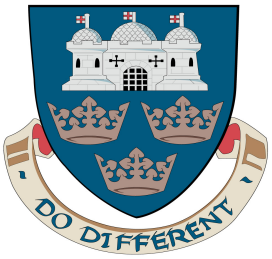
Audio-Visual Speaker Separation

Faheem Khan

A thesis submitted for the degree of
Doctor of Philosophy

School of Computing Sciences
University of East Anglia



August 2016

Dedicated to my parents, teachers and my wife and children.....

Abstract

Communication using speech is often an audio-visual experience. Listeners hear what is being uttered by speakers and also see the corresponding facial movements and other gestures. This thesis is an attempt to exploit this bimodal (audio-visual) nature of speech for speaker separation. In addition to the audio speech features, visual speech features are used to achieve the task of speaker separation. An analysis of the correlation between audio and visual speech features is carried out first. This correlation between audio and visual features is then used in the estimation of clean audio features from visual features using Gaussian Mixture Models (GMMs) and Maximum a Posteriori (MAP) estimation.

For speaker separation three methods are proposed that use the estimated clean audio features. Firstly, the estimated clean audio features are used to construct a Wiener filter to separate the mixed speech at various signal-to-noise ratios (SNRs) into target and competing speakers. The Wiener filter gains are modified in several ways in search for improvements in quality and intelligibility of the extracted speech. Secondly, the estimated clean audio features are used in developing visually-derived binary masking method for speaker separation. The estimated audio features are used to compute time-frequency binary masks that identify the regions where the target speaker dominates. These regions are retained and form the estimate of the target speaker's speech. Experimental results compare the visually-derived binary masks with ideal binary masks which shows a useful level

of accuracy. The effectiveness of the visually-derived binary mask for speaker separation is then evaluated through estimates of speech quality and speech intelligibility and shows substantial gains over the original mixture. Thirdly, the estimated clean audio features and the visually-derived Wiener filtering are used to modify the operation of an effective audio-only method of speaker separation, namely the soft mask method, to allow visual speech information to improve the separation task. Experimental results are presented that compare the proposed audio-visual speaker separation with the audio-only method using both speech quality and intelligibility metrics. Finally, a detailed comparison is made of the proposed and existing methods of speaker separation using objective and subjective measures.

Acknowledgements

I am very much thankful to my supervisor Dr Ben Milner for his supervision, guidance, support and patience.

Contents

List of Figures **ix**

List of Tables **xiii**

Statement of Originality **xv**

Acronyms **xvi**

1 Introduction **1**
1.1 Aims . . . 2
1.2 Introduction . . . 2
1.2.1 Bimodality of speech . . . 2
1.2.2 Speaker separation . . . 4
1.3 Motivations . . . 8
1.4 Thesis structure . . . 9

2 Audio and Visual Feature Extraction and Correlation Analysis **12**
2.1 Speech production . . . 13
2.1.1 Lungs . . . 13
2.1.2 Larynx and vocal folds . . . 14
2.1.3 Vocal tract . . . 14
2.2 Classification of speech sounds . . . 15
2.2.1 Phonemes . . . 16
2.2.2 Visemes . . . 16
2.3 Databases used in this work . . . 18
2.3.1 Messiah database . . . 19
2.3.2 LIPS2008 database . . . 20
2.3.3 GRID database . . . 21
2.4 Audio features . . . 22
2.4.1 Mel-Scale filterbank features . . . 22
2.4.2 MFCC . . . 23

2.5 Visual features . . . 24
2.5.1 Two-Dimensional Discrete Cosine Transform . . . 27
2.6 Correlation measurement . . . 30
2.6.1 Audio-visual correlation . . . 33
2.6.2 Audio-visual correlation Messiah database . . . 33
2.6.3 Audio-visual correlation LIPS2008 database . . . 33
2.6.4 Audio-visual correlation GRID database . . . 34
2.7 Summary . . . 39

3 Estimation of Clean Audio Speech Features from Visual Features 40
3.1 Introduction . . . 41
3.2 Estimation of audio features from visual features . . . 41
3.2.1 Augmenting audio and visual feature vectors . . . 42
3.2.2 GMM training . . . 43
3.2.3 MAP estimation of audio features . . . 44
3.3 Interpolation of filterbank features . . . 45
3.4 Experiments . . . 46
3.4.1 Audio and visual features and databases . . . 46
3.4.2 Results . . . 47
3.4.3 Filterbank estimation errors . . . 48
3.4.4 Log power spectral estimation errors . . . 51
3.5 Summary . . . 56

4 Speaker Separation Using Wiener Filtering 57
4.1 Introduction . . . 58
4.2 Visually-derived Wiener filtering for speaker separation . . . 63
4.2.1 Perceptual gain transformation . . . 64
4.3 Estimation of audio features from video . . . 66
4.4 Implementation . . . 67
4.4.1 Perceptual gain calculation . . . 67
4.4.2 Speaker separation . . . 68
4.5 Experimental results . . . 70
4.5.1 Audio-visual data . . . 70
4.5.2 Speech quality . . . 71
4.5.3 Speech intelligibility . . . 80
4.6 Summary . . . 83

5 Speaker Separation using Visually-derived Binary Masks 85
5.1 Introduction . . . 86
5.2 Visually-derived binary masks . . . 89
5.2.1 Mixing model . . . 89

5.2.2 Estimation of binary mask . . . 90
5.2.3 Time-domain reconstruction . . . 92
5.3 Estimation of audio features from video . . . 92
5.3.1 Audio and visual features . . . 93
5.4 Experimental results . . . 93
5.4.1 Audio-visual data . . . 93
5.4.2 Mask accuracy . . . 94
5.4.3 Effect of number of channels on visually-derived and ideal binary masks . . . 96
5.4.4 Filterbank estimation accuracy . . . 97
5.4.5 Speech quality . . . 99
5.4.6 Speech intelligibility . . . 103
5.5 Summary . . . 108

6 Exploiting audio and visual information for single-channel speaker separation 110
6.1 Introduction . . . 111
6.2 Audio-only speaker separation . . . 113
6.2.1 Relation to binary masking . . . 118
6.3 Audio-visual speaker separation . . . 119
6.3.1 AV-Alpha . . . 120
6.3.2 AV-Beta . . . 122
6.3.3 AV-VW . . . 123
6.4 Estimation of audio features from video . . . 125
6.5 Experimental results . . . 125
6.5.1 Experimental set up . . . 126
6.5.2 AV-Alpha . . . 127
6.5.3 AV-Beta . . . 130
6.5.4 AV-VW . . . 132
6.5.5 Comparison of A-only, AV-Alpha and AV-VW methods . . . 136
6.6 Summary . . . 137

7 Comparsions of proposed and existing methods 142
7.1 Introduction . . . 143
7.1.1 Proposed methods . . . 144
7.1.2 Existing methods . . . 145
7.2 Experimental Results . . . 145
7.2.1 Audio and visual data . . . 145
7.2.2 Experimental set up for subjective tests . . . 146
7.2.3 Objective measures . . . 150
7.2.4 Subjective measures . . . 156

7.3 Summary 158

8 Conclusions and future work **161**
8.1 Review 161
8.2 Conclusions 164
8.3 Future work 166

References **168**

List of Figures

1.1 The single-channel speaker separation task 5
1.2 Speaker separation task using a single audio channel and visual speech information 8

2.1 Speech production system in humans [69] 13
2.2 Sketch of vocal folds, looking down the larynx, in two states: (a) voicing and (b) breathing [69] 15
2.3 The classification of the phonemes in the English language [1]. . . . 17
2.4 Phoneme to viseme mapping for the Messiah database [1]. 18
2.5 Phoneme to viseme mapping for the TIMIT database [10]. 19
2.6 Frames showing the variability of speech articulators during the articulation of /t/ in different contexts [94]. 20
2.7 Example frames from the Messiah database. 20
2.8 Example frames from the LIPS2008 database. 21
2.9 Filterbank extraction process [1] 23
2.10 Mel-frequency versus linear frequency 24
2.11 First four filterbanks of the Messiah database first utterance 25
2.12 First four filterbanks of the LIPS2008 database first utterance . . . 26
2.13 MFCC extraction process [1]. 26
2.14 Example from the Messiah database:(a) 180×100 pixel ROI, (b) Two Dimensional-Discrete Cosine Transform (2D-DCT) of ROI, (c) zigzag ordering of 2D-DCT feature vector 28
2.15 Example from the LIPS2008 database:(a) 120×100 pixel ROI, (b) 2D-DCT of ROI, (c) zigzag ordering of 2D-DCT feature vector . . . 29
2.16 Reconstruction of an image using different percentages of 2D-DCT coefficients in the inverse 2D-DCT process, (a) 100 %, (b) 40 %, (c) 20 %, (d) 10 % 29
2.17 First four (2 to 5) 2D-DCT features of the Messiah database first utterance 30
2.18 First four (2 to 5) 2D-DCT features of the LIPS2008 database first utterance 31

2.19 First four filterbank features of the Messiah database first utterance (top left), First four 2D-DCT features of Messiah database first utterance (top right), First four filterbank features of the LIPS2008 database first utterance (bottom left), First four 2D-DCT features of LIPS2008 database first utterance (bottom right) 32
2.20 Correlation of 15 dimensional 2D-DCT features to each channel of 23 dimensional filterbank features of the Messiah database 34
2.21 Correlation of 15 dimensional 2D-DCT features to each channel of 23 dimensional filterbank features of the LIPS2008 database 35

3.1 GMM creation and training. 42
3.2 Time domain waveform and filterbanks channel 10 and channel 15 of the utterance '*Ada aims to serve chump chops with chips cooked in pure oil, with ice-cream to follow*', of the Messiah database: (a) reference waveform, (b) filterbank channel 10, (c) filterbank channel 15 54
3.3 Time domain waveform and filterbanks channel 10 and channel 15 of the utterance '*Ada aims to serve chump chops with chips cooked in pure oil, with ice-cream to follow*', of the LIPS2008 database: (a) reference waveform, (b) filterbank channel 10, (c) filterbank channel 15 55

4.1 Speaker separation using visually derived Wiener filtering including the training stage. 62
4.2 Perceptual gain functions 65
4.3 The process of extracting a target speaker frame from the mixed speech using visually derived Wiener filtering: a) estimated speaker 1 filterbank, b) estimated speaker 2 filterbank, c) interpolated estimated speaker 1 filterbank, d) interpolated estimated mixed (noisy) filterbank, e) Wiener filter gain, f) Mixed magnitude spectrum, g) extracted and reference magnitude spectrums. 69
4.4 *Output SIR variations with Input SIR for the target speaker of: top) Messiah database, bottom) GRID database.* 73
4.5 *SDR variations with SNR for the target speaker of: top) Messiah database, bottom) GRID database.* 75
4.6 *SAR variations with SNR for the target speaker of: top) Messiah database, bottom) GRID database.* 77
4.7 Spectrograms of the utterance '*Bin blue at e nine please*': a) target speaker (male), b) mixed with competing speaker at an SIR of 0dB, c) visually-derived speaker separation using $H2$ with α=0.4 78

4.8 Spectrograms of the utterance '*Set white with v four soon*': a) competing speaker (female), b) mixed with target speaker at an SIR of 0dB, c) visually-derived speaker separation using $H2$ with α=0.4 . . 79
4.9 *Accuracy variations (%) with SNR for the target speaker of: top) Messiah database, bottom) GRID database.* 81

5.1 Speaker separation using visually derived binary masking including the training stage. 88
5.2 *Binary masks: a) 128-channel ideal, b) 2-channel ideal, c) 2-channel visually-derived.* 100
5.3 *Binary masks: a) 128-channel ideal, b) 23-channel ideal, c) 23-channel visually-derived.* 101
5.4 *Binary masks: a) 128-channel ideal, b) 50-channel ideal, c) 50-channel visually-derived.* 102
5.5 *Spectrograms showing: a) target speaker saying* 'Higher oil prices may amaze those thinking of investing their money*', b) interfering speaker saying* 'Zulu warriors have sure ideas when watching a video yeti eat pure nectarines*' c) target speaker mixed with interfering speaker at an SIR of 0dB , d) target speaker extracted using D=2 channels.* 104
5.6 *Spectrograms showing: a) target speaker saying* 'Higher oil prices may amaze those thinking of investing their money*', b) interfering speaker saying* 'Zulu warriors have sure ideas when watching a video yeti eat pure nectarines*' c) target speaker mixed with interfering speaker at an SIR of 0dB , d) target speaker extracted using D=23 channels.* 105
5.7 *Spectrograms showing: a) target speaker saying* 'Higher oil prices may amaze those thinking of investing their money*', b) interfering speaker saying* 'Zulu warriors have sure ideas when watching a video yeti eat pure nectarines*' c) target speaker mixed with interfering speaker at an SIR of 0dB , d) target speaker extracted using D=50 channels.* 106

6.1 *SIR when varying α from 0 to 1 in Equation 6.20. The small red circles are showing the peak values.* 128
6.2 *SDR when varying α from 0 to 1 in Equation 6.20. The small red circles are showing the peak values.* 129
6.3 *SAR when varying α from 0 to 1 in Equation 6.20.* 130
6.4 *SIR, SDR and recognition accuracy when $\alpha = 0.35$ and varying β from 0 to 1 in Equation 6.21.* 131

6.5 *SIR when varying α from 0 to 1 in Equation 6.22. The small red circles are showing the peak values.* . . . 133
6.6 *SDR when varying α from 0 to 1 in Equation 6.22. The small red circles are showing the peak values.* . . . 134
6.7 *SAR when varying α from 0 to 1 in Equation 6.22.* . . . 135
6.8 *Comparisons of SIR gains for A-only, AV-Alpha and AV-VW methods, for optimal values of α, for SNRs of -20dB to +20dB.* . . . 137
6.9 *Comparisons of SDR gains for A-only, AV-Alpha and AV-VW methods, for optimal values of α, for SNRs of -20dB to +20dB.* . . . 138
6.10 *Comparisons of SAR gains for A-only, AV-Alpha and AV-VW methods, for optimal values of α, for SNRs of -20dB to +20dB.* . . . 139
6.11 *Comparisons of word recognition accuracy (%) for A-only, AV-Alpha and AV-VW methods, for optimal values of α, for SNRs of -20dB to +20dB.* . . . 140

7.1 Speaker separation task: To extract the target male speaker and suppress the competing female speaker. . . . 149
7.2 Graphical User Interface (GUI) for the training phase of the listening test. . . . 149
7.3 GUI for the evaluation phase of the listening test. . . . 150
7.4 *SIR gains comparisons for the various methods at the shown input SNRs for the target speaker.* . . . 151
7.5 *SDR comparisons for the various methods at the shown input SNRs for the target speaker.* . . . 153
7.6 *SAR comparisons for the various methods at the shown input SNRs.* 154
7.7 *Word recognition accuracy for the various methods at the shown input SNRs for the target speaker.* . . . 155
7.8 *Speech quality rating in terms of the suppression of the competing speaker for the different methods at various input SNRs.* . . . 157
7.9 *Speech quality rating in terms of the preservation of the target speaker for the different methods at various input SNRs.* . . . 158
7.10 *Overall quality ratings of the target speaker for the different methods at various input SNRs.* . . . 159

List of Tables

2.1 *GRID database sentence grammar.* 21
2.2 *Average correlation across all the channels for different sizes of visual vector and different number of clusters for speaker 6 of GRID database. ASize and VSize are representing the sizes of audio and visual features vectors respectively in the augmented AV-vector.* . . . 36
2.3 *Average correlation across all the channels for speaker 4 of GRID database for filterbank features.* 37
2.4 *Average correlation across all the channels for different sizes of visual vector and and different number of clusters in the GMM for speaker 6 of GRID database.* 38
2.5 *Average correlation across all the channels for speaker 4 of GRID database for log power spectral features.* 39

3.1 *Mean percentage filterbank estimation errors for different sizes of visual vector and different number of clusters for speaker 6 of GRID database. ASize and VSize are representing the sizes of audio and visual features vectors respectively in the augmented AV-vector.* . . . 49
3.2 *Mean percentage filterbank estimation errors for speaker 4 of the GRID database.* 50
3.3 *Mean percentage filterbank estimation errors for different number of clusters for the Messiah database.* 50
3.4 *Mean percentage filterbank estimation errors for clusters sizes from 16 to 32 for the LIPS2008 database.* 51
3.5 *Mean percentage LPS estimation errors for different sizes of visual vector and and different number of clusters in the GMM for speaker 6 of the GRID database.* 52
3.6 *Mean percentage LPS estimation errors for speaker 4 of the GRID database.* 53

5.1 *Visually-derived mask estimation accuracy (%) at SIRs from -10dB to +20dB and filterbank sizes from 2 to 50 channels for Messiah database* 95
5.2 *Comparison of the accuracy (%) of the visually-derived binary masks and ideal binary masks subject to filterbank quantisation, for filterbank sizes from 2 to 50 channels at an SIR of 0dB for Messiah database.* 96
5.3 *Comparison of the filterbank estimation errors (%) of the visually-derived filterbank audio features and ideal filterbank audio features subject to filterbank quantisation, for filterbank sizes from 2 to 50 channels at an SIR of 0dB for Messiah database.* 97
5.4 *Comparison of input and output SIRs for filterbank sizes from 2 to 50 channels.* 99
5.5 *Comparison of input and output SIRs for the target speaker of GRID database for $D = 23$ channels.* 103
5.6 *Target speaker monophone recognition accuracy (%) at SIRs from -10dB to +20dB for filterbank sizes from 2 to 50 channels.* 107
5.7 *GRID database target speaker word accuracy (%) at SIRs from -10dB to +20dB for filterbank sizes of 23 channels.* 108

Statement of Originality

This thesis is an attempt to exploit the bimodal (audio-visual) nature of speech for speaker separation and has contributed three speaker separation techniques. The following are the publications that has resulted from the work in this thesis:

- F. Khan and B. Minler. Single-channel audio speaker separation using visual speech features. In *Interspeech 2013, Lyon France*
- F. Khan and B. Minler. Speaker separation using visually-derived binary masks. In *AVSP 2013, Annecy France*
- F. Khan and B. Minler. Using audio and visual information for single channel speaker separation. In *Interspeech 2015, Dresden Germany*

Acronyms

MFCC	Mel-Frequency Cepstral Coefficients
2D-DCT	Two Dimensional-Discrete Cosine Transform
ROI	Region of Interest
DCT	Discrete Cosine Transform
AAM	Active Appearance Model
GMM	Gaussian Mixture Model
MAP	Maximum a Posteriori
AV	Audio-Visual
VAD	Voice Activity Detection
EM	Expectation Maximization
PDF	Probability Density Function
PCHIP	Piecewise Cubic Hermite Interpolating Polynomial
EI	Early Integrated
LI	Late Integrated
SDR	Signal-to-Distortion Ratio
SNR	Signal-to-Noise Ratio
SIR	Signal-to-Interference Ratio

SAR Signal-to-Artefact Ratio

CASA Computational Auditory Scene Analysis

ASA Auditory Scene Analysis

ITD Interaural Time Differences

ILD Interaural Level Differences

ASR Automatic Speech Recognition

BSS Blind Source Separation

CMU Carnegie Mellon University

TIMIT Texas Instruments-Massachusetts Institute of Technology

ICA Independant Component Analysis

GUI Graphical User Interface

Chapter 1

Introduction

Preface

This chapter starts with a brief explanation of the aims of this thesis. Then the bimodal (audio-visual) nature of speech and its importance in human speech perception and in speech processing applications such as speech enhancement and Automatic Speech Recognition (ASR) is highlighted. This is followed by a categorization of the speaker separation methods based on their purpose, number of channels and the underlying algorithms. Then an overview of speaker separation methods both in audio only and audio-visual domains is presented. Then the motivations that has led to the undertaking of this work are explained. And finally a sketch of the thesis structure is presented.

1.1 Aims

The main and final aim of this thesis is to use the bimodal nature of speech for speaker separation. This main aim is achieved by dividing it into several sub aims. The first sub aim is to determine and maximise the correlation between the audio and visual features of different speakers. The second sub aim is to exploit this audio-visual correlation in the estimation of clean audio features from the visual features. The third sub aim is to use the estimated clean audio features in the construction of visually-derived Wiener filter and visually-derived binary masking for speaker separation. The fourth sub aim is to improve the performance of the visually-derived Wiener filter by using perceptual gain functions. The fifth sub aim is to enhance the accuracy of the estimation of visually-derived binary masks for speaker separation. The sixth sub aim is to use the visual information in the form of the estimated clean audio features to improve the performance of an audio-only soft mask method of speaker separation.

1.2 Introduction

This section introduces first the bimodal nature of speech. Then speaker separation which is the aim of this thesis is introduced. An overview of speaker separation systems is presented both in audio and audio-visual domains.

1.2.1 Bimodality of speech

Communication using speech is often an audio-visual experience. Listeners hear what is being uttered by speakers and also see their corresponding facial move-

ments and other gestures. The facial movements include the movements of lips, jaws, tongue and eyes and sometimes the entire head. Other movements can be the gestures made with hands or sometimes the entire body. Humans use these visual cues in addition to the audio to enhance their understanding of the uttered speech. In this way the speech perception system in humans exploits this bimodal nature of speech by integrating the audio and visual streams of information to form a better perception of what is being uttered [90]. The visual stream becomes more important when the SNR is low because the audio stream is susceptible to acoustic noise but the visual stream is not affected by acoustic noise. It was reported in [90] that it was shown in [93], that the visual stream can cause an increase in the intelligibility of audio speech that could be caused by a 16dB decrease in acoustic noise.

On one hand visual speech is sometimes ambiguous as one set of facial movements can be interpreted for different words as many words have similar visual appearance such as in the words 'bob', 'bop', and 'pop'. But on the other hand, visual speech also has the ability to differentiate between many acoustically ambiguous word pairs like 'met' and 'net'. As this pair is acoustically ambiguous but visually it is clearly different. The two nasal consonants /n/ and /m/, at the beginning of the words, are visually different, as the lips are closed while uttering /m/ but for for /n/ the lips remain open [90], [89]. In the McGurk effect [62], the effect of vision upon speech perception was explained. The utterance 'ba' was dubbed on to the visual cues for 'ga', normal adults reported it as 'da'. These adults recognised these sounds accurately when they listened to the audio only or when they watched the unmodified film.

1.2.2 Speaker separation

Speech enhancement and ASR systems are becoming more and more helpful and handy in everyday life. The performance of these systems drop considerably in the presence of noise. This noise can be of different forms such as car noise, babble noise, background music and competing speakers. Competing speech is considered to be the most challenging type of noise in automatic speech recognition and speech enhancement systems because high correlation in the temporal structure of the target and competing speech exists and the acoustic features of the target speech can be easily confused with that of the competing speech [31],[89],[77]. Another reason that adds to this challenge is the highly non-stationary nature of the competing speech that can vary instantaneously and results in the variation of the noise estimate and the reliability of the target speech [90]. This task becomes more challenging when only one channel recording of the mixed speech is available because information regarding the source location is missing in this case [31].

The performance of ASR and speech enhancement systems is much reduced in the presence of competing speech as compared to human listeners. Human listeners have excellent abilities to either mask the unwanted speech or extract the target speech or both to have a better perception of the target speech in the presence of different unwanted sources [31]. Speaker separation is the process of extracting a target speaker from a mixture of sounds that comprises other speakers and acoustic noise. Single-channel speaker separation where only single-microphone recording of the mixture is available is the aim of this thesis. The task of single-channel speaker separation is shown in Figure 1.1. Although multiple microphones are preferred for speaker separation as in this case spatial information can be ex-

ploited but existence of multiple microphones is not always the case. For example, in ASR of radio broadcasts and teleconferencing. Therefore, single-channel systems have a role to play [31]. Auditory Scene Analysis (ASA) explains the speech

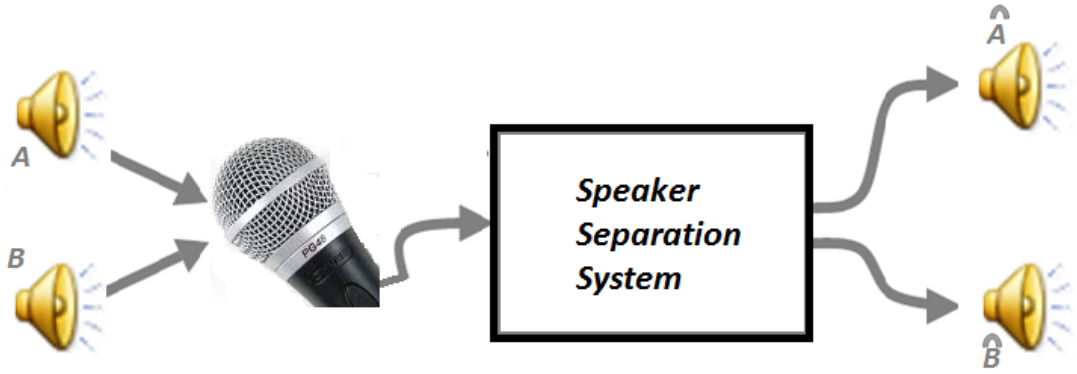


Figure 1.1: The single-channel speaker separation task

separation process in humans in the presence of competing speech [11]. Computational Auditory Scene Analysis (CASA) techniques [47], [12], [16], [22], [63],, [35], imitate the ASA principles by using computers. CASA is summarised as a two-step process: segmentation and regrouping. The mixed speech is decomposed into smaller units and then the similar units are grouped together to be used in the construction of the target speech. CASA techniques were reviewed in [102].

In [31], speaker separation systems were classified based on their characteristics as: single-channel-based or multiple-channel based, knowledge-based or statistical-based (model-based) and recognition-based or speech enhancement based. Multiple-channel based systems use multiple microphones for the collection of audio signals and use the spatial cues such as Interaural Time Differences (ITD) and Interaural Level Differences (ILD) to distinguish between the wanted and unwanted sources [56]. Most of the time, only one microphone is available and in this case, no spatial

cues are available. Single-channel systems rely on the information carried by the speech signal itself. Model-based systems [83], [30], [44], [71], [73], [72], [45], need enough training data to learn the corresponding speech statistics. Knowledge-based (CASA) systems implement the knowledge known about human speech perception using machines. The performance of model-based systems is often better than CASA systems because model-based systems use training data and CASA systems replicate the human speech perception process which is still not exactly known [31],[102]. The objective of speech-enhancement-based systems is to improve the quality of the wanted speech while the objective of recognition-based systems is to improve the intelligibility of ASR systems. Model-based single-channel speaker separation (SCSS) techniques can be considered similar to model-based single-channel speech enhancement (SCSE) techniques [24], [92], [60]. The difference is the non-stationary nature of both the target and interfering sources in the case of SCSS [70].

Audio-only speaker separation is well established when multiple microphone channels are available. Techniques such as deconvolution and Blind Source Separation (BSS) make the assumption that the various signals in the mixture are independent and exploit the set of input signals to extract individual audio sources [102], [103], [67], [85], [98], [21], [66], [86], [37], [87], [57]. Other work has considered the more difficult problem of speaker separation from a single audio channel. In this instance prior statistical knowledge of the speakers is utilised to enable extraction of the target speaker. Methods using spectral masking have been effective at solving this problem and use either hard or soft masks to identify time-frequency regions that belong to a target speaker [77], [84].

Visual speech information from a target speaker's mouth region has also been

used in multiple channel speaker separation to supplement audio-based methods of extracting a target speaker [78], [50], [51], [52], [54]. For example, in [50] a target speaker is first extracted from a speech mixture using audio BSS. Visual information from speakers is then used to address permutation and scaling ambiguities present after BSS. The method still uses multiple audio channels but supplements this information with visual information that increases the quality of the extracted target speech. Visual speech has also been used to aid single-channel speaker separation [33] by improving the accuracy of hidden Markov model (HMM) decoding of input speech signals, with the HMMs providing statistics on the speech to be separated.

In [81], an overview of the key methodologies of the audio-visual speaker separation methods was provided and the research activities in this area were broadly categorized as

- To robustly model the Audio-Visual (AV) correlation [78], [54].
- To combine the AV correlation with time-frequency (T-F) masking or Independant Component Analysis (ICA) [55].
- To use the AV correlation in resolving the permutation and scaling ambiguities present after BSS [78], [50], [51], [52], [54], [80].
- To use the visual information in Voice Activity Detection (VAD) algorithms [79], [91],[3], [49], [6], [53].
- To use the visual information in determining the position, direction of arrival and velocity of the moving sources [41], [64], [65].

This work examines whether the problem of speaker separation can be achieved through the use of visual speech information. When humans listen to audio sounds that comprise a mixture of different speakers, they are very good at extracting a target speaker from the various interfering speakers. Having two ears improves the situation but humans also exploit other cues such as observing visual speech information from the speakers. This work considers the scenario of a single-channel audio input and examines whether visual speech information can provide information to allow extraction of a target speaker from this mixture of sounds. The task is shown in Figure 1.2.

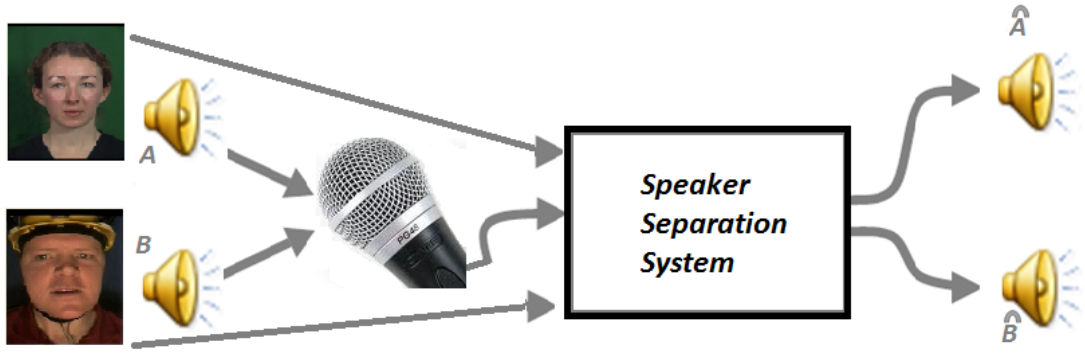


Figure 1.2: Speaker separation task using a single audio channel and visual speech information

1.3 Motivations

Although it has been shown that the speech signal is of bimodal nature and the visual component does carry information which improves speech perception [13], [62], [88], yet traditional speech processing applications such as speech recognition, speech enhancement, speech coding and speaker identification, all focus on

the audio component of the speech signal. To use this bimodal nature of the speech signal for speaker separation is the main motivation for this thesis. Secondly, the audio component of the speech signal is vulnerable to acoustic noise while the visual component does has the advantage that it is not contaminated by acoustic noise. Thus when the audio component becomes unreliable in noisy conditions, the visual component can play an important role. Thirdly, audio-visual speech processing algorithms can be a part of modern communication devices such as PCs, laptops and mobile phones as they all now have built-in cameras and their computational processing powers and memory capacities are increasing, making the deployment of these algorithms possible in real time.

