameters $\mathbf{c}_s$ for speaker s obtained from the AAM are used. Instead a dimensionally reduced vector $\mathbf{v}_s$ is used. It is obtained by performing PCA on the final set of appearance parameters $\mathbf{c}_s$, and retaining a percentage of the total energy, as discussed in Chapter 3. The probability distribution of the set of vectors $\mathbf{u}_s$ is modelled using either a GMM, or an HMM when their time dynamics are considered. For completeness of the experiments, the results of using the GMM or HMM to model the audio-visual coherence are compared. Before

the audio-visual BSS algorithm can be implemented, models of training data must be built. To train the models, the following steps are followed:

- Calculate the MFCCs from the training speech sequence to obtain the audio feature $\mathbf{a}_s$, as described in Chapter 3.
- Obtain the appearance parameters $\mathbf{c}_s$, and apply PCA to obtain the reduced visual features $\mathbf{v}_s$.
- Concatenate the audio and visual features to obtain the joint audio-visual features $\mathbf{u}_s$
- Train the GMM or HMM to obtain the model parameters as described in Chapter 3.

Next the audio-visual information must be integrated into a BSS algorithm. The BSS algorithm used is the penalty function based frequency domain BSS algorithm [127]. In the original algorithm of [127] the learning rate $\mu_{J_c}(\omega)$ is controlled by a function of the penalty value at that iteration of the algorithm. In the current work it is controlled by a function of the audio-visual coherence.

$$\mu_{Jc}(\omega) = \frac{\xi}{\zeta + f'(P_{av})} \tag{4.23}$$

where P_{av} is the joint audio-visual probability (a measure of the coherence [126]). The values for ξ and ζ are empirically chosen constants so that the steady state separation performance of all the algorithms is identical and f' is a logarithmic mapping of the model output. It is necessary to calculate P_{av} using a different method when using a GMM or HMM to model the training data. For the case of a GMM:

$$p(\mathbf{u}_s) = \sum_{i=1}^{K} \mathbf{w}_i \frac{exp\{-1/2(\mathbf{u}_s - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{u}_s - \boldsymbol{\mu}_i)\}}{\sqrt{(2\pi)^K \mid \boldsymbol{\Sigma}_i \mid}} \tag{4.24}$$

where $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$, $\mathbf{w}_i$ and K are the mean vector, covariance matrix, kernel weights and the number of Gaussian kernels respectively. P_{av} is then found by summing the log of (4.24) and for the HMM the log probabilities were calculated using the method in [93].

The AVSS algorithm is performed as follows:

1. Estimate the source signals from the current estimate of the unmixing matrix $\mathbf{W}$ and calculate the audio features (MFCCs).

2. Concatenate the audio feature with the visual features to form a new joint audio-visual feature.

3. Calculate the audio-visual coherence output P_{av} using either the GMM or HMM model.

4. Calculate a new value for (4.23).

5. Update the unmixing matrix $\mathbf{W}$ until converged.

The algorithm is said to have converged when the change in value of μ_{Jc} falls below a chosen threshold.

4.7 Simulations

The statistical models (GMM, HMM) were trained on the audio-visual features extracted from a video of a subject in an office environment with low level acoustic noise and artificial front on lighting. Audio and video were recorded for two speakers, the video data were captured

using a digital video camera with a resolution of 720 by 576 pixels, at 25fps and the audio was captured using a directional microphone, sampled at 32KHz, 16-bit mono. For both speakers the lip region in the video was tracked using an AAM using 17 landmarks to provide a joint model of shape and texture information with 10 appearance parameters per frame. The speech features were extracted using Mel-cepstral analysis with a 20ms Hamming window, providing 12 MFCCs per frame. The appearance parameters (40ms) were then interpolated in order to retain one-to-one correspondence with the audio parameters (20ms). The number of Gaussian kernels for the GMM and the number of states for the HMM were set to 10 and the audio-visual feature space had 22 dimensions, 10 video plus 12 audio and remained the same size during separation.

The experiments were conducted on each speaker separately. Only 2×2 mixtures (2 speakers, 2 microphones) were considered, where the speech signal of the speaker present in the video was artificially mixed with another speaker in a convolutive system with 9 taps. The mixing filters are the same as those used in the experiments by Wang [127]. The values of ξ and ζ for the audio only experiments were the same for both speakers, 0.2 and 0.05 respectively. For the audio-visual BSS algorithm, the values differed for each statistical model used. For speaker one, the values when using the GMM were $\xi = 28$, $\zeta = 3.1 \times 10^4$, and for the HMM $\xi = 212$, $\zeta = 6.45 \times 10^4$. The values for speaker two were $\xi = 6$, $\zeta = 3.13 \times 10^4$ for the GMM and $\xi = 105$, $\zeta = 7.5 \times 10^4$ for the HMM. It should be noted that the performance of the method is not sensitive to minor changes in the values for ξ and ζ.

Figures 4.3 and 4.4 show the results of the simulations. It can

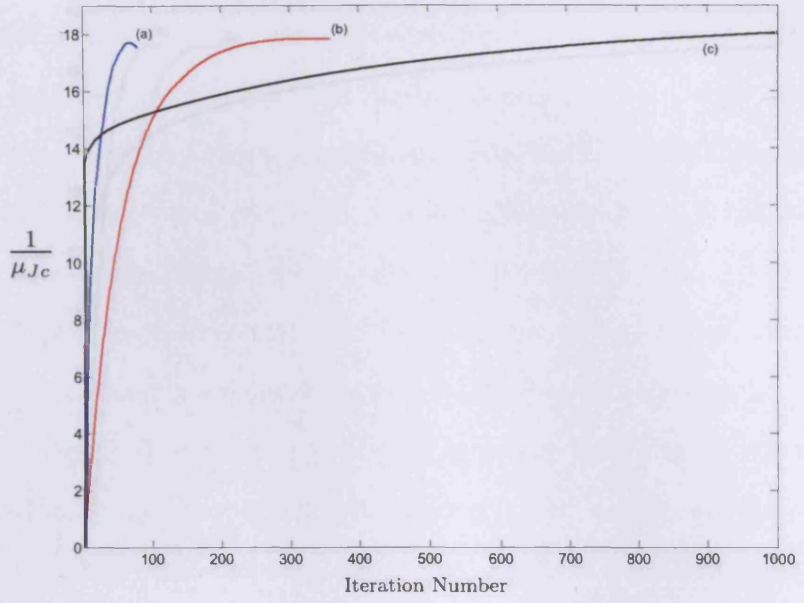


Figure 4.3. Comparison of learning rate for speaker one using (a)HMM, (b)GMM, (c)audio only to control the step size.

be seen that the audio-visual model requires fewer iterations to converge (the end of the curve denotes convergence of the BSS algorithm), which is very likely to be useful in a non-stationary environment for example when the speaker is moving. Furthermore, the advantage of using an HMM compared to a GMM was also observed. This could be contributed to the fact that HMMs are better able to capture the coarticulation[2] of speech. The quality of the reconstructed sources was judged subjectively by listening tests to be essentially identical for the three methods.

[2]Coarticulation is defined to be the interaction of speech articulators (lips, jaw, tongue etc)over time, during the production of speech [51].

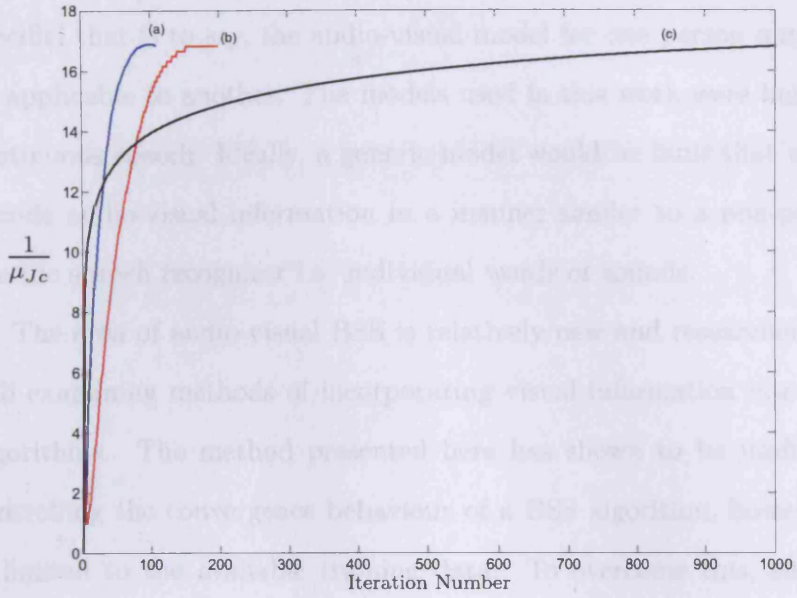


Figure 4.4. Comparison of learning rate for speaker two using (a)HMM, (b)GMM, (c)audio only to control the step size.

4.8 Conclusion

The results of the proposed technique shown here are promising. The experiment indicates that by combining audio and visual information the convergence behaviour of a BSS algorithm can be improved. The results of the proposed method are also compared to the BSS algorithm which uses raw audio information alone (i.e. no model) to control the convergence. Figures 4.3 and 4.4 indicate that a significant improvement in convergence behaviour can be obtained. It was shown that when using a statistical model of audio-visual coherence to control the convergence behaviour, the number of iterations required for convergence was reduced from over 1000 to only a few hundred, depending on which statistical model (HMM or GMM) was used. As was stated previously, the outputs were judged subjectively using listening tests and

found to be almost identical. Yet, the audio-visual models are person specific; that is to say, the audio-visual model for one person may not be applicable to another. The models used in this work were built on continuous speech. Ideally, a generic model would be built that would encode audio-visual information in a manner similar to a non-person specific speech recognizer i.e. individual words or sounds.

The area of audio-visual BSS is relatively new and researchers are still examining methods of incorporating visual information into BSS algorithms. The method presented here has shown to be useful for controlling the convergence behaviour of a BSS algorithm, however it is limited to the available training data. To overcome this, simpler video based speech cues are sought than the ones used here, and this is the topic of the next chapter.

Chapter 5

VISUAL-VOICE ACTIVITY DETECTION USING AAM

5.1 Introduction

In Chapter 4, a novel audio-visual BSS method was presented. The outcome of simulations showed an advantage in using a statistical audio-visual speech model over raw audio data alone. However, the algorithm had a high computation time. This led to the investigation of an alternative method to use visual information to improve the performance of BSS methods.

Sodoyer et al. [113] proposed a visual voice activity detector (V-VAD) to classify a speaker as speaking (active) or not speaking (inactive) using information of the speaker's lips. This was subsequently used in the BSS algorithm of Rivet et al. [100] to help solve the permutation problem inherent to BSS. However, the V-VAD proposed by Sodoyer obtains shape information of the speaker's lips using a chroma-key system, but this is impractical in a natural environment as it requires blue make-up to be applied to the speaker's lips.

Therefore, in this chapter a novel V-VAD is proposed where the visual descriptor is the set of AAM parameters of the speaker's lips. To

detect voice activity it is necessary only to distinguish between speech and non-speech lip motion. Note that the non-speech periods are not defined as silence because while the speaker may be silent, there could still be motion of the lips, e.g. a smile. Thus the two categories are defined as speech and non-speech related lip motion. For modelling the dynamics of the lip appearance parameters during non-speech phases a Hidden Markov Model (HMM) is used. The justification for this is that the experiments in Chapter 4 have provided evidence that using an HMM to model speech features performed better than using a GMM. Once the HMM of some training data is built, the probability of new (test) data belonging to that model is calculated. If this probability is above some threshold value (i.e. a high probability) then it is classed as non-speech data, if it is below this threshold (low probability) then it is classed as speech data. Simulations show that a high rate of correct classification can be achieved. Moreover, it is shown that the proposed method has a similar performance to a retinal filter based method [4,95].

The organization of this chapter is as follows: in Section 5.2, some background on previous V-VADs is provided. The proposed V-VAD is presented in Section 5.2.1 with simulation results given in Section 5.4. The results are discussed in Section 5.5 and Section 5.6 concludes this work.

5.2 Background

Voice activity detectors (VADs) are used to detect the presence or absence of speech in an acoustic environment. As VAD methods traditionally rely on acoustic information, their accuracy is highly dependent on the acoustic environment (e.g. the presence of competitive sources or

highly non stationary noise). However, as speech is a bi-modal signal (with both audio and visual aspects), being able to see a speaker's face, especially the lips, can provide additional information regarding speech activity. The most visible aspect of speech production is the movement of lips; in the past it has been shown that there is a high coherence between the speaker's lips and the resulting acoustic signal [130]. This characteristic is regularly used to improve speech recognition [92] and speech enhancement [46]; as well as more recently in blind speech separation [3,96]. Recently, VAD based on visual data as opposed to audio data has been developed [59, 67, 113, 122]. Visual voice activity detection (V-VAD) has an advantage over audio based VAD in that it is not susceptible to the problems associated with the acoustic environment (e.g. noise, simultaneous speakers and reverberations).

There have been several approaches to V-VAD. Iyengar and Neti [59] developed a V-VAD which was used for deciding a person's intent to speak. Their V-VAD uses a head pose and lip motion detector to switch a microphone on and off in a speech recognition system. This is achieved by extracting the mouth region and calculating the average illumination in the region. The idea being that an open mouth will have a lower illumination than a closed mouth, and a decision is made by comparing the illumination to a threshold value. The drawback of this method is that it does not distinguish between speech and non-speech movement of the lips. Liu and Wang [67] proposed a V-VAD that used statistical models of speech and non-speech activities. Visual information relating to non-speech activity was modelled using a single Gaussian distribution, and visual information relating to speech activity was modelled using multiple Gaussian distributions. New data

were classified on the basis of a likelihood calculation. However, their method does not model the dynamics of lip motion and they assume the lips are essentially stationary during non speech periods. More recently, Sodoyer et al. [113] proposed a method for V-VAD that uses temporal smoothing of dynamical lip motion. Unfortunately, their method relies on a high computational cost chroma-key system, which is impractical for a natural environment. An audio-visual solution was recently proposed [122] that uses a correlation between audio vowel frequencies and the shape of the lips for those respective vowels. In particular, the roundness of the lips shape was used to aid with ambiguous decisions. However, they do not mention if their data contains lip motion during non speech periods, nor do they provide the percentage of true/false classifications.

5.2.1 V-VAD using Appearance Parameters

Chapter 3 discussed how to obtain a set of appearance parameters from an image, and those readers unfamiliar with the subject are referred to this chapter.