1.4 Thesis structure

This section gives a short description of the thesis.

Chapter 2

This chapter starts with a description of the human speech production system. The relation between audio and visual speech units is described then in terms of phonemes to visemes mapping. This is followed by the description of the audio-visual databases used in this work and then the methods of audio and visual feature extraction are described. Then correlation between these audio and visual features is also described and analysed.

Chapter 3

This chapter exploits the audio-visual correlation for the estimation of clean audio features from visual features using a Gaussian Mixture Model (GMM) and Maximum a Posteriori (MAP) estimation. The accuracy of the estimation is also measured.

Chapter 4

This chapter explains the construction of a Wiener filter and perceptual gain functions for speaker separation using the clean speech estimates for the target and competing speakers made in Chapter 3. The speaker separation tasks are evaluated using different quality and intelligibility measures.

Chapter 5

This chapter explains the derivation of binary masks for speaker separation using visual features. The accuracy of the estimation of binary masks along with the factors affecting it is discussed. The quality and intelligibility of the extracted speech using visually derived binary masking is also measured along with the study of the factors affecting these measures.

Chapter 6

This chapter combines audio-only soft mask speaker separation with visual information with the aim to improve the quality and intelligibility of the extracted speech.

Chapter 7

This chapter provides a comparison of the proposed methods in this thesis with the existing methods using both objective and subjective measures for the evaluation of the quality and intelligibility of the extracted speech.

Chapter 8

This chapter presents a summary of the work and the conclusions derived from this thesis. Some directions for the future work are also suggested.

Chapter 2

Audio and Visual Feature Extraction and Correlation Analysis

Preface

This chapter gives a description of the human's speech production process along with the functioning of the main organs involved. The relation between audio and visual speech units is described in terms of phonemes to visemes mapping. The three AV speech databases used in this work are then described. Then methods for the extraction of the audio and visual speech features from these databases are discussed. Then correlation between these audio and visual features is discussed. And finally the correlation results for these three databases used are presented.

2.1 Speech production

The speech production process in humans involves several organs shown in Figure 2.1 [69].

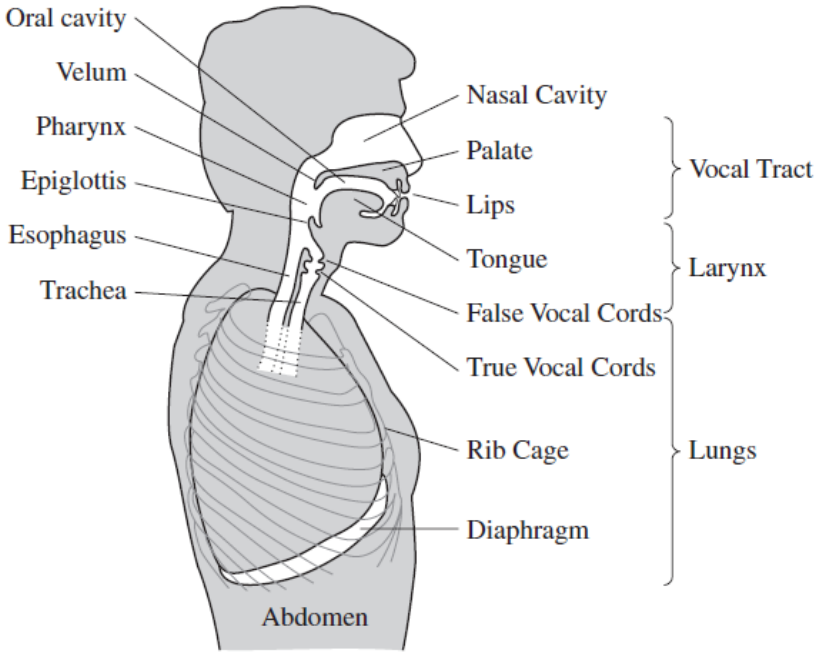


Figure 2.1: Speech production system in humans [69]

2.1.1 Lungs

Lungs oxygenate the blood in human body by inhaling and exhaling the air. This inhalation and exhalation of air also lead to the production of speech. During inhalation, the chest cavity expands, which decreases the air pressure in the lungs and allows air to flow in through vocal tract, trachea (windpipe) and into the lungs. During exhalation, the chest cavity contracts, which increases the air pressure in the lungs and allows air to flow out of the body through larynx and vocal tract

[56].

2.1.2 Larynx and vocal folds

The working of the two vocal cords (vocal folds) is controlled by the larynx. The vocal folds are shown in Figure 2.2 [69]. The gap between the two cords is called the glottis [56]. The vocal cords can have one of three states: voiced, unvoiced and breathing. When breathing, the vocal folds do not offer any resistance to the air flowing from the lungs through the glottis that is wide open. When in the voicing state, the vocal cords come closer to each other, a decrease and increase of tension of the cords, along with a decrease and increase in pressure of glottis, opens and closes the glottis periodically. When in the unvoiced state, the vocal folds do not vibrate. They are tenser and come closer to each other, that causes the air to be turbulent while it passes through the glottis. This air turbulence is known as aspiration, and occurs in normal speech when producing sounds like /h/ in "house" or when whispering [56]. Based on the states of vocal folds, speech is divided into voiced and unvoiced sounds. Sounds produced during the voicing state are called voiced sounds while those produced during unvoiced state are called unvoiced or voiceless sounds [56].

2.1.3 Vocal tract

The vocal tract consist of the pharynx cavity, nasal cavity and oral cavity, and extends from the lips and nose to the larynx as shown in Figure 2.1. Depending on the position of speech articulators namely the jaws, lips, teeth, tongue and velum, the oral cavity can have different cross-sectional areas and shapes. The

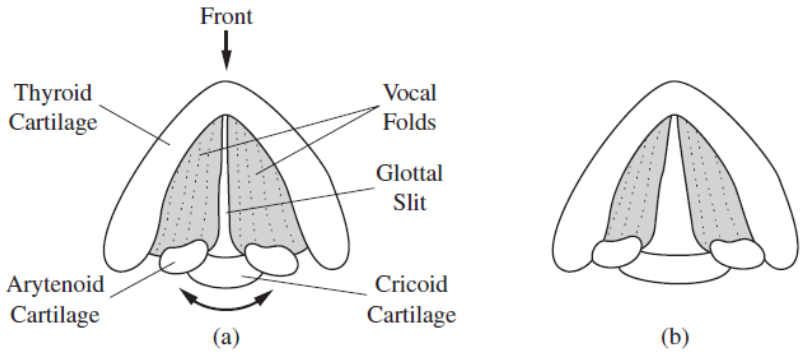


Figure 2.2: Sketch of vocal folds, looking down the larynx, in two states: (a) voicing and (b) breathing [69]

vocal tract functions as a physical linear filter and shapes the input into different sounds. The vocal tract resonates in response to sounds that contain frequencies that match the natural resonant frequencies of the volume of air. These resonances of the vocal tract are called formants and the frequencies at which they resonate are known as formant frequencies [56].

2.2 Classification of speech sounds

Speech sounds are described by the nature of the source such as voiced or unvoiced. The place and way of articulation is also used to describe them. The place and way of articulation of speech sounds is basically determined by the location of the tongue in the oral cavity and the related expansion and contraction in the vocal tract [56]. Speech sounds can be classified into smallest units called phonemes based on these descriptions i.e. the nature of the source and the place and way of articulation. In the visual domain, phonemes are represented by visemes.

2.2.1 Phonemes

Speech sounds can be represented in terms of phonemes. A phoneme is the smallest unit of a language that a listener can perceive [56]. As an example, the word "tan" can be represented in terms of three phonemes that are /t/, /ae/ and /n/ and these phonemes belong to three different classes called plosives or stops consonants, vowel class and nasal class respectively [56]. The British English language has 40-44 phonemes which are grouped as vowels, semi-vowels, consonants and diphthongs, as shown in Figure 2.3 [1].

2.2.2 Visemes

Visemes are used to represent phonemes in the visual domain [94]. They describe the facial positions and movements of speech articulators during the audition of a phoneme. The mapping between visemes and phonemes is not always one-to-one. Usually, several phonemes are mapped to a single viseme, because these phonemes have the same visible facial positions and movements of the speech articulators during their audition. Also some facial positions and movements of the speech articulators are not visible during the audition of several phonemes. This partial visibility of the speech articulators also pushes acoustically distinct phonemes to have common visemes [1].

It is difficult to have universally accepted phoneme to viseme mapping. Several studies have produced different phoneme to viseme mappings. As in [1], the 45 phonemes of the Carnegie Mellon University (CMU) phoneme set, for the Messiah database, were mapped into 14 visemes as shown in Figure 2.4 [1]. While in [10], the 46 phonemes of American English, for the Texas Instruments-Massachusetts

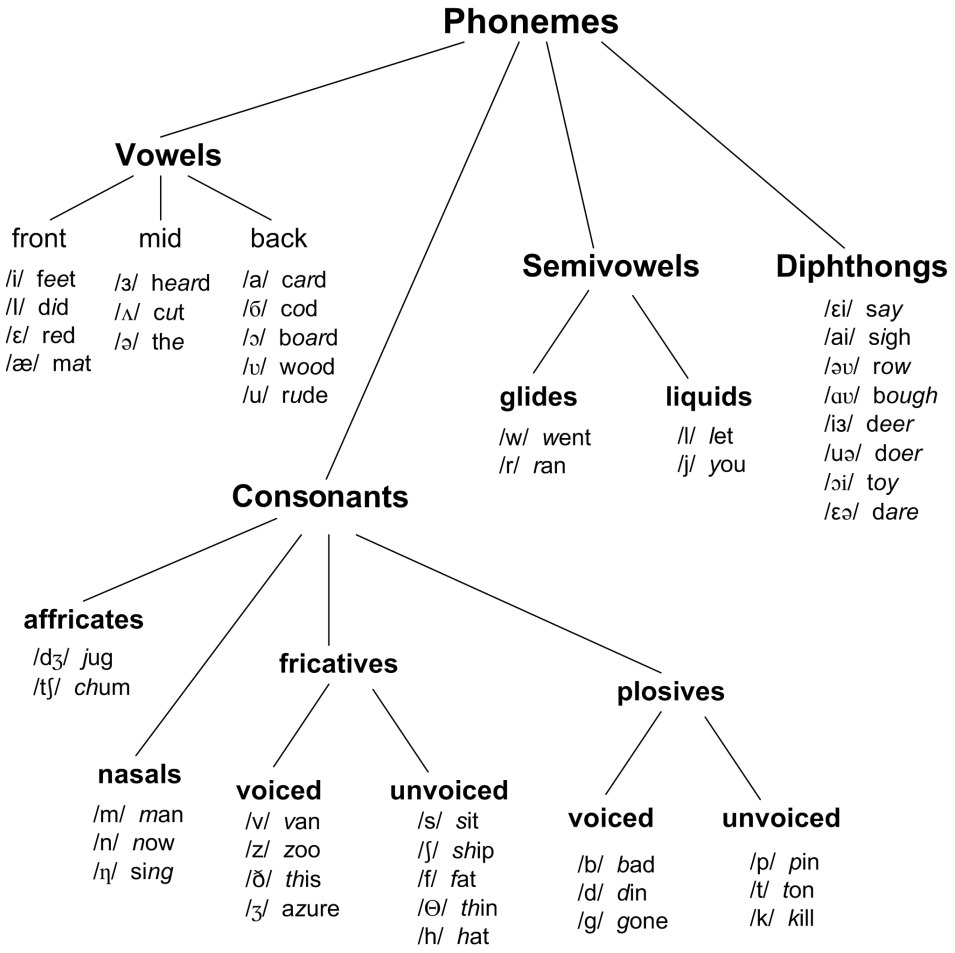
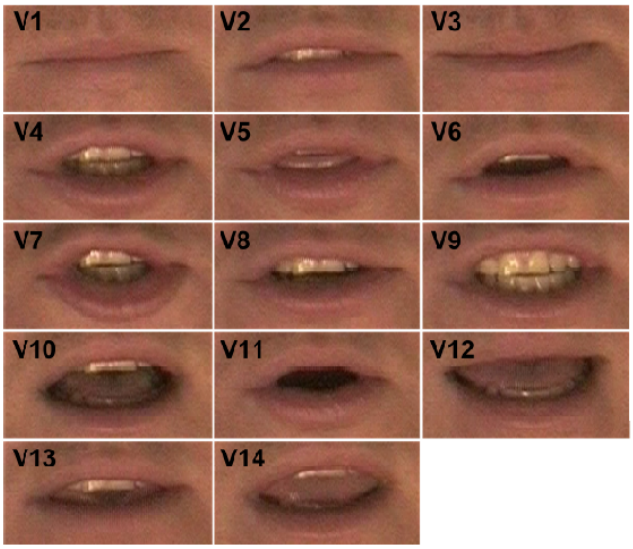


Figure 2.3: The classification of the phonemes in the English language [1].

Institute of Technology (TIMIT) speech database, were mapped into 16 visemes as shown in Figure 2.5 [10].

The phoneme to viseme mapping is badly affected by coarticulation, specially in continuous speech. Coarticulation means that neighbouring visemes affect the current viseme [1]. The effect of coarticulation on phoneme to viseme mapping can be seen clearly in Figure 2.6 [94].

Despite the shortcomings of visemes, they do help in having a good percep-



Mapping of 45 phonemes to 14 visemes

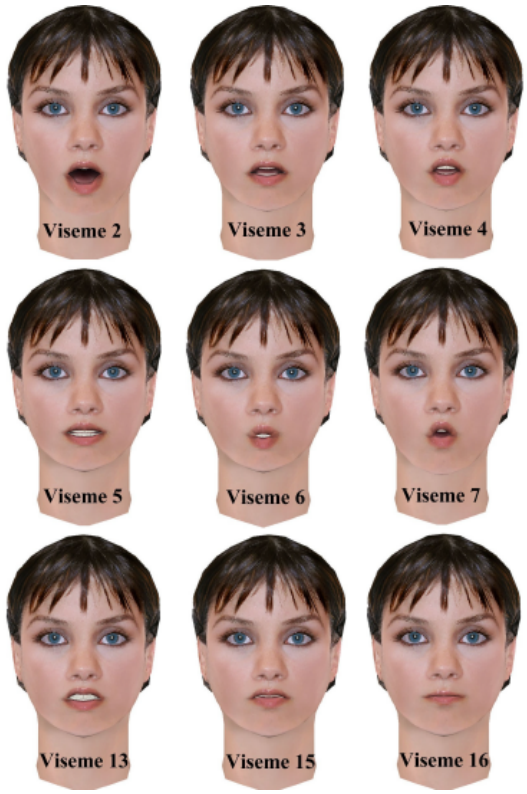
Viseme	Description	Phonemes
V1	silence	sil
V2	labiodental	f v
V3	bilabial	p b m
V4	alveolar	t d n
V5	dental	θ ð
V6	velar	k g w ŋ
V7	palato-alveolar	ch dʒ ʃ ʒ
V8	alveolar-semivowels	l r y
V9	alveolar-fricatives	s z
V10	*lip-rounding*	ɔ ɑ ʌ o oɪ aʊ h ɜˠ
V11	*and*	ʊ oʊ u ʊə
V12	*lip-spreading*	æ ai ei ɛ
V13	*based vowels*	ɪ i ə
V14		ea ia

Figure 2.4: Phoneme to viseme mapping for the Messiah database [1].

tion of speech. They have the ability to differentiate between many acoustically ambiguous word pairs like 'met' and 'net'. As this pair is acoustically ambiguous but visually it is clearly different. The two nasal consonants /n/ and /m/, at the beginning of the words, are visually different, as the lips are closed while uttering /m/ but for for /n/ the lips remain open [90], [89]. As shown in Figure 2.4 and Figure 2.5, these phonemes: /m/ and /n/ are mapped into different viseme classes.

2.3 Databases used in this work

Throughout this work, three audio-visual databases: Messiah database, LIPS2008 database and the GRID database, have been used. The following sections discuss these databases.



Mapping of 46 phonemes to 16 visemes

Viseme Classes	Timit Phoneset	Examples
1	pau	-
2	ay, ah	bite, but
3	ey, eh, ae	bait, bet, bat
4	er	bird
5	ix, iy, ih, ax, axr,y	debit, beet, bit, about, butter, yacht
6	uw, uh, w	boot, book, way
7	ao, aa, oy, ow	bought, bott, boy, boat
8	aw	bout
9	g, hh, k, ng	gay, hay, key, sing
10	r	ray
11	l, d, n, en, el, t	lay, day, noon, button, bottle, tea
12	s, z	sea, zone
13	ch, sh, jh, zh	choke, she, joke, azure
14	th, dh	thin, then
15	f, v	fin, van
16	m, em, b, p	mom, bottom, bee, pea

Figure 2.5: Phoneme to viseme mapping for the TIMIT database [10].

2.3.1 Messiah database

This audio-visual speech database consists of a frontal recording of a single British male speaker. The sentences are phonetically balanced and were spoken in a natural way with no emotions and no over articulation. The recordings were made using a camera mounted on a helmet worn by the speaker. It was tried to keep the head still throughout the recording. The video frames were recorded with a frame rate of 25 frames/sec and a resolution of 576×720 pixels. The audio signal was recorded using a camera built-in microphone at a sampling rated of 11025 Hz and a resolution of 16 bits/sample [95]. Throughout the experiments, the audio was down-sampled to 8 KHz and the frame rate of 100 frames per second was

Figure 2.6: Frames showing the variability of speech articulators during the articulation of /t/ in different contexts [94].

used. Some example frames from the Messiah database are shown in Figure 2.7.

Figure 2.7: Example frames from the Messiah database.

2.3.2 LIPS2008 database

This audio-visual speech database consists of a frontal recording of a single British female speaker. The sentences are phonetically balanced and were spoken in a natural way with no emotions and no over articulation. The recordings were made using a front camera not head mounted as in the case of the Messiah database. It was tried to keep the head still through out the recording. The video frames were recorded with a frame rate of 50 frames/sec and a resolution of 576×720 pixels.

The audio signal was recorded using a microphone placed near the speaker at a sampling rate of 44.1 KHz and a resolution of 16 bits/sample [96]. Throughout our experiments, we have used the audio down sampled to 8 KHz and the frame rate of 100 frames per second. Some example frames from the LIPS2008 database are shown in Figure 2.8.

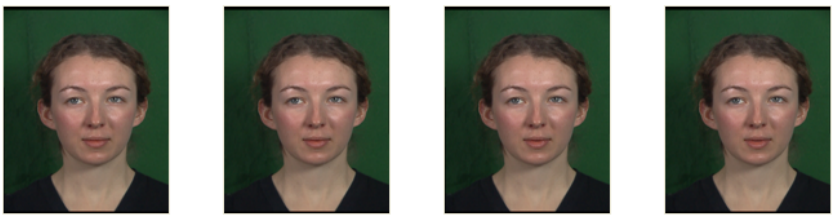


Figure 2.8: Example frames from the LIPS2008 database.

2.3.3 GRID database

This audio-visual speech database [15] contains the high quality frontal recordings of thirty four speakers. Each speaker has spoken 1000 simple sentences in a natural way with no emotions. Each utterance is of three seconds duration and consists of six words of the structure

command→colour→preposition→letter→digit→adverb.

The grammar of these sentences is shown in Table 2.1.

command	colour	preposition	letter	digit	adverb
bin	blue	at	A-Z	1-9,0	again
lay	green	by	excluding w		now
place	red	in			please
set	white	with			soon

Table 2.1: *GRID database sentence grammar.*

Speaker 6 (male) and speaker 4 (female) were chosen for the experiments in this work because of their clear articulation and minimal error rates [15]. The audio for both the speakers was down-sampled to a sampling frequency of 8KHz from 50KHz and the video was up-sampled to 100 frames per second from 25 frames per second to match the audio frame rate.

2.4 Audio features

Raw speech data are transformed into feature vectors in speech processing applications. These features extract the important aspects of the speech signal. Various methods have been developed for feature extraction from the speech signal. Some of the well-known methods are Mel-Frequency Cepstral Coefficients (MFCC), filterbank, formants and energy based methods. Through out this work filterbank and MFCC features are used. MFCC features are the standard in automatic speech recognition applications. Log filterbank features were shown to have higher correlation with the visual 2D-DCT features [1]. These features are extracted from 8 KHz sampled audio at a rate of 100 vectors per second.

2.4.1 Mel-Scale filterbank features

Filterbank features are a coarse spectral envelope type representation of speech that is obtained by quantising the power spectrum or magnitude spectrum across frequency. Linear or a perceptual scale can be used to place the filterbank overlapping channels. This perceptual scaling resolves frequencies non-linearly across the spectrum and is based upon human hearing. The filterbank feature extraction method is shown in Figure 2.9 [1]. In filterbank extraction the first step is

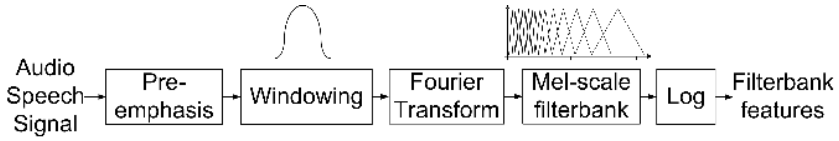


Figure 2.9: Filterbank extraction process [1]

pre-emphasis. In this step the regions of frequencies greater than 1 KHz are emphasized and the speech power spectrum is flattened [1]. In the 2nd step, a suitable window function is applied. The Hamming window function was used throughout this work. In the 3rd step, the Fourier transform is applied to get the magnitude spectrum. In the 4th step, mel-scale filterbank channels are used to filter the magnitude spectrum. These channels model the human ear's frequency response and make the lower frequencies more sensitive. The non-linear mel-frequency is given by the equation

$$mel(f) = 2595 \times \log_{10}(1 + \frac{f}{700}) \tag{2.1}$$

The non-linear mel-frequency is plotted against the linear frequency in Figure 2.10. In the last step a log is applied, which non-linearly compresses the amplitudes of the filterbank coefficients. The first four filterbank channels of utterance "Look out of the window and see if it is raining" for both Messiah and LIPS2008 databases are shown in Figure 2.11 and Figure 2.12 respectively.

2.4.2 MFCC

MFCC audio features are considered to be the standard in ASR applications. MFCC features represent the spectral envelope in a compact manner. The MFCC

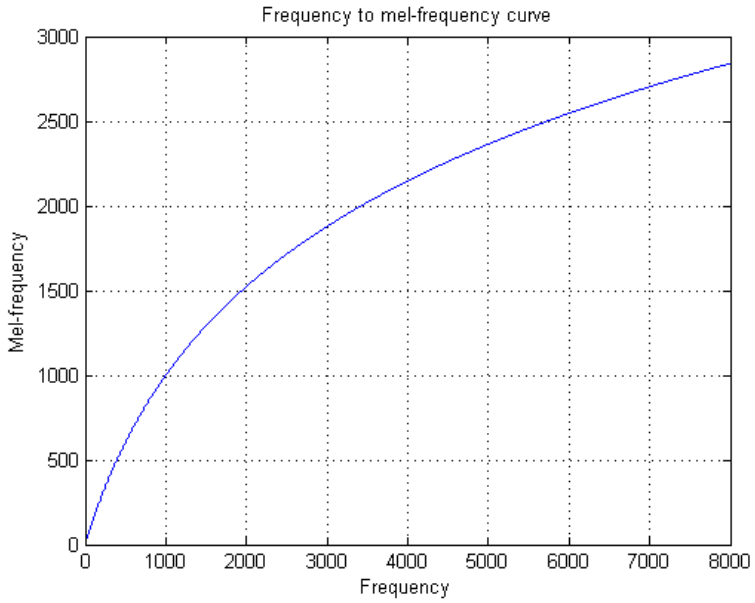


Figure 2.10: Mel-frequency versus linear frequency

extraction is standardized in [27]. The extraction process is shown in Figure 2.13 [1].

The extraction of log filterbank features was explained in the previous section. Adding the steps of Discrete Cosine Transform (DCT), truncation and log-energy, the filterbank features are converted into MFCC features as shown in Figure 2.13. The DCT step, separates the vocal tract and source components and the truncation step discards the source components. The resulting MFCC feature vector consists of 13 cepstral coefficients and a log-energy coefficient.

2.5 Visual features

Before visual feature extraction, a region is targeted from where to extract the visual features. This region is called as the Region of Interest (ROI). In this work,

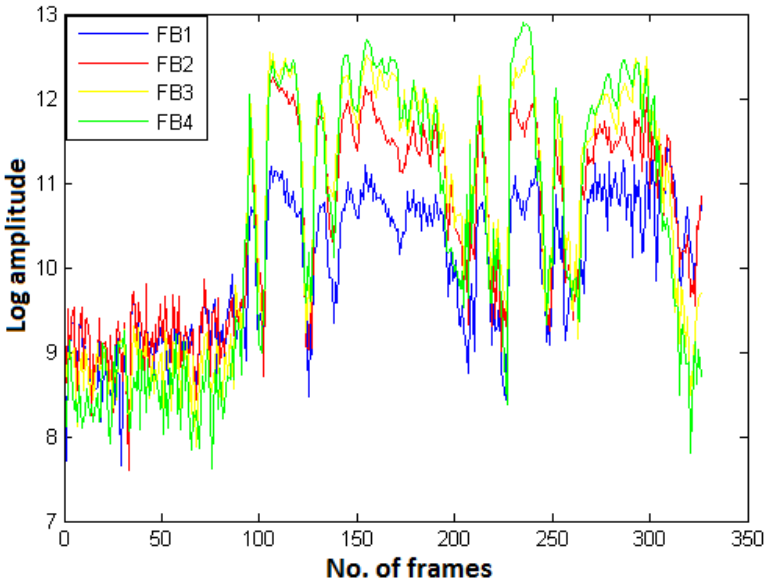


Figure 2.11: First four filterbanks of the Messiah database first utterance

the ROI for the Messiah database is a 180×100 pixels rectangle centred across the mouth and for the LIPS2008 database the ROI is a 120×100 pixels rectangle centred across the speaker's mouth. These sizes for the ROIs were selected based on the mouth sizes of the two speakers. The ROI can include the cheeks, jaws, forehead eyebrows, eyes and even the complete face. Eyes, eyebrows and forehead movements and gestures usually carry information about emotions. But in this work, the emotions of the speakers are not considered and the databases used are of emotionless data. That is why the ROI in this work is limited to the mouth area only. The ROI needs to be tracked. In this work, all the databases have already been tracked and landmarks indicating the lips of the speakers have been provided.

Visual feature extraction methods are broadly categorized as appearance-based

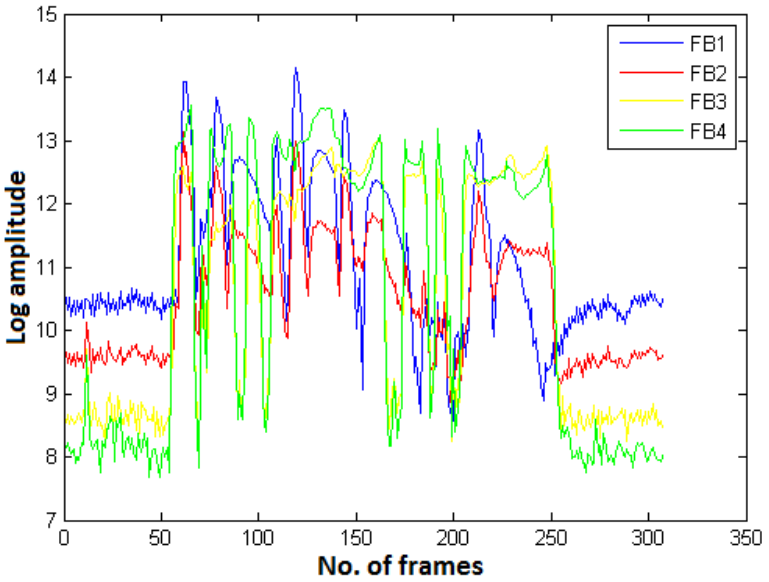


Figure 2.12: First four filterbanks of the LIPS2008 database first utterance

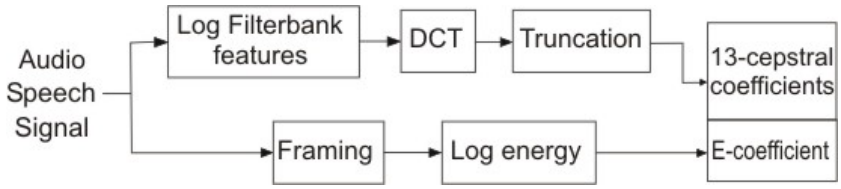


Figure 2.13: MFCC extraction process [1].

(image transform based), shape-based (model based), and the combination of both. In appearance-based methods, the image pixels in the ROI are directly processed for the extraction of features. In shape-based methods, some pre-determined model is fitted to the data [61].

It was shown in [61], in a comparison study of transform-based and model-based visual features for large vocabulary continuous audio-visual speech recognition, that the image transform methods particularly DCT perform better than

model-based methods and even better than Active Appearance Model (AAM) which is a combination of both the methods. however, it is worth mentioning that the training data for the AAMs was not sufficient. Correlation between different audio and visual features was measured in [2],[5], and it was found that AAM visual features have slightly higher correlation to audio filterbank features as compared to the correlation between 2D-DCT visual features and audio filterbank features. On the basis of the above findings combined with the relative simplicity of computing 2D-DCT, it is decided to use 2D-DCT visual features in this work.