5.3 Dynamic Modelling of Appearance Parameters for V-VAD

Typically, audio based solutions for VAD detect the presence of speech, however in this work silence periods are specifically sought, the reason for this will be discussed in the following section. Much of the previous work on V-VAD has failed to account for motion of the lips during non speech periods. While this motion is not complex per se, motion of the lips that is more than just small movements is defined to be complex. For example, where the lips are more or less static is described as little

to no movement, but for something like a smile or biting/licking the lips, the movements are defined to be complex. It is this complex motion that can cause ambiguities in a V-VAD, when classifying lip movement as speech or not. One of the novel aspects of the work contained in this chapter is accounting for and dealing with complex lip motion to reduce its effect on lip motion classification.

5.3.1 V-VAD using an HMM

Given a set of appearance parameters $\mathbf{c}(j)_{1\leq j\leq T, j\in\mathbb{N}}$ sampled over time, dynamical changes can be modelled over time using an HMM. HMMs have been used extensively in the past to model the dynamics of speech (e.g. [93]) and more recently to model joint audio-visual features [3]. For training an HMM, the standard Baum-Welch algorithm [93] is used, which provides the model $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$, where π is a vector of the initial state probabilities, $\mathbf{A}$ is the state transition matrix and $\mathbf{B}$ is the state probability distribution.

The task is to determine if a person is speaking or silent in a given period of time, *i.e.* in a given sequence of appearance parameters. For this the likelihood that a sequence of appearance parameters is generated is calculated by the HMM λ . The likelihood $P(\mathbf{O}|\lambda)$ for a sequence of consecutive frames $\mathbf{O}$ ($\mathbf{O} = \mathbf{c}(k)\ldots\mathbf{c}(l)$) is calculated, where the number of frames between k and l is unchanged for all sequences. Each observation $\mathbf{O}$ generates an associated likelihood value P. As mentioned earlier, detecting the presence of non-speech lips motion is the focus. To make this possible, an HMM was built using examples of lips motion resulting from non-speech activity. Early experiments showed it was necessary to smooth the likelihood values to remove mi-

nor false detections. For this purpose an averaging filter is used and the filtered likelihood is denoted as P_f. The final step is to classify the observation $\mathbf{O}$ by comparing the filtered probability P_f to a threshold value β (where the value of β is found experimentally). If $P_f < \beta$ then the current sequence of frames is classified as speech, if $P_f > \beta$ then the sequence of frames is classified as non-speech.

The proposed method for V-VAD can be described with the following steps:

- Obtain appearance parameters using the method described in Chapter 3.
- Build an HMM (using training data) on non-speech appearance parameters.
- Calculate $P(\mathbf{O}|\lambda)$ using unseen (new) appearance parameters.
- Classify the observation as speech or non-speech data by comparing to a threshold value β.

5.4 V-VAD Simulations

In this section the results of classifying visual speech data as speech or non-speech using the method described above are presented. The data collected are also described, and the visual features used to conduct the numerical evaluation. The reason for recording a new database is that no existing audio-visual database where there is significant lip motion during silence sections of continuous speech is available. In addition, the AAM based V-VAD is compared to the retinal filter based approach, proposed by Rivet et al. [95].

The results of the V-VAD are given in the form of Receiver Operating Characteristic (ROC) curves. They represent the ratio of correct silence detection to false silence detection. The correct silence detection (CSD) is defined as the ratio between the number of actual silence frames correctly detected as silence ($N_{\text{Sil|Sil}}$) and the number of actual silence frames (N_{Sil}):

$$\text{CSD} = \frac{N_{\text{Sil|Sil}}}{N_{\text{Sil}}}. \tag{5.1}$$

The false silence detection (FSD) is defined as the ratio between the number of actual speech frames detected as silence ($N_{\text{Sil|Spe}}$) and the number of actual speech frames (N_{Spe}):

$$\text{FSD} = \frac{N_{\text{Sil|Spe}}}{N_{\text{Spe}}}. \tag{5.2}$$

The ROC curve was then produced by varying the threshold β between the maximum and the minimum values of P_f and calculating the CSD and FSD at each value.

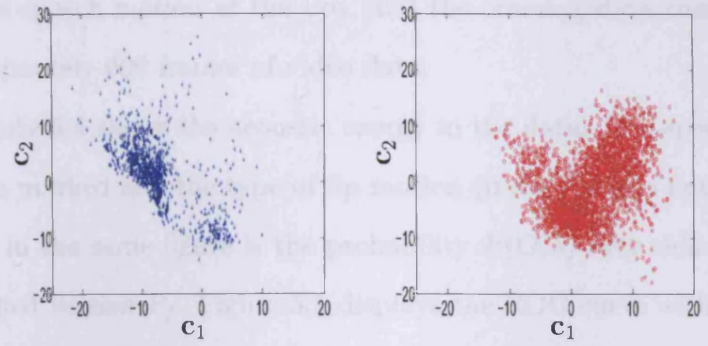
5.4.1 Audio-Visual Corpus

The dataset collected for use in these experiments consists of audio and video recordings of two speakers, one male, one female, reciting a poem in English. Each recording is approximately 2.5 minutes in length (both audio and video), with the video recorded at 30fps and the audio at 44.1KHz, and where the resolution of each video frame is 640x480 pixels. While the speakers are male and female, the results of the male speaker are not specific to all male speakers and likewise for the female speaker. The terms are used only to distinguish between the two speakers. To rigourously test the capabilities of the V-VAD,

the speaker's lips during the silence phases, which in this case were breaks between verses of the poem, were not always stationary. In fact, as previously mentioned, complex lip movements were performed such as smiling, biting or licking lips. As stated earlier, people naturally perform such movements during silence phases. Example images of the female speaker from the dataset are shown in Figure. 5.1 and the male shown in Figure 5.2.

Figure 5.1. Frames from the dataset of the female speaker saying the word 'much'. Frames read, top to bottom, left to right.

Figure 5.2. Frames from the dataset of the male speaker saying the word 'about'. Frames read, top to bottom, left to right.



(a) Non-speech (b) Speech

Figure 5.3. Distribution of the first two dimensions $\mathbf{c}_1$ and $\mathbf{c}_2$ (relating to the two highest eigenvalues) of non-speech (Figure 5.3(a)) and speech (Figure 5.3(b)) appearance parameters.

5.4.2 Visual Features

The visual features produced using the active appearance model for both speakers are of similar dimensionality, and in this case produce 400 dimensional vectors $\mathbf{c}(j)$, which are too large for numerical calculations. To reduce the dimensionality, $\mathbf{c}$ is only composed of the parameters associated with the N most important eigenvalues. However, there is a large overlap between the speech and non speech features (Fig. 5.3). Thus, N is a trade off between the size of the appearance parameter vector and the ability to separate speech and non-speech events. The experiments will detail the number of eigenvalues N retained for the experiment in question.

Results of V-VAD with the Female Speaker

As stated above, the vectors $\mathbf{c}(j)$ have 400 dimensions and this is reduced to retain a percentage of the total energy. In this experiment the first ten eigenvectors $N = 10$ were retained, which contained 75% of the original appearance energy. The HMM was trained solely us-

ing non-speech motion of the lips, and the training data consisted of approximately 600 frames of video data.

Figure 5.4 shows the acoustic energy in the data. Non-speech periods are marked and the type of lip motion (if any) is also noted. Also shown in the same figure is the probability $P(\mathbf{O}|\lambda)$ over time and the smoothed version P_f. Figure 5.5 displays the ROC curve with varying values of number of consecutive frames.

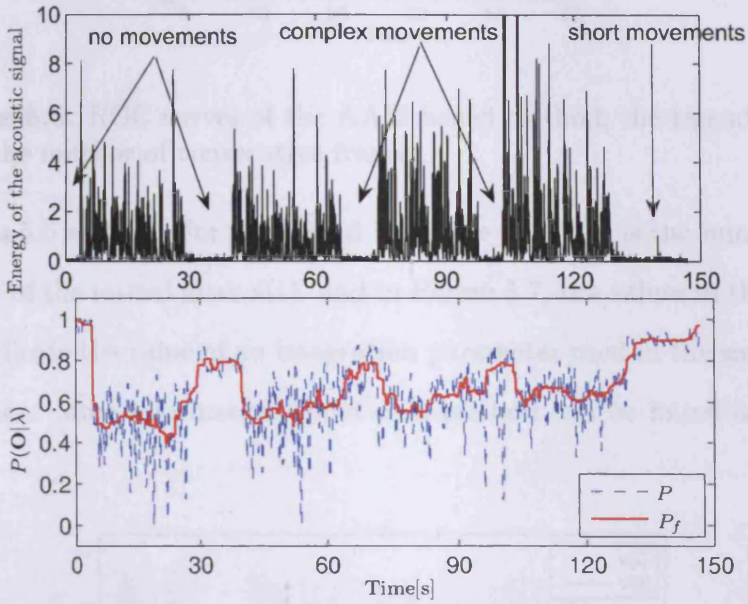


Figure 5.4. Temporal results. From top to bottom : energy of the acoustic signal, silence probability obtained from the AAM based method.

Simulations using different size observation windows are provided in Figure 5.5. The optimum number of frames in an observation $\mathbf{O}$ was found to be 10 consecutive frames. Poor results are obtained with observation sizes of 6 and 15 frames. However, the observation size of 15 frames does provide a similar response to the optimum observation size for an FSD $\leq 5\%$.

The results of the retinal filter based method [95] are also provided in

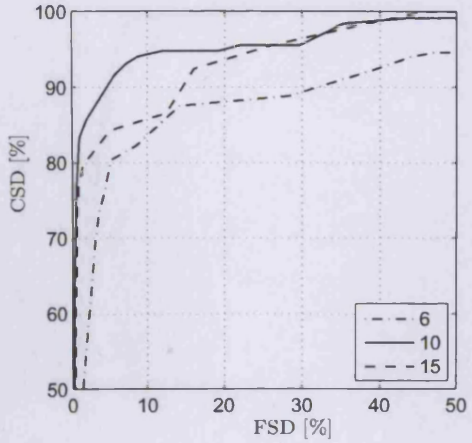


Figure 5.5. ROC curves of the AAM based method, the legend indicates the number of consecutive frames.

Figures 5.6 and 5.7. For the legend in Figure 5.6, V(t) is the smoothed output of the retinal filter v(t), and in Figure 5.7, the values in the legend indicate the value of an integration parameter used in the smoothing filter. More information about this method can be found in [95].

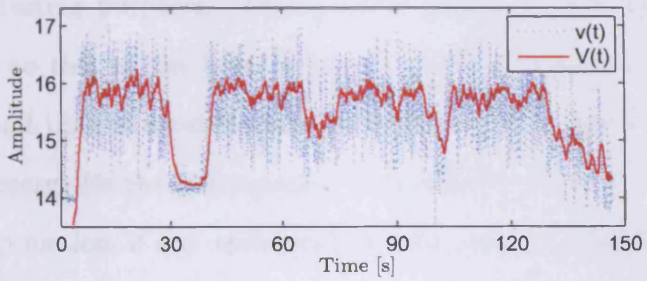


Figure 5.6. Original (v(t)) and smoothed (V(t)) filter outputs of the retinal filter based approach.

Results of V-VAD with the Male Speaker

Similar experiments were conducted on the data of the male speaker and the results of the AAM based method are given in Figures 5.9 and 5.10.

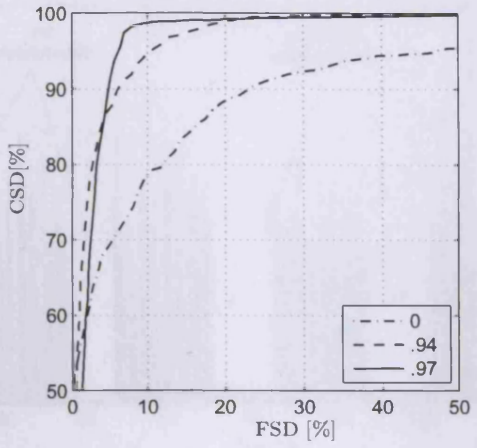


Figure 5.7. Retinal filtering based method, the legend indicates the integration parameter.

Again the original dimensionality of the appearance parameters was reduced to retain the first ten eigenvectors ($N = 10$), which contain 80% of the original appearance energy. The HMM was again trained using solely non-speech motion of the lips, and the training data consisted of approximately 1000 frames of video data, which left 4000 frames of data for testing purposes. The acoustic energy of the male speaker is similar to that of the female speaker given in Figure 5.4, and the number and time of occurrence of the silence periods are similar. The acoustic energy for the male speaker is provided in Figure 5.8. Also the type of lip motion, if any, contained in each silence period is the same as noted in Figure 5.4.

Figure 5.9 shows the output P of the AAM based V-VAD, along with the smoothed version P_f. The ROC curve for the AAM method is given in Figure 5.10. As with the female speaker, the best results were obtained using an observation $\mathbf{O}$ of 10 consecutive frames.

Figures 5.11 and 5.12 show the results of the retinal filter method

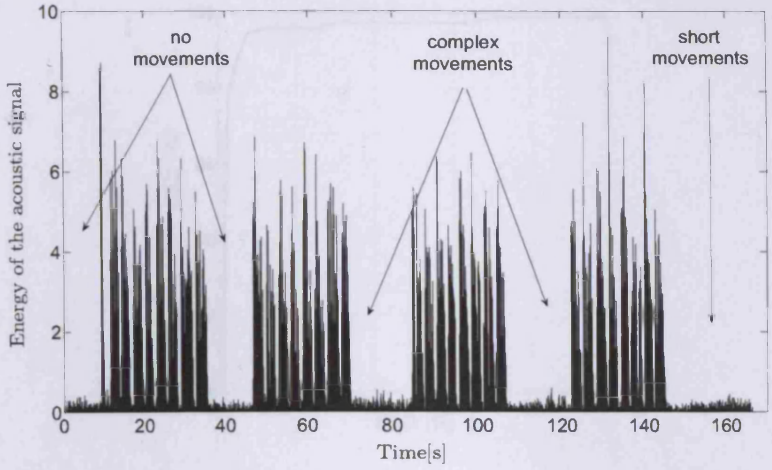


Figure 5.8. Energy of the acoustic signal for the male speaker.

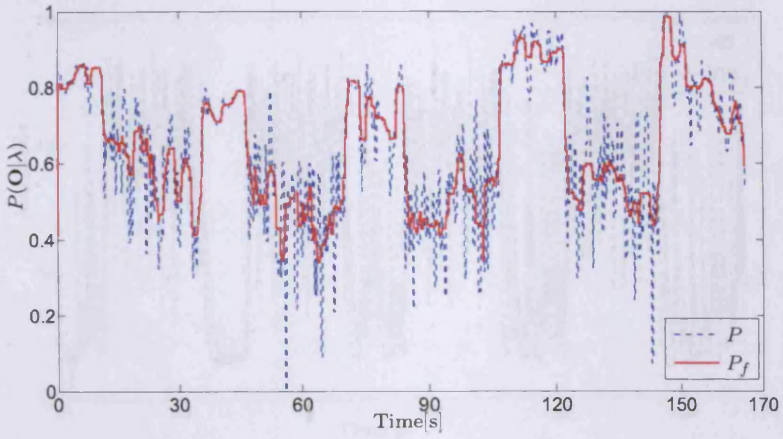


Figure 5.9. Silence probability obtained from the AAM based method for the Male speaker.

[4, 95]. For the legend in Figure 5.11, V(t) is the smoothed output of the retinal filter v(t) and the value of the integration parameter used is $h = 0.97$.

Comparing the results of the proposed method (Figures 5.5 and 5.10) with those of the retinal filter (Figures 5.7 and 5.12), it can be seen that for the Female speaker both methods achieve a high CSD (90%) for a low FSD (5%), and for the Male speaker a CSD of 97%

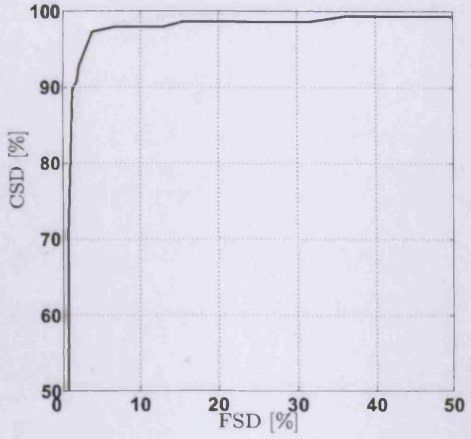


Figure 5.10. ROC curve of the AAM based method, with an observation size of ten frames.

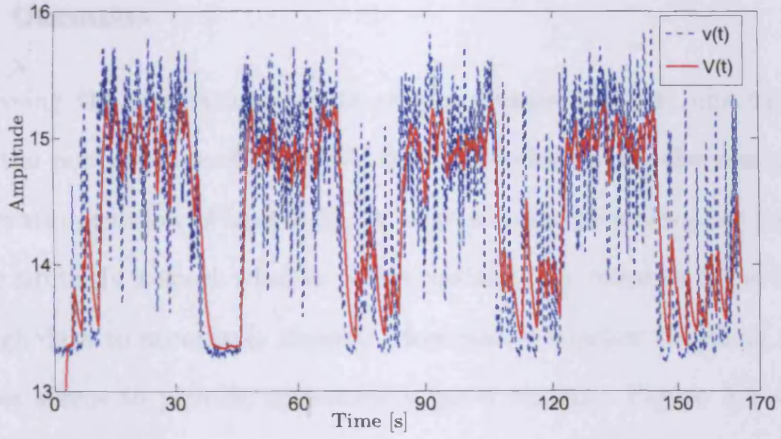


Figure 5.11. Original (v(t)) and smoothed (V(t)) filter outputs of the retinal filter based approach of the Male speaker.

for an FSD of 5% can be obtained. What can also be seen is that the proposed method provides a higher CSD for an FSD range of $0\% \rightarrow 5\%$ for the Female speaker. The proposed method also shows an advantage for the Male speaker, albeit a minor one, for a CSD of 90% the AAM method achieves an FSD of 1% while the retinal filter method achieves an FSD of 2%.

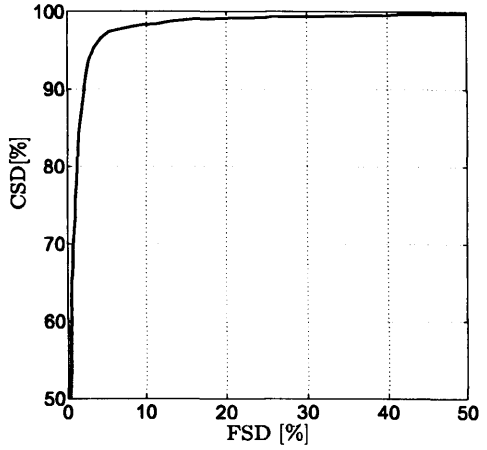


Figure 5.12. ROC curve of the retinal filtering based method for the Male speaker.

5.5 Discussion

Analysing the simulation results of the female speaker, one can see that the performance of the V-VAD is dependent upon the size of the observation window (Figure 5.5). A large window provides poor results, while similarly a short window is not satisfactory because there is not enough data to accurately classify. However, a window length of 10-12 frames seems to provide consistently good results. Figure 5.4 shows that the method is able to identify accurately the periods containing little to no motion, such as between 30s and 40s. The method is also able to identify the periods containing complex lip motion, even though this is a more difficult task, thus being able to achieve a CSD of 90% for an FSD of 5% for the Female speaker and a CSD of 97% for an FSD of 5%.

5.6 Conclusion

The results of the proposed method were compared to another existing V-VAD [95] and both techniques were shown to have similar performance. However the proposed method has an advantage over the retinal filter method: both methods use a window of observations, and for the retinal filter approach the size of this window determines the smallest silence period it is able to detect. The same can be said for the proposed method; because a statistical model is used, a smaller observation window can be used, thus smaller non-speech periods can be detected. For the simulations presented here, the number of consecutive frames used in the retinal filter method was 20 (in [95] this is denoted as T_F), while the proposed method works best with a window size of around 10-12 frames.

However, there is a drawback to using appearance parameters as the visual feature. The HMM is not easily applied to a speaker who is not included in the training data. Work on building a generic set of appearance parameters has been published [48] but limited success has been reported. What is needed is a visual speech representation method that provides a good representation of the lips motion but is also not specific to one (or a few) person (people). The next chapter in this thesis suggests such a method.

Chapter 6

VOICE ACTIVITY DETECTION USING COMPLEX WAVELETS

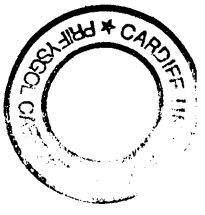
6.1 Introduction

The work described in this chapter is a natural progression on the results presented in the previous chapter. In Chapter 5 a novel visual voice activity detector (V-VAD) was presented that is able to detect the silence periods of speech using only visual information of the speaker's mouth. The method described in Chapter 5 was based on the use of Active Appearance Models (AAMs) to obtain shape and texture information of the movement of a speaker's lips over time. Hidden Markov Models (HMMs) were used to decide if the speaker was active or not. The disadvantage with the AAM approach is that it is not speaker independent. That is to say, the data used to build the HMM does not easily lend itself to another person from outside that data set, i.e. the AAM is more or less person specific. Building a model based on the dataset obtained from several people does partly overcome this problem but at the cost of lowering the accuracy of classification. In addition,

building a model like this is time consuming, and the tracking of the lips is not easy due to the speed with which the lower lip moves. However, the authors of [42] presented a method that reportedly overcomes the tracking problem. The work presented in this chapter addresses the issue of speaker dependent models encountered in the previous chapter, and negates the tracking issue by using a motion flow field to represent the motion contained in the mouth region. The dynamics of the motion field are captured using an HMM, and used to classify the speaker's activity.

In this chapter the above issues are addressed by the use of wavelets to estimate the motion of the lips. The wavelets are not person specific and can be used instead of the AAM to obtain the lip motion. To obtain the motion field of a speakers lip region, the complex discrete wavelet transform (CDWT) motion estimation algorithm proposed by Magarey and Kingsbury [74] is used. The advantage of the CDWT over the standard discrete wavelet transform (DWT) is that the CDWT filters provide better directional distinction. The standard DWT has only 3 directions, vertical, horizontal and diagonal, where as the CDWT has 6 discrete directions oriented at $\pm 15°$, $\pm 45°$ and $\pm 75°$. The phase information provided by the CDWT together with the additional directionality enhances the lip contour, thus allowing the motion of the lips to be found more accurately.

The remainder of the chapter is set out as follows: the following section provides an introduction to wavelets, Sections 6.3 and 6.4 contain an overview of the CDWT and the motion estimation algorithm proposed by Magarey and Kingsbury [74]. The method for voice activity detection is given in Section 6.5 together with results of simulations



and the chapter is concluded in Section 6.7.

6.2 The Wavelet Transform

The wavelet transform is based on analysing a signal with a 'mother' wavelet, and translated and scaled versions of that wavelet. They are useful in analysing non-periodic or non-stationary signals and the practical implementation is closely related to filterbank theory [117]. Wavelet transforms (WTs) are most often compared with the Short-Term Fourier transform (STFT). They were developed to address the need to find an alternative to Fourier based transforms (FT) in analysing non-periodic/stationary signals because the FT cannot localise a signal in both time and frequency simultaneously. When an FT is applied to non-stationary signals the frequency content is described in a similar manner as for stationary signals. That is to say, the frequencies contained in the signal are shown to occur at all times. While this is true for stationary signals, the same cannot be said for non-stationary signals. The STFT was developed to overcome the localisation issue. The STFT uses a (sliding) time localised window ($w(t)$) of fixed length to analyse a signal $x(t)$ for a 'short' period of time, for which during that time the signal is considered stationary.

$$\text{STFT}(\tau, \omega) = \sum_{t=-\infty}^{\infty} x(t)w(t-\tau)e^{-j\omega t} \tag{6.1}$$

However, the STFT has its own limitation. The window is of fixed length, and because of this the time frequency resolution is also fixed for the length of the signal.

The continuous wavelet transform (CWT) decomposes a signal $x(r)$

using a set of basis functions that are obtained from scaling and translating a *'mother wavelet'* ψ, where r is used to denote continuous time:

$$CWT_x(s,\tau) = \int_{-\infty}^{+\infty} x(r)\psi^*_{s,\tau}(r)dr \tag{6.2}$$

where

$$\psi_{s,\tau}(r) = \frac{1}{\sqrt{s}}\psi\left(\frac{r-\tau}{s}\right) \tag{6.3}$$

where s and τ are the scaling and translating parameters respectively and (.*) denotes complex conjugate. Continuous values of s and τ mean that the CWT is a very redundant representation of the signal $x(r)$. Constraining these to discrete values to form the DWT reduces this redundancy. Unlike the STFT, which has a fixed window length, the wavelet transform uses a varying window length to analyse a signal. This results in a time frequency decomposition where the maximum frequency resolution has the minimum time resolution, and vice versa. This is illustrated in Figure 6.1.

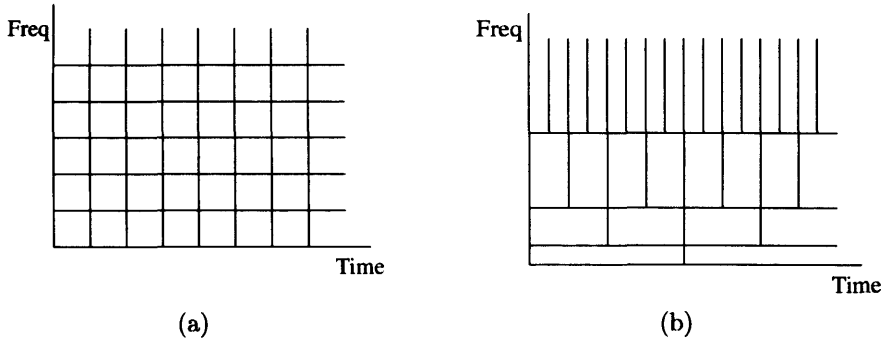


Figure 6.1. Time-frequency cells of the (a) STFT and (b)wavelet transform.

The purpose of using wavelets in this work is to analyse digital images. Therefore a discrete wavelet transform is required which can be applied to such signals.

6.3 Complex Discrete Wavelet Transform

Mallat [75] discussed how a practical implementation of a discrete wavelet transform (DWT) can be realised, using a multilayer filterbank containing high and low pass filters shown in Figure 6.2.

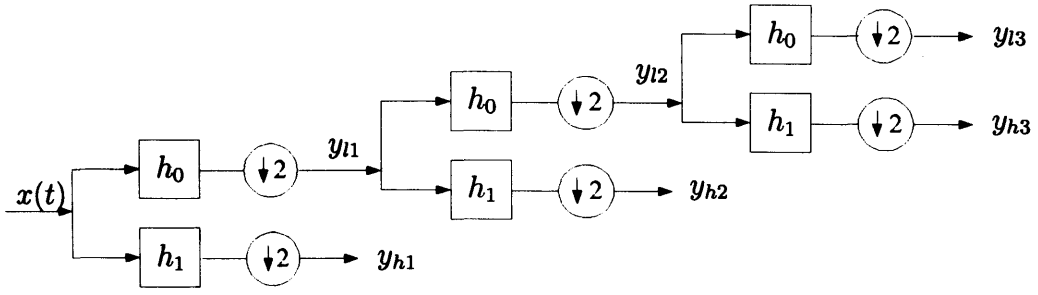


Figure 6.2. A three level 1-D DWT filterbank implementation.

$x(t)$ is the discretely sampled signal, h_0 and h_1 are lowpass and highpass filters respectively. The $\downarrow 2$ block denotes a downsample by two operation [75]. The high pass filter provides the detail (wavelet) information $y_{hp}, p = 1, 2, 3$, of the input signal $x(t)$ and the low pass filter provides the approximation (scaling) information $y_{lp}, p = 1, 2, 3$. Multiple levels of analysis are possible by filtering the lowpass results of each stage.

Mallat [75] also discussed how to use the 1-D filterbank to obtain the wavelet transform of 2-D data. The top processing path of Figure 6.3 illustrates the standard 2-D DWT. Again, filtering the lowpass result of each filtering stage results in a hierarchy of subimages. For the standard 2-D DWT three subimages $\mathbf{D}^{(n,m)}, n = 1, 2, 3, m = 1, \ldots, m_{max}$ are obtained at each level m of the transform, and will contain detail in the horizontal, vertical and diagonal directions. A fourth subimage is also produced, $\mathbf{A}^{(m)}$ which is the (lowpass) course approximation of the original image $\mathbf{A}$.

Magarey and Kingsbury [74] built on Mallat's work to develop a DWT that utilizes 1-D complex filters to provide the complex data required for a hierarchial phase based motion estimation method [71, 73, 74]. The Complex DWT (CDWT) is similar to a standard DWT, except that the low and highpass filters are a complex valued pair of FIR filters with Gabor like characteristics [71, 74].