2.5.1 Two-Dimensional Discrete Cosine Transform

2D-DCT is an appearance-based feature extraction method which extracts visual features from an ROI. In this work the ROI is the region across the speaker's mouth. If the ROI is an $M \times N$ pixel image centred across the mouth, represented by a matrix Z, and $z_{m,n}$ is the grey-scale value of m, n^{th} pixel, then the computation 2D-DCT coefficient matrix C_{pq} for the input image Z of size $M \times N$ is given as [42]

$$C_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} Z_{mn} \cos \frac{\pi(2m+1)p}{2M} \cos \frac{\pi(2n+1)q}{2N} \tag{2.2}$$

where the range of p and q is $(0 \leq p \leq M-1)$, $(0 \leq q \leq N-1)$, and

$$\alpha_p = \begin{cases} \frac{1}{\sqrt{M}}, & p = 0 \\ \sqrt{\frac{2}{M}}, & 1 \leq p \leq M-1 \end{cases}$$

and

$$\alpha_q = \begin{cases} \frac{1}{\sqrt{N}}, & q = 0 \\ \sqrt{\frac{2}{N}}, & 1 \leq q \leq N-1 \end{cases}$$

Equation (2.2) can be written as

$$C_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \cos \frac{\pi(2m+1)p}{2M} \sum_{n=0}^{N-1} Z_{mn} \cos \frac{\pi(2n+1)q}{2N} \tag{2.3}$$

Equation (2.3) can be written in simple terms as

$$C_{pq} = 1D - DCT_M \{1D - DCT_N \{Z_{mn}\}\} \tag{2.4}$$

This property is called 'separability' [42] and means that the 2D-DCT can be computed in two steps. In the first step, to apply the 1D-DCT vertically and in the second step to apply the 1D-DCT horizontally to the resultant of the first step. These two steps can be reversed as well, by applying first the 1D-DCT to the rows and then to the columns of the resultant of the first step.

In this work, the ROI for the Messiah database is a 180×100 pixels rectangle centred across the mouth shown in Figure 2.14 and for the LIPS2008 database the ROI is a 120×100 pixels rectangle centred across the speaker's mouth shown in Figure 2.15. The 2D-DCT concentrates the energy of the ROI image in the

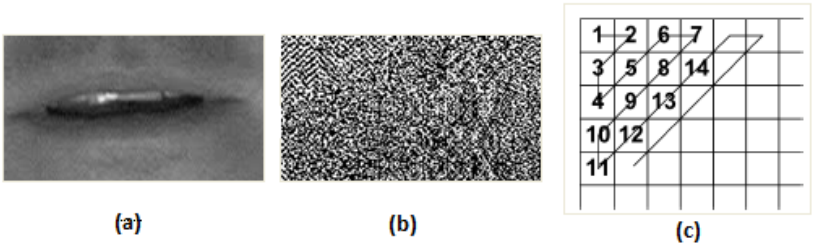


Figure 2.14: Example from the Messiah database:(a) 180×100 pixel ROI, (b) 2D-DCT of ROI, (c) zigzag ordering of 2D-DCT feature vector

upper-left corner of the output matrix [42] as shown in Figure 2.14 (b) and Figure

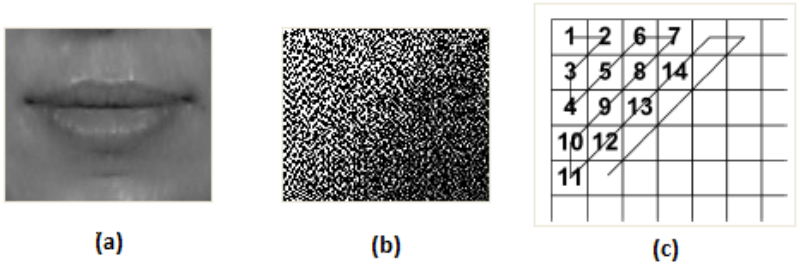


Figure 2.15: Example from the LIPS2008 database:(a) 120×100 pixel ROI, (b) 2D-DCT of ROI, (c) zigzag ordering of 2D-DCT feature vector

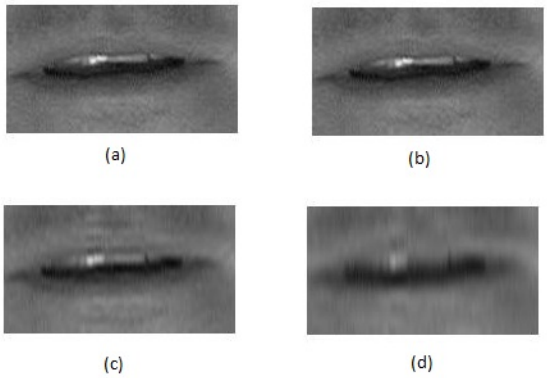


Figure 2.16: Reconstruction of an image using different percentages of 2D-DCT coefficients in the inverse 2D-DCT process, (a) 100 %, (b) 40 %, (c) 20 %, (d) 10 %

2.15 (b) for the Messiah database and the LIPS2008 database frames respectively. The final 2D-DCT visual features vector contains the first K (K = 15, in this work) coefficients selected from the 2D-DCT output matrix in a zigzag manner as shown in Figure 2.14 (c) and 2.15 (c). The zigzag scanning puts the high energy (low frequency) coefficients at the top of the output vector. The low energy coefficients can be discarded without introducing a notable distortion into the reconstructed image using inverse 2D-DCT. The effect of discarding the low energy coefficients on reconstruction, is shown in Figure 2.16.

The first four 2D-DCT features of utterance "Look out of the window and see if it is raining" for both Messiah and LIPS2008 database are shown in Figure 2.17 and Figure 2.18 respectively.

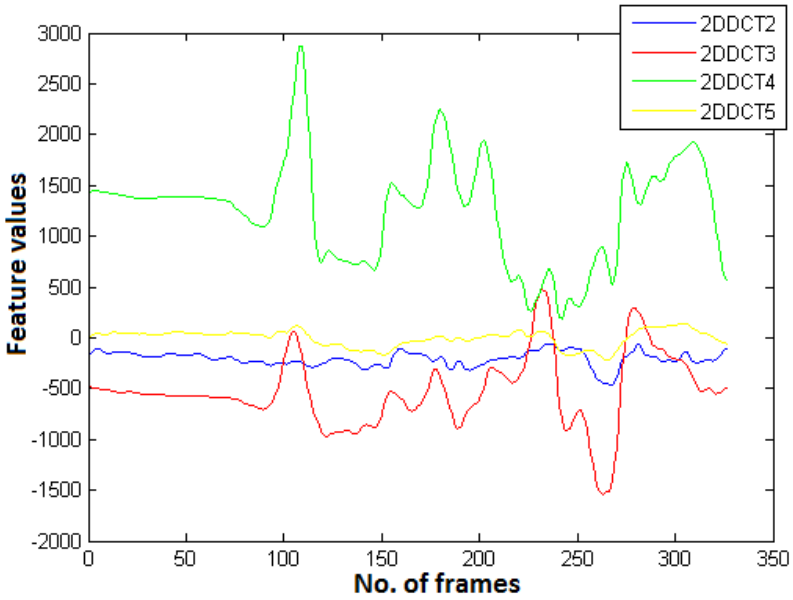


Figure 2.17: First four (2 to 5) 2D-DCT features of the Messiah database first utterance

Figure 2.19 is showing the first four filterbank and 2D-DCT features for the first utterance of both the databases.

2.6 Correlation measurement

Correlation measure gives an indication that how much two variables are related or associated. A statistical technique called least squares multiple linear regression is used to study this correlation between one dependant variable and one or several independent variables [1]. The multiple correlation coefficient is computed to

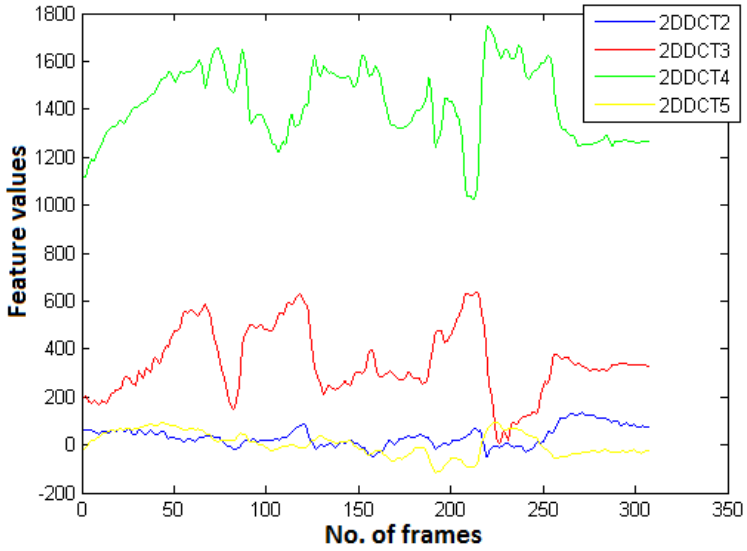


Figure 2.18: First four (2 to 5) 2D-DCT features of the LIPS2008 database first utterance

determine the multiple correlation of every individual component of the audio feature vector (the dependant variable) to the entire visual feature vector (the independent variables). The multiple correlation coefficient also called as the Pearson product moment correlation coefficient can be calculated as [1]

$$R(i) = \frac{\sum_{t=1}^{T}(a_t(i) - \bar{a}_t(i))(\hat{a}_t(i) - \bar{\hat{a}}_t(i))}{\sqrt{\sum_{t=1}^{T}(a_t(i) - \bar{a}_t(i))^2 \sum_{t=1}^{T}(\hat{a}_t(i) - \bar{\hat{a}}_t(i))^2}} \tag{2.5}$$

where T is the total number of vectors, $\hat{a}_t(i)$ is the predicted i^{th} component of the t^{th} audio vector and $\bar{\hat{a}}_t(i)$ is the mean of the predicted i^{th} component of the t^{th} audio vector. The prediction of the audio vectors from the visual vectors is the topic of the next chapter.

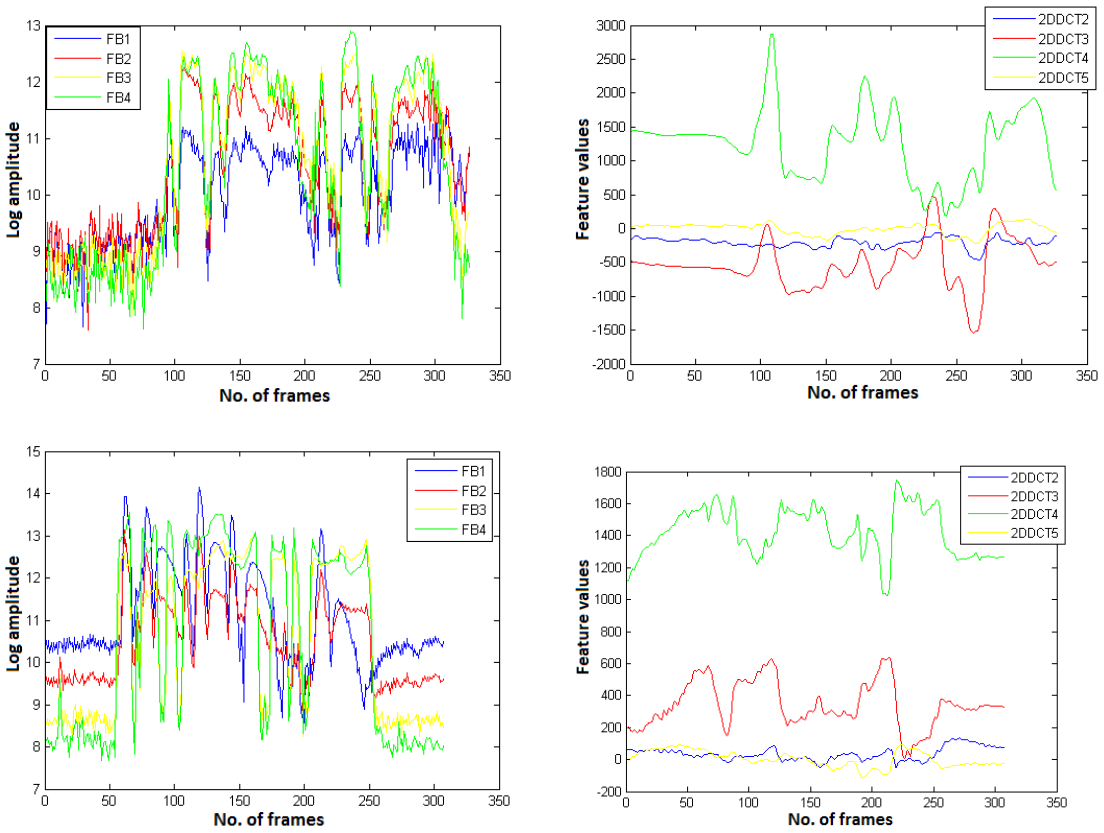


Figure 2.19: First four filterbank features of the Messiah database first utterance (top left), First four 2D-DCT features of Messiah database first utterance (top right), First four filterbank features of the LIPS2008 database first utterance (bottom left), First four 2D-DCT features of LIPS2008 database first utterance (bottom right)

2.6.1 Audio-visual correlation

Based on the findings of [61] and [2], it is decided to use the 2D-DCT as visual features and filterbank features as the audio features in this work. And following from [70], another set of audio features used in the experiments with the GRID database is the log power spectral vectors.

2.6.2 Audio-visual correlation Messiah database

The multiple correlation across the Messiah database was measured. Fifteen (15) dimensional 2D-DCT visual features and 23 dimensional filterbank features were used in the experiments. The initial 200 utterances were used for the training of the models using 16 clusters. The correlation results of 15 dimensional 2D-DCT visual vectors to each channel of the 23 dimensional filterbank vectors are shown in Figure 2.20. The average correlation across all the channels was 0.7655.

2.6.3 Audio-visual correlation LIPS2008 database

The multiple correlation across the LIPS2008 database was measured. Fifteen (15) dimensional 2D-DCT visual features and 23 dimensional filterbank features were used in the experiments. The initial 200 utterances were used for the training of the models using 16 clusters while the remaining 79 utterances were used for the testing. The correlation results of 15 dimensional 2D-DCT visual vectors to each channel of the 23 dimensional filterbank vectors are shown in Figure 2.21. The average correlation across all the channels was 0.5784.

The results in Figure 2.20 and Figure 2.21, show that the average correlation is higher for the Messiah database as compared to the LIPS2008 database. The

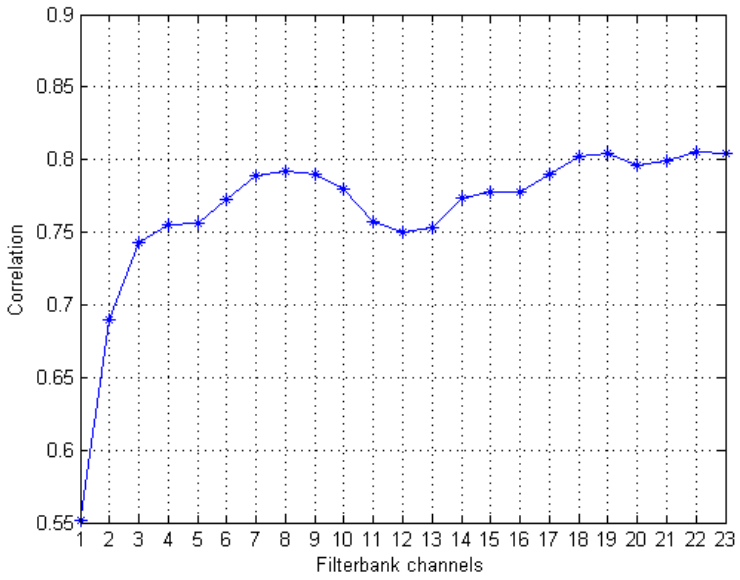


Figure 2.20: Correlation of 15 dimensional 2D-DCT features to each channel of 23 dimensional filterbank features of the Messiah database

reason for this is the less visible articulators and less clear articulation of speech by the LIPS2008 speaker as compared to the Messiah database speaker.

2.6.4 Audio-visual correlation GRID database

The audio-visual correlations for the two speakers (speaker 6 and speaker 4)of the GRID database were measured. These two speakers were selected because of the lower word error rates reported for them [15]. The correlation results for audio filterbank features and log power spectral features are presented in the following sections.

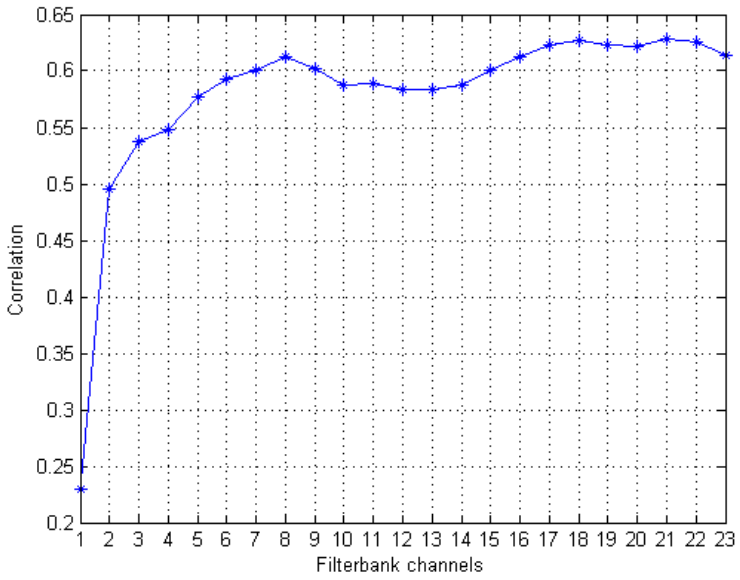


Figure 2.21: Correlation of 15 dimensional 2D-DCT features to each channel of 23 dimensional filterbank features of the LIPS2008 database

Audio-visual correlation for filterbank features

Table 2.2, shows the average correlation across all the channels for the speaker 6 (male) of GRID database. The results show that increasing the size of visual features vector results in the increase of correlation. But this increase in correlation is very small. In the same way, increasing the number of clusters in the GMM, gives improvements in the correlation but again these improvements are small. The best correlation is given by a visual vector of size 50 and the number of clusters being 64.

ASize	VSize	No of Clusters	Average correlation
23	15	16	0.78
23	20	16	0.80
23	25	16	0.80
23	30	16	0.82
23	40	16	0.82
23	50	16	0.82
23	15	32	0.80
23	20	32	0.81
23	25	32	0.81
23	30	32	0.82
23	40	32	0.82
23	50	32	0.83
23	15	64	0.80
23	20	64	0.82
23	25	64	0.82
23	30	64	0.83
23	40	64	0.83
23	50	64	**0.84**
23	15	128	0.81
23	20	128	0.83
23	25	128	0.83
23	30	128	0.83
23	40	128	0.83
23	50	128	0.83

Table 2.2: *Average correlation across all the channels for different sizes of visual vector and different number of clusters for speaker 6 of GRID database. ASize and VSize are representing the sizes of audio and visual features vectors respectively in the augmented AV-vector.*

Table 2.3, shows the average correlation across all the channels for the speaker 4 (female) of GRID database for a visual vector of size 25 and the number of clusters being 64 in the GMM.

ASize	VSize	No of Clusters	Average correlation
23	25	64	0.77

Table 2.3: *Average correlation across all the channels for speaker 4 of GRID database for filterbank features.*

The results in Table 2.2 and Table 2.3, for the two speakers of the GRID database show that the average correlation is slightly less in case of speaker 4 (female) as compared to speaker 6 (male). This finding is in accordance with the study reported in [15] where more error rates were reported for the speech of speaker 4 as compared to that of speaker 6.

Audio-visual correlation for log power spectral features

Table 2.4, shows the average correlation across all the channels for the speaker 6 (male) of the GRID database. The results show that increasing the size of the visual features vector results in the increase of average correlation. But this increase in accuracy is very small. In the same way, increasing the number of clusters in the GMM, gives improvements in the average correlation across all the channels but again these improvements are small. The best average correlation across all the channels is given by the visual vectors of sizes 25 and 30 and the number of clusters being 128.

ASize	VSize	No of Clusters	Average correlation
128	15	16	0.76
128	20	16	0.77
128	25	16	0.78
128	30	16	0.78
128	40	16	0.79
128	15	32	0.76
128	20	32	0.78
128	25	32	0.78
128	30	32	0.79
128	40	32	0.80
128	15	64	0.77
128	20	64	0.78
128	25	64	0.79
128	30	64	0.80
128	40	64	0.80
128	15	128	0.78
128	20	128	0.79
128	25	128	**0.80**
128	30	128	**0.80**

Table 2.4: *Average correlation across all the channels for different sizes of visual vector and and different number of clusters in the GMM for speaker 6 of GRID database.*

Table 2.5, shows the average correlation across all the channels for the speaker

4 (female) of the GRID database for a visual vector of size 30 and the number of clusters being 32 in the GMM.

ASize	VSize	No of Clusters	Average correlation
128	30	32	0.73

Table 2.5: *Average correlation across all the channels for speaker 4 of GRID database for log power spectral features.*

The results in Table 2.4 and Table 2.5, for the two speakers of the GRID databse show that the average correlation is slightly less in case of speaker 4 (female) as compared to speaker 6 (male).

2.7 Summary

This chapter discussed the speech production process, both from audio and visual perspectives. The audio-visual databases used in the work were introduced along with the audio and visual feature extraction methods. Correlation between the audio and visual features was measured. It was shown that the size of the visual feature vector and the number of clusters used in the modelling process affect the correlation levels. The LIPS2008 database shows significantly lower correlation levels as compared to the Messiah database which can be attributed to the distance from the camera factor, which results in the lower resolution of the articulators.

For the two speakers of the GRID database, speaker 6 (male) shows slightly higher correlation as compared to speaker 4 (female). This finding is in accordance with the study reported in [15] where more error rates were reported for the speech of speaker 4 as compared to that of speaker 6.

Chapter 3

Estimation of Clean Audio Speech Features from Visual Features

Preface

This chapter discusses the estimation of clean acoustic speech features from visual speech features. 2D-DCT features are used as the visual features and log filterbank and log power spectral features are used as the audio speech features. The joint density of the audio-visual vectors of each speaker is modelled using a GMM with various number of clusters. Then using these trained models, a MAP estimate of the acoustic speech features from the visual speech features is made. The accuracy of the estimation is measured in terms of mean percentage filterbank estimation errors and mean percentage log power spectral estimation errors. The results show that the number of dimensions in the visual vector and the number of clusters in

the GMM affect the accuracy of estimation.

3.1 Introduction

This chapter explains a statistical method to exploit the correlation between audio speech features and visual speech features. This method can be divided into two stages: firstly, to train the GMM to model the joint density of AV vectors. Secondly, to estimate audio features from visual features using the trained GMMs. The basis for this method is the correlation between audio and visual speech features, discussed in the previous chapter. The creation of training data pools is described in Section 3.2.1. The GMM training is described in Section 3.2.2 and the estimation of audio features from visual features in Section 3.2.3. The filterbank interpolation process is explained in Section 3.3. The experimental results in terms of estimation errors are presented in the Section 3.4.2 for the three databases: Messiah, LIPS2008 and GRID.

3.2 Estimation of audio features from visual features

As was said previously this method is a two step process: To train and to estimate. The process of creating a GMM which models the joint density of the AV vectors is shown in Figure 3.1. The process consist of augmenting the audio and visual vectors, pooling them across the training data set and then training the GMM on the pooled augmented AV vectors.

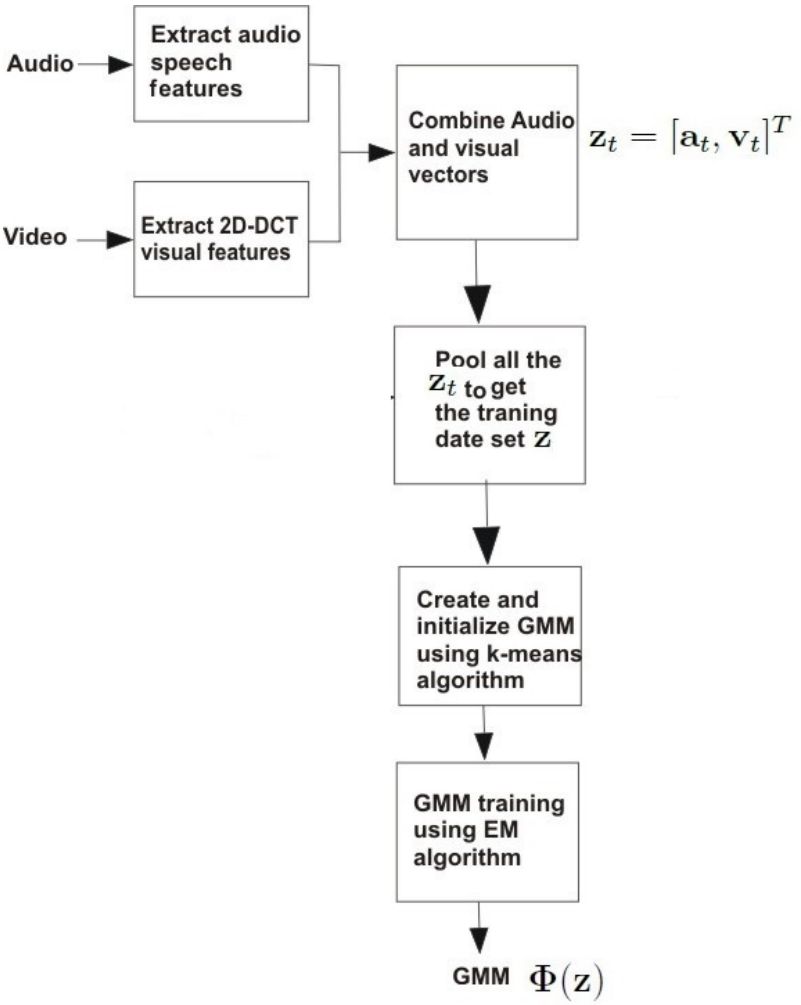


Figure 3.1: GMM creation and training.

3.2.1 Augmenting audio and visual feature vectors

The I-dimensional audio feature vectors and J-dimensional 2D-DCT visual feature vectors are augmented together to form an $I + J$ dimensional AV feature vector as

$$\mathbf{z}_t = [\mathbf{a}_t, \mathbf{v}_t]^T \tag{3.1}$$

where $\mathbf{a}_t$ and $\mathbf{v}_t$ are the audio and visual feature vectors respectively. The audio features are either filterbank features or log power spectral features. These feature vectors are extracted from the clean speech at time frame t. All the AV vectors are pooled together across the training data to give a training data set $\mathbf{z}$. This training data set $\mathbf{z}$, is to be used at the training step.

3.2.2 GMM training

The training data set $\mathbf{z}$, of the augmented AV vectors is used to create a GMM, $\mathbf{\Phi}(\mathbf{z})$. This GMM models the correlation between audio and visual features. To determine the initial cluster positions of the feature vectors, the K-means algorithm [32] is used which is then further refined by using the Expectation Maximization (EM) algorithm [20]. The created GMM, $\mathbf{\Phi}(\mathbf{z})$, is given as

$$\mathbf{\Phi}(\mathbf{z}) = \sum_{k=1}^{K} \alpha_k \phi_k(\mathbf{z}) = \sum_{k=1}^{K} \alpha_k N(\mathbf{z}; \mu_k^{\mathbf{z}}, \mathbf{\Sigma}_k^{\mathbf{z}}) \tag{3.2}$$

where α_k is the prior probability of the k^{th} cluster, and $\phi_k(\mathbf{z})$ is the k^{th} Gaussian Probability Density Function (PDF) of the GMM. This PDF is defined by the covariance matrix $\mathbf{\Sigma}_k^{\mathbf{z}}$ and the mean vector $\mu_k^{\mathbf{z}}$, where

$$\mathbf{\Sigma}_k^{\mathbf{z}} = \begin{pmatrix} \mathbf{\Sigma}_k^{\mathbf{aa}} & \mathbf{\Sigma}_k^{\mathbf{av}} \\ \mathbf{\Sigma}_k^{\mathbf{va}} & \mathbf{\Sigma}_k^{\mathbf{vv}} \end{pmatrix} \tag{3.3}$$

and

$$\mu_k^{\mathbf{z}} = \begin{pmatrix} \mu_k^{\mathbf{a}} \\ \mu_k^{\mathbf{v}} \end{pmatrix} \tag{3.4}$$

For the k^{th} cluster of the GMM, the mean vector $\mu_k^{\mathbf{z}}$, comprises of the I-dimensional mean audio vector, $\mu_k^{\mathbf{a}}$, and a J-dimensional mean visual feature vector, $\mu_k^{\mathbf{v}}$. The covariance matrix for the k^{th} cluster $\mathbf{\Sigma}_k^{\mathbf{z}}$, has four components: the $I \times I$ dimensional covariance matrix of the audio vectors $\mathbf{\Sigma}_k^{\mathbf{aa}}$, the $I \times J$ dimensional covariance matrix of the audio and visual vectors $\mathbf{\Sigma}_k^{\mathbf{av}}$, the $J \times I$ dimensional covariance matrix of the visual and audio vectors $\mathbf{\Sigma}_k^{\mathbf{va}}$, and the $J \times J$ dimensional covariance matrix of the visual vectors $\mathbf{\Sigma}_k^{\mathbf{vv}}$. As was stated previously, initial cluster positions are determined using the K-means algorithm, which is then refined by the EM algorithm. The EM iterations are carried out until no change in the cluster positions occur in further iterations or when the number of iterations exceed a specific number. The number of clusters, K, is determined experimentally and is based on minimizing the estimation errors.

3.2.3 MAP estimation of audio features

The trained GMM can be used to estimate the audio filterbank feature vector, $\hat{\mathbf{a}}_t$ from the input visual feature vector $\mathbf{v}_t$. This estimation can be carried out from the most probable cluster k^*, in the GMM. The most probable cluster, k^*, for the input vector, $\mathbf{v}_t$, is given as [18], [17]

$$k^* = \arg\max_k \{\, p(\mathbf{v}_t | \phi_k(\mathbf{z}))\} \tag{3.5}$$

using the most probable cluster k^*, and the t^{th} visual vector $\mathbf{v}_t$, a MAP estimate of the audio vector, $\hat{\mathbf{a}}_t$, can be made as [74]

$$\hat{\mathbf{a}}_t = \arg\max_{\mathbf{a}_t} \{\, p(\mathbf{a}_t | \mathbf{v}_t, \phi_{k^*})\} \tag{3.6}$$

where $p(\mathbf{a}_t|\mathbf{v}_t, \phi_{k^*})$, represent the probability of the audio vector given the visual vector using the most probable cluster k^*. The estimate of the audio features vector for the corresponding visual vector from the most probable cluster is given as [17], [97]

$$\hat{\mathbf{a}}_t = \mu_{k^*}^{\mathbf{a}} + \mathbf{\Sigma}_{k^*}^{\mathbf{av}}(\mathbf{\Sigma}_{k^*}^{\mathbf{vv}})^{-1}(\mathbf{v}_t - \mu_{k^*}^{\mathbf{v}}) \tag{3.7}$$

It was shown in [19], [17], that for better accuracy to reduce the estimation errors, the estimation can be made from a weighted combination of all the clusters, K. The weighting factor is the posterior probability $h_k(\mathbf{v}_t)$, of the t^{th} visual vector, belonging to the k^{th} cluster. So the weighted MAP estimate of the audio vector from all the K clusters is given by

$$\hat{\mathbf{a}}_t = \sum_{k=1}^{K} h_k(\mathbf{v}_t)\{\mu_k^{\mathbf{a}} + \mathbf{\Sigma}_k^{\mathbf{av}}(\mathbf{\Sigma}_k^{\mathbf{vv}})^{-1}(\mathbf{v}_t - \mu_k^{\mathbf{v}})\} \tag{3.8}$$

This weighting factor $h_k(\mathbf{v}_t)$, representing the posterior probability is determined as [1], [17]

$$h_k(\mathbf{v}_t) = \frac{\alpha_k p(\mathbf{v}_t|\phi_k^{\mathbf{v}})}{\sum_{k=1}^{K} p(\mathbf{v}_t|\phi_k^{\mathbf{v}})} \tag{3.9}$$

where $p(\mathbf{v}_t|\phi_k^{\mathbf{v}})$ is representing the visual vector's marginal distribution for the k^{th} cluster of the GMM specific to the t^{th} vector [4].

3.3 Interpolation of filterbank features

The estimated audio filterbank features are to be used in the construction of visually-derived Wiener filter for speaker separation in Chapter 4 and in the derivation of binary masks in Chapter 5. But these estimated 23-dimensional filterbank

features need to be interpolated to 128 dimensions. The 128 dimensions are for the reason to match the number of spectral bins in the mixed speech in Chapter 4 and Chapter 5 respectively. To interpolate the filterbank features, the 23 dimensional filterbank feature vectors are arranged in mel-scale positions in an interval of 1 to 128. Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) interpolation gives less interpolation errors [1], and was applied to these 23-dimensional filterbank feature vectors to be interpolated into 128 dimensions.

3.4 Experiments

The aim of the experiments in this section is to evaluate the accuracy of the estimation of acoustic features from 2D-DCT visual features for the three databases: Messiah, LIPS2008 and GRID. Firstly, the audio and visual features and the databases used, are described. Secondly, the experimental results are presented that evaluate the accuracy of the estimation accuracy in terms of mean percentage estimation errors.

3.4.1 Audio and visual features and databases

In the case of experiments with the Messiah (male) and LIPS2008 (female) databases, the first 200 utterances of each database were used for training while the remaining 79 utterances were used for the evaluation. The audio in both databases was down-sampled to a sampling frequency of 8 kHz. The video was up sampled to 100 frames per second to match the audio frame rate. In the case of experiments with the GRID database, the data of speaker 6 (male) and speaker 4 (female) were used. Out of the 1000 utterances of each speaker, 800 were used for training and

the remaining 200 for the evaluation.

To accurately estimate audio features from visual features it is necessary to select features that exhibit high levels of audio-visual correlation. As such, based on [4], 23 channel mel-scale filterbank vectors, $\mathbf{a_t}$, are used as the audio features. The dimension 23 for filterbank features is based on the experimental results in Chapter 4 and Chapter 5. Another type of audio features used is the log power spectral features of 128 dimensions [70]. The two types of audio features are extracted from 20ms duration frames of audio at 10ms intervals in accordance with the ETSI XAFE standard [28]. The 2D-DCT Visual features, $\mathbf{v}_t$, are extracted from an ROI centred on the speaker's mouth. A 2D-DCT is applied to the ROI and the first J coefficients retained as the visual vector. For a detailed discussion of the audio and visual features and the databases used, please refer to Chapter 2.

3.4.2 Results

The estimation accuracy for the log filterbank and log power spectral features is measured in terms of mean percentage absolute estimation errors, E, as

$$E = \frac{1}{TI} \sum_{t=0}^{T-1} \sum_{i=0}^{I-1} \mid \frac{a_t(i) - \hat{a}_t(i)}{a_t(i)} \mid \times 100 \tag{3.10}$$

where $a_t(i)$ and $\hat{a}_t(i)$ are non-negative and are the reference and estimated values of the i^{th} channel and t^{th} frame of the audio features. T is the total number of frames and I is the number of channels. The estimation errors E, for the various arrangements of the three databases are shown in Table 3.1 - 3.6.

3.4.3 Filterbank estimation errors

Table 3.1, shows the mean filterbank estimation errors for the speaker 6 (male) of GRID database. The results show that increasing the size of visual features vector results in the reduction of estimation errors thus increasing the estimation accuracy. But this increase in accuracy is very small. In the same way, increasing the number of clusters in the GMM, gives improvements in the estimation accuracy but again these improvements are small. The best accuracy is given by a visual vector of size 50 and the number of clusters being 64.

ASize	VSize	No of Clusters	Mean % FB estimation errors
23	15	16	10.80
23	20	16	10.47
23	25	16	10.32
23	30	16	10.04
23	40	16	10.05
23	50	16	9.90
23	15	32	10.29
23	20	32	9.99
23	25	32	10.01
23	30	32	9.81
23	40	32	9.76
23	50	32	9.71
23	15	64	10.00
23	20	64	9.74
23	25	64	9.61
23	30	64	9.50
23	40	64	9.55
23	50	64	**9.44**
23	15	128	9.67
23	20	128	9.52
23	25	128	9.43
23	30	128	9.51
23	40	128	9.52
23	50	128	9.62

Table 3.1: *Mean percentage filterbank estimation errors for different sizes of visual vector and different number of clusters for speaker 6 of GRID database. ASize and VSize are representing the sizes of audio and visual features vectors respectively in the augmented AV-vector.*

Table 3.2, shows the mean filterbank estimation errors for the speaker 4 (female) of GRID database for a visual vector of size 25 and the number of clusters being 64 in the GMM.

ASize	VSize	No of Clusters	Mean % FB estimation errors
23	25	64	10.41

Table 3.2: *Mean percentage filterbank estimation errors for speaker 4 of the GRID database.*

Table 3.3, shows the mean filterbank estimation errors for the Messiah database. The results show that increasing the number of clusters in the GMM, gives improvements in the estimation accuracy but again these improvements are small. The best accuracy is given by the number of clusters being 16 and further increase in the number of clusters decreases the estimation accuracy.

ASize	VSize	No of Clusters	Mean % FB estimation errors
23	15	1	9.49
23	15	2	8.72
23	15	4	8.50
23	15	8	8.38
23	15	16	**8.12**
23	15	32	8.45
23	15	64	9.02

Table 3.3: *Mean percentage filterbank estimation errors for different number of clusters for the Messiah database.*

Table 3.4, shows the mean filterbank estimation errors for the LIPS2008 database. The results show that the best accuracy is given by the number of clusters being 16 and further increase in the number of clusters decreases the estimation accuracy.

The results in Table 3.1 and Table 3.2, for the two speakers of the GRID databse show that the estimation errors are almost of the same level. But the re-

ASize	VSize	No of Clusters	Mean % FB estimation errors
23	15	16	**12.85**
23	15	32	13.60

Table 3.4: *Mean percentage filterbank estimation errors for clusters sizes from 16 to 32 for the LIPS2008 database.*

sults in Table 3.3 and Table 3.4, show that the estimation errors are more for the LIPS2008 database as compared to the Messiah database. This can be attributed to the low correlation between the audio and visual features of the LIPS2008 database. As measured in Chapter 2, the average correlation between the audio and visual features of the Messiah database was 0.7655 and 0.5784 for LIPS2008 database. The reason for this is the less visible articulators and less clear articulation of speech by the LIPS2008 speaker as compared to the Messiah database speaker.

3.4.4 Log power spectral estimation errors

The Log power spectral estimation errors for the speaker 6 (male) of the GRID database were calculated using Equation 3.10 and are shown in Table 3.5. The results show that increasing the size of visual features vector results in the reduction of estimation errors thus increasing the estimation accuracy. But this increase in accuracy is very small. In the same way, increasing the number of clusters in the GMM, gives improvements in the estimation accuracy but again these improvements are small. The best accuracy is given by a visual vector of size 30 and the number of clusters being 128.

ASize	VSize	No of Clusters	Mean % LPS estimation errors
128	15	16	17.66
128	20	16	17.16
128	25	16	16.98
128	30	16	16.89
128	40	16	16.50
128	15	32	17.35
128	20	32	16.59
128	25	32	16.67
128	30	32	16.30
128	40	32	16.21
128	15	64	16.93
128	20	64	16.48
128	25	64	16.21
128	30	64	16.04
128	40	64	15.99
128	15	128	16.61
128	20	128	16.20
128	25	128	15.94
128	30	128	**15.91**

Table 3.5: *Mean percentage LPS estimation errors for different sizes of visual vector and and different number of clusters in the GMM for speaker 6 of the GRID database.*

Table 3.6, shows the Log power spectral estimation errors for the speaker 4

(female) of the GRID database for a visual vector of size 30 and the number of clusters being 32 in the GMM.

ASize	VSize	No of Clusters	Mean % LPS estimation errors
128	30	32	17.18

Table 3.6: *Mean percentage LPS estimation errors for speaker 4 of the GRID database.*

The results in Table 3.5 and Table 3.6, for the two speakers of the GRID databse show that the estimation errors are almost of the same level.

Figure 3.2 and Figure 3.3 are showing the comparison of two estimated filterbanks channel 10 and channel 15 centred at 860 Hz and 1613 Hz with the reference filterbanks of a sentence from the Messiah and LIPS2008 databases respectively. The mismatch between the reference and estimated filterbanks in the initial silence region of Figure 3.2 is because of the mouth opening of the speaker which can be verified from the video.

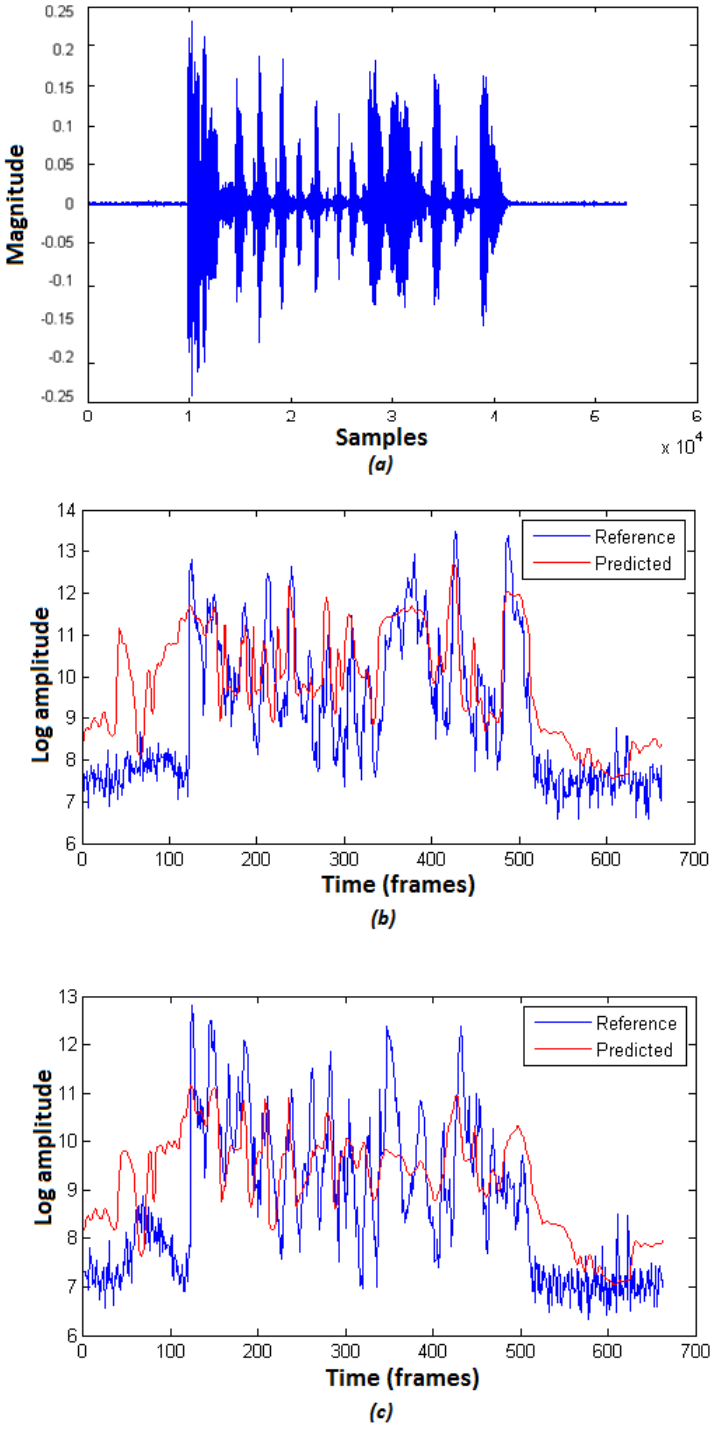


Figure 3.2: Time domain waveform and filterbanks channel 10 and channel 15 of the utterance '*Ada aims to serve chump chops with chips cooked in pure oil, with ice-cream to follow*', of the Messiah database: (a) reference waveform, (b) filterbank channel 10, (c) filterbank channel 15

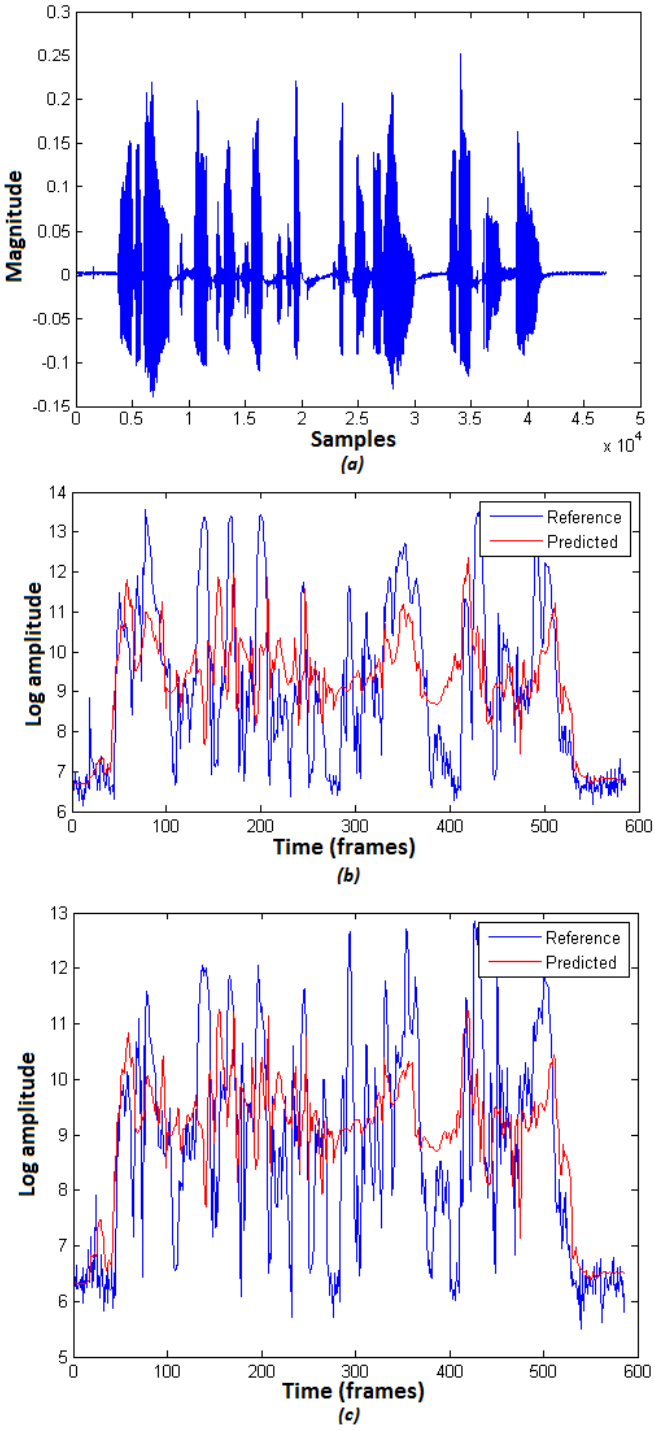


Figure 3.3: Time domain waveform and filterbanks channel 10 and channel 15 of the utterance '*Ada aims to serve chump chops with chips cooked in pure oil, with ice-cream to follow*', of the LIPS2008 database: (a) reference waveform, (b) filterbank channel 10, (c) filterbank channel 15

3.5 Summary

In this chapter, it was shown that the higher levels of correlation between audio and visual speech features can be exploited to estimate filterbank and log power spectral audio features from visual features. The estimation process consists of two stages. Firstly, the audio and visual speech feature vectors were augmented across the training dataset and these training datasets were used to train the GMMs using K-means and EM algorithms. Secondly, using MAP estimation and visual feature vectors, the corresponding audio features were estimated.

The estimation accuracy was measured in terms of estimation errors and the results show that the audio features can be estimated from the visual features with a good accuracy. The size of the visual feature vector and the number of clusters in the GMM has an effect on the estimation accuracy and these are determined experimentally. The accuracy of the estimation is directly proportional to the levels of correlations existing between the audio and video. The accuracy levels for the two speakers of the GRID database are almost the same but not for the Messiah and LIPS2008 databases. The accuracy levels for the LIPS2008 database are lower as compared to the Messiah database. This is attributed to the lower correlation between the audio and visual features of the LIPS2008 database because of the recording conditions.

It was also shown that visual features for the open mouth in the silence region can mislead the estimation process and the non-speech regions are taken as speech regions because of the mouth opening.

Chapter 4

Speaker Separation Using Wiener Filtering

Preface

This chapter proposes a method of single-channel audio speaker separation that uses visual speech information to extract a target speaker's speech from a mixture of speakers. The method requires a single audio input and visual inputs from each speaker in the mixture. The visual information from speakers is used to create a visually-derived Wiener filter. The Wiener filter gains are then non-linearly adjusted by a perceptual gain transform to improve the quality and intelligibility of the target speech. Experimental results are presented that measure the quality and intelligibility of the extracted target speaker and a comparison is made of the different perceptual gain transforms. These show that significant gains are achieved by the application of the perceptual gain functions.

4.1 Introduction

This chapter considers Wiener filtering for speaker separation using visual speech information. Wiener filtering has been used extensively in speech enhancement [56],[100]. In the time domain, if the clean speech signal $x(n)$ is contaminated by noise $d(n)$, then the noisy signal $y(n)$, is given as

$$y(n) = x(n) + d(n) \tag{4.1}$$

The Wiener filter coefficients are computed to minimise the mean-squared estimation error between the estimated $\hat{x}(n)$ and the desired (clean) signal $x(n)$ [56].

$$\hat{x}(n) = \sum_{k=0}^{M-1} w(k)y(n-k) \qquad n = 0, 1, 2, ... \tag{4.2}$$

where $w(k)$ is the filter coefficient and M is the total number of coefficients. The estimation error in the mean-square sense is given as

$$E[e^2(n)] = E[(x(n) - \hat{x}(n))^2] \tag{4.3}$$

where $E[.]$ is representing the expectation operator. By applying the Fourier Transform to Equation 4.1 we get

$$Y(f) = X(f) + D(f) \tag{4.4}$$

where $X(f)$ is the clean signal complex spectrum and $D(f)$ is the noise signal complex spectrum. By taking element-wise squared magnitude and assuming that the two signals are uncorrelated [77],[76], it can be written in the power spectrum

domain as

$$|Y(f)|^2 = |X(f)|^2 + |D(f)|^2 \tag{4.5}$$

For additive noise, the Wiener filter, $W(f)$, in the discrete Fourier transform (DFT) domain is defined as [100]

$$W(f) = \frac{P_{XX}(f)}{P_{XX}(f) + P_{DD}(f)} \tag{4.6}$$

where $P_{XX}(f)$ and $P_{DD}(f)$ represent the clean speech and noise power spectra, respectively.

The main challenge in the implementation of the Wiener filter is the estimation of the power spectrum of the desired signal because the desired signal is usually observed in noise [100] or along with competing speakers [39]. In the case of Wiener filtering for speech enhancement, the noise power spectrum can be estimated from noise only regions and at regular periods the estimate of the noise is updated. These noise only regions can be identified using VAD techniques. These techniques usually use features such as short-time energy, zero-crossings [38] and periodicity [99]. Other VAD techniques have exploited visual information as well [3], [49], [91], [6], [53]. In this case the noise sources are assumed to be stationary, examples of which can be such as train noise, office noise from fans and computers and car engine noise. The desired signal power spectrum can be obtained using the noise power spectrum and the noisy signal power spectrum using a method such as spectral subtraction [100], [56]. The noise estimate is critical for the performance of speech enhancement algorithms. Noise will remain there in the enhanced

speech, if the noise estimate is too low. On the other hand, if the noise estimate is too high, the enhanced signal will lose intelligibility and will be distorted [56]. In [58], [59], minimum statistics noise estimation algorithms were used to determine the estimate of noise. In these algorithms the noise estimate is made from the minimum power spectrum of the noisy speech in each frequency band. In [25], HMMs trained on clean speech were used to provide the clean speech estimates for the implementation of the Wiener filter.

In the case of speech separation, the speech sources are highly non-stationary. In order to use the Wiener filter for speech separation, the clean speech estimate and competing speech estimates must be made at frame level as speech is considered to be quasi-stationary at frame level. In this case, the Wiener filter is called block-adaptive or segment-adaptive [100]. In [9], [70], an extended form of Wiener filter was used for audio source separation. In this case each audio source was characterized by a GMM. The Wiener filter coefficients were calculated using the trained GMMs, the statistics of the mixtures of audio sources available at the training stage and the single-channel observed audio mixture.

For application to speaker separation the Wiener filter for speech enhancement of Equation 4.6 is modified so that the clean speech is replaced by the target speaker, X_1, and the contaminating noise is replaced by the competing speaker, X_2. So the Wiener filter for extracting the target speaker for a two speaker problem is given as

$$W_1(f) = \frac{P_{X_1X_1}(f)}{P_{X_1X_1}(f) + P_{X_2X_2}(f)} \tag{4.7}$$

This equation can be rearranged to extract the competing speaker by replacing $P_{X_1X_1}(f)$ in the numerator by $P_{X_2X_2}(f)$, and the corresponding Wiener filter is

given as

$$W_2(f) = \frac{P_{X_2X_2}(f)}{P_{X_2X_2}(f) + P_{X_1X_1}(f)} \tag{4.8}$$

The proposed method of visually-derived Wiener filtering for speaker separation is described in Section 4.2. This requires audio estimates of the target and competing speakers which are estimated from visual speech features and this is discussed in Section 4.3. As a further processing stage, several perceptual gain transforms are applied to the Wiener filter gains that improve both speech quality and intelligibility and these transforms are discussed in Section 4.2.1. Details of the implementation in terms of creating the time-domain target speaker's speech are explained in Section 4.4. Experimental results are presented in Section 6.5 to evaluate the proposed method in terms of speech quality and intelligibility. The whole process of speaker separation using visually derived Wiener filtering is shown in Figure 4.1 where the top panel shows the 'separation process' while the bottom panel shows the 'training process'.

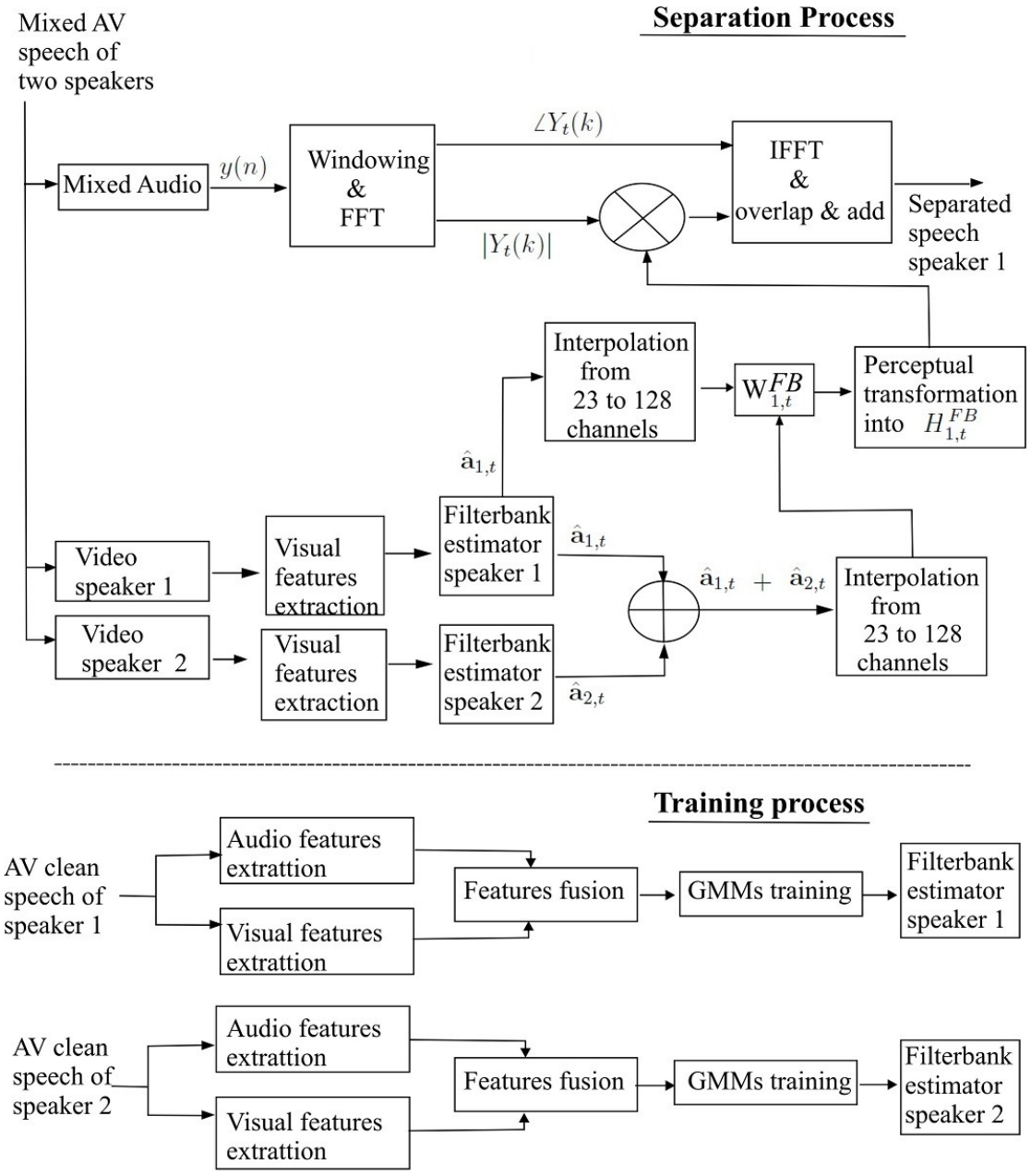


Figure 4.1: Speaker separation using visually derived Wiener filtering including the training stage.