The 1-D complex filters can be utilized to implement a separable 2-D complex wavelet analysis in a similar way as described by Mallat [75]. The subtle difference is that the complex filters can only detect information in the first quadrant (positive horizontal frequency, positive vertical frequency) of the 2-D unit frequency cell. However, images will contain useful information in both the first and second (negative horizontal frequency, positive vertical frequency) quadrants.

To obtain the extra information a parallel processing path is added to the 2-D DWT to use the complex conjugates h_0^* and h_1^* when row filtering [74]. Magarey also noted that the filters for the first level of the transform must be modified by some prefilter f (denoted by $h_0 f$ and $h_1 f$ in Figure 6.3) to maintain a uniform frequency response for all levels of the CDWT [71, 74]. The complete 2-D CDWT is illustrated in Figure 6.3.

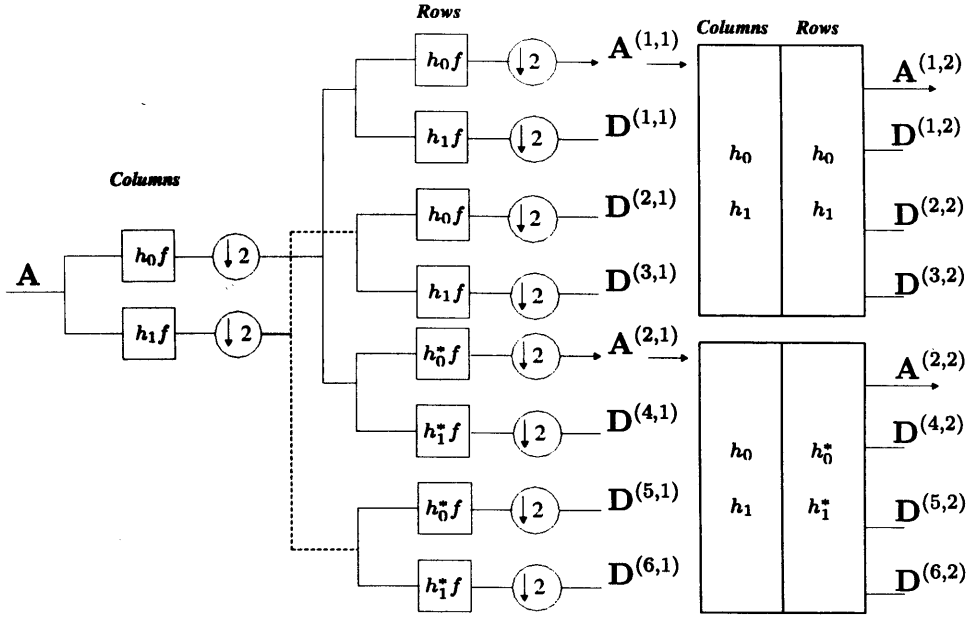


Figure 6.3. Two level ($m = 2$) CDWT implementation.

Notice that there are eight subimages at each level. Two coarse approximations, $\mathbf{A}^{(1,m)}$ and $\mathbf{A}^{(2,m)}$, and six subimages $\mathbf{D}^{(n,m)}$, $n = 1, \ldots, 6$. The equations for the level m subimages can be expressed as [74]:

$$\mathbf{A}^{(m)}(\mathbf{n}) = \sum_{\mathbf{k}} \mathbf{A}(\mathbf{k}) \phi^{(m)}(2^m \mathbf{n} - \mathbf{k}) \tag{6.4}$$

$$\mathbf{D}^{(n,m)}(\mathbf{n}) = \sum_{\mathbf{k}} \mathbf{A}(\mathbf{k}) \psi^{(n,m)}(2^m \mathbf{n} - \mathbf{k}) \tag{6.5}$$

where $\phi^{(m)}$ is the level m scaling filter, $\psi^{(n,m)}$ is the wavelet filter for subband (n, m) and $\mathbf{n} = (n_1, n_2)^T$ are the spatial coordinates with vertical listed first, and down and right being the positive directions.

Figures 6.4 and 6.5 are examples of the DWT and CDWT of the Lenna image respectively. The CDWT provides twice as much information for analysis compared to the DWT. The phase information given by the CDWT should also provide a better representation of the shape of the lips, thus allowing their changing shape to be estimated more

accurately.

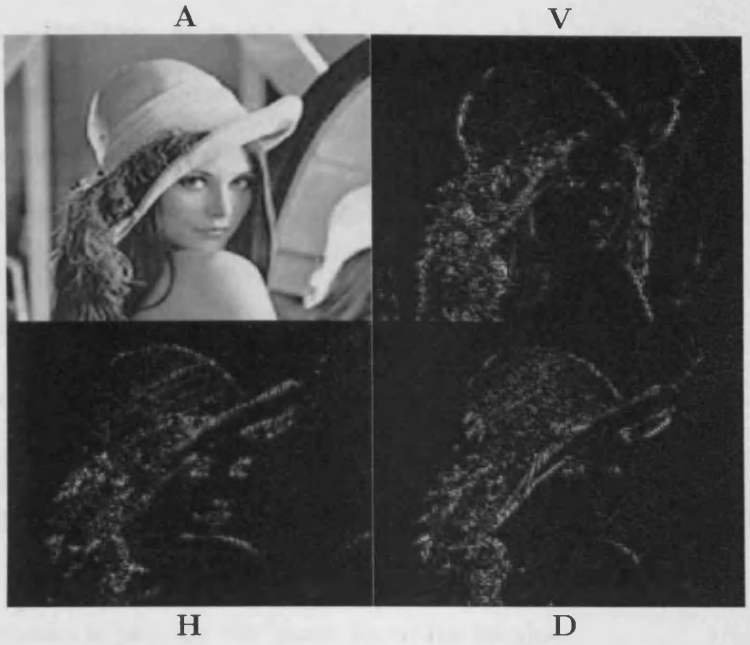


Figure 6.4. Single level DWT decomposition of Lenna, where **V**, **H**, and **D** note the vertical, horizontal and diagonal filter results, and **A** is the lowpass approximation of the original image.

Figure 6.5. Single level Complex DWT decomposition of Lenna, where $\mathbf{D}^{(n,m)}$ are the results of the highpass filtering and $\mathbf{A}^{(m)}$ are the lowpass approximations.

6.4 Motion Estimation

Motion estimation is commonly used for video coding such as the MPEG or H.26x families of video coding standards [9]. Motion estimation involves finding the motion vector that describes the displacement of a pixel between consecutive video frames, and can be either whole or sub pixel vertical and horizontal values. Obtaining a displacement value for each pixel results in an optical flow field of vectors that describes the motion of each pixel from the current frame to the next. Optical flow methods can be simplified into three main categories: gradient, block matching and phase based methods. Barron et al. [6] compared several optical flow methods, and report that of those tested, phase based methods provide the more accurate results.

The optical flow method employed here is a hierarchical phase based motion estimation algorithm. The technique uses the complex discrete wavelet transform (CDWT) [71, 74] described in the previous section. The method will not be described in significant detail as there are already several papers [74] (and references contained therein) and a PhD thesis [71] discussing it as well as comparing it to other techniques.

6.4.1 Motion Estimation using the CDWT

The motion estimation (ME) algorithm used in this chapter was originally developed by Magarey for video coding [71, 74]. Several papers have since been published which use this algorithm for moving target detection [16], or stereo image matching [72, 102]. Here it is used to obtain information about the motion of a speaker's lips.

To calculate the motion between two consecutive frames it is necessary to find the displacement of each pixel $\mathbf{n}$ in the current frame from

the corresponding pixel in the previous frame. If $\mathbf{A}_2$ is the current frame and $\mathbf{A}_1$ the previous frame then the displacement $\mathbf{d(n)}$ of pixel $(\mathbf{n})$ is $\mathbf{A}_1(\mathbf{n}+\mathbf{d(n)}) = \mathbf{A}_2(\mathbf{n})$, with $\mathbf{n} = (n_1, n_2)^T$ where down and right are the positive directions and n_1 being vertical and n_2 horizontal [71].

The subimages $\mathbf{D}^{(n,m)}$ from the CDWT provide the input data for Magarey's ME algorithm [74]. As stated previously, for a single image, the CDWT produces six bandpass subimages at each level (m) of the decomposition. The algorithm starts at the coarsest level $(m = m_{max})$ and finds the vertical and horizontal displacement for each pixel. This produces a motion field at level m. This field is then interpolated by two in each direction (rows and columns), and along with the twelve subimages (six each from the images $\mathbf{A}_1$ and $\mathbf{A}_2$) is used as the initial estimate for motion estimation at the next finest level $(m-1)$. There are several steps to the algorithm performed at each level of the decomposition and continuing until level $m = m_{min}$ is reached. The matching criterion at each subband (n, m) and pixel $\mathbf{n}$ is the squared difference (SD) of two consecutive frames and is expressed as:

$$\mathrm{SD}^{(n,m)}(\mathbf{n}_1 + \mathbf{f}, \mathbf{n}_2) = |\mathbf{D}_1^{(n,m)}(\mathbf{n}_1 + \mathbf{f}) - \mathbf{D}_2^{(n,m)}(\mathbf{n}_2)|^2 \qquad (6.6)$$

where $\mathbf{n}_1, \mathbf{n}_2$ are the corresponding pixels from the subband images $\mathbf{D}_1^{(n,m)}$ and $\mathbf{D}_2^{(n,m)}$, which themselves are obtained from the CDWT of frames $\mathbf{A}_1$ and $\mathbf{A}_2$, and $\mathbf{f}$ is a fractional offset to allow sub-pixel accuracy. The six subband differences are then summed up to form the subband squared difference:

$$\mathrm{SSD}^{(m)} = \sum_{n=1}^{6} \mathrm{SD}^{(n,m)}(\mathbf{n}_1 + \mathbf{f}, \mathbf{n}_2) \qquad (6.7)$$

The motion estimate $\mathbf{f}$ for each pixel at level m is that which minimises $\text{SSD}^{(m)}$. Summing over the six subbands and finding the overall minimum leads to a robust motion estimate for that pixel. The SSD can be expressed as an elliptical surface, composed of five parameters [71,74]:

- Surface curvature parameters, α, β, γ.
- Surface minimum coordinate vector $\mathbf{f}_0 = (f_{10}, f_{20})$.
- Surface minimum value δ.

The multiresolution structure of the CDWT allows the motion estimation to work from course to fine levels of information. It incorporates information from the previous level estimate into the current level in order to refine the motion estimate. However, before any motion estimation on the current level can begin, it is necessary to warp the current (finer) CDWT subimages using the course level estimate of $\mathbf{f}_0$ [74]. Once this warping has been carried out, the motion estimates of the current level are calculated and combined with the results of the previous level by adding the quadratic surface parameters of related pixels from the levels to form the *cumulative subband squared difference* (CSSD):

$$\text{CSSD}^{(m)}(\mathbf{n}, \mathbf{f}) = \begin{cases} \text{CSSD}'^{(m+1)}(\mathbf{n}, \mathbf{f}) + \text{SSD}^{(m)}(\mathbf{n}, \mathbf{f}) & m_{min} \leq m \leq m_{max} \\ \text{SSD}^{(m)}(\mathbf{n}, \mathbf{f}) & m = m_{max} \end{cases} \tag{6.8}$$

Although, before adding the surfaces, the information from the course level must be scaled and interpolated (indicated by the prime $(.')$) as the motion field density is of a lower resolution than the current level estimates. This refinement continues until level $m = m_{min}$ is

reached and Figure 6.6 is a block diagram of the hierarchical structure of the motion estimation process. The result of this hierarchial process is a robust subpixel accurate motion field [74]. Figure 6.7 contains two consecutive frames from the dataset and the resulting motion field.

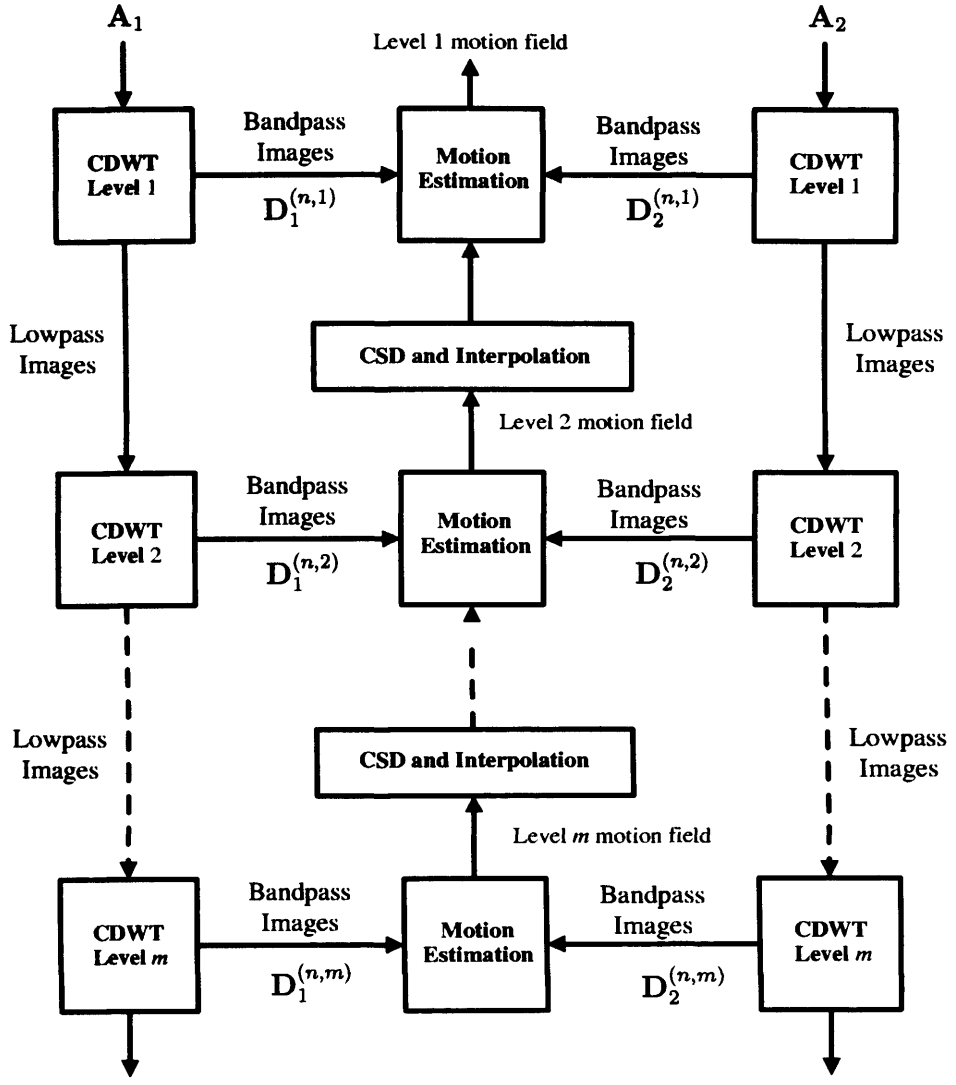


Figure 6.6. Block Diagram of the CDWT based motion estimation algorithm, with $m_{min} = 1$ and $m_{max} = m$, adapted from [74].