4.2 Visually-derived Wiener filtering for speaker separation

In this work, the audio-visual correlation between a speaker's mouth shape and the resulting audio signal is exploited. To obtain the power spectra statistics for the target and competing speakers, it is proposed to estimate these from visual speech features taken from the two speakers. Analysis into the correlation of audio and visual speech features and the estimation of audio speech features from visual speech features were discussed in detail in Chapter 2 and Chapter 3. However, the analysis also revealed that insufficient audio-visual correlation is present to make a fine resolution estimate, although estimation of a less spectrally detailed filterbank vector is possible. As a consequence the speaker separation Wiener filter to extract speaker 1, W_1, from Equation 4.7, is modified to operate in the filterbank domain and can be defined as

$$W_{1,t}^{FB}(i) = \frac{\widehat{a}_{1,t}(i)}{\widehat{a}_{1,t}(i) + \widehat{a}_{2,t}(i)} \tag{4.9}$$

where $\widehat{a}_{1,t}(i)$ and $\widehat{a}_{2,t}(i)$ are filterbank estimates for the target speaker and competing speaker, i indicates the filterbank channel and t represents the time frame. In the same way to extract speaker 2, W_2, from Equation 4.8, is modified to operate in the filterbank domain and can be defined as

$$W_{2,t}^{FB}(i) = \frac{\widehat{a}_{2,t}(i)}{\widehat{a}_{2,t}(i) + \widehat{a}_{1,t}(i)} \tag{4.10}$$

4.2.1 Perceptual gain transformation

A series of perceptually-motivated transformations of the Wiener gains are now considered. The motivation behind these perceptual gain functions is to have different gain functions resulting in different suppression behaviours at different SNR levels. When SNR is high, lower suppression is expected but when SNR is low, higher levels of suppression are expected. However, the higher suppression of the competing speaker might come at the cost of distortion in the target speaker [56]. The gain functions introduced here aim ideally to both reduce distortion of the target speaker and improve the suppression of the competing speaker and are implemented as a perceptual gain transform, $\Pi(.)$. This can be considered a non-linear transformation of the Wiener filter gains and gives a perceptual gain $H^{FB}(i)$. Note that for clarity subscripts have been dropped from notation.

$$H(i) = \Pi(W^{FB}(i)) \tag{4.11}$$

Four different perceptual gain transformations have been investigated and these can broadly be described as piecewise or parametric. Equations (4.12) to (4.15) define the resulting gain functions, $H1$ to $H4$, and these are also plotted in Figure 4.2.

$$H1(i) = W^{FB}(i) \tag{4.12}$$

$$H2(i) = \begin{cases} W^{FB}(i) & W^{FB}(i) > \alpha \\ 0 & W^{FB}(i) \leq \alpha \end{cases} \tag{4.13}$$

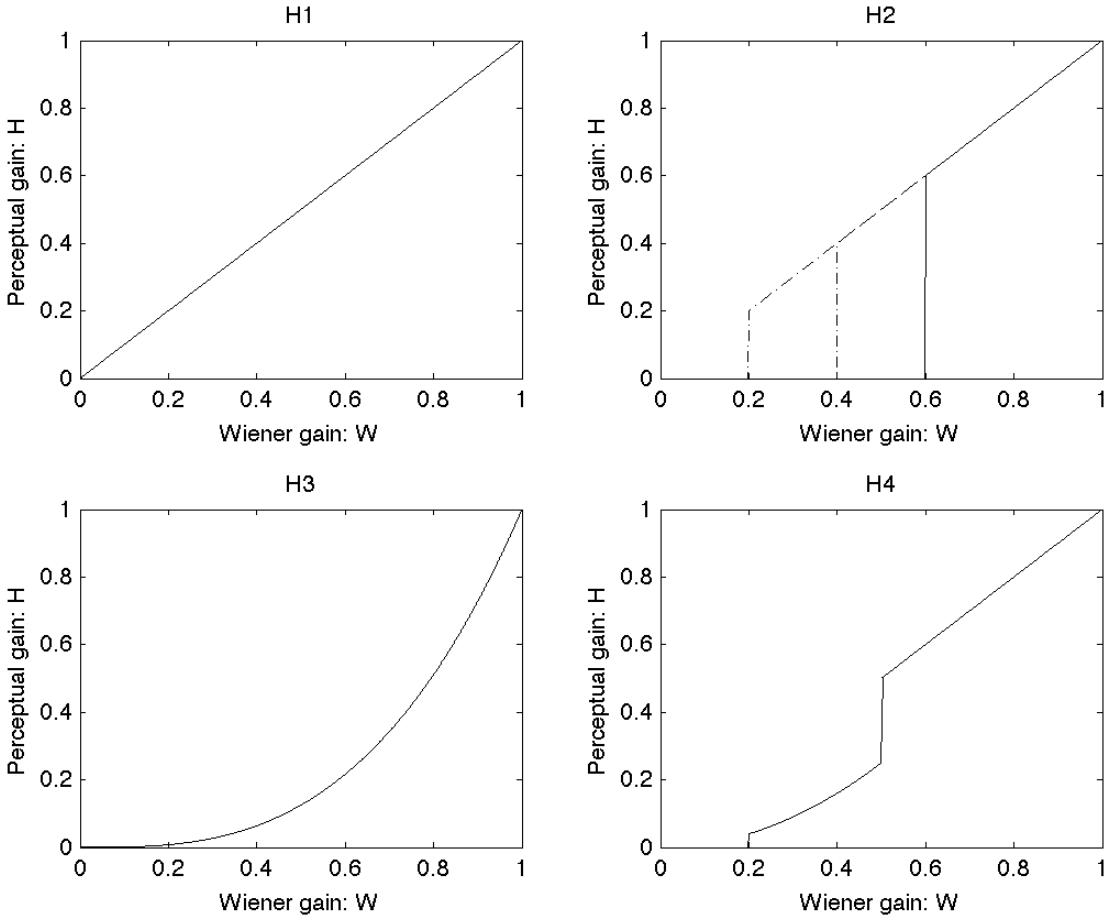


Figure 4.2: Perceptual gain functions

$$H3(i) = (W^{FB}(i))^3 \tag{4.14}$$

$$H4(i) = \begin{cases} 0 & W^{FB}(i) < \beta_1 \\ (W^{FB}(i))^2 & \beta_1 \leq W^{FB}(i) \leq \beta_2 \\ W^{FB}(i) & W^{FB}(i) > \beta_2 \end{cases} \tag{4.15}$$

Gain function $H1$ serves as a baseline and is set equal to the Wiener filter gain, W^{FB}. The second function, $H2$, restricts the gain so that if it falls below a

threshold, α, then it is set to zero. This has the effect of removing any time-frequency region where the SNR falls below a certain threshold and can be likened to the binary mask method of speech enhancement [77], but now with the mask estimated from visual features. Instead of removing regions with local SNRs below 0dB (corresponding to a gain of 0.5), gain cut-off values of $\alpha = 0.2$, 0.4 and 0.6 have been tested in this work. Gain function $H3$ is the cube of the Wiener gain and this has the effect of non-linearly reducing the Wiener gain. Lower gain values experience a considerable downscaling while higher gains are reduced by a smaller factor. The fourth gain function, $H4$, is a piecewise function that aims to capture properties of the previous gain functions by dividing the gain into three regions with zero gain, a squared Wiener gain and linear Wiener gain, respectively. Two variables, β_1 and β_2, define these regions and have been set to 0.2 and 0.5 for this work, based on preliminary test results.

4.3 Estimation of audio features from video

In the case of experiments with the Messiah and LIPS200 databases, $I = 23$ channel mel-scale filterbank vectors, $\mathbf{a}_{1,t}$ and $\mathbf{a}_{2,t}$ are used as the audio features for speaker 1 and speaker 2 respectively. This dimensionality of $I = 23$ was found to be optimal in [40]. Theses vectors are extracted from 20ms duration frames of audio at 10ms intervals in accordance with the ETSI XAFE standard [28]. Visual features, $\mathbf{v}_{1,t}$ and $\mathbf{v}_{2,t}$, for speaker 1 and speaker 2 respectively, are extracted from an ROI centred on a speaker's mouth at a rate of 100 frames per second. A 2D-DCT is applied and the first J coefficients are scanned in a zigzag manner and retained as the visual vector. The estimation process involves first training

a K clusters GMM to model the joint density of augmented audio-visual feature vectors for each speaker. MAP estimation can then be applied to estimate the audio features from visual features. The above same procedure of audio and visual feature extraction, training of GMMs and estimation of audio features from visual features is repeated for the two speakers from the GRID database. For a detailed discussion of audio and visual features, the training and the estimation process, please refer to Chapter 2 and Chapter 3.

4.4 Implementation

This section outlines the stages involved in applying visually-derived speaker separation to extract a target speaker. These stages were shown in the upper part 'Separation process' of Figure 4.1.

4.4.1 Perceptual gain calculation

The first stage involves utilising the visual speech features to calculate the perceptual gain, $H(i)$, and is summarised below:

1. Extract visual vectors, $\mathbf{v}_{1,t}$ and $\mathbf{v}_{2,t}$, from the two video sequences corresponding to target and competing speakers.

2. Estimate audio filterbank vectors, $\hat{\mathbf{a}}_{1,t}$ and $\hat{\mathbf{a}}_{2,t}$, for the two speakers from the visual features using MAP estimation.

3. Construct visually-derived Wiener filter of Equation (4.9).

4. Apply perceptual gain transforms to the Wiener filter from equations (4.12) to (1.15) to give perceptual gain, $H(i)$.

This gives a 23-D filterbank-domain perceptual gain function. These perceptual gains are used in the next section for the extraction of the target speaker from the mixed speech.

4.4.2 Speaker separation

From the single-channel audio input that comprises the mixed speech, short duration frames of speech are extracted and the magnitude spectrum, $|Y_t(k)|$ and phase, $\angle Y_t(k)$, are computed. The perceptual gain can now be applied to the magnitude spectrum of the mixed speech to extract the target speaker. However, before this can be applied the 23-D filterbank-domain perceptual gain must be interpolated up to the dimensionality of the magnitude spectrum, which in this work is K=128 spectral bins. The magnitude spectrum estimate of the target speaker, $|\widehat{X}_{1,t}(k)|$, can now be computed as

$$|\widehat{X}_{1,t}(k)| = H_t(k)|Y_t(k)| \tag{4.16}$$

The magnitude spectrum estimate of the target speaker is now combined with the phase of the mixed speech, $\angle Y_t(k)$, and an inverse Fourier transform is applied to obtain a short-duration frame of time-domain samples.

$$\widehat{x}_{1,t}(n) = \text{IFFT}(|\widehat{X}_{1,t}(k)|\angle Y_t(k)) \tag{4.17}$$

Overlap and adding of frames gives the final estimate of the target speaker. Figure 4.3, explains the extraction process of a single frame of the target speaker 1 from the mixed speech. The similarity between the extracted and the reference

magnitude spectrums in Figure 4.3 (g), shows the effectiveness of the method.

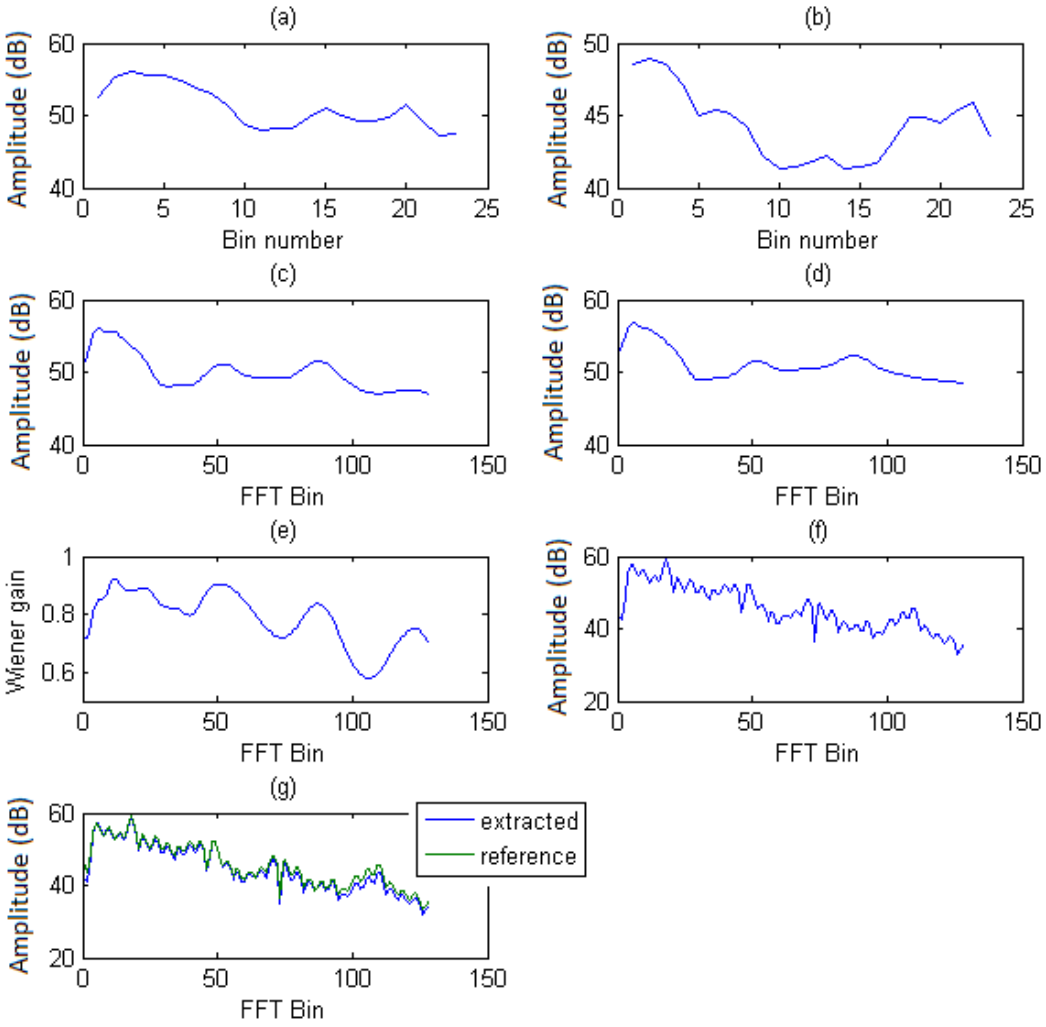


Figure 4.3: The process of extracting a target speaker frame from the mixed speech using visually derived Wiener filtering: a) estimated speaker 1 filterbank, b) estimated speaker 2 filterbank, c) interpolated estimated speaker 1 filterbank, d) interpolated estimated mixed (noisy) filterbank, e) Wiener filter gain, f) Mixed magnitude spectrum, g) extracted and reference magnitude spectrums.

4.5 Experimental results

This section evaluates the effectiveness of the proposed method of visually-derived speaker separation. First the audio-visual data and experimental set up used, are described. Secondly, two sets of experimental results are presented that evaluate the quality and the intelligibility of the target speaker's speech following visually-derived speaker separation.

4.5.1 Audio-visual data

In the case of experiments with the Messiah (male) and LIPS2008 (female) databases, the first 200 utterances of each database were used for training while the remaining 79 utterances were used for the evaluation. The audio in both databases was down sampled to a sampling frequency of 8 kHz and filterbank vectors extracted at 10 ms intervals as discussed in Section 4.3. The video was up sampled to 100 frames per second to match the audio frame rate. For both speakers, video was captured from the front of the face and the ROI was centred on the speaker's mouth.

The experimental scenario investigated is of two speakers talking simultaneously and being located close together in space, with the male speaker the target and the female the competing speaker. The mixed speech was created by mixing the speech utterances from the two databases to get the mixed signals (noisy speech). The Lips2008 utterances are scaled and added to the Messiah database utterances in such a way that the resulting mixed utterances are having a signal-to-interference ratio (SIR) of -10dB, -5dB, 0dB, 05dB, 10dB and 20dB. The SIRs are calculated only over speech periods by removing the initial and end silence from the utterances.

In the case of experiments with the GRID database, the data of speaker 6 (male) and speaker 4 (female) was used. Out of the 1000 utterances, 800 were used for training and the remaining 200 for the evaluation. The female speaker utterances are scaled and added to the male speaker's utterances in such a way that the resulting mixed utterances are having a signal-to-interference ratio (SIR) of -10dB, -5dB, 0dB, 05dB, 10dB and 20dB. The rest of the experimental set up was kept the same as in the case of the other two databases. For a detailed discussion of the audio-visual speech databases used, please refer to Chapter 2.

4.5.2 Speech quality

The extracted speech $\hat{\mathbf{s}}$ (either speaker 1 or speaker 2) in the time domain, is decomposed as [101]

$$\hat{\mathbf{s}} = \mathbf{s}_{target} + \mathbf{e}_{interf} + \mathbf{e}_{noise} + \mathbf{e}_{artif} \tag{4.18}$$

where $\mathbf{s}_{target}$ represents the part of $\hat{\mathbf{s}}$, perceived as to be coming from the target speaker, $\mathbf{e}_{interf}$ represents the part coming from the competing speaker, $\mathbf{e}_{noise}$ represents the part coming from the sensor noises (additive noise), and $\mathbf{e}_{artif}$ represents the part introduced by the algorithmic processing such as musical noise. To measure the quality of the extracted target speech, various energy ratios expressed in decibles (dB) are calculated from the above four components of the extracted speech. These ratios are Signal-to-Interference Ratio (SIR), Signal-to-Distortion Ratio (SDR) and Signal-to-Artefact Ratio (SAR), and these ratios are defined as [101]

$$SIR = 10\log_{10}\frac{\|\mathbf{s}_{target}\|^2}{\|\mathbf{e}_{interf}\|^2} \tag{4.19}$$

$$SDR = 10\log_{10}\frac{\|\mathbf{s}_{target}\|^2}{\|\mathbf{e}_{interf} + \mathbf{e}_{noise} + \mathbf{e}_{artif}\|^2} \tag{4.20}$$

$$SAR = 10\log_{10}\frac{\|\mathbf{s}_{target+\mathbf{e}_{interf}+\mathbf{e}_{noise}}\|^2}{\|\mathbf{e}_{artif}\|^2} \tag{4.21}$$

Tests were carried out at initial SIRs of -10dB, -5dB, 0dB, 5dB, 10dB and 20dB. Visually-derived Wiener filter speaker separation was applied to the mixtures using the four perceptual gain functions introduced in Section 4.2.1 and the resulting SIRs, SDRs and SARs were computed using the 'BSS evaluation' toolbox [29] and the results are shown in Figure 4.4, Figure 4.5, Figure 4.6.

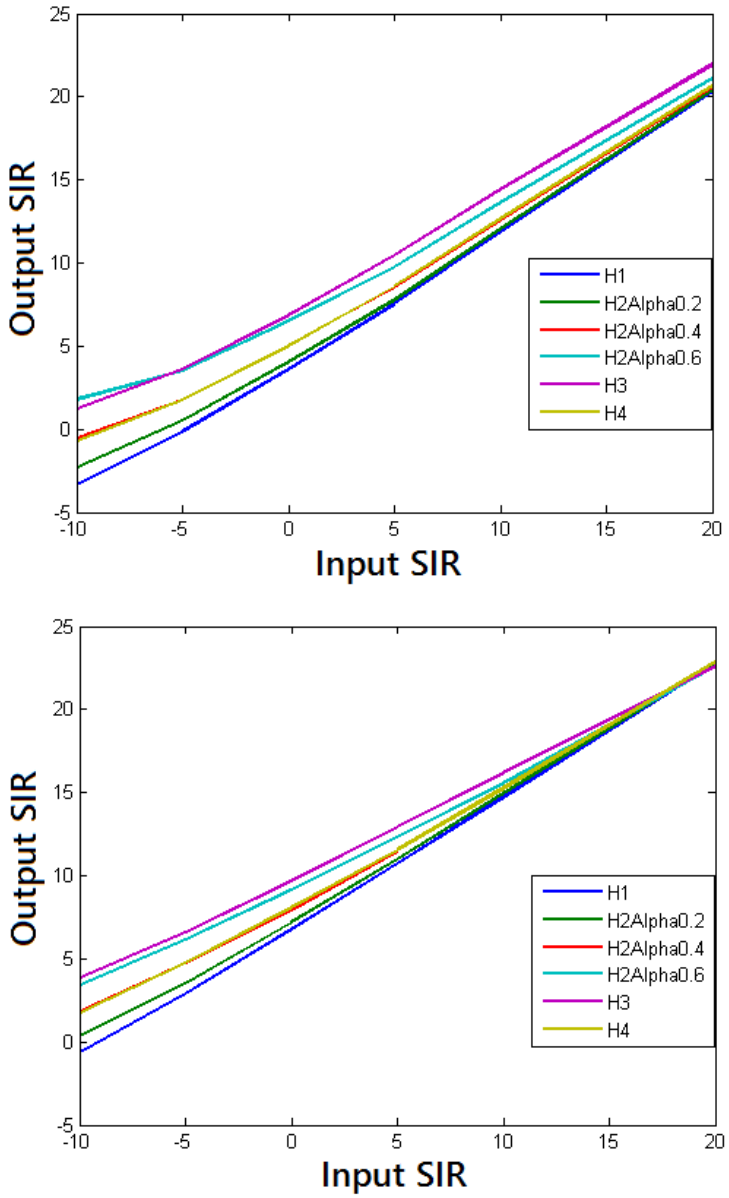


Figure 4.4: *Output SIR variations with Input SIR for the target speaker of: top) Messiah database, bottom) GRID database.*

The SIR results in Figure 4.4, show that using the Wiener gain, $H1$, gives a good increase in quality in terms of the suppression of the competing speaker,

particularly at the lower SIRs. Applying a perceptual gain transform gives further increases in the output SIR. The cube gain function $H3$ gives best performance, with gain function $H2$ (with α=0.6) being very close. The perceptual gain function $H2$ (with α=0.6), corresponds to the Wiener gain with spectral masking below an SIR of 1.8dB. Lowering the point of spectral masking reduces speech quality in terms of the output SIR. The output SIR gains decrease with increase in input SIRs and very little gains are obtained at 20dB.

The SDR results in Figure 4.5, show that using the Wiener gain and the various perceptual transforms gives a good increase in SDR at lower SIRs (from -10dB to 0dB). The SDR gains decrease with an increase in input SIRs. The performance of the various perceptual gain functions does not vary by a considerable amount in this lower SIR region. The performance of gain function $H2$ (with α=0.6) and $H3$ in terms of SDR gains deteriorates rapidly in the higher input SIR regions i.e. above 10dB. Gain function $H1$ and $H2$ (with α=0.2) performs the best in this higher input SIR region as these functions do not introduce much distortion into the extracted speech by spectrally masking their parts.

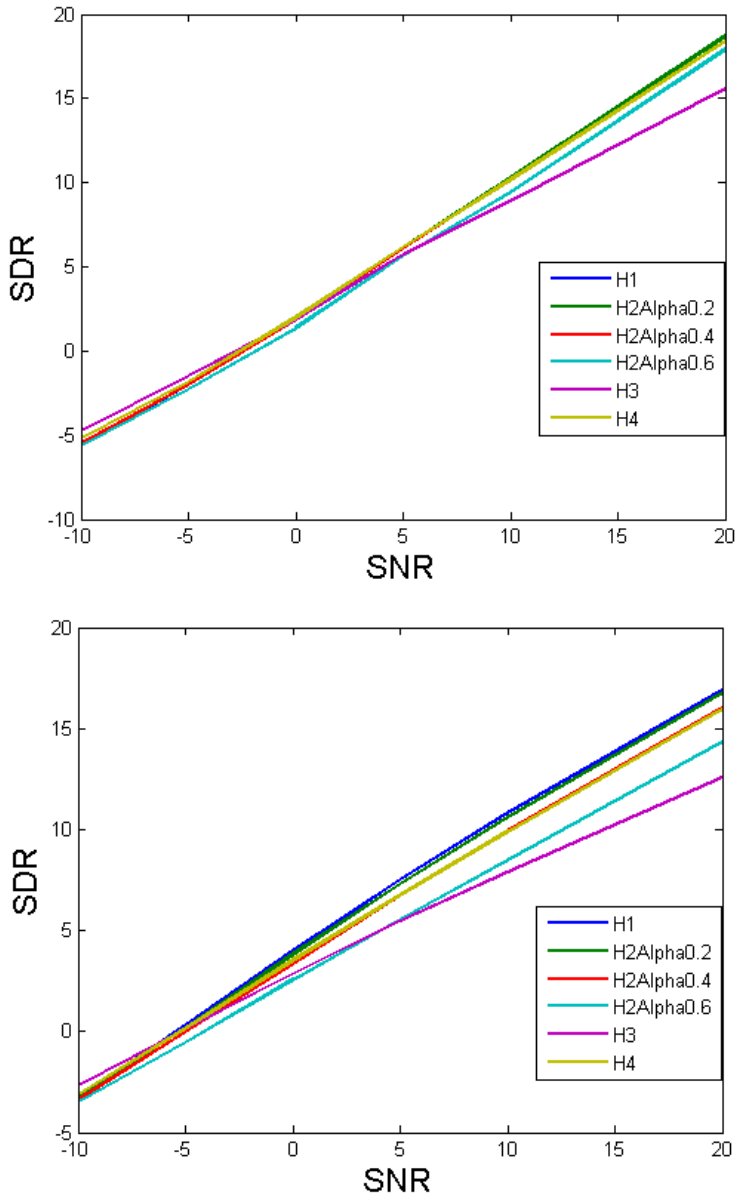


Figure 4.5: *SDR variations with SNR for the target speaker of: top) Messiah database, bottom) GRID database.*

The SAR results in Figure 4.6, show that the cube gain function $H3$ and $H2$ (with α=0.6), give the worst performance in terms of output SAR as these func-

tions introduce more distortion by spectrally masking more parts of the extracted speech. $H1$ and $H2$ (with α=0.2), give the best performance in terms of SAR as these functions introduce less processing distortion.

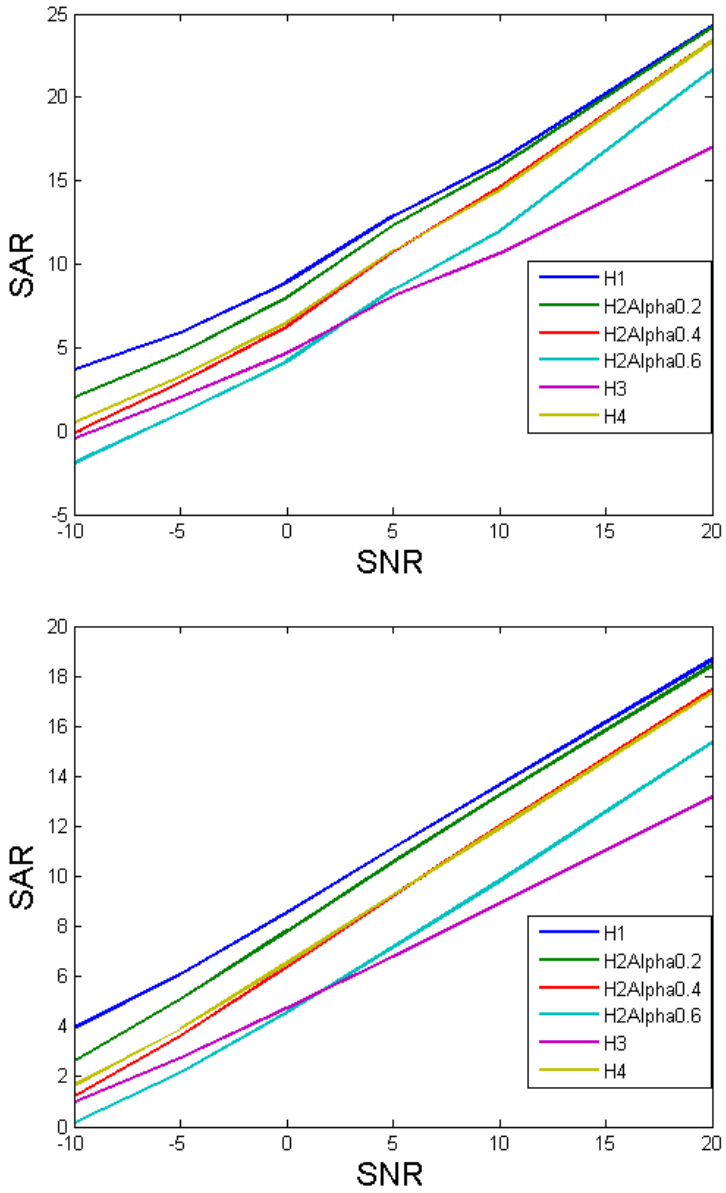


Figure 4.6: *SAR variations with SNR for the target speaker of: top) Messiah database, bottom) GRID database.*

The effectiveness of the speaker separation is illustrated in Figure 4.7 and Figure 4.8. Figure 4.7 shows the spectrograms of an utterance from the target

speaker (male), the resulting mixture with a competing speaker at an SIR of 0dB and finally the result of visually-derived speaker separation using $H2$ with α=0.4. This shows that the mixture has been processed successfully to remove most of the unwanted components from the wanted source.

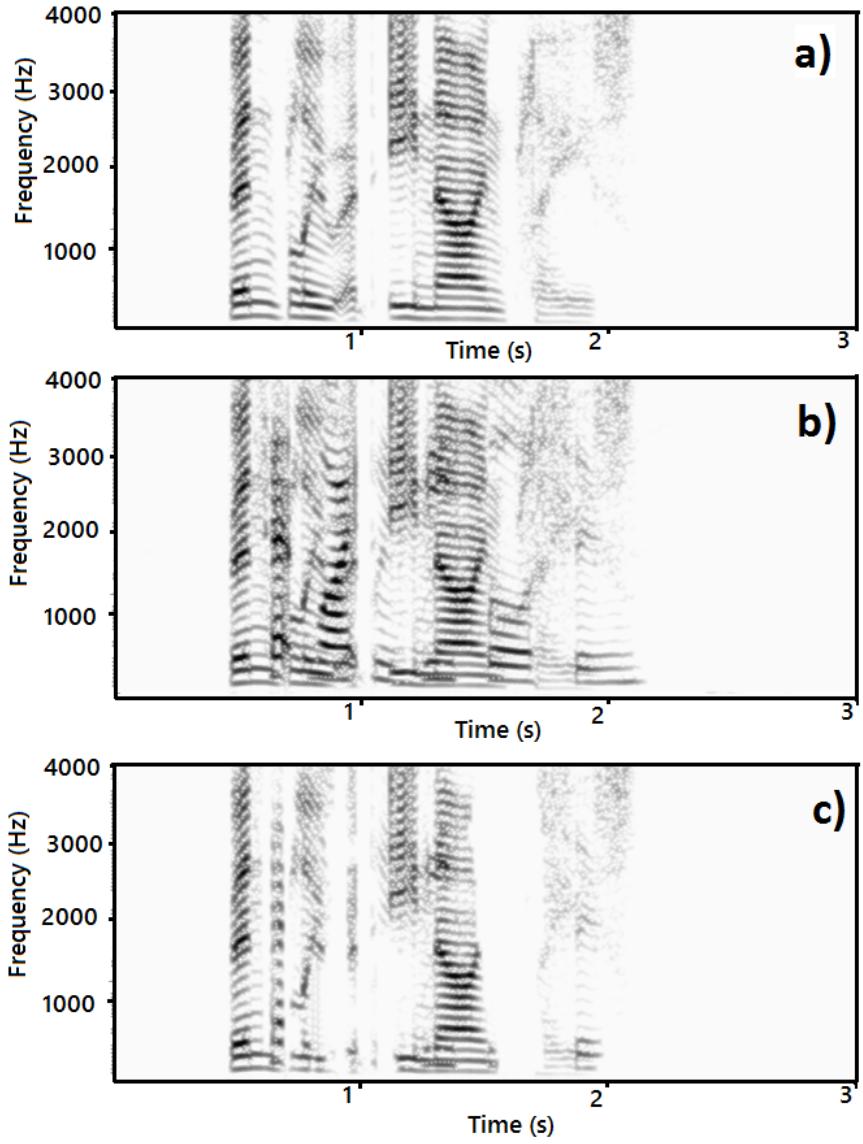


Figure 4.7: Spectrograms of the utterance '*Bin blue at e nine please*': a) target speaker (male), b) mixed with competing speaker at an SIR of 0dB, c) visually-derived speaker separation using $H2$ with α=0.4

In the same way, in Figure 4.8, the competing speaker (female) used in the mixture in the Figure 4.7, has been extracted using the same configuration as in Figure 4.7.

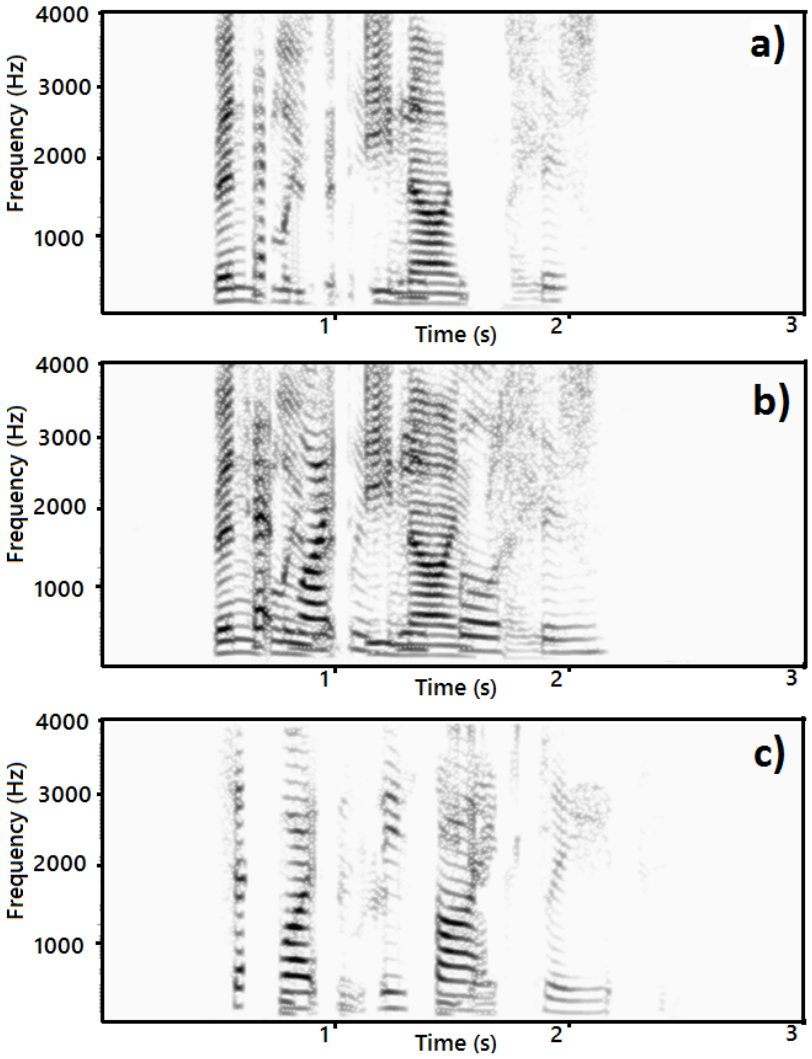


Figure 4.8: Spectrograms of the utterance '*Set white with v four soon*': a) competing speaker (female), b) mixed with target speaker at an SIR of 0dB, c) visually-derived speaker separation using $H2$ with α=0.4

4.5.3 Speech intelligibility

In addition to speech quality it is useful to know whether the proposed visually-derived speaker separation is able to improve the intelligibility of the target speaker. For the Messiah database target speaker, to measure the intelligibility an unconstrained mono-phone speech recogniser was employed. This comprised a set of 45 mono-phone HMMs including silence that were arranged in a fully connected grammar. The total number of states per HMM are five including the two non-emitting states 1 and 5. For the GRID database, a whole word speech recogniser is used. Each utterance follows a grammar containing six words of the following structure

command→*colour*→*preposition*→*letter*→*digit*→*adverb*.

The total number of models including silence is 52 with eight active states per HMM.

For the Messiah and LIPS2008 databases, the initial 200 utterances were used for training while the remaining 79 utterance were used in the testing. Similarly for the GRID database, out of the 1000 utterance, 800 were used for training and 200 for testing. For both sets of results, 23-dimensional filterbank vectors of 20ms duration at 10ms intervals from the estimates of the target speaker's speech were extracted to be used in the recognisers. Figure 4.9 shows mono-phone and word recognition accuracy for the target speakers of the two sets of results using various perceptual gain functions at SIRs from -10dB to +20dB. The entry named 'Mixture', shows results when no speaker separation has been applied to the mixture.

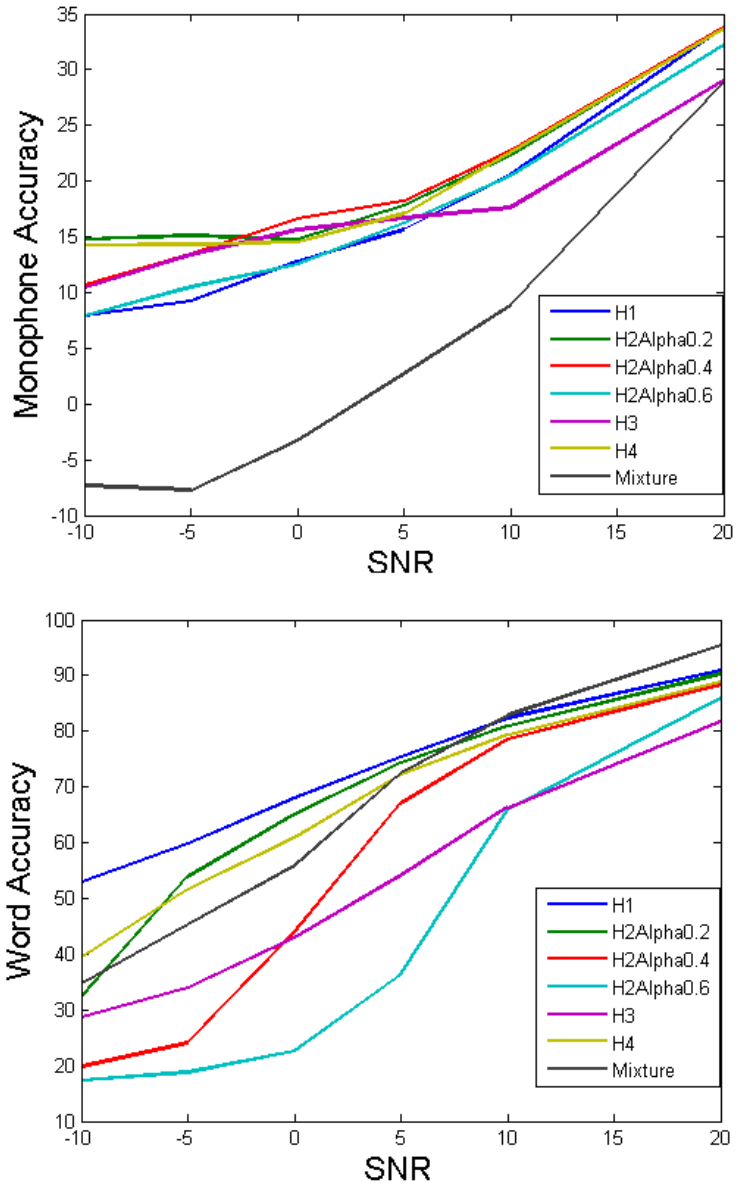


Figure 4.9: *Accuracy variations (%) with SNR for the target speaker of: top) Messiah database, bottom) GRID database.*

The unconstrained mono-phone accuracy for the original target speaker in clean conditions is 49.22%. The results in Figure 4.9, for the Messiah database

speaker (top panel), show that with no speaker separation, recognition accuracy falls significantly as SNRs reduce with a sizeable drop observed below 20dB. Applying speaker separation using the Wiener gain (i.e. $H1$) gives a good increase in recognition accuracy for the target speaker over the uncompensated case. The perceptual gain functions give further increases in recognition accuracy. The gain functions $H2$ with α=0.6 and $H3$ perform the worst at many times although these functions were performing as the best in case of SIR gains. The reason for this poor intelligibility performance is that these gain functions introduce heavy distortions into the wanted speech signals due to heavy spectral masking. It is a kind of trade-off between the unwanted signal's suppression and the wanted signal's distortion. Those gain function who remove lesser parts of the processed speech, perform better on intelligibility but worse on SIR gains. Consistently best performance is given by $H2$ α=0.2 at the very low SNRs of -10dB and -5dB, while at higher SNRs, $H2$ with α=0.4 is better. The piecewise gain function of $H4$ also performs well and has highest recognition accuracy when averaged across all SNRs. $H1$ performs the worst in the lower SIR region because of the presence of unwanted speech of the competing speaker.

The word accuracy for the target speaker of the GRID database in clean conditions is 99%. The results for the GRID database target speaker, Figure 4.9 (bottom panel), almost follows the same pattern, with the exception of gain function $H1$, which performs the best all the times for the GRID target speaker. The difference is because of the insertion errors. $H1$'s, performance is heavily degraded by insertion errors for Messiah database speaker. These insertion errors can be attributed to the nature of the recogniser (mono-phone) and the lower levels of audio-visual correlation in the LIPS2008 database, and also to the nature of the

two data sets, as the GRID data set is consisting of speech of isolated words while the Messiah and LIPS2008 data sets consist of continuous speech. Also for GRID target speaker, the intelligibility results are better for the unprocessed mixture as compared to $H2$ with α=0.4 and α=0.6 and $H3$ as these gains functions spectrally mask huge portions of the mixed speech causing the loss of wanted segments of the target speaker.

During the speaker separation process, some useful segments of the target speaker are lost. The speech recognition accuracy can be improved by using either AV-recogniser or by reconstructing the lost segments by using a suitable technique such as missing data techniques [7]. But here the purpose is not the recognition accuracy but speaker separation and the recognition results are presented just to show the effectiveness of the method.

4.6 Summary

This chapter provided an overview of Wiener filtering for speech enhancement and speaker separation. Clean speech and noise estimates are required for the construction of Wiener filter for speech enhancement. While for the construction of Wiener filter for speaker separation, the clean speech estimates of target and competing speakers are required. Speaker separation is a more difficult problem because of the similarity between the acoustic features of the target and competing speakers. To exploit the higher levels of correlation between the audio and visual speech features, it was proposed to obtain the power spectra statistics for the target and competing speakers from the visual speech features taken from the two speakers. The Wiener filter was constructed in the filterbank domain because the

audio-visual correlation is found to be insufficient to give a fine resolution estimate, although estimation of a less spectrally detailed filterbank vector is possible. The Wiener filter gains were modified using several perceptual gain functions, in the search for improved speech quality and intelligibility.

The results were shown for two sets of data: Messiah and LIPS2008 data set and the GRID data set, and for different SIRs of -10dB to +20dB. The results showed that the proposed method of Wiener filtering and the subsequent perceptual gain functions, improve the speech quality in terms of the suppression of the competing speaker and the intelligibility of the extracted target speaker. The selection of the perceptual gain functions is a trade-off between the quality and intelligibility. Those gain functions that try to suppress heavily the competing speaker by spectrally masking more parts of the mixed speech, introduce more distortion into the extracted target speech. So to keep a balance between the quality and intelligibility, the gain functions $H2$ with α=0.2 and with α=0.4 and $H4$ seems to be doing better.

Chapter 5

Speaker Separation using Visually-derived Binary Masks

Preface

This chapter proposes a solution for the problem of single-channel speaker separation and exploits visual speech information to aid the separation process. Audio from a mixture of speakers is received from a single microphone and to supplement this, video from each speaker in the mixture is also captured. The visual features are used to create a time-frequency binary mask that identifies regions where the target speaker dominates. These target dominant regions are retained and form the estimate of the target speaker's speech. While the regions where the competing speaker is dominant, are masked and discarded. Experimental results compare the visually-derived binary masks with ideal binary masks which shows a useful level of accuracy. The effect of the number of filterbank channels on mask accuracy is also studied . The effectiveness of the proposed method of speaker sep-

aration using visually-derived binary masks is then evaluated through estimates of speech quality and speech intelligibility. These results show substantial gains in quality and intelligibility for the processed speech over the original mixture.

5.1 Introduction

Most methods of speaker separation exploit the masking property of human speech perception. Humans have this inborn capability to either suppress the unwanted speakers and noises or extract the target speaker or both at the same time [31]. Most of the speaker separation methods aim to identify and extract time-frequency regions of the speech mixture that are dominated by the target speaker and mask out the other regions. These masks are known as binary masks and each time-frequency component is set to either one or zero depending on whether the region is dominated by the target speaker or is to be masked. The challenge is to estimate accurately the mask and identify time-frequency components to be retained and those which are to be masked. Various approaches have been employed to find the mask and these typically operate by grouping time-frequency regions according to various criteria. One of the most effective is computational auditory scene analysis (CASA) which groups regions perceptually, making use of cues such as harmonicity and onset and offset times [102]. Alternative approaches have used statistical approaches whereby dependencies between time-frequency regions are established and used to form the mask. An extension of the binary mask is the soft mask, where instead of a binary decision as to whether a time-frequency component is masked a probability of masking is computed which thereby allows some uncertainty to exist in the mask [77], [84].

This work proposes using visual speech information from each speaker in the mixture to estimate the binary mask. Significant correlation exists between audio and visual speech features extracted from a speaker and this can be exploited to enable audio features to be estimated from visual features [105], [4]. Given audio feature estimates for the speakers in the mixture an estimate of the binary mask can be made from which the target speaker can be extracted. The proposed system uses a single microphone as the audio input which receives the mixture of speech from the speakers. Information to enable separation of speakers is provided by visual speech features that are extracted from the mouth region of each speaker in the mixture.

The proposed method of visually-derived binary masks estimation for speaker separation is described in Section 5.2. To compute the binary masks, audio estimates of the target and competing speakers are required and these estimates are made from the corresponding visual speech features and this estimatition process is discussed in Section 5.3. Experimental results are presented in Section 5.4 which first examine the accuracy of the visually-derived binary masks and then evaluate the extracted target speaker's speech in terms of speech quality and intelligibility. The entire process of speaker separation using visually derived binary masking is shown in Figure 5.1, where the top panel shows the 'separation process' while the bottom panel shows the 'training process'.

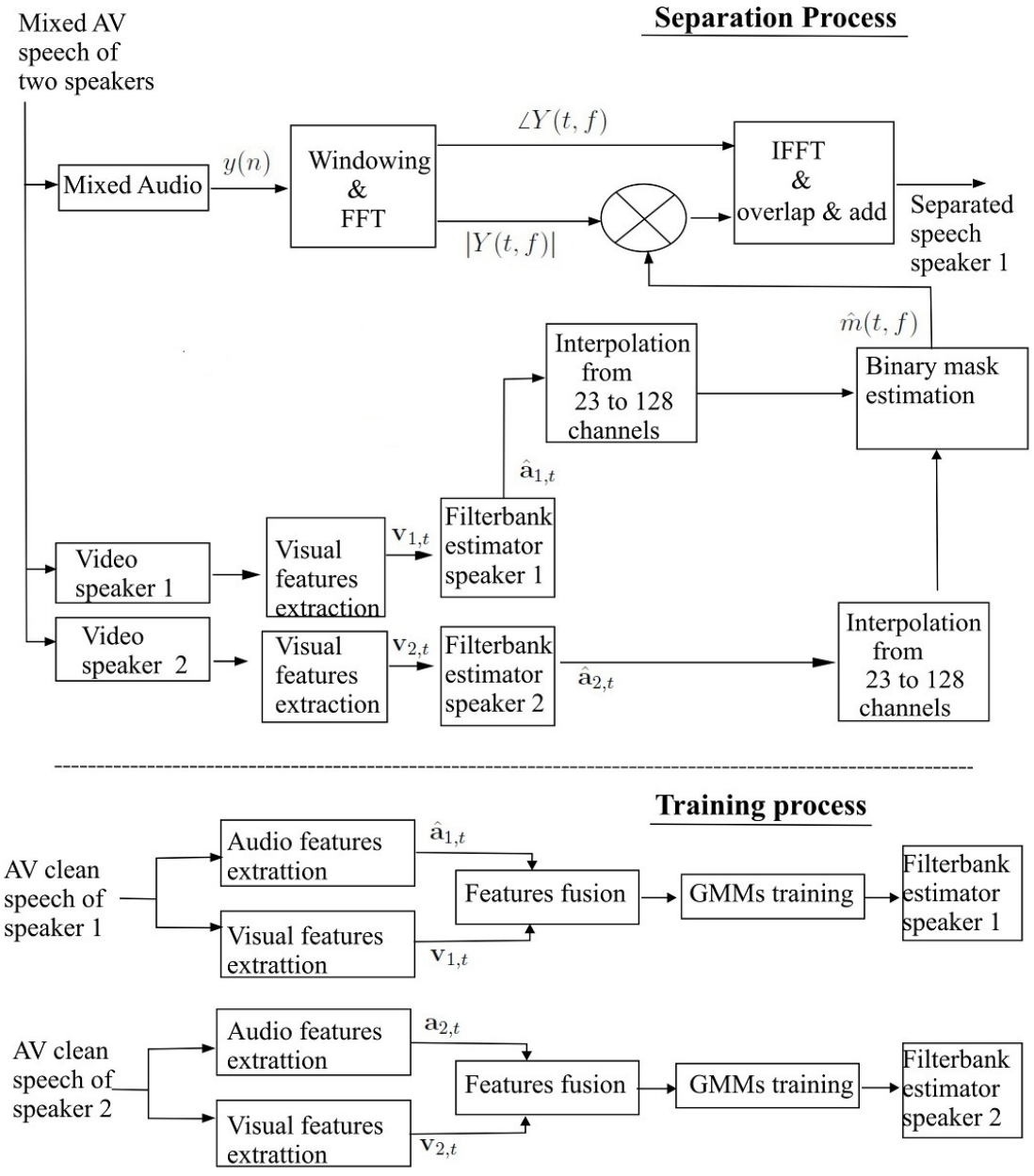


Figure 5.1: Speaker separation using visually derived binary masking including the training stage.

5.2 Visually-derived binary masks

In this chapter, the audio-visual correlation between a speaker's mouth shape and the resulting audio signal is exploited for deriving binary masks to be used for the single channel speaker separation task. Speaker separation using binary masks involves first the estimation of a time-frequency mask where each component signifies whether that time-frequency component is dominated by either the target speaker or interfering speakers. Areas where the binary mask indicates the region is target-dominated are retained, while regions that are dominated by interfering speakers are masked and discarded. This work exploits audio-visual correlation and proposes a method of estimating the binary mask using visual speech information.

5.2.1 Mixing model

In the time-domain it is assumed that a mixed signal, $y(n)$, is made from the addition of speech from a target speaker and an interfering speaker, $x_1(n)$ and $x_2(n)$, where

$$y(n) = x_1(n) + x_2(n) \tag{5.1}$$

By applying the Fourier Transform we get

$$Y(f) \;\; = \;\; X_1(f) + X_2(f) \tag{5.2}$$

where $X_1(f)$ and $X_2(f)$ are the complex spectrums of speaker 1 and speaker 2 respectively. By taking element-wise squared magnitude and assuming that the two signals are uncorrelated [77],[75], it can be written in the power spectrum

domain as

$$|Y(f)|^2 \;=\; |X_1(f)|^2 + |X_2(f)|^2 \tag{5.3}$$

where $|Y(f)|^2$, $|X_1(f)|^2$ and $|X_2(f)|^2$ are the power spectra of the mixture and the two speech signals respectively, where f represents the spectral bin.

5.2.2 Estimation of binary mask

The proposal in this work is to use information from visual speech features taken from both the target speaker and interfering speaker to estimate the binary mask. Analysis of audio and visual speech features has shown that significant correlation exists between the two, enabling audio speech features to be estimated from visual speech features [4]. In particular, broad spectral envelope features such as log filterbank or MFCC features can be estimated from 2D-DCT or AAM visual features with good accuracy [2]. An advantage of such a visually-derived estimate is that the resulting audio features are free from any interference from other speakers or any other sound sources. Estimation of fine spectral detail, such as harmonic frequencies, is not possible from the visual features as they do not contain source information but a smoothed spectral representation is attainable.

From the target and interfering speaker, visual features, $\mathbf{v}_1(t)$ and $\mathbf{v}_2(t)$ are extracted at each time frame, t. From the visual features, estimates of corresponding audio features, $\hat{\mathbf{a}}_1(t)$ and $\hat{\mathbf{a}}_2(t)$, are made using MAP estimation

$$\begin{aligned} \hat{\mathbf{a}}_1(t) &= \mathrm{MAP}(\mathbf{v}_1(t)) \\ \hat{\mathbf{a}}_2(t) &= \mathrm{MAP}(\mathbf{v}_2(t)) \end{aligned} \tag{5.4}$$

where the estimation is shown by the function $MAP()$. The process of estimating audio features from visual features is explained in Section 5.3. In this work the visual features are formed from a 2D-DCT of an ROI centred around each speaker's mouth, while the audio features are from a D-dimensional log filterbank.

To compute the binary mask, the estimated D-dimensional log filterbank vector must be interpolated to the dimensionality of the power spectral features which in this work is F=128, and $D < 128$. This is achieved by cubic spline interpolation to give time-frequency spectral representations for the target and interfering speakers, $A_1(t, f)$ and $A_2(t, f)$

$$\begin{aligned} A_1(t, f) &= interp(\hat{\mathbf{a}}_1(t)) \\ A_2(t, f) &= interp(\hat{\mathbf{a}}_2(t)) \quad 1 \leq t \leq T, 1 \leq f \leq F \end{aligned} \tag{5.5}$$

where T is the number of time frames in the utterance. The binary mask, $\hat{m}(t, f)$, is now computed by comparing the spectral values at the corresponding T-F units of the target and competing speakers as shown in Equation 5.6, where the value of binary mask is set to 1 when the target speaker's energy is greater than that of the interfering speaker, or in other words when the local signal-to-noise ratio (SNR) is greater than 0dB typically.

$$\hat{m}(t, f) = \begin{cases} 1 & A_1(t, f) \geq A_2(t, f) \\ 0 & A_1(t, f) < A_2(t, f) \end{cases} \tag{5.6}$$

This is based on the log-max assumption which assumes that in any particular frequency band at any time, the energy contribution of one speaker in the mixture is dominant and masks the other speakers in the mixture [84].

5.2.3 Time-domain reconstruction

From the time-frequency representation of the mixed signal magnitude spectrum, $|Y(t,f)|$, an estimate of the magnitude spectrum of the target speaker, $|\hat{X}_1(t,f)|$, can be made using the estimated binary mask

$$|\hat{X}_1(t,f)| = \hat{m}(t,f)|Y(t,f)| \quad 1 \leq t \leq T, 1 \leq f \leq F \tag{5.7}$$

The sequence of magnitude spectral frames of the filtered target speech must now be transformed into a discrete time-domain speech signal, $\hat{x}_1(n)$. This is achieved by first combining each magnitude spectrum estimate with the phase of the original mixed speech signal, $\angle Y(t,f)$, and then applying an inverse Fourier transform to obtain a short-duration frame of time-domain samples.

$$\widehat{x}_{1,t}(n) = \text{IFFT}(|\widehat{X}_1(t,f)|\angle Y(t,f)) \tag{5.8}$$

These frames are then overlapped by 50% and added together to create the estimate of the target speaker's speech.

5.3 Estimation of audio features from video

The correlation between audio and visual features is exploited to estimate audio features from visual features. The estimation process involves first training a GMM to model the joint density of audio and visual speech features. MAP estimation can then be applied to estimate audio features, $\widehat{\mathbf{a}}_1(t)$, from visual features. For the details of the estimation process please refer to Chapter 3.

5.3.1 Audio and visual features

D-channel mel-scale filterbank features, $\mathbf{a}_t$, are used as the audio features. These are extracted from 20ms duration frames of audio at 10ms intervals in accordance with the ETSI XAFE standard [28]. The dimensionality of the filterbank, D, is an important parameter and its effects on mask estimation accuracy and filterbank estimation errors are examined in Section 5.4.2 and Section 5.4.4 respectively. Visual features, $\mathbf{v}_t$, are extracted from an ROI centered on a speaker's mouth. A 2D-DCT is then applied to the ROI and the first 15 coefficients are retained in a zigzag manner as the 2D-DCT visual vector. For a detailed discussion of these audio and visual features, please refer to Chapter 2.

5.4 Experimental results

An evaluation of the effectiveness of the visually-derived binary masks for speaker separation is made in this section. First, the audio-visual data and experimental set up used, are described. Second, an analysis of the accuracy of the visually-derived binary masks is presented. Finally, experimental results are presented on the quality and the intelligibility of the target speaker's speech following visually-derived speaker separation.

5.4.1 Audio-visual data

In the case of experiments with the Messiah (male) and LIPS2008 (female) databases, the first 200 utterances of each database were used for training while the remaining 79 utterances were used for the evaluation. The audio in both databases was

down sampled to a sampling frequency of 8 kHz and he video was up sampled to 100 frames per second to match the audio frame rate.

The experimental scenario investigated is of two speakers talking simultaneously and being located close together in space, with the male speaker the target and the female the competing speaker. The mixed speech was created by mixing the speech utterances from the two databases to get the mixed signals (noisy speech). The LIPS2008 utterances are scaled and added to the Messiah database utterances in such a way that the resulting mixed utterances are having a signal-to-interference ratio (SIR) of -10dB, -5dB, 0dB, 05dB, 10dB and 20dB. The SIRs are calculated only over speech periods by ignoring the initial and end silence from the utterances.

In the case of experiments with the GRID database, the data of speaker 6 (male) and speaker 4 (female) were used. Out of the 1000 utterances, 800 were used for training and the remaining 200 for the evaluation. The female speaker utterances are scaled and added to the male speaker's utterances in such a way that the resulting mixed utterances are having a signal-to-interference ratio (SIR) of -10dB, -5dB, 0dB, 05dB, 10dB and 20dB. The rest of the experimental set up was kept the same as in the case of the other two databases. For a detailed discussion of the audio-visual speech databases used, please refer to chapter 2.

5.4.2 Mask accuracy

The accuracy of the visually-derived binary mask is evaluated by comparing it with the ideal binary mask. The ideal binary mask is computed from the actual energy levels in the target and interfering speakers at each time-frequency point

for the clean reference data before mixing. The metric used for evaluation is the percentage of components in the visually-derived mask that were estimated correctly and is defined as

$$Accuracy = \frac{\hat{m}_{total}(t,f) - \hat{m}_{incorrect}(t,f)}{\hat{m}_{total}(t,f)} \times 100 \tag{5.9}$$

where $\hat{m}_{incorrect}(t,f)$ and $\hat{m}_{total}(t,f)$ are the incorrectly identified and total binary masks respectively and the 'Accuracy' is representing the percentage of correctly identified binary masks.

The experiments examine the effect of different numbers of filterbank channels (from $D = 2$ to $D = 50$) and at SIRs from -10dB to +20dB, which are reported in Table 5.1. The results show that mask accuracy improves slightly with increasing numbers of filterbank channels but this improvement varies only by at most at around 3%.

SIR	-10dB	-5dB	0dB	5dB	10dB	20dB
D=2	71.57	66.27	67.07	70.37	74.53	82.79
D=6	72.06	67.49	67.60	69.86	74.49	83.20
D=12	73.05	67.43	67.74	70.08	73.95	83.20
D=18	**73.76**	68.33	67.96	70.39	74.13	83.14
D=23	72.03	66.88	68.30	69.32	74.03	82.04
D=27	73.21	68.44	68.42	70.96	74.80	**83.23**
D=30	73.19	68.38	68.32	71.54	**75.57**	83.04
D=50	72.95	**68.66**	**68.96**	**71.93**	75.30	83.09

Table 5.1: *Visually-derived mask estimation accuracy (%) at SIRs from -10dB to +20dB and filterbank sizes from 2 to 50 channels for Messiah database .*

5.4.3 Effect of number of channels on visually-derived and ideal binary masks

To investigate further the effect of varying the number of filterbank channels, an artificial test was carried out that took the ideal binary masks calculated from the D-dimensional ideal filterbank features interpolated to 128 dimensions. The accuracy of these ideal binary masks from the D-dimensional ideal filterbank features was measured by comparing it with the ideal binary masks computed from the ideal 128-dimensional filterbank features. Table 5.2 compares the accuracy of these ideal filtered binary masks to the visually-derived binary masks extracted at an SIR of 0dB. The results for the filtered ideal mask show that the process

Number of channels	Accuracy of Visually-derived	Accuracy of Filtered ideal
D=2	67.07	82.01
D=6	67.60	84.94
D=12	67.74	86.97
D=18	67.96	87.70
D=23	68.30	88.62
D=27	68.42	88.84
D=30	68.32	88.94
D=50	68.96	91.06

Table 5.2: *Comparison of the accuracy (%) of the visually-derived binary masks and ideal binary masks subject to filterbank quantisation, for filterbank sizes from 2 to 50 channels at an SIR of 0dB for Messiah database.*

of filterbank quantisation introduces a substantial reduction in mask accuracy – with quantisation to 2 channels, accuracy is reduced by almost 18%. However, accuracy of the filtered ideal mask does recover rapidly as more filterbank channels are introduced. In comparison, recovery of the visually-derived binary mask is much less – by only 2% in comparison to 10% when moving from 2 to 50 channels

in case of ideal binary masks. This suggests that there is a fairly low limit on the amount of spectral detail that can be extracted from visual features.

5.4.4 Filterbank estimation accuracy

The effect of varying the number of filterbank channels was further studied using filterbank estimation errors for the filterbank audio features estimated from the visual features. The mean filterbank estimation error in percentage, E, was defined in Chapter 3 in Equation 3.10. These estimation errors are shown in Table 5.3 along with the filterbank estimation errors for the ideal filterbank audio features for various number of channels. The results show huge estimation errors for both

Number of channels	Errors for visually derived	Errors for Filtered ideal
D=2	44.22	39.77
D=6	36.61	31.68
D=12	29.13	23.77
D=18	24.69	19.20
D=23	22.45	16.32
D=27	21.00	14.59
D=30	20.08	13.48
D=50	16.56	8.33

Table 5.3: *Comparison of the filterbank estimation errors (%) of the visually-derived filterbank audio features and ideal filterbank audio features subject to filterbank quantisation, for filterbank sizes from 2 to 50 channels at an SIR of 0dB for Messiah database.*

the visually-derived and ideal filterbank audio features at the lower number of filterbank channels. The decrease in estimation errors is very rapid as the number of channels are increased from $D = 2$ to $D = 23$ channels. For the visually derived filterbank features, the estimation errors drop from 44.22% for $D = 2$ channels to 16.56% for $D = 50$ channels, giving a decrease of 27.66% in estimation errors. For

the ideal filterbank features, the estimation errors drop from 39.77% for $D = 2$ channels to 8.33% for $D = 50$ channels, giving a decrease of 31.44% in estimation errors.

By comparing the results for mask accuracy and filterbank estimation errors, from Table 5.2 and Table 5.3, an observation is drawn. The comparison show that by increasing the numbers of filterbank channels from 2 to 50 channels for the visually derived cases, the mask accuracy improves slightly by at most at around 3% but the reduction in filterbank estimation errors is huge and is 27.66%. The reason for this is that the reduction of estimation errors require an accurate estimate of the filterbank features which is not possible from just 2 channels and as the number of channels are increased, the estimate of the filterbank features becomes more accurate and results in less estimation errors. While the estimation of binary masks only requires a rough estimate of the energy levels of the estimated features, and an even lower number of filterbank channels also provide good information to estimate the binary masks with a reasonable accuracy.

Figure 5.2, Figure 5.3 and Figure 5.4, provide further insight into mask estimation and show the ideal binary masks and then binary masks computed for 2, 23 and 50 channel filterbanks, with each showing the ideal and visually-derived masks. White regions indicate regions that are dominated by the target speaker and are to be retained. Examination reveals that at low numbers of channels the entire time frame is often classed as either target or interfering speaker due to the lack of spectral details available. As the number of channels increases, spectral details improve and so more frequency discrimination is possible. This is certainly evident in the filtered ideal masks, but less discrimination is available from the visually-derived masks as fine spectral details are not present in the visual features.