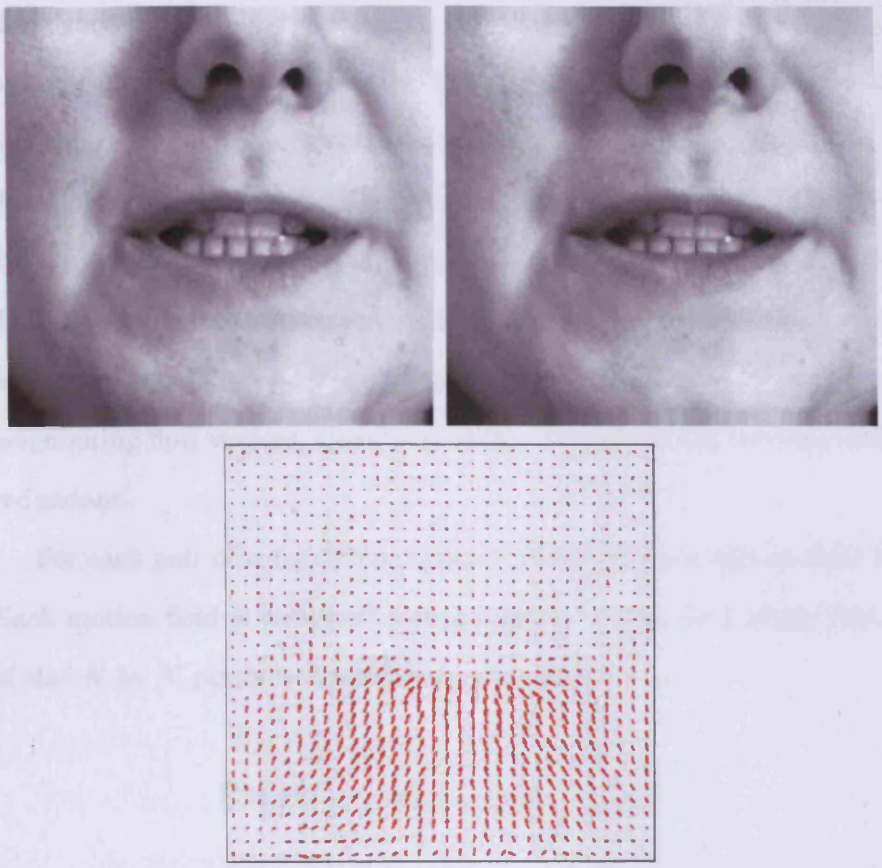


Figure 6.7. Frames 1923 (top left) and 1924 (top right) and the flow field (bottom) describing the motion between the two frames, taken from the female speaker dataset.

6.5 Voice Activity Detection

The method employed here for visual voice activity detection again uses HMMs, but in this instance they are used to model the dynamics of the optical flow field. To obtain the optical flow of the data, first the images were cropped and resized to allow the wavelet decomposition to be performed. The need to resize the images before calculating the flow fields is because of the downsampling by two in each direction operation that occurs in between every level of the wavelet decomposition. For

the algorithm to operate correctly, the dimensions of the images must be multiples of $2^{m_{max}}$. Secondly, the complex wavelet ME algorithm was run on the data for a total of $1 : m_{max}$ levels, and the level $m = 1$ flow fields were downsampled from the original image size to reduce the dimensionality of the data before constructing the required HMMs. This is possible because the optical flow field has been smoothed, and as such there should not be large variations in direction and magnitude of neighboring flow vectors, therefore, some of the data can be considered redundant.

For each pair of consecutive video frames there is a motion field $\mathbf{F}$. Each motion field is reshaped into a column vector, so a single frame of size N by N pixels is described by the vector:

$$\mathbf{F} = [f_{(1,1)} \cdots f_{(1,N)}, f_{(2,1)} \cdots f_{(2,N)}]^T$$

where the motion vector $\mathbf{f}$ at pixel $\mathbf{n}$ is described as $\mathbf{f(n)} = (f_1, f_2)^T$ where f_1 is the vertical motion, and f_2 is the horizontal motion, with down and right being the positive directions [74]. This results in the data now being described as the set of vectors $\mathbf{F}_{(j)}$, with j being the number of video frames. Even with the downsampled flow field the dimensionality of the data was still high, so PCA was used to reduce the dimensionality even further. Early experiments showed that performing PCA on the vertical and horizontal coordinates of each vector simply meant that the most significant principal components were almost (if not entirely) composed of purely vertical motion vectors. While this is a sensible result for speech, the same cannot be said for a smile. As such the results obtained had only a 60% to 70% accuracy in the detection of silence frames. Also, the dimensionality of the data had doubled

as there were two motion values associated with each pixel (horizontal and vertical). Therefore it was decided to use the magnitude of each motion vector, and reduce the dimensionality of these.

6.5.1 Simulations

Results were obtained using the same video data as described in Chapter 5. Two datasets were used for the experiments, one of a female speaker, and one of a male speaker. The motion estimation algorithm was applied to the data and the V-VAD was applied in the same manner as in Chapter 5. As was noted in Chapter 5, the terms male and female are only used to distinguish between the two speakers, the results do not indicate the performance of the V-VAD for just male speakers or just female speakers.

The original frames were 480 by 640 pixels, and were reshaped into 512 by 512 pixels per frame. For the wavelet decomposition m_{max} was set to 6 and m_{min} to 1. The resultant flow fields were then downsampled to a size of 32 by 32 vectors and the magnitude of each vector obtained. This resulted in each vector in the set $\mathbf{F}_{(j)}$ being of length 1024. PCA was then used to reduce the dimensionality, and retaining the first ten eigenvectors accounted for 90% of the total variance. Therefore each frame was reduced to only ten dimensions. To build the HMM, again only silence data were used, and the training data consisted of around 800 frames and the remainder of the dataset used as unseen data. As in Chapter 5, the output of the V-VAD was smoothed with an averaging filter, where the length of the filter (either 5 or 10 samples) is indicated in the legend of each figure. The above details are the same for both the male and female speakers.

Female Speaker V-VAD Results

Figures 6.8 and 6.9 show the results of the V-VAD applied to the female speaker. These show results of using different silence data in the building of the HMM. Initial classification results of the female speaker were not as expected. Further investigation found that during the third silence period in which complex lip motion occurs, there is non-negligible movement of the speaker's head. There are also small movements during the fourth silence period, but not as significant as those in the third period. Figure 6.8 shows the results of using the fourth and fifth (final) silence periods to build the HMM. Comparing these to Figure 6.9 where the third and fifth silence periods are used to build the HMM, it can be seen that including the third silence period into the data to build the HMM improves the results by around 10%. The female speaker results in Figure 6.9 show that a CSD of 88% for a FSD of 5% can be achieved. However, for an FSD $\geq$ 10% the results are poorer compared to those of the AAM and retinal filter methods given in Chapter 5, although an FSD of $\geq$ 10% is most likely undesirable as for large datasets this would result in a significant amount of error.

Male Speaker V-VAD Results

The results of the male speaker are provided in Figures 6.10, 6.11 and 6.12. The figures show the results of V-VAD when different silence datasets were used to build the HMM. As can be seen, there is little difference in the results. Further experiments verified that provided the motion estimates from one of the silence periods with complex lip motion were used in the HMM, the V-VAD results were similar.

Figure 6.12 shows the best results obtained. For these results, the

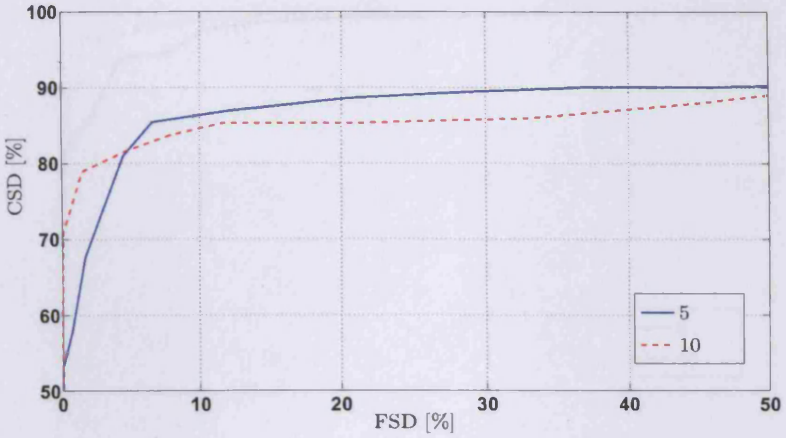


Figure 6.8. ROC curves of silence detection for the female speaker using the fourth and fifth silence periods.

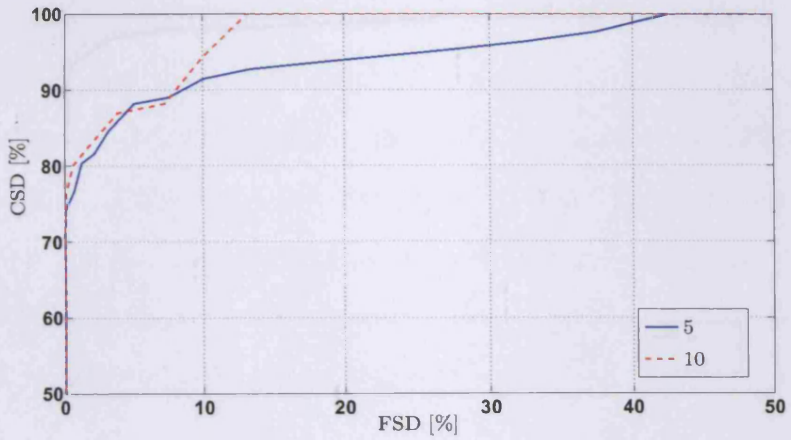


Figure 6.9. ROC curves of silence detection for the female speaker using the third and fifth silence periods.

HMM was built using the first and fifth silence periods, and it can be seen that a CSD of 98% for a FSD of 5% can be achieved. Furthermore, based on the classification results of several different HMMs of the male speaker, an average CSD of 94% for a FSD of 5% can be achieved.

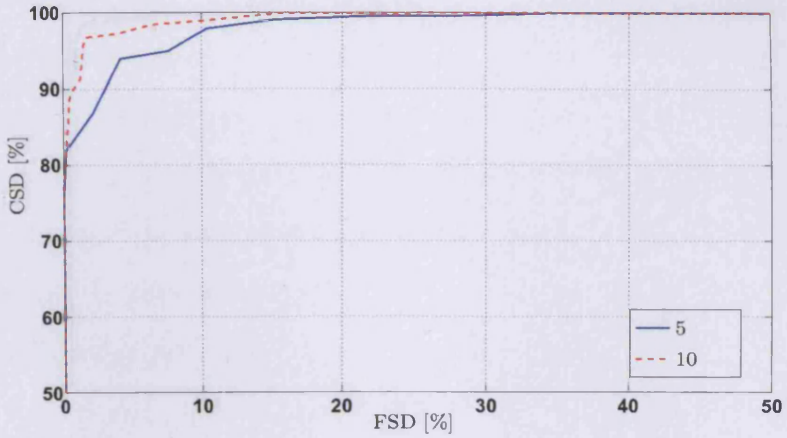


Figure 6.10. ROC curves of silence detection for the male speaker using the first and third silence periods.

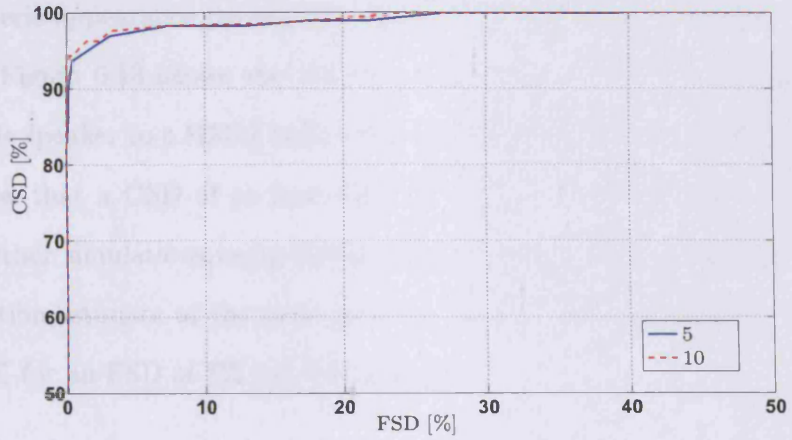


Figure 6.11. ROC curves of silence detection for the male speaker using the fourth and fifth silence periods.

A generic V-VAD

The motivation behind the work described in this chapter was to find a method of representing the motion of a speaker's lips in such a way that a generic V-VAD could be built on the basis of these descriptors. As mentioned in Chapter 5, the AAMs obtained for each speaker are almost person specific. While work has been published on building

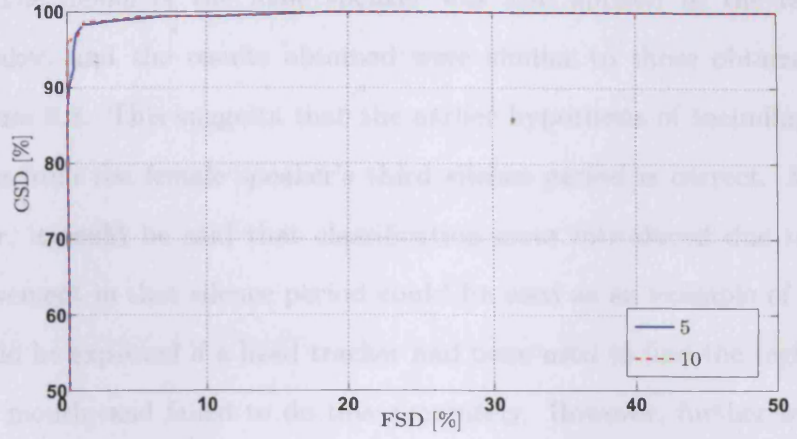


Figure 6.12. ROC curves of silence detection for the male speaker using the first and fifth silence periods.

generic appearance parameters, limited success has been reported [48].