5.4.5 Speech quality

To estimate the quality of the target speaker's speech, SIR is used as the measures. In case of experiments with the Messiah and LIPS2008 databases, tests used the set of 79 mixed sentences and were carried out at initial SIRs of -10dB, -5dB, 0dB, 5dB, 10dB and 20dB. The visually-derived binary masks were applied to the mixtures and the resulting SIRs computed using the BSS toolbox [29]. The SIR results are shown in Table 5.4. The results show that the visually-derived

Input SIR	-10dB	-5dB	0dB	5dB	10dB	20dB
D=2	**0.06**	**2.19**	4.82	8.07	11.73	**20.36**
D=6	-0.11	1.97	**5.03**	**8.13**	**11.91**	20.19
D=12	-0.78	1.16	4.47	7.81	11.53	19.95
D=18	-0.54	1.49	4.29	7.68	11.41	19.82
D=23	-0.19	1.77	3.50	8.03	**11.91**	19.86
D=27	-2.46	-0.03	3.41	7.38	10.94	19.29
D=30	-2.30	-0.26	3.11	7.41	11.34	19.43
D=50	-3.32	-1.02	2.70	6.75	10.83	19.48

Table 5.4: *Comparison of input and output SIRs for filterbank sizes from 2 to 50 channels.*

binary masks are able to extract the target speaker from the mixture and thereby increase the SIRs. Largest gains in SIR occur at the lower input SIRs. The results also show that the number of filterbank channels does not have a large effect on the output SIR which is supported by the findings in Table 5.1, that showed little differences in binary masks accuracy for varying the number of channels. The results also show that higher SIR gains are obtained at lower number of channels.

Tests were also carried out using the GRID database. The set of 200 mixed

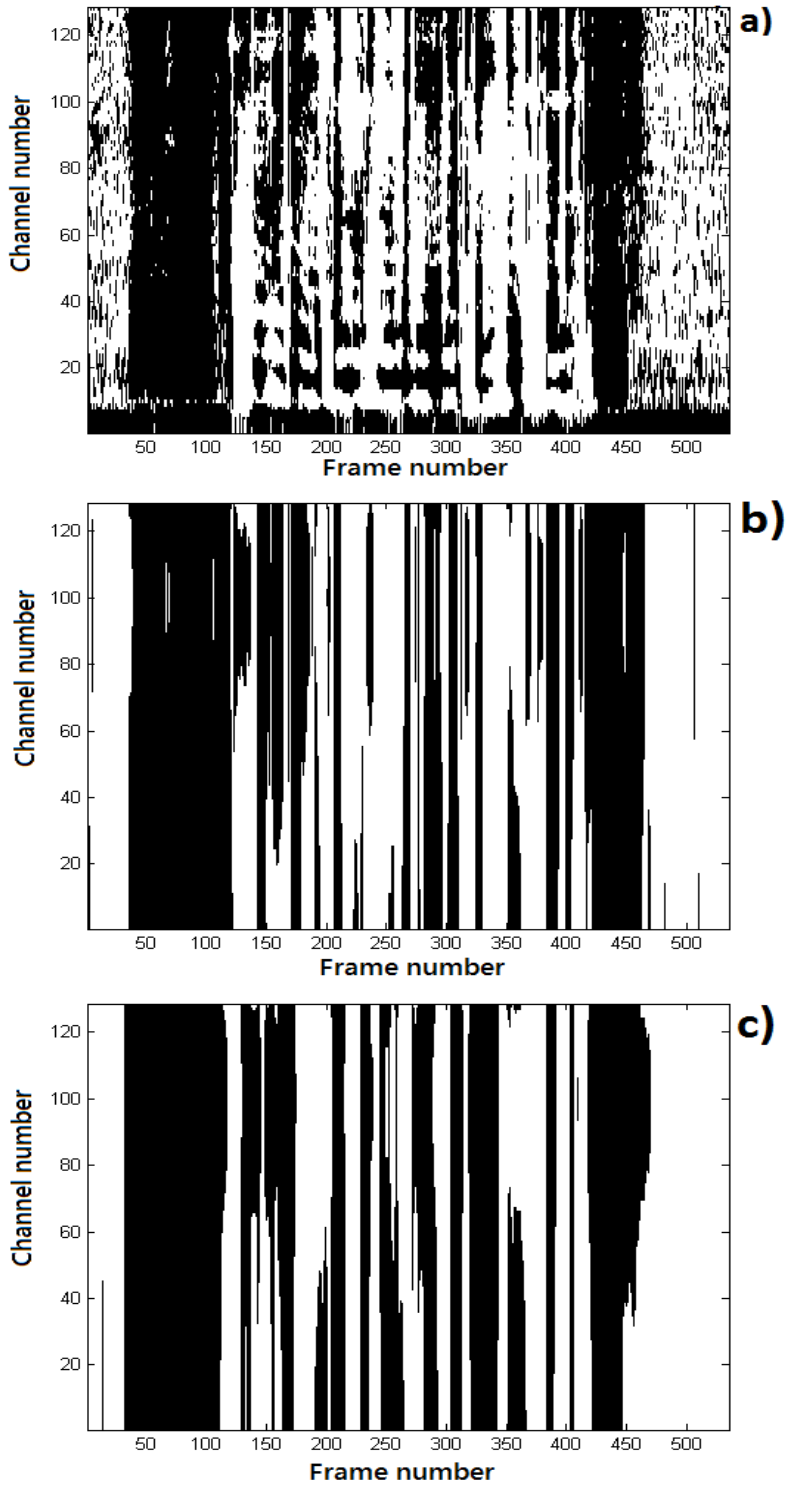


Figure 5.2: *Binary masks: a) 128-channel ideal, b) 2-channel ideal, c) 2-channel visually-derived.*

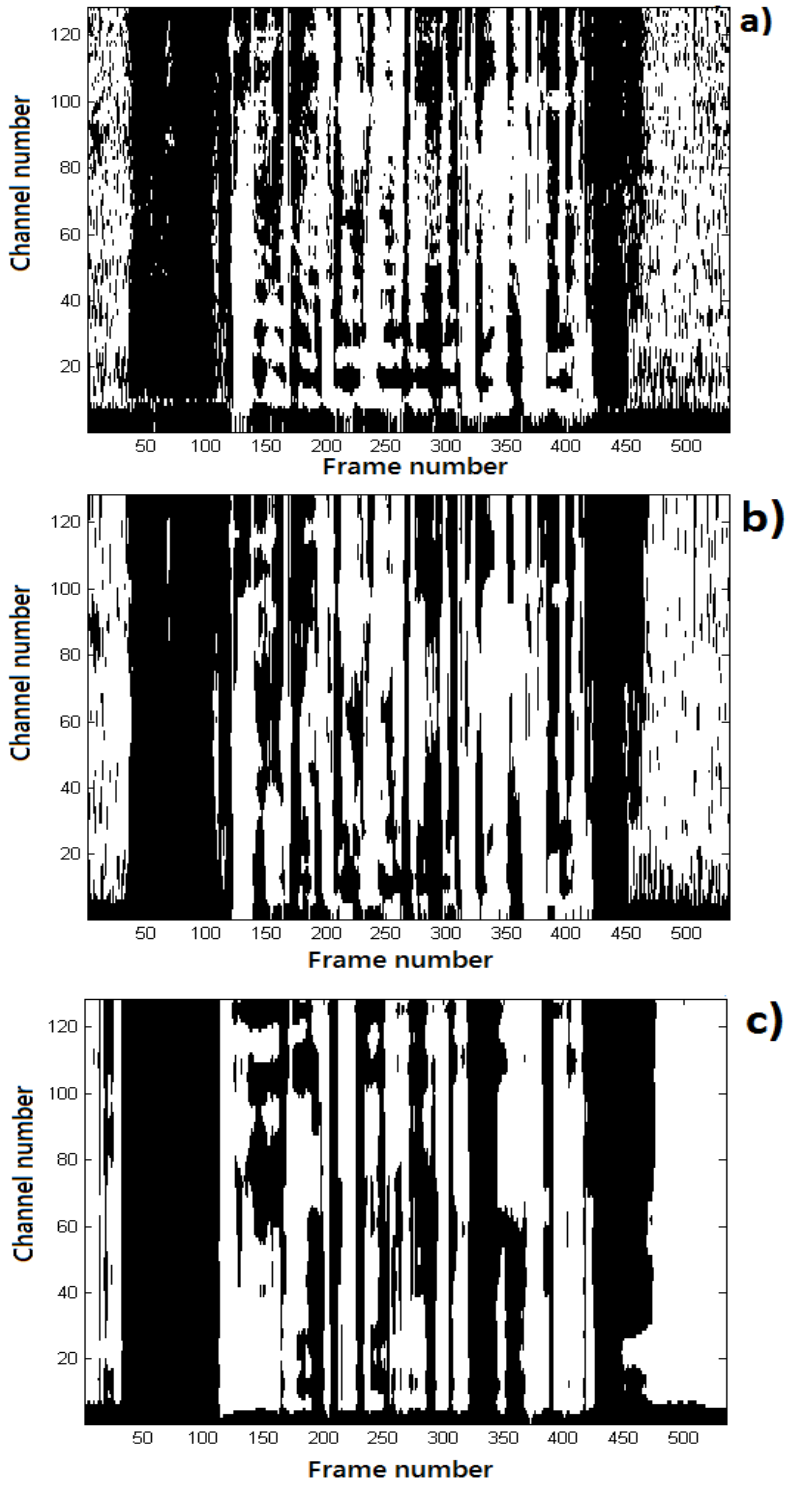


Figure 5.3: *Binary masks: a) 128-channel ideal, b) 23-channel ideal, c) 23-channel visually-derived.*

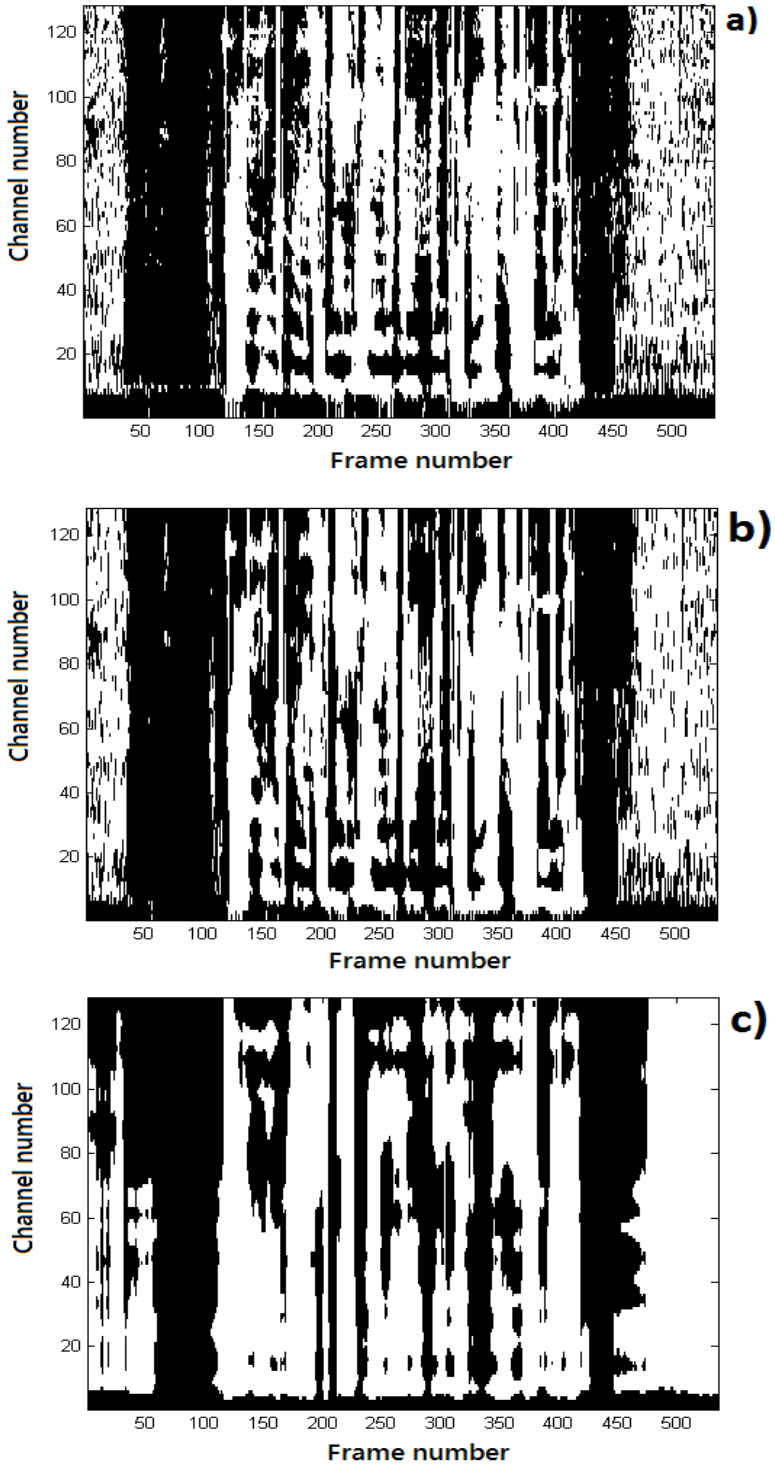


Figure 5.4: *Binary masks: a) 128-channel ideal, b) 50-channel ideal, c) 50-channel visually-derived.*

sentences was used and the tests carried out at initial SIRs of -10dB, -5dB, 0dB, 5dB, 10dB and 20dB using $D = 23$ channels. As the results in Table 5.1 show that higher SIR gains are obtained at lower number of channels and the results in Table 5.6, show that higher recognition accuracy is achieved at higher number of filterbank channels, therefore to keep a balance between quality and intelligibility, $D = 23$ channels is a good option. The SIR results for the GRID database target speaker are shown in Table 5.5

Input SIR	-10dB	-5dB	0dB	5dB	10dB	20dB
Output SIR(dB)	1.67	4.51	7.64	10.90	14.60	22.34

Table 5.5: *Comparison of input and output SIRs for the target speaker of GRID database for $D = 23$ channels.*

The effectiveness of the speaker separation is illustrated in Figure 5.5, Figure 5.6 and Figure 5.7, which show spectrograms of an utterance from the target speaker, the interfering speaker, the resulting mixture at an SIR of 0dB, and finally the results of the visually-derived binary masking using 2, 23 and 50 filterbank channels. The results show many of the attributes of the target speaker to have been successfully extracted from the mixture.

5.4.6 Speech intelligibility

This section investigates the effectiveness of speaker separation using the visually-derived binary masks in terms of speech intelligibility. In this work an estimate of speech intelligibility is made using an unconstrained monophone speech recogniser. This comprised a set of 44 monophone HMMs that were arranged in a fully connected grammar. From the masked time-domain estimates of the target speaker's

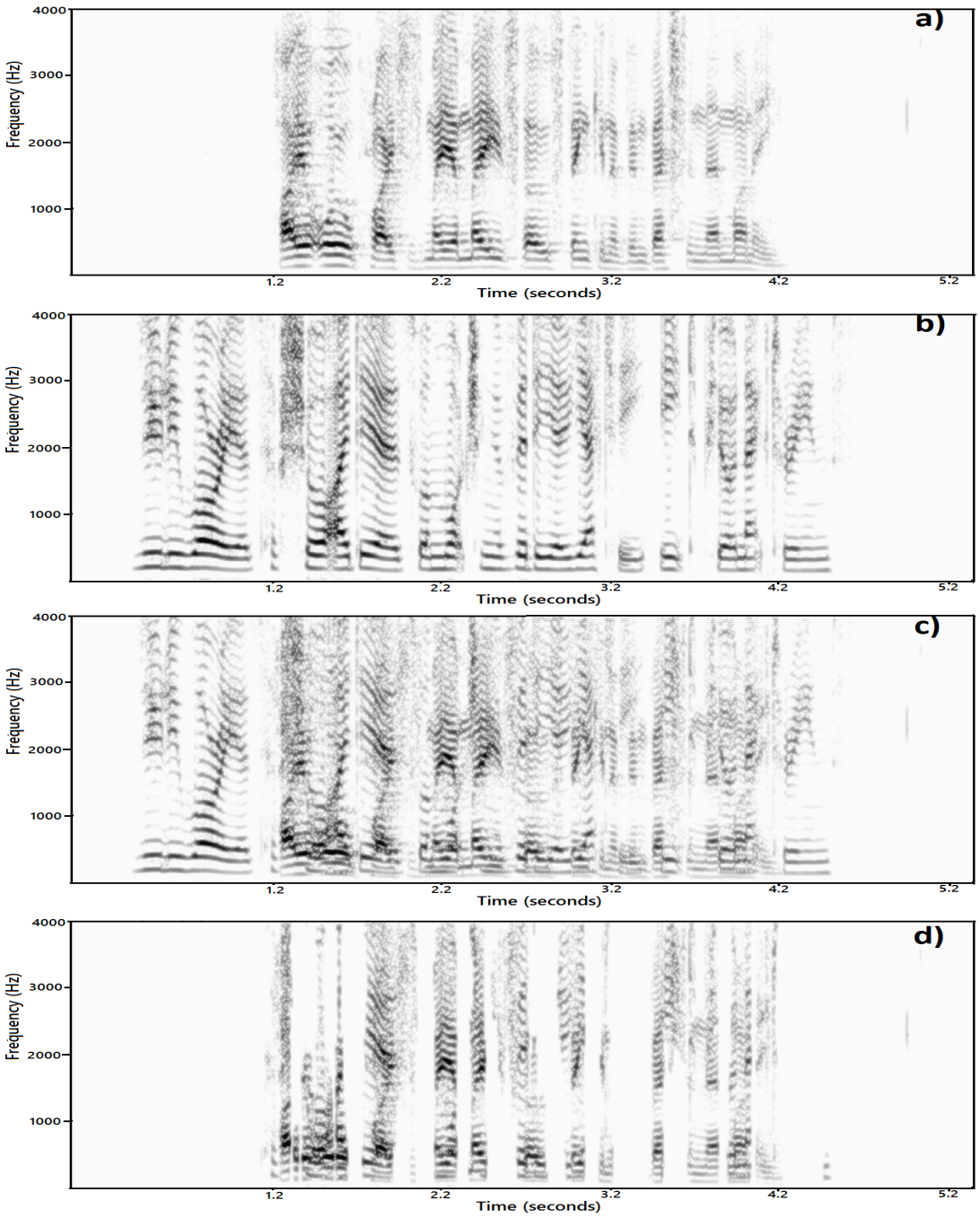


Figure 5.5: *Spectrograms showing: a) target speaker saying* 'Higher oil prices may amaze those thinking of investing their money*', b) interfering speaker saying* 'Zulu warriors have sure ideas when watching a video yeti eat pure nectarines*' c) target speaker mixed with interfering speaker at an SIR of 0dB , d) target speaker extracted using D=2 channels.*

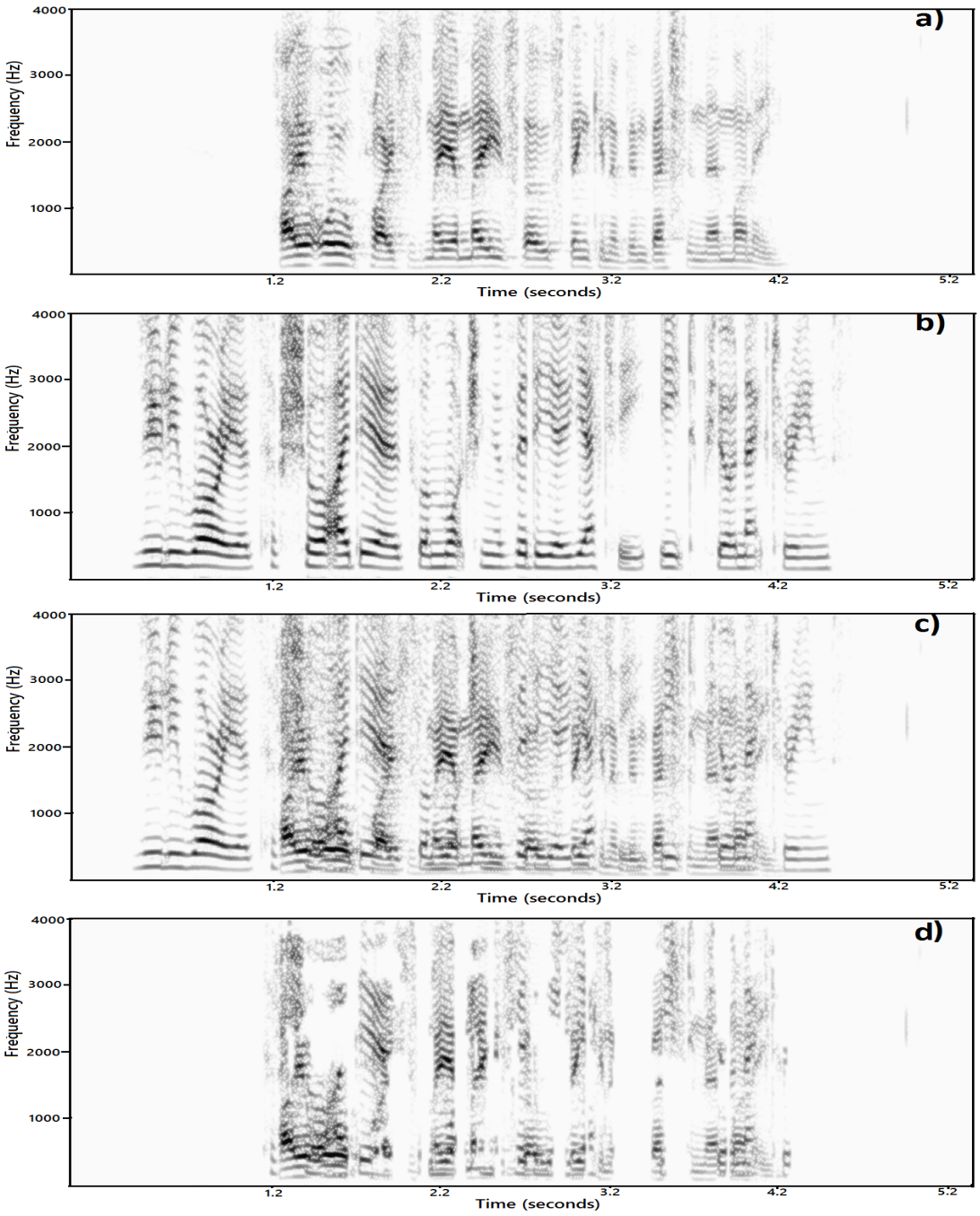


Figure 5.6: *Spectrograms showing: a) target speaker saying* 'Higher oil prices may amaze those thinking of investing their money*', b) interfering speaker saying* 'Zulu warriors have sure ideas when watching a video yeti eat pure nectarines*' c) target speaker mixed with interfering speaker at an SIR of 0dB , d) target speaker extracted using D=23 channels.*

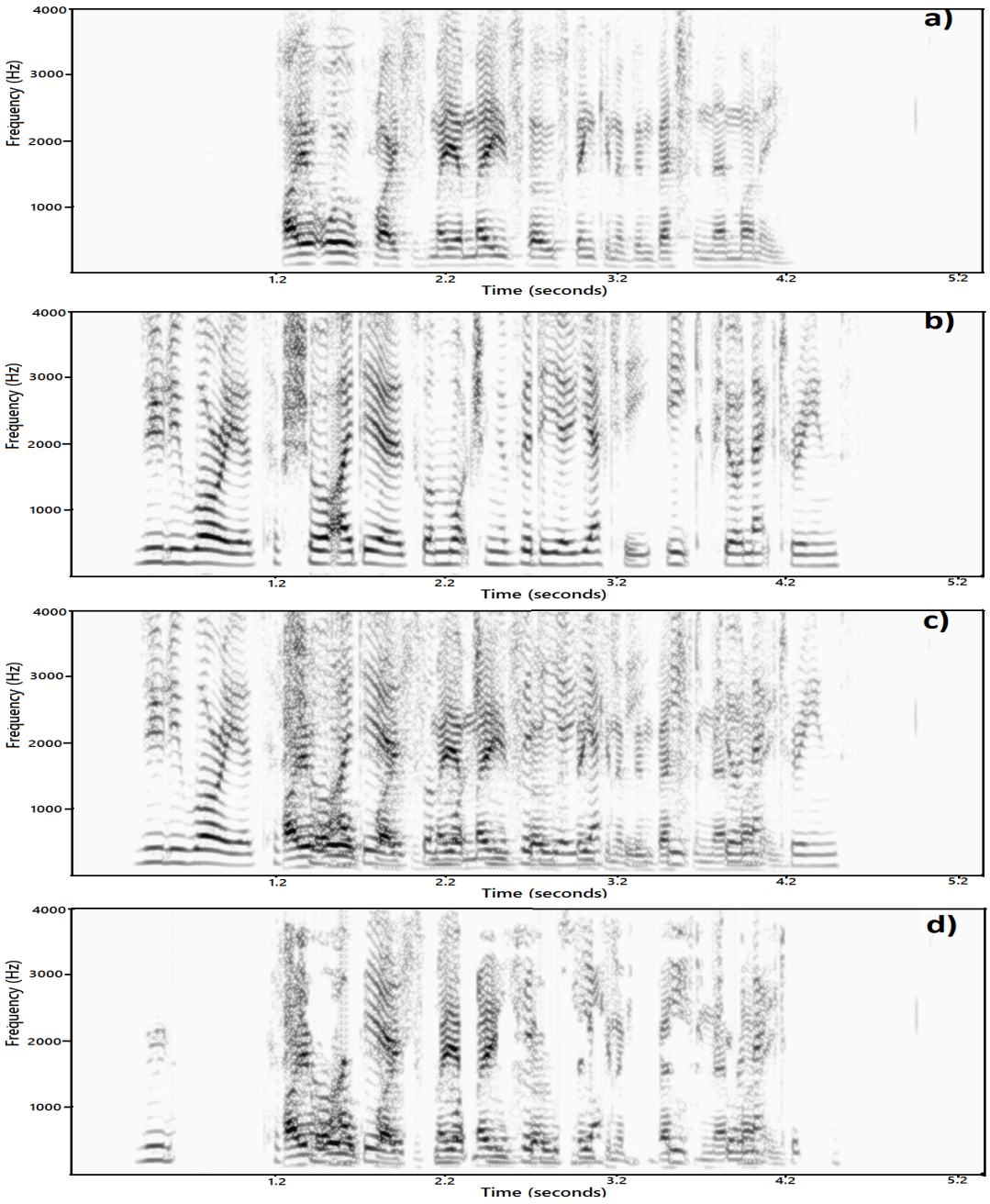


Figure 5.7: *Spectrograms showing: a) target speaker saying* 'Higher oil prices may amaze those thinking of investing their money*', b) interfering speaker saying* 'Zulu warriors have sure ideas when watching a video yeti eat pure nectarines*' c) target speaker mixed with interfering speaker at an SIR of 0dB , d) target speaker extracted using D=50 channels.*

speech, MFCC vectors were extracted in accordance with the ETSI XAFE standard [28]. Table 5.6 shows recognition accuracy for the target speaker's speech extracted using from 2 to 50 channel filterbanks and at SIRs from -10dB to +20dB. The table also shows baseline performance when no speaker separation (NSS) is applied. Unconstrained monophone accuracy for the original target speaker in clean conditions is 49.22%. These speech recognition tests are included to provide an indication of intelligibility and not as a proposed method of speaker separation for speech recognition. For this task, effective methods have been developed that operate on the features themselves without reconstructing an audio signal [8]. With no speaker separation (NSS), recognition accuracy falls significantly as

SIR	-10dB	-5dB	0dB	5dB	10dB	20dB
NSS	-7.34	-7.73	-3.30	2.71	8.88	28.84
D=2	6.81	8.82	11.83	15.10	21.50	35.00
D=6	7.17	10.79	12.42	15.07	21.68	33.88
D=12	7.99	9.97	13.18	16.18	21.82	34.95
D=18	8.20	10.23	13.71	17.16	23.83	35.06
D=23	9.70	12.53	14.57	18.67	23.27	35.06
D=27	9.73	12.33	15.92	18.87	**24.59**	35.03
D=30	9.35	13.24	16.16	**19.43**	24.30	34.97
D=50	**10.97**	**13.74**	**16.90**	18.76	24.39	**35.21**

Table 5.6: *Target speaker monophone recognition accuracy (%) at SIRs from -10dB to +20dB for filterbank sizes from 2 to 50 channels.*

SIRs reduce with a sizeable drop observed below 20dB. Applying speaker separation using the visually-derived binary mask improves recognition accuracy for the target speaker over the uncompensated case. Recognition accuracy consistently increases with larger numbers of filterbank channels up to 27, but in most of the cases best recognition accuracy was achieved with 50 channels.

Intelligibility tests were also carried out using the GRID database data. The

set of 200 mixed sentences was used and the tests were carried out at initial SIRs of -10dB, -5dB, 0dB, 5dB, 10dB and 20dB using $D = 23$ channels. For the GRID database, a whole word speech recogniser is used. Each utterance follows a grammar containing six words of the following structure

command→colour→preposition→letter→digit→adverb.

The total number of models including silence is 52 with eight active states per HMM. The results are shown in Table 5.7 where NSS represents when no speaker separation is applied and SS represents when speaker separation using binary masking is applied. The intelligibility results for the GRID database in Table

SIR	-10dB	-5dB	0dB	5dB	10dB	20dB
NSS	34.67	45.25	55.92	72.5	82.92	95.42
SS	18.50	21.00	31.75	61.25	79.33	90.33

Table 5.7: *GRID database target speaker word accuracy (%) at SIRs from -10dB to +20dB for filterbank sizes of 23 channels.*

5.7 show larger decrease in word recognition accuracy at lower SNRs and smaller decrease in word recognition accuracy at higher SNRs. This is because of the reason that larger segments are dominated by the competing speaker at lower SNRs and are discarded. The results in Table 5.6 and Table 5.7 also prove that the word recognition accuracy is much more sensitive to the discarding of spectral segments than the monophone recognition accuracy.

5.5 Summary

This chapter provided an overview of speaker separation using binary masking. Instead of using purely audio information for deriving the binary masks, visual

speech features were proposed to provide the information for the derivation of the binary masks. The experimental results in terms of masks accuracy, extracted speech quality and the intelligibility, confirmed that visual speech features can provide sufficient spectral information that can be used to create binary masks for speaker separation purposes. It is observed that the number of filterbank channels does not affect significantly either the mask estimation accuracy or the output SIRs following speaker separation. However, in terms of speech recognition accuracy and especially the filterbank estimation errors, the method is more sensitive to the number of filterbank channels. The reason for this is that the reduction of estimation errors require an accurate estimate of the filterbank features which is not possible from just 2 channels and as the number of channels are increased, the estimate of the filterbank features becomes more accurate and results in less estimation errors. While the estimation of binary masks only requires a rough estimate of the energy levels of the estimated features and even lower number of filterbank channels also provide good information to estimate the binary masks with reasonable accuracy.