Figure 6.13 shows the results of applying the motion data of the male speaker to a HMM built solely on the female speaker. The results show that a CSD of at least 92% for a FSD of 5% can be achieved. Further simulations using HMMs of the female speaker to classify the motion estimate of the male speaker, showed that an average CSD of 90% for an FSD of 5% can be obtained.

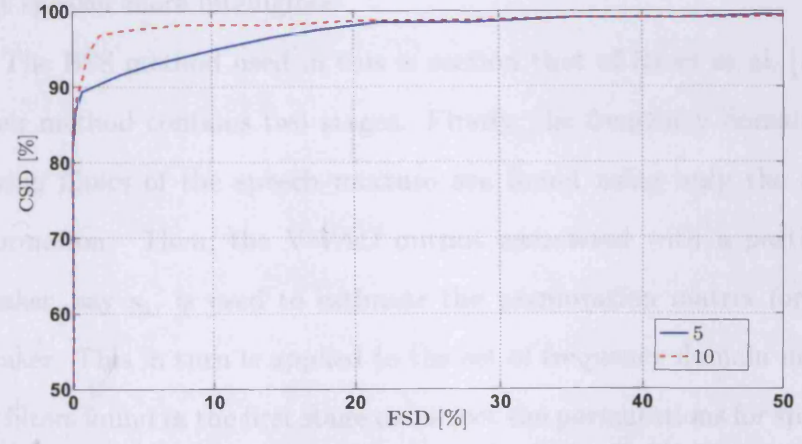


Figure 6.13. ROC curves of silence detection for the male speaker using a HMM built on silence data from the female speaker.

The model of the male speaker was also applied to the female speaker, and the results obtained were similar to those obtained in Figure 6.8. This suggests that the earlier hypothesis of including the data from the female speaker's third silence period is correct. Moreover, it could be said that classification error introduced due to the movement in that silence period could be used as an example of what could be expected if a head tracker had been used to find the region of the mouth, and failed to do this accurately. However, further testing would be needed to verify this.

6.6 Using a V-VAD to Regularise the Permutations in Convolutive BSS

This penultimate section brings the thesis back to where it started; the cocktail party problem. The experiments in this section show that the output of the V-VADs described in this chapter and Chapter 5 can be used in an effective manner to correct the permutations and hence separate a chosen speech signal from a convolutive mixture to make that speaker more intelligible.

The BSS method used in this is section that of Rivet et al. [100][1]. Their method contains two stages. Firstly, the frequency domain unmixing filters of the speech mixture are found using only the audio information. Then, the V-VAD output associated with a particular speaker, say s_1, is used to estimate the permutation matrix for that speaker. This in turn is applied to the set of frequency domain unmixing filters found in the first stage to correct the permutations for speaker s_1. To correct the permutations for any of the other speakers in the

[1]The code for this method was kindly supplied by Dr Bertrand Rivet

mixture, the output of the V-VAD associated with that speaker must be available. The validity of this method has already been discussed by Rivet et al. [100]. The purpose of these experiments is to compare the outputs of the AAM and motion estimation V-VADs proposed in this thesis when used to aid in solving the cocktail party problem.

6.6.1 Simulation Results of BSS Using a V-VAD to Correct the Permutation Problem

The results given in this section are based on two speakers, s_1 and s_2 mixed artificially (in a convolutive manner) in a room $10 \times 10 \times 10m^3$ in size [2]. The FIR mixing filters had approximately 1000 lags, and the audio recordings were approximately 70 seconds in length. The speech signal for speaker s_1 is the audio component of the male speaker data used in the previous section, as well as the experiments in Chapter 5. Each of the V-VADs were applied to the whole video signal, and approximately 70 seconds of audio data and corresponding V-VAD output removed from the middle of the recording and used for testing here.

Figures 6.14 and 6.15 show the original speech signals and mixed signals respectively. As can be seen in Figure 6.14 there are clearly defined periods of speech and silence for s_1, and continuous speech for s_2.

The estimates $\hat{s}_1$ of the original signal s_1 when using the output of a V-VAD are given in Figure 6.17. The dashed line represents the manual indexation of the speech and silence periods, while the solid line represents the output of the respective V-VAD used to regularise

[2] The Matlab file simroommix.m file found at http://sound.media.mit.edu/ica-bench/ was used for this purpose.

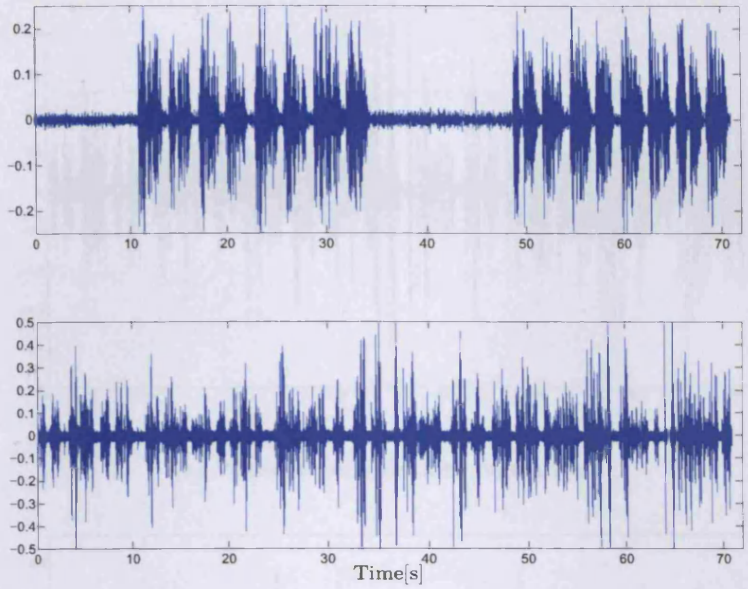


Figure 6.14. Original speech signals for speakers 1 (top) and 2 (bottom).

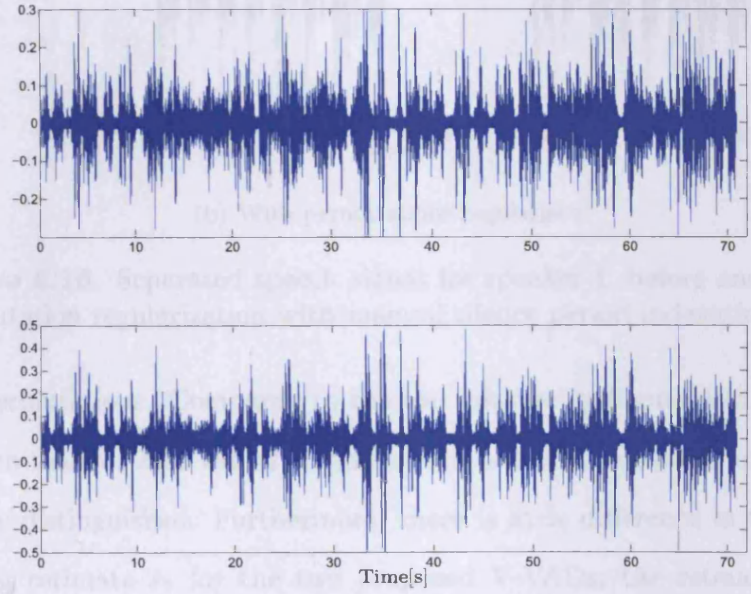
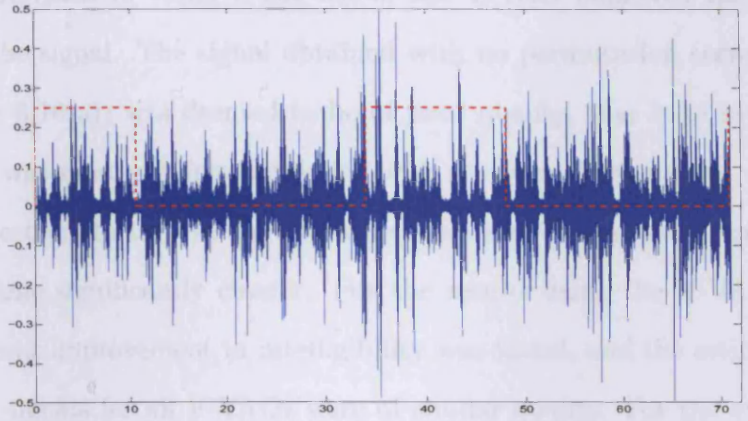
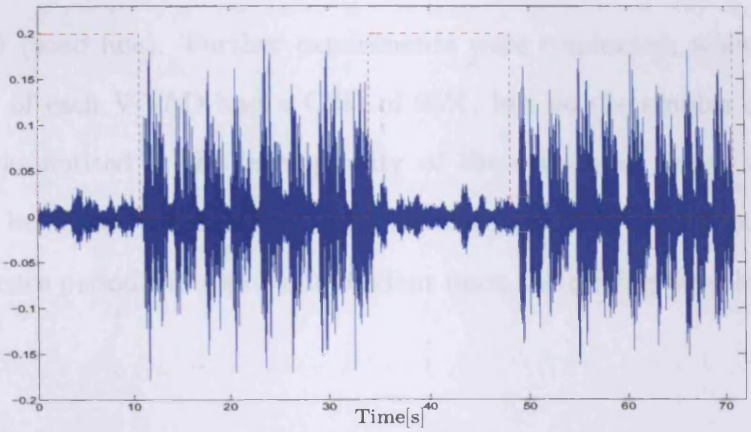


Figure 6.15. Mixed speech signals.



(a) Without permutations regularised



(b) With permutations regularised

Figure 6.16. Separated speech signal for speaker 1, before and after permutation regularization with manual silence period indexation.

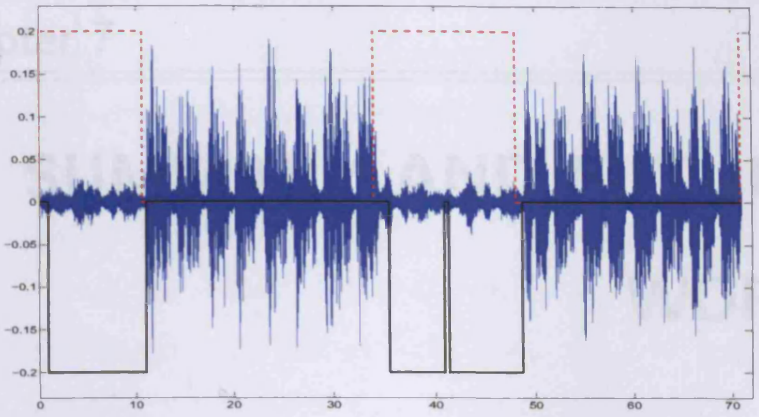
the permutations. Compared to the top mixture in Figure 6.15 it can be seen that for each signal $\hat{s}_1$, the speech and silence periods are now clearly distinguished. Furthermore, there is little difference in the resulting estimate $\hat{s}_1$ for the two proposed V-VADs, the retinal filter V-VAD and the manually indexed silence periods (Figure 6.16(b)). A subjective comparison in the form of listening tests of the estimated signal $\hat{s}_1$ for the experimental results given in Figures 6.16 and 6.17

was performed to verify if the use of the V-VAD improved the clarity of the signal. The signal obtained with no permutation correction (Figure 6.16(a)) was deemed to be of poor quality, that is to say, the speech was mostly unintelligible for both speakers. However, as would be expected, during the periods when speaker s_1 was silent, speaker s_2 became significantly clearer. For the results using the V-VADs, a significant improvement in intelligibility was found, and the estimated speech signals for all V-VADs were of similar quality. For the experiments given in Figure 6.17 the output of the V-VADs corresponded to a CSD of 90%. What is also shown is the silence estimation of each V-VAD (solid line). Further experiments were conducted, where the output of each V-VAD had a CSD of 95%, but no discernable difference was noticed in the intelligibility of the estimated signal $\hat{s}_1$. It should be noted that in a real room environment, the useable portion of a silence period will also be dependent upon the mixing filter length.

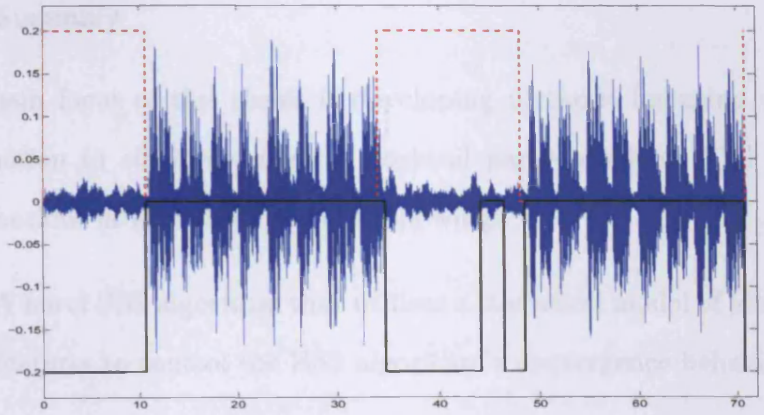
6.7 Conclusion

A novel V-VAD was presented in this chapter, based on using the optical flow of a speaker's mouth region. Comparing the proposed method to the AAM and retinal filter V-VAD presented in Chapter 5, it can be seen that similar results can be obtained. However, the method proposed in this chapter has advantages over the previous methods. The optical flow based V-VAD has the same advantage over the retinal filter approach as the AAM based method; the window length used for the optical flow and AAM based methods is shorter than that of the retinal filter method. As explained at the end of Chapter 5, the window size determines the smallest silence period that can be detected, as the optical flow and AAM methods use statistical models they are able to achieve similar results to the retinal filter method using a smaller window size. It has also been shown that using the optical flow V-VAD, a generic method is possible while still achieving a high rate of classification, with an average CSD of 90% for an FSD of 5%. A generic model was not possible with the AAM based approach. Moreover, the optical flow method does not require the lips to be explicitly tracked.

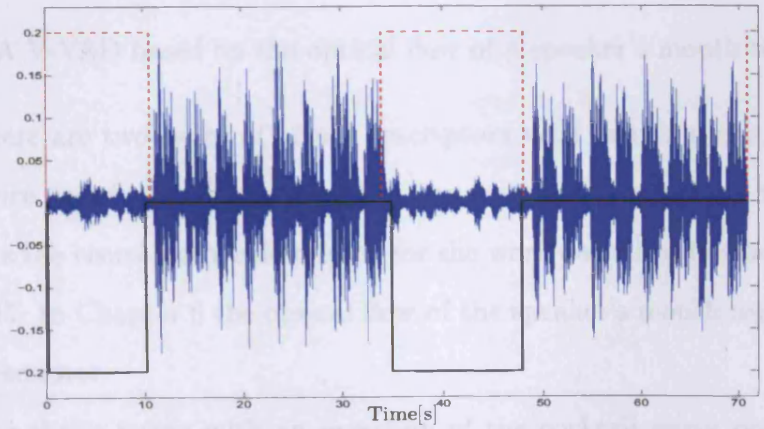
The final experiments in this thesis show that the output of the V-VADs described herein can be used in the audio-visual BSS algorithm of Rivet et al. [100] to improve the intelligibility of a chosen speaker (the speaker associated with the V-VAD output) when mixed in a convolutive manner with another speaker. The results of listening experiments show that a significant improvement in speech intelligibility can be obtained for the speaker associated with the V-VAD output.



(a) Retinal filter based V-VAD



(b) AAM based V-VAD



(c) CDWT motion estimation based V-VAD

Figure 6.17. Separated speech signal for speaker 1 ($\hat{s}_1$), using the (a) retinal filter, (b) AAM and (c) CDWT motion estimation based V-VADs to perform the permutation regularisation.

Chapter 7

SUMMARY AND FUTURE WORK

7.1 Summary

The main focus of this thesis is developing methods for using visual information to aid in solving the cocktail party problem. The novel contributions of this thesis are the following:

- A novel BSS algorithm that utilises a statistical model of joint AV features to control the BSS algorithm's convergence behaviour.
- A V-VAD in which the visual descriptors are appearance parameters obtained from an AAM.
- A V-VAD based on the optical flow of a speaker's mouth region.

There are two types of visual descriptors used in this thesis. Appearance parameters of the speaker's lips, obtained from an AAM, are used as the visual feature descriptor for the work described in Chapters 4 and 5. In Chapter 6 the optical flow of the speaker's mouth region is the descriptor.

The thesis began with an overview of the cocktail party problem and a review of current literature on the subject. Next, an in depth

discussion of the visual and audio features used in this thesis together with an overview of the statistical modelling methods used in this study is given. The remaining chapters provided an in depth discussion of the novel contributions of this study.

The first contribution is an AV-BSS algorithm that employs a joint statistical model of audio and visual features to control the convergence behaviour of the BSS algorithm of Wang et al. [127]. The motivation behind this work was to improve the convergence behaviour of the BSS algorithm in non-stationary environments. To this end an AV model of one of the speakers was built and used at each iteration of the algorithm for this purpose. The statistical models used to model the AV features were the GMM and HMM. The results of this work show that when integrating a joint AV model into the BSS algorithm, a significant improvement of convergence rate is obtained when compared to using the raw audio information (no model). Furthermore, the use of an HMM to model the joint AV features provides an advantage over using the GMM. This can be attributed to the transition probabilities between states contained in the HMM capturing the time dynamics of the features, so while an observation of AV features may be valid, the transition between them may not. However, the models are limited to the training data used in this study. Therefore, the remaining chapters focus on simpler video based speech cues, specifically methods for visual-voice activity detection (V-VAD).

The video data for the V-VADs in this thesis are different to data used previously in existing V-VADs [59,67,113]. During silence periods -the speakers were asked to perform a variety of naturally occurring mouth expressions such as smiling, biting lips and licking lips which

are labelled as complex lip motion. Then models were built on training data from silence periods and used to find the silence periods in a dataset.

The second main contribution presented in this thesis is a V-VAD based on using AAM parameters to describe the speaker's lips. They were chosen as they had already been successfully used for the AV-BSS method. The dynamics of AAM parameters during silence periods are captured with an HMM that was then used to classify unseen AAM parameters as silence or speech. The results of experiments show that appearance parameters can successfully be used for visual voice activity detection and that the complex silence periods are also successfully detected. Use of appearance parameters as the visual descriptor has the drawback that the HMMs built using them are essentially person specific. This could be overcome by building a model on several people but the dataset used in this study was limited to only two people. Previous research has shown limited success in building generic AAMs [48].

This issue led to the choice of a visual descriptor which would allow a generic HMM to be built for use in a V-VAD. The final contribution to this study is given in Chapter 6 and is a V-VAD that can accurately classify the lip motion of a speaker as speech or silence using a model that was not built exclusively on that speaker's data. In fact results are given that show the V-VAD can accurately classify the lip motion of a speaker using a model that does not include any of their data. For this last contribution, the visual descriptor was an optical flow field of the speaker's mouth region. The flow field was obtained using a CDWT based motion estimation algorithm. It is known that due to the speed

at which the lips move during speech, in particular the lower lip, that the lip shape is not easy to automatically track with a high degree of accuracy. This particular method was chosen as the motion estimation method uses the phase information from the CDWT to calculate the flow field. This meant that the edges (contour) of the lips could be emphasised, and the motion over several frames found. The method was thoroughly tested by comparing the results of models built using different training data. Results showed a consistently high degree of correct silence detection.

The performances of the V-VADs presented in this study were also compared to a previously published V-VAD [4] that uses a retinal filter to obtain the visual descriptor. The methods in this study were found not only to perform as well as the retinal filter method, but also have the advantage of being able to detect shorter periods of silence (approximately 10-15 frames) compared to the retinal filter approach (20-25 frames).

Final experiments are performed with the AV-BSS algorithm of Rivet et al. [100]. The algorithm utilises the output of a V-VAD to regularise the permutations of the estimated speech signal (corresponding to the V-VAD output), thus mostly solving the permutation problem inherent to BSS. Experiments on convolutive mixtures of speech are conducted using the V-VADs presented in this thesis, and results are provided that show the advantage of using such V-VADs in a BSS scenario.

7.2 Future Work

The future of audio visual solutions to the cocktail party problem has a promising outlook. With the increase in computing power and miniaturising of microprocessors producing powerful consumer electronics such as digital photo cameras that have the ability to find faces in the viewing area to better focus the camera lens on the subjects, a real time solution to the BSS problem is conceivable.

The work presented in this thesis comprises of two main elements. The first is directly linked to the overall theme of the thesis; the cocktail party problem. The AV-BSS method described herein is restricted to the training data used to build the AV models. However, simpler AV models could remove this restriction.

The visual information used herein is obtained from the speaker's lips, but this is not the only visual information that could be of use. It has already been mentioned in the review chapter that beamforming based BSS methods rely on the location of the speakers. Their locations are traditionally found using audio information alone. However, obtaining these locations with video information should prove to be more accurate, especially in acoustically noisy environments or non-stationary environments containing several speakers. Using the location of a speaker to enhance their speech has recently been investigated by Maganti et al. [70].

What can be said about the integration of visual information into BSS methods is that the intended application will determine what visual information is available. For example, where there are several speakers who are continually moving around, such as in an airport or train station, detecting a speaker's face will probably be difficult due to

occlusions from other people and objects in the environment. Detecting the locations of the speakers visually could be simpler. Alternatively, in a video conferencing situation, a face on view of the speaker would be available, making the tracking and extraction of facial features possible.

The remaining contributions of this study are focused on the development of a V-VAD that can be applied to any sequence of video where the mouth region can clearly be viewed. A V-VAD could easily be integrated into an AVSR framework, as the video camera would already be processing information of the speaker's mouth area, so very little extra processing would be incurred. In fact it may be the case that in audio noisy situations that unnecessary speech recognition processing could be eliminated by the V-VAD. With regards to the cocktail party problem, a V-VAD could be used to reduce the amount of processing required in that situation as well, by knowing when a person is speaking or not could allow a better estimate of the unmixing matrix.

With regards to modelling the visual feature, a natural path to proceed upon is possibly one used by speech recognition. That is to model each non-speech mouth action separately and collect them in multiple or multi layered HMMs, and actually recognise each observation of the speaker's lips as a mouth action to better class them as speech or not speech. The inclusion of audio information in acoustically low noise conditions would also allow for improved VAD.

Whilst reading through the literature for the review section, it became apparent that there appears to be very little dialogue between researchers in the fields of CASA, SE and BSS. In fact very little comparison between solutions to the cocktail party problem from each area had been done. Increased dialogue between these areas could result in

an improved solution.

Indeed, a single method cannot solve this problem alone, a system is required that utilises several methods. Cherry [21] has already given his thoughts on what should be considered. His thoughts are echoed by Haykin and Chen [52] who also propose their own framework. However, in [52] their active audition solution is based solely on information derived from audio. But where available, humans use both audio and visual information. It is the combination of these senses that allow us to converse in noisy environments and a solution that exploits this should provide a better performance than one that relies solely on audio information.

The cocktail party problem is unlikely to ever be completely solved by any method. What can be achieved is a satisfactory solution, and the intended application will determine what constitutes a satisfactory solution.

BIBLIOGRAPHY

[1] S. Amari, S.C. Douglas, A. Cichocki, and H.H. Yang. Multichannel blind deconvolution and equalization using the natural gradient. *First IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications*, pages 101–104, April 1997.

[2] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari. The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *Speech and Audio Processing, IEEE Transactions on*, 11(2):109–116, Mar 2003.

[3] A. Aubrey, J. Lees, Y. Hicks, and J. Chambers. Using the Bimodality of Speech for Convolutive Frequency Domain Blind Source Separation. In *IMA 7th International Conference on Mathematics in Signal Processing*, December 2006.

[4] A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, and C. Jutten. Two novel visual voice activity detectors based on appearance models and retinal filtering. In 15^{th} *European Signal Processing Conference (EUSIPCO)*, 2007.

[5] A.V. Barbosa and H.C. Yehia. Measuring the relation between speech acoustics and 2D facial motion. *IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP)*, 2001.

[6] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *Int. J. Comput. Vision*, 12(1):43–77, 1994.

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

[8] A. Belouchrani, K.A. Meraim, J.F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans Signal Processing*, 45:434–444, 1997.

[9] A. Bovik, editor. *Handbook of Image & Video Processing.* Academic Press, 2000.

[10] C. Bregler and Y. Konig. Eigenlips for robust speech recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, Apr 1994.

[11] A.S. Bregman. *Auditory Scene Analysis.* MIT Press, 1990.

[12] G.J. Brown and D Wang. Separation of speech by computational auditory scene analysis. In J. Benesty, S. Makino, and J. Chen, editors, *Speech Enhancement*, pages 371–402. Springer, 2005.

[13] C. Busso, S. Hernanz, C.W. Chu, S. Kwon, S. Lee, P.G. Georgiou, I. Choen, and S. Narayanan. Smart room: Participant and speaker localization and identification. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, 2005.

[14] J.P. Campbell. Speaker recognition: a tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, Sep 1997.

[15] J. F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. *Proceeedings of the IEE*, 140(6):362–370, Dec 1993.

[16] G. Castellano, J. Boyce, and M. Sandler. Moving target detection in infrared imagery using a regularized CDWT optical flow. In *Proc. IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications (CVBVS '99)*, pages 13–22, 21–22 June 1999.

[17] H.E. Cetingul, Y. Yemez, E. Erzin, and A.M. Tekalp. Robust lip-motion features for speaker identification. *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, March 2005.

[18] H.E. Cetingul, Y. Yemez, E. Erzin, and A.M. Tekalp. Discriminative analysis of lip motion features for speaker identification and speech-reading. *IEEE Transactions on Image Processing*, 15(10):2879–2891, Oct. 2006.

[19] N. Checka, K.W. Wilson, M.R. Siracusa, and T. Darrell. Multiple person and speaker activity tracking with a particle filter. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004.

[20] Y. Chen and Y. Rui. Real-time speaker tracking using particle sensor fusion. *Proceedings of the IEEE*, Vol 92, issue 3:485–494, March 2004.

[21] C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal Of The Acoustical Society Of America*, 25(5):975–979, September 1953.

[22] C. Cherry and W.K. Taylor. Some further experiments upon the recognition of speech, with one and with two ears. *The Journal Of The Acoustical Society Of America*, 26(4):554–559, July 1954.

[23] Y. Chu, H. Ding, and X. Qiu. Quasi-blind source separation algorithm for convolutive mixture of speech. *Digital Signal Processing Workshop, 12th - Signal Processing Education Workshop, 4th*, pages 233–238, Sept. 2006.

[24] A. Cichocki and S Amari. *Adaptive Blind Signal and Image Processing - Learning Algorithms and Applications.* John Wiley and Sons, 2005.

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

[26] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models: their training and application. *Comput. Vis. Image Underst.*, 61(1):38–59, 1995.

[27] T. F. Cootes and C.J. Taylor. An algorithm for tuning an active appearance model to new data. *Proc. British Machine Vision Conference*, 3:919–928, 2006.

[28] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active Appearance Models. *Proc. Fifth European Conf. Computer Vision, H. Burkhardt and B. Neumann, eds.*, 2:484–498, 1998.

[29] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active Appearance Models. *IEEE Trans on Pattern Analysis and Machine Intelligence (PAMI)*, 23(6):681–685, 2001.

[30] T.F. Cootes and C.J. Taylor. Statistical models of apperance for computer vision. *Available from T.F.Cootes webpage, http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/.*

[31] D. Cosker. *Animation and Perceptual Evaluation of a Hierarchical Appearance Based Facial Model.* PhD thesis, Cardiff University, 2005.

[32] D. Cosker, D. Marshall, P. Rosin, and Y. Hicks. Speech driven facial animation using a hierarchical model. *IEE Vision, Image and Signal Processing*, 151(4):314–321, 2004.

[33] D. Cosker, D. Marshall, P. Rosin, and Y. Hicks. Video realistic talking heads using hierarchical non-linear speech-appearance models. *Proceedings of Mirage, INRIA Rocquencourt*, France, 2003.

[34] D. Cosker, D Marshall, P.L. Rosin, and Y. Hicks. Speech and expression driven animation of a video-realistic appearance based hierarchical facial model. *IEEE Computer Vision and Pattern Recognition (CVPR) Workshop on Learning, Representation and Context for Human Sensing in Video*, 2006.

[35] D. Cristinacce and T.F. Cootes. Facial feature detection and tracking with automatic template selection. *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 429–434, 10-12 April 2006.

[36] T. Darrell, J.W. Fisher III, and W. Freeman. Audio visual segmentation and the cocktail party effect. *International Conference on Multimodal Interfaces*, 2000.

[37] J. R. Deller, J. H. L. Hansen, and J. G. Proakis. *Discrete-Time Processing of Speech Signals.* Wiley, 1999.

[38] P. Delmas and M. Lievin. From face features analysis to automatic lip reading. *7th International Conference on Control, Automation, Robotics and Vision,(ICARCV)*, 2002.

[39] P. Divenyi(Ed.). *Speech Separation by Humans and Machines.* Kluwer Academic, 2004.

[40] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification.* John Wiley and Sons, 2nd edition, 2001.

[41] G.J. Edwards, C.J. Taylor, and T.F. Cootes. Face recognition using active appearance models. *Proc. of 5th European Conference on Computer Vision (ECCV)*, pages 582–595, 1998.

[42] N. Eveno, A. Caplier, and P.-Y. Coulon. Accurate and quasi-automatic lip tracking. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(5):706–715, May 2004.

[43] D. Focken and R. Stiefelhagen. Towards vision-based 3-d people tracking in a smart room. *Fourth IEEE International Conference on Multimodal Interfaces*, October 2002.

[44] Y Freund and R-E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences.*, 55:119–139, August 1997.

[45] L. Girin, A. Allard, and J.L. Schwartz. Speech signals separation: A new approach exploiting the coherence of audio and visual speech. *IEEE Int. Workshop on Multimedia Signal Processing (MMSP), Cannes, France*, 2001.

[46] L. Girin, J.L. Schwartz, and G. Feng. Audio-visual enhancement of speech in noise. *Journal of the Acoustical Society of America*, Vol 109(06):3007–3020, 2001.

[47] R. Goecke, G. Potamianos, and C. Neti. Noisy audio feature enhancement using audio-visual speech data. *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP).*, 2:2025–2028, 2002.

[48] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(11):1080–1093, November 2005.

[49] R. Gross, I. Matthews, and S. Baker. Active appearance models with occlusion. *Image and Vision Computing*, 24(6):593–604, 2006.

[50] J. Hérault, C. Jutten, and B. Ans. Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. *Proceedings of GRETSI, Nice, France*, pages 1017–1022, May 1985.

[51] W. J. Hardcastle and J. Laver, editors. *The Handbook of Phonetic Sciences.* Blackwell Publishing, 1997.

[52] S. Haykin, J. Principe, T. Sejnowski, and J McWhirter(Eds). *New Directions in Statistical Signal Processing: From Systems to Brains.* MIT Press, 2006.

[53] J. Hershey and M. Casey. Audiovisual sound separation via hidden markov models. *In Proc. Advances in Neural Information Processing Systems (NIPS'02)*, 2002.

[54] Y. Huang, J. Benesty, and J. Chen. A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment. *IEEE Trans on Speech and Audio Processing*, Vol 13(05):882–895, September 2005.

[55] Y. Huang, J. Benesty, and J. Chen. Speech acquisition and enhancement in a reverberant, cocktail-party-like environment. *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, May 2006.

[56] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley and Sons, 2001.

[57] M.Z. Ikram and D.R. Morgan. Permutation inconsistency in blind speech separation: Investigation and solutions. *IEEE Trans Speech and Audio Processing*, Vol 13(01):1 – 13, January 2005.

[58] M.Z. Ikram and D.R. Morgan. Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment. *IEEE International Conference on Acoustics, Speech and Signal Processing,(ICASSP)*, Turkey, June 2000.

[59] G. Iyengar and C. Neti. A Vision based Microphone Switch for Speech Intent Detection. In *Recognition, Analysis and Tracking of Face and Gestures in Real Time Systems (RATFG-RTS) Workshop at ICCV*, Vancouver, Canada, 2001.

[60] J. Jiang, A. Alwan, L.E. Berstein, P. Keating, and E. Auer. On the correlation between facial movements, tongue movements and speech acoustics. *ICSLP, Beijing*, 2000.

[61] M Kass, A Witkin, and D Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, V1(4):321–331, January 1988.

[62] R. Kjeldsen and J. Kender. Finding skin in color images. In *Proc. Second International Conference on Automatic Face and Gesture Recognition*, pages 312–317, 1996.

[63] R. Lambert. *Multichannel Blind Deconvolution: FIR matrix algebra and separation of multipath mixtures.* PhD thesis, University of Southern California, Dept of Electrical Engineering, 1996.

[64] T.W. Lee. *Independent Component Analysis. Theory and Applications.* Kluwer Academic, 2000.

[65] T.W. Lee, A. Bell, and R. Lambert. Blind separation of delayed and convolved sources. *Advances in Neural Information Processing Systems 9*, pages 758–764, 1997.

[66] F. Liu, X. Lin, S. Li, and Y. Shi. Multi-modal face tracking using Bayesian network. *IEEE International Workshop on Analysis and Modelling of Faces and Gestures*, 2003.

[67] P. Liu and Z. Wang. Voice Activity Detection Using Visual Information. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, Montreal, Canada, 2004.

[68] J. Luettin, N.A. Thacker, and S. Beet. Locating and tracking facial speech features. *In Proc. of International Conference on Pattern Recognition ICPR*, 1996.

[69] J. Luettin, N.A. Thacker, and S.W. Beet. Speaker identification by lipreading. *Proceedings Fourth International Conference on Spoken Language, ICSLP*, pages 62–65 vol.1, Oct 1996.

[70] H. K. Maganti, D. Gatica-Perez, and I. McCowan. Speech enhancement and recognition in meetings with an audio visual sensor array. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2257–2269, Nov. 2007.

[71] J. Magarey. *Motion Estimation Using Complex Wavelets.* PhD thesis, Cambridge University, 1997.

[72] J. Magarey and A. Dick. Multiresolution stereo image matching using complex wavelets. In *Proc. Fourteenth International Conference on Pattern Recognition*, volume 1, pages 4–7, 16–20 Aug. 1998.

[73] J. Magarey and N. Kingsbury. Motion estimation using complex wavelets. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP).*, volume 4, pages 2371–2374, 7–10 May 1996.

[74] J. Magarey and N. Kingsbury. Motion estimation using a complex-valued wavelet transform. *Signal Processing, IEEE Transactions on*, 46(4):1069–1084, Apr 1998.

[75] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 11(7):674–693, July 1989.

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

[77] I. McCowan, M.H. Krishna, D. Gatica-Perez, D. Moore, and S. Ba. Speech acquisition in meetings with an audio-visual sensor array. *IEEE International Conference on Multimedia and Expo, ICME.*, pages 1382–1385, July 2005.

[78] H. McGurk and J. McDonald. Hearing lips and seeing voices. *Nature.*, 264:746–748, 1976.

[79] N. Mitianoudis and M.E. Davies. Audio source separation of convolutive mixtures. *IEEE Transactions on Speech and Audio Processing*, 11(5):489–497, Sept. 2003.

[80] H. Nait-Charif and S.J. McKenna. Head tracking and action recognition in a smart meeting room. *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance,(PETS)*, Austria, March 2003.

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

[82] H.L. Nguyen Thi and C. Jutten. Blind source separation for convolutive mixtures. *Signal Processing*, Vol 45:209–229, 1995.

[83] R.M. Nickel and A.N. Iyer. A novel approach to automated source separation in multispeaker environments. *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP).*, May 2006.

[84] T. Nishikawa, H. Saruwatari, and K. Shikano. Blind source separation of acoustic signals based on multistage ICA combining fequency-

domain ICA and time-domain ICA. *IEICE Trans Fundamentals*, E86-A(4):846–858, 2003.

[85] H. Okuno, Y. Nakagawa, and H. Hitano. Incorporating visual information into sound source separation. *In Working Notes of IJCAI Workshop on Computational Auditory Scene Analysis (CASA'99)*, 1999.

[86] L.C. Parra and C.V. Alvino. Geometric source separation: merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, 10(6):352–362, Sep 2002.

[87] L.C. Parra and C. Spence. Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, Vol 8(3):320–327, May 2000.

[88] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra. A survey of convolutive blind source separation methods. In *Springer Handbook of Speech Processing*. Benesty, J. and Sondhi, M.M. and Huang, Y. (Eds), Springer Press, 2007.

[89] E.D. Petajan. *Automatic Lipreading to Enhance Speech Recognition.* PhD thesis, University of Illinois, 1984.

[90] D.T. Pham, C. Serviere, and H. Boumaraf. Blind separation of convolutive audio mixtures using nonstationarity. *ICA*, April, 2003.

[91] G. Potamianos, C. Neti, and S. Deligne. Joint audio-visual speech processing for recognition and enhancement. *Proceedings of the Auditory-Visual Speech Processing Tutorial and ResearchWorkshop (AVSP), France*, pages 95–104, September 2003.

[92] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. *Audio-Visual Automatic Speech Recognition: An Overview.* E. Vatikiotis-Bateson, G. Bailly, and P. Perrier (Eds.), MIT Press, 2008. ISBN: 0-26-222078-4.

[93] L.R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, Vol 77(Issue 2):pp. 257–286, 1989.

[94] S. Rajaram, A.V. Nefian, and T.S. Huang. Bayesian separation of audio-visual speech sources. *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP).*, Montreal, Canada, 2004.

[95] B. Rivet, A. Aubrey, L. Girin, Y. Hicks, C. Jutten, and J. Chambers. Development and comparison of two approaches for visual speech analysis with application to voice activity detection. *Proc Int Auditory-Visual Speech Processing (AVSP)*, 2007.

[96] B. Rivet, L. Girin, and C. Jutten. Solving the Indeterminations of Blind Source Separation of Convolutive Speech Mixtures. In *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP).*, Philadelphia, 2005.

[97] B. Rivet, L. Girin, and C. Jutten. Visual voice activity detection as a help for speech source separation from convolutive mixtures. *Speech Communication.*, 49(7-8):667–677, 2007.

[98] B. Rivet, L. Girin, and C. Jutten. Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. *IEEE Transactions on Audio, Speech and Language Processing*, 15(1):96–108, Jan. 2007.

[99] B. Rivet, L. Girin, C. Jutten, and J-L. Schwartz. Using audiovisual speech processing to improve the robustness of the separation of convolutive speech mixtures. *IEEE 6th Workshop on Multimedia Signal Processing*, pages 47–50, 2004.

[100] B. Rivet, L. Girin, C. Serviere, D-T. Pham, and C. Jutten. Using a visual voice activity detector to regularize the permutations in blind separation of convolutive speech mixtures. pages 223–226, July 2007.

[101] E. Robledo-Arnuncio and B.H. Juang. Issues in frequency domain blind source separation- a critical revisit. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, 2005.

[102] A. Samani, J. Winkler, and M. Niranjan. Automatic face recognition using stereo images. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[103] S. Sanei, S.M. Naqvi, J.A. Chambers, and Y. Hicks. A geometrically constrained multimodal approach for convolutive blind source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP).*, April 2007.

[104] M.E. Sargin, E. Erzin, Y. Yemez, and A.M. Tekalp. Lip feature extraction based on audio-visual correlation. *EUSIPCO, Antalya, Turkey.*, 2005.

[105] H. Saruwatari, S. Kurita, and K. Takeda. Blind source separation combining frequency domain ICA and beamforming. *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP).*, 2001.

[106] H. Sawada, M. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutaion problem of frequency domain blind source separation. *IEEE Trans Speech and Audio Processing*, Vol 12(05), September 2004.

[107] H. Sawada, R. Mukai, S. Araki, and S. Makino. Frequency domain blind source separation. In J. Benesty, S. Makino, and J. Chen, editors, *Speech Enhancement*, pages 299–327. Springer, 2005.

[108] M. Siracusa, L-P. Morency, K. Wilson, J. Fisher, and T. Darrell. A multi-modal approach for determining speaker location and focus. *In Proceedings of the International Conference on Multi-modal Interfaces (ICMI), 2003.*

[109] P. Smaragdis. Information theoretic approaches to source separation. Master's thesis, MAS Department, Massachusetts Institute of Technology, 1997.

[110] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22:21–34, 1998.

[111] D. Sodoyer, L. Girin, C. Jutten, and J.L. Schwartz. Further experiments on audio-visual speech source separation. *International Conference on Audio-Visual Speech Processing*, September 2003.

[112] D. Sodoyer, L. Girin, C. Jutten, and J.L. Schwartz. Developing an audio-visual speech source separation algorithm. *Speech Communication*, 44(1-4):113–125, 2004.

[113] D. Sodoyer, B. Rivet, L. Girin, J.L. Schwartz, and C. Jutten. An Analysis of Visual Speech Information Applied to Voice Activity De-

tection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP).*, pages 601–604, Toulouse, France, 2006.

[114] D. Sodoyer, J.L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten. Separation of audio-visual speech sources: A new approach exploiting theaudio-visual coherence of speech stimuli. *Eurasip Journal on Applied Signal Processing*, pages 1165–1173, 2002.

[115] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision.* Brooks/Cole Publishing Company, 2nd edition, 1999.

[116] S. Stevens, J. Volkman, and E. Newman. A scale measurement of the psychological magnitude of pitch. *Journal of the Acoustical Society of America*, 8:185–190, 1937.

[117] G. Strang and T. Nguyen. *Wavelets and Filter Banks.* Wellesley Cambridge Press, 1996.

[118] W. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *Journal Acoustical Society of America.*, 26:212–215, 1954.

[119] R. Taagepera, K.C. Yow, and R. Cipolla. Feature-based human face detection. *Image and Vision Computing*, 15, Number 9, September 1997 , pp. (23)(9):713–735, 1997.

[120] S. Tamura, K. Iwano, and S. Furui. Multi-modal speech recognition using optical-flow analysis for lip images. *J. VLSI Signal Process. Syst.*, 36(2-3):117–124, 2004.

[121] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.

[122] K.C. van Bree and H.J.W. Belt. The use of a formant diagram in audiovisual speech activity detection. 15^{th}*European Signal Processing Conference (EUSIPCO)*, 2007.

[123] A.J.W. van der Kouwe, D. Wang, and G.J. Brown. A comparison of auditory and blind separation techniques for speech segregation. *IEEE Transactions on Speech and Audio Processing*, 9(3):189–195, 2001.

[124] P. Viola and M.J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, 2004.

[125] W. Wang, J. Chambers, and S. Sanei. A joint diagonalization method for convolutive blind separation of nonstationary sources in the frequency domain. *Proc, ICA, Nara, Japan*, April 2003.

[126] W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. Chambers. Video assisted speech source separation. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, 2005.

[127] W. Wang, S. Sanei, and J. Chambers. Penalty function based joint diagonalization approach for convolutive blind source separation of nonstationary sources. *IEEE Transactions on Signal Processing*, 53(05):1654–69, May 2005.

[128] G. Yang and T.S. Huang. Human face detection in a complex background. *Pattern Recognition*, 27(1):53–63, 1994.

[129] M.H. Yang, D.J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans on Pattern Analysis and Machine Intelligence (PAMI)*, Vol 24(No 1), January, 2002.

[130] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behaviour. *Speech Communication*, 26(1):23–43, 1998.