At present the proposed method uses speaker-dependent models, and while this seems typical of single channel speaker separation methods, it would be desirable to have a speaker-independent system. The high levels of speaker variability in the visual domain make this challenging, but methods of speaker adaptation and speaker-independent visual features are currently being investigated [46]. At present the requirement of speaker-specific GMMs is necessary to attain good audio feature estimates as speaker variability is high for visual features [46].

Chapter 6

Exploiting audio and visual information for single-channel speaker separation

Preface

This chapter proposes a method to exploit both audio and visual speech information to extract a target speaker from a mixture of competing speakers. The chapter begins by taking an effective audio-only method of speaker separation, namely the soft mask method, and modifying its operation to allow visual speech information to improve the separation process. The audio input is taken from a single channel and includes the mixture of speakers, and a separate set of visual features is extracted from each speaker. This allows modification of the separation process to include not only the audio speech but also visual speech from each speaker in the mixture. Experimental results are presented that compare the pro-

posed audio-visual speaker separation method with the audio-only method using both speech quality and intelligibility metrics.

6.1 Introduction

This chapter addresses the problem of single-channel speaker separation by using information from both audio and visual sources. Humans are very good at extracting a target speaker from a mixture of interfering speakers. Having two ears is beneficial but humans also exploit other cues such as observing visual speech information from a target speaker. Many audio-only methods of speaker separation have been proposed and have varying levels of success [102],[77],[84],[70]. A smaller number of visual-only methods of speaker separation have also been proposed [39],[40],[33]. However, few approaches have examined whether the audio and visual information can be combined to further improve separation of speakers.

Audio-only speaker separation can be very effective when multiple microphones are used. Techniques such as deconvolution and blind source separation (BSS) make assumptions that the signals in the mixture are independent and exploit the input signals to extract the individual audio sources [68],[102]. Speaker separation from just a single audio channel is substantially more difficult making it necessary to employ knowledge of the way humans perceive speech and to make various assumptions about the speech signals. Most methods exploit the masking property of human speech perception and aim to identify and extract time-frequency regions of the speech mixture that are dominated by the target speaker and mask or attenuate other regions. Binary masking involves determining whether each time-frequency component represents the target speaker or not and is subsequently

retained or removed [82],[48]. Soft masking [77], [84], [104], can be better as uncertainty in the mask is allowed, where rather than retaining or removing a time-frequency component, a fraction of the component is retained, generally in proportion to the local SNR. With both methods a major challenge is to estimate accurately the mask and identify time-frequency components to be retained and those which are to be masked. Many approaches have been employed and these typically operate by grouping time-frequency regions according to various criteria. One of the most effective is computational auditory scene analysis (CASA) which groups regions perceptually, making use of cues such as harmonicity and onset and offset times [102]. Alternative approaches have used statistical approaches whereby dependencies between time-frequency regions are established and used to form the mask [70].

There are substantially fewer visual-only methods of speaker separation. These rely on correlation existing between the visual and audio speech features to provide an estimate of the audio feature given a visual feature [105],[4]. Visually-derived audio feature estimates have been used to form a perceptually motivated filter that can extract a target speaker from the mixture [39]. An alternative method uses visually-derived audio features from both speakers in a mixture to estimate a binary mask that extracts the target speaker from the audio mixture [40]. In other applications visual features have been used to improve hidden Markov model (HMM) decoding of input speech signals where the HMMs provide statistics on the speech to be separated [33].

Some work on using both audio and visual speech information for speaker separation has been reported although this is applied to multiple audio channels rather than to a single channel which is the focus of this work. In [50] a target speaker

is first extracted from a speech mixture using audio BSS. Visual information from speakers is then used to address permutation and scaling ambiguities present after BSS.

This work proposes combining the audio-only soft mask method with visual speech information to improve speaker separation. A review of the soft-mask method of speaker separation is presented in Section 6.2. The combination of this audio only method with the visual speech information is presented in Section 6.3. Section 6.4 explains how the necessary audio features are estimated from visual features. Experimental results in terms of quality and intelligibility are presented in Section 6.5.

6.2 Audio-only speaker separation

In this section a review is presented of the soft mask audio-only method of speaker separation [70]. The experimental results produced by this method have been shown to outperform both audio-only binary masking and audio-only Wiener filtering methods for single-channel speaker separation. Consequently, this method forms the basis for the proposed combined audio-visual method of speaker separation.

In the time-domain, speech from the target speaker, $x_1(n)$, and competing speaker, $x_2(n)$, are assumed to be additive to create the time-domain mixture, $y(n)$.

$$y(n) = x_1(n) + x_2(n) \tag{6.1}$$

From the time-domain signals, short-time log spectral vectors are extracted. By applying Fourier Transform to the above equation we get

$$Y(f) = X_1(f) + X_2(f) \tag{6.2}$$

where $X_1(f)$ is the complex spectrum of speaker 1 and $X_2(f)$ is the complex spectrum of speaker 2. By taking element-wise squared magnitude and assuming that the two signals are uncorrelated [77],[75], it can be written in the power spectrum domain as

$$|Y(f)|^2 = |X_1(f)|^2 + |X_2(f)|^2 \tag{6.3}$$

If $Y^{'}(f)$, $X_1^{'}(f)$ and $X_2^{'}(f)$ represent the logarithm of $|Y(f)|^2$, $|X_1(f)|^2$ and $|X_2(f)|^2$, then Equation 6.3 can be written as

$$Y^{'}(f) = X_1^{'}(f) + X_2^{'}(f) \tag{6.4}$$

Adopting the same notation as in [70], Equation 6.4 can be written as

$$y_d = x_{1d} + x_{2d} \qquad d = 1, ..., D \tag{6.5}$$

where x_{1d}, x_{2d} and y_d are the d^{th} elements in the D=128 dimensional vectors of speaker 1, speaker 2 and the mixture of the two speakers respectively in the log spectral domain. The extraction process of the log spectral vectors was described in detail in Chapter 2.

The soft mask method makes an element-wise mixture-maximisation assump-

tion of the log spectral vectors from the speakers in the mixture [14] and Equation 6.5 can be written as

$$y_d = \max\left(x_{1d}, x_{2d}\right) + e_d \tag{6.6}$$

where e_d is the error in the mix-max approximation.

An MMSE estimate of each element of the target speaker's log spectral vector, $\hat{x}_{1d}$, is made from the conditional expectation given the mixed signal $\boldsymbol{y}$

$$\hat{x}_{1d} = E\left(x_{1d}|\boldsymbol{y}\right) = \int_{x_{1d}} x_{1d} p(x_{1d}|\boldsymbol{y}) dx_{1d} \quad d = 1, ..., D \tag{6.7}$$

The log spectral features of each speaker are modelled using a GMM that comprises I Gaussian subsources for speaker 1, s_1, and J subsources for speaker 2, s_2. Each subsource from the target speaker has a prior probability, $p_{s_1}(s_1 = i|i = 1, 2, \ldots, I)$ and for the competing speaker $p_{s_2}(s_2 = j|j = 1, 2, \ldots, J)$. The subsources are modelled using Gaussian distributions as

$$p_{\boldsymbol{x}_1|s_1}\left(\boldsymbol{x}_1|s_1 = i\right) = \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{x}_1, \mu_{1d}^i, \Sigma_{1d}^i\right) \tag{6.8}$$

$$p_{\boldsymbol{x}_2|s_2}\left(\boldsymbol{x}_2|s_2 = j\right) = \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{x}_2, \mu_{2d}^j, \Sigma_{2d}^j\right) \tag{6.9}$$

where μ_{1d}^i, μ_{2d}^j, Σ_{1d}^i and Σ_{2d}^j are the means and variances of speakers 1 and 2 and subsources i and j respectively. Again, adopting the same notation as in [70], Σ_{1d}^i and Σ_{2d}^j are written as σ_{1d}^{2i} and σ_{2d}^{2j}.

Modelling the subsources allows the MMSE estimate of Equation 6.7 to be

conditioned on each combination of the subsources, i and j.

$$\hat{x}_{1d} = \sum_{i,j} \underbrace{\int_{x_{1d}} x_{1d} p(x_{1d}|\boldsymbol{y}, s_1 = i, s_2 = j) d_{x_{1d}}}_{Factor I} \times \underbrace{p(s_1 = i, s_2 = j|\boldsymbol{y})}_{Factor II} \quad (6.10)$$

This comprises of two factors. Factor I is an MMSE estimate of x_{1d} given $\boldsymbol{y}$ for a particular combination, i and j, of the subsources. The second factor, Factor II, is the posterior probability of the two subsources given $\boldsymbol{y}$. This can be viewed as a weighted summation, according to the probability of each pair of subsources, of the conditional estimate of x_{1d} from $\boldsymbol{y}$ according to the subsources i and j which, following [70] is evaluated as

$$\hat{x}_{1d} = \sum_{i,j} p(s_1 = i, s_2 = j|\boldsymbol{y}) \times \begin{cases} \frac{\sigma_{1d}^{2i}}{\sigma_{1d}^{2i}+\sigma_d^2} y_d + \frac{\sigma_d^2}{\sigma_{1d}^{2i}+\sigma_d^2} \mu_{1d}^i & \text{if } \mu_{1d}^i \geq \mu_{2d}^j \\ \mu_{1d}^i & \text{if } \mu_{1d}^i < \mu_{2d}^j \end{cases} \quad (6.11)$$

where σ_{1d}^{2i} is the variances of speakers 1 for subsource i and σ_d^2 is the variance of the mixture. The variance of the mixture is calculated across the training data set. Following from Equation 6.6

$$e_d = y_d - \max(x_{1d}, x_{2d}) \quad (6.12)$$

If T represent the total number of frames in the training data set then

$$e_{d,T} = \sum_{t=1}^{T} (y_{d,t} - \max(x_{1d,t}, x_{2d,t})) \quad (6.13)$$

$$e_{d,T}^2 = \sum_{t=1}^{T} (y_{d,t} - \max(x_{1d,t}, x_{2d,t}))^2 \quad (6.14)$$

The variance of the mixture σ_d^2 is given as (note the subscript T is dropped for simplicity purpose)

$$\sigma_d^2 = \mu_{e_{d,T}^2} - (\mu_{e_{d,T}})^2 \tag{6.15}$$

where $\mu_{e_{d,T}^2}$ and $\mu_{e_{d,T}}$ represent the means of $e_{d,T}^2$ and $e_{d,T}$.

For the reduction of computational complexity, it was further shown in [70] following [26] that instead of using the weighted summation of all the subsources, the MMSE estimate can instead be made from the two most probable subsources that maximize $p(s_1 = i, s_2 = j|\boldsymbol{y})$ and is computed as

$$\hat{x}_{1d} = \begin{cases} \frac{\sigma_{1d}^{2i^*}}{\sigma_{1d}^{2i^*}+\sigma_d^2} y_d + \frac{\sigma_d^2}{\sigma_{1d}^{2i^*}+\sigma_d^2} \mu_{1d}^{i^*} & \text{if } \mu_{1d}^{i^*} \geq \mu_{2d}^{j^*} \\ \mu_{1d}^{i^*} & \text{if } \mu_{1d}^{i^*} < \mu_{2d}^{j^*} \end{cases} \tag{6.16}$$

where i^* and j^* are representing the two most probable subsources that maximize $p(s_1 = i, s_2 = j|\boldsymbol{y})$.

$$\{i^*, j^*\} = \arg\max_{i,j} p(s_1 = i, s_2 = j|\boldsymbol{y}) \tag{6.17}$$

It was further shown in [70] that the most probable subsources can be determined as

$$\begin{aligned} \{i^*, j^*\} \quad &= \arg\min_{i,j} \tfrac{1}{2} \textstyle\sum_d \frac{(y_d - max(\mu_{1d}^i, \mu_{2d}^j))^2}{\sigma_{dmax}^2} \\ &+ \; log\sigma_{dmax} - logp(s_1 = i) - logp(s_2 = j) \end{aligned} \tag{6.18}$$

Thus Equation 6.16, in conjunction with Equation 6.18 is used to estimate $\hat{x}_{1d}$.

6.2.1 Relation to binary masking

The conditional estimate is computed in two ways depending on whether the mean component of the target speaker from the i^*th subsource, μ^{i*}_{1d}, is greater or less than the mean of the competing speaker from the j^{*th} subsource, μ^{j*}_{2d}. If the target mean, μ^{i*}_{1d}, is greater than the competing mean, μ^{j*}_{2d}, then the soft mask method assumes that the target speaker can be extracted from the mixed audio. But when target mean is less than the competing mean, the soft mask method assumes that no information about the target can be obtained from the audio mixture.

This can be likened to binary masking which would set the output to zero when the target mean is less than than the competing mean. However, with this soft mask method the output is set to the mean of the target rather than zero. Similarly, when the target mean is greater than the competing mean in binary masking the output is set to y_d. Instead, with this soft masking method, the output is set to a weighted combination of y_d and the target mean using a variant of a Wiener filter. In the same way an estimate of the competing speaker (speaker 2) can be made as

$$\hat{x}_{2d} = \begin{cases} \frac{\sigma^{2j^*}_{2d}}{\sigma^{2j^*}_{2d}+\sigma^2_d} y_d + \frac{\sigma^2_d}{\sigma^{2j^*}_{2d}+\sigma^2_d}\mu^{j^*}_{2d} & \text{if } \mu^{i^*}_{1d} \leq \mu^{j^*}_{2d} \\ \mu^{j^*}_{2d} & \text{if } \mu^{i^*}_{1d} > \mu^{j^*}_{2d} \end{cases} \tag{6.19}$$

The method using Equation 6.16 and Equation 6.19 for the estimation of the speakers is referred to as 'audio-only' (A-only) method in the rest of the work.

6.3 Audio-visual speaker separation

The audio-only soft mask method presented in Section 6.2, can be extended to utilise visual information with the aim of improving the estimation of the speakers' spectral components, $\hat{x}_{1d}$ and $\hat{x}_{2d}$, from the mixture. Following on from Equations 6.16 and 6.19, the visual information is introduced in three different ways.

1. The visual speech information is introduced only when the target mean component is less than the competing speaker mean component (i.e. $\mu^{i*}_{1d} < \mu^{j*}_{2d}$). In this case, a weighting term, α is also introduced to control the contribution of visual and audio speech information in the estimation process. This variation of the soft mask method is to be referred as 'audio-visual-Alpha' (AV-Alpha) method in the rest of the work.

2. In addition to the variation discussed for AV-Alpha method, visual information and a weighting term β, are introduced in the situation when the target mean component is greater than the competing speaker mean component (i.e. $\mu^{i*}_{1d} \geq \mu^{j*}_{2d}$). This second variation of the soft mask method is to be referred as 'audio-visual-Beta' (AV-Beta) method in the rest of the work.

3. A visually derived Wiener filter as described in Chapter 4, is introduced along with the weighting term α, in the situation when the target mean component is less than the competing speaker mean component (i.e. $\mu^{i*}_{1d} < \mu^{j*}_{2d}$). This variation is to be referred as 'audio-visual visual Wiener' (AV-VW) method in the rest of the work.

These three methods are discussed in detail in the next sections.

6.3.1 AV-Alpha

In Equation 6.16, of the A-only method, the target speaker estimate $\hat{x}_{1d}$, is determined differently for the two conditions: When the target mean is less that the competing mean and when the target mean is greater than the competing mean. In the same way, the introduction of the visual information in the two conditional parts of the AV-Alpha method, is discussed separately.

Target mean less than competing mean : $\mu^{i*}_{1d} < \mu^{j*}_{2d}$

In binary masking when the target mean is less than the competing mean ($\mu^{i*}_{1d} < \mu^{j*}_{2d}$), it is assumed that no information about the target can be obtained from the audio mixture and the estimate is set to zero. The A-only soft mask method, improves on this limitation of binary masking by setting the estimate equal to the target mean, μ^{i*}_{1d} as shown in Equation 6.16.

In the condition when the target mean is less than the competing mean ($\mu^{i*}_{1d} < \mu^{j*}_{2d}$), although the A-only method gives an improvement over binary masking by setting the target estimate equal to the target mean, μ^{i*}_{1d}, instead of setting it equal to zero. But again this has drawback as the observed mixed signal is not used/filtered directly to determine the estimate of the target instead the estimate of the target is set equal to the target mean, μ^{i*}_{1d}, which is determined from the trained GMMs using the observed mixed signal for determining the most likely subsource. Thus the observed mixed signal is contributing to the final target estimate only indirectly by contributing in the determination of the most likely

subsource.

In AV-Alpha method, the visual information is introduced in the condition when the target mean is less than the competing mean ($\mu^{i*}_{1d} < \mu^{j*}_{2d}$). The aim is to improve on the drawback of the A-only method by using the multi-modal observed mixed signal in making the final target estimate. This is achieved by modifying the A-only method of Equation 6.16, by making the estimate a weighted combination of the target mean μ^{i*}_{1d}, and an estimate of the target audio in log spectral domain, $\hat{a}_{1d}$, that is derived from a visual speech feature, v_{1d}, extracted from the video of the target speaker's mouth. Thus in the AV-Alpha method, the audio-visual mixed signal is contributing in the following two ways to the final estimate of the target speaker.

1. The mixed audio only signal is used in determining the most likely subsource, whose means μ^{i*}_{1d}, is to be used in the calculation of the target estimate.

2. The corresponding visual signal v_{1d}, for the target speaker in the mixture, is used in determining the audio log spectral estimate of the target speaker, $\hat{a}_{1d}$.

The final target estimate is given by Equation 6.20

$$\hat{x}_{1d} = \begin{cases} \frac{\sigma^{2i^*}_{1d}}{\sigma^{2i^*}_{1d}+\sigma^2_d} y_d + \frac{\sigma^2_d}{\sigma^{2i^*}_{1d}+\sigma^2_d}\mu^{i^*}_{1d} & \text{if } \mu^{i^*}_{1d} \geq \mu^{j^*}_{2d} \\ \alpha\mu^{i^*}_{1d} + (1-\alpha)\hat{a}_{1d} & \text{if } \mu^{i^*}_{1d} < \mu^{j^*}_{2d} \end{cases} \tag{6.20}$$

The weighting term, α, adjusts the contributions made by the target mean and the visual component in the estimate, $\hat{x}_{1d}$. The procedure for obtaining the audio estimate $\hat{a}_{1d}$, from visual speech feature, v_{1d}, extracted from the videos of the speaker's mouth region, is explained in section 6.4.

Target mean greater than competing mean : $\mu^{i*}_{1d} \geq \mu^{j*}_{2d}$

In binary masking when the target mean is greater than the competing mean ($\mu^{i*}_{1d} \geq \mu^{j*}_{2d}$), the estimate is set equal to the input mixture, y_d, without processing it. The audio-only soft mask method again improves on this and the estimate of the target is made from a Wiener-type weighting of the target mean and input mixture of speakers, y_d as shown in Equation 6.16. In the AV-Alpha method, it is assumed that when the target mean is greater than the competing mean, then the information contained in the audio component is sufficient for the estimation of the target speaker and the visual speech information is not needed. Hence no modifications are introduced in this conditional part of the of the A-only method.

6.3.2 AV-Beta

The introduction of the visual information in the two conditional parts of the AV-Beta method, is discussed separately in the next two sections.

Target mean less than competing mean : $\mu^{i*}_{1d} < \mu^{j*}_{2d}$

In AV-Alpha method of Equation 6.20, for the condition when the target mean is less than the competing mean, the target estimate was determined using the target mean and the estimate $\hat{a}_{1d}$, from the visual features and their contribution was set by the weighting term α. In the same way for AV-Beta method, this conditional part is kept the same as in AV-Alpha method with the only difference that the optimal constant value of α is used that was determined for AV-Alpha method.

Target mean greater than competing mean : $\mu^{i*}_{1d} \geq \mu^{j*}_{2d}$

When the target mean is greater than the competing mean, then in the A-only method of Equation 6.16, the estimate of the target is made from a Wiener-type weighting of the target mean and input mixture of speakers, y_d. In the AV-Alpha method, it was assumed that when the target mean is greater than the competing mean, then the information contained in the audio component is sufficient for the estimation of the target speaker and the visual speech information is not needed. Hence no modifications were introduced in this conditional part of the of the AV-Alpha method. In AV-Beta method, the visually-derived estimate of the target, $\hat{a}_{1d}$, is introduced in this conditional part also using another weighting term β as shown in Equation 6.21

$$\hat{x}_{1d} = \begin{cases} \beta\left(\frac{\sigma^{2i^*}_{1d}}{\sigma^{2i^*}_{1d}+\sigma^2_d} y_d + \frac{\sigma^2_d}{\sigma^{2i^*}_{1d}+\sigma^2_d}\mu^{i^*}_{1d}\right) + (1-\beta)\hat{a}_{1d} & \text{if } \mu^{i^*}_{1d} \geq \mu^{j^*}_{2d} \\ \alpha\mu^{i^*}_{1d} + (1-\alpha)\hat{a}_{1d} & \text{if } \mu^{i^*}_{1d} < \mu^{j^*}_{2d} \end{cases} \tag{6.21}$$

6.3.3 AV-VW

The introduction of the visual information in the two conditional parts of the AV-VW method, is discussed in the next two sections.

Target mean less than competing mean : $\mu^{i*}_{1d} < \mu^{j*}_{2d}$

The previous three methods: A-only, AV-Alpha and AV-Beta, do not use the observed mixed signal directly in the final estimate of the target speaker. Instead, the observed audio mixture, y_d, is used to identify the most likely subsource whose mean is used in the calculation of the final estimate of the target speaker or the

corresponding visual information is used to give a visually derived estimate, $\hat{a}_{1d}$, that is used in the final estimate of the target,$\hat{x}_{1d}$, through weighting terms α and β. In AV-VW method, the observed mixture is directly used in the final estimate of the target speaker. The A-only method of the Equation 6.16, is modified by making the estimate a weighted combination of the target mean and an estimate of the target from the mixture using visually derived Wiener filter described in Chapter 4 and the final estimate of the target is given as in Equation 6.22

$$\hat{x}_{1d} = \begin{cases} \frac{\sigma_{1d}^{2i^*}}{\sigma_{1d}^{2i^*}+\sigma_d^2} y_d + \frac{\sigma_d^2}{\sigma_{1d}^{2i^*}+\sigma_d^2}\mu_{1d}^{i^*} & \text{if } \mu_{1d}^{i^*} \geq \mu_{2d}^{j^*} \\ \alpha\mu_{1d}^{i^*} + (1-\alpha)W_{1d}\ y_d & \text{if } \mu_{1d}^{i^*} < \mu_{2d}^{j^*} \end{cases} \tag{6.22}$$

where W_{1d}, is defined as

$$W_{1d} \quad = \quad \frac{\hat{a}_{1d}}{\hat{a}_{1d} + \hat{a}_{2d}} \tag{6.23}$$

where $\hat{a}_{1d}$ and $\hat{a}_{2d}$ are the log spectral estimates for the target speaker 1 and competing speaker 2 from their corresponding visual features. The estimation of the audio features from visual features is explained in section 6.4.

Target mean greater than competing mean : $\mu^{i*}_{1d} \geq \mu^{j*}_{2d}$

When the target mean is greater than the competing mean in Equation 6.22, the estimate of the target is made from a Wiener-type weighting of the target mean and input mixture of speakers, y_d, as in the A-only method where no visual information is introduced in this conditional part. This decision of not introducing the visual information in this conditional part is based on the findings of the AV-Beta method. Hence in the AV-VW method, it is concluded that when the target

mean is greater than the competing mean, then the information contained in the audio component is sufficient for the estimation of the target speaker and the visual speech information is not needed. Hence no modifications are introduced in this conditional part of the of the AV-VW method as shown in Equation 6.22.

6.4 Estimation of audio features from video

$D = 128$ channel log spectral vectors, $\mathbf{x}_1$ and $\mathbf{x}_2$ are used as the audio features for speaker 1 and speaker 2 respectively. These are extracted from 20ms duration frames of audio at 10ms intervals in accordance with the ETSI XAFE standard [28]. Visual features, $\mathbf{v}_1$ and $\mathbf{v}_2$, for speaker 1 and speaker 2 respectively, are extracted from an ROI centred on a speaker's mouth at a rate of 100 frames per second. A 2D-DCT is applied and the first J coefficients are scanned in a zigzag manner and retained as the visual vector. The estimation process involves first training a K-cluster GMM to model the joint density of augmented audio-visual feature vectors for each speaker. MAP estimation can then be applied to estimate the audio features, $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$, from the visual features, $\mathbf{v}_1$ and $\mathbf{v}_2$. For a detailed discussion of the audio and visual features, the training and the estimation processes, please refer to Chapter 2 and Chapter 3.

6.5 Experimental results

The performance of audio-visual speaker separation for the proposed methods, is evaluated in this section. First, the experimental set up is described. Second, the speaker separation results in terms of quality and intelligibility, are presented,

compared and discussed, for the different methods. Three metrics; SIR, SDR and SAR, are used to measure the speech quality while the intelligibility is measured in terms of word accuracy, of the separated speech. To measure the intelligibility a whole word speech recogniser trained on the GRID database [15] is used. In GRID database, each utterance follows a grammar containing six words of the structure

command→*colour*→*preposition*→*letter*→*digit*→*adverb*.

From the estimates of the target speaker's speech, MFCC vectors were extracted and the resulting word accuracy used as an estimate of intelligibility. It should be noted that these recognition tests are used to provide an indication of intelligibility. The methods presented in this work are not the proposed methods of speaker separation for speech recognition. For this task, effective methods have been developed that operate on the features themselves without the need to reconstruct an audio signal [8].

6.5.1 Experimental set up

The GRID audio-visual speech database is used in these experiments [15]. A male speaker (speaker 6) is used as the target and a female speaker (speaker 4) as the competing speaker. Of the 1000 utterances spoken by each speaker, 800 are used for training and the remaining 200 for testing. The audio for both the speakers was down-sampled to a sampling frequency of 8KHz and log spectral vectors extracted at 10ms intervals. The video was up-sampled to 100 frames per second to match the audio frame rate. For both speakers, 2D-DCT visual features were captured from the mouth region centred on the speaker's mouth. The extraction process of

the audio and visual features, and the GRID database were described in detail in Chapter 2.

The test scenario assumes that the two speakers are talking simultaneously and are located close together. Video is captured from each speaker with a separate camera. The mixed audio is created by taking speech from the target speaker and mixing it with the speech from the competing speaker that is scaled to create the desired SNR levels of -20dB, -10dB, -5dB, 0dB, 5dB, 10dB and 20dB. For the tests reported, the male speaker is the target and the female the competing speaker. The 200 test utterances from the male speaker were mixed with the 200 utterances from the female speaker with the restriction that no mixture used the same two sentences. Similar results were obtained when considering the female as the target and the male as the competing speaker.

6.5.2 AV-Alpha

This section examines the results for AV-Alpha method where visual information is introduced into the A-only soft mask method when the target mean is less than the interfering mean as described by Equation 6.20. The variable α controls the ratio of target mean, μ^{i*}_{1d}, to the visually derived log spectral estimate, $\hat{a}_{1d}$. When $\alpha = 1$, no visual information is used and so the estimate is purely the A-only soft mask result. While when $\alpha = 0$ the output is purely the visual estimate in the conditional part when when the target mean is less than the interfering mean.

Figure 6.1, shows the SIR variations when varying α from 0 to 1 for different SNRs of -20dB to +20dB. For the lower SNRs (-20dB to -5dB), SIR peaks when most of the contributions are made by the visual components i.e. 80%. As the SNR

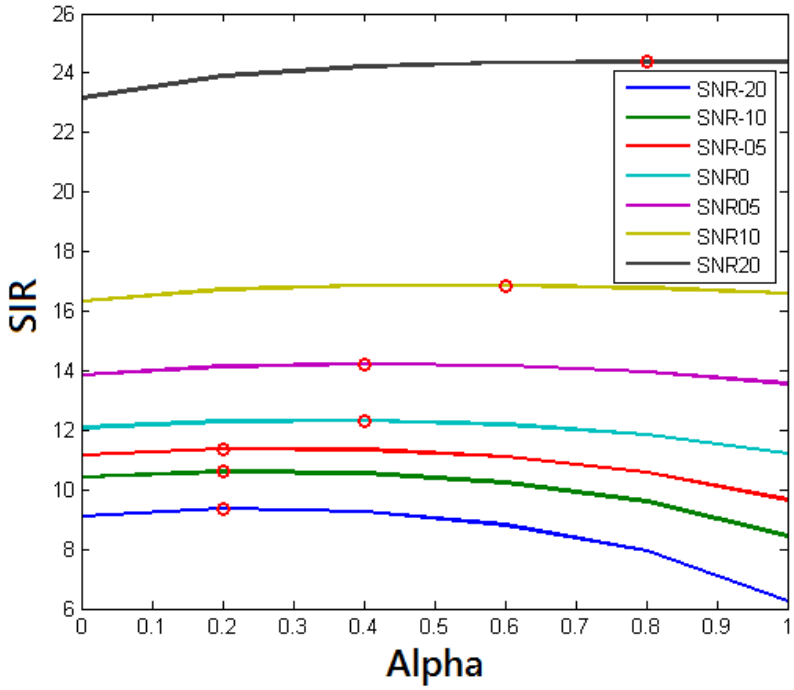


Figure 6.1: *SIR when varying α from 0 to 1 in Equation 6.20. The small red circles are showing the peak values.*

increases, the SIR gains are increasing when decreasing the visual contributions although at 0dB and +5dB the visual contribution for the maximum SIRs are 60% i.e $\alpha = 0.4$. The figure is showing that AV-Alpha method is giving significant gains in SIR over the A-only method i.e. when α is 1.

Figure 6.2, shows the SDR variations when varying α from 0 to 1. These variations also follow the same trends as in SIR's case i.e. as the SNR increases the visual contributions need to be decreased to obtain maximum gains in SDR. Contrary to the SIR gains, for SDR gains, the audio and visual contributions are almost balanced from -20dB to 0dB, while in case of SIR, most of the contributions were made by the visual component in this region. At higher SNRs of +10dB and +20dB, the visual component's contributions towards SDR gains are zero.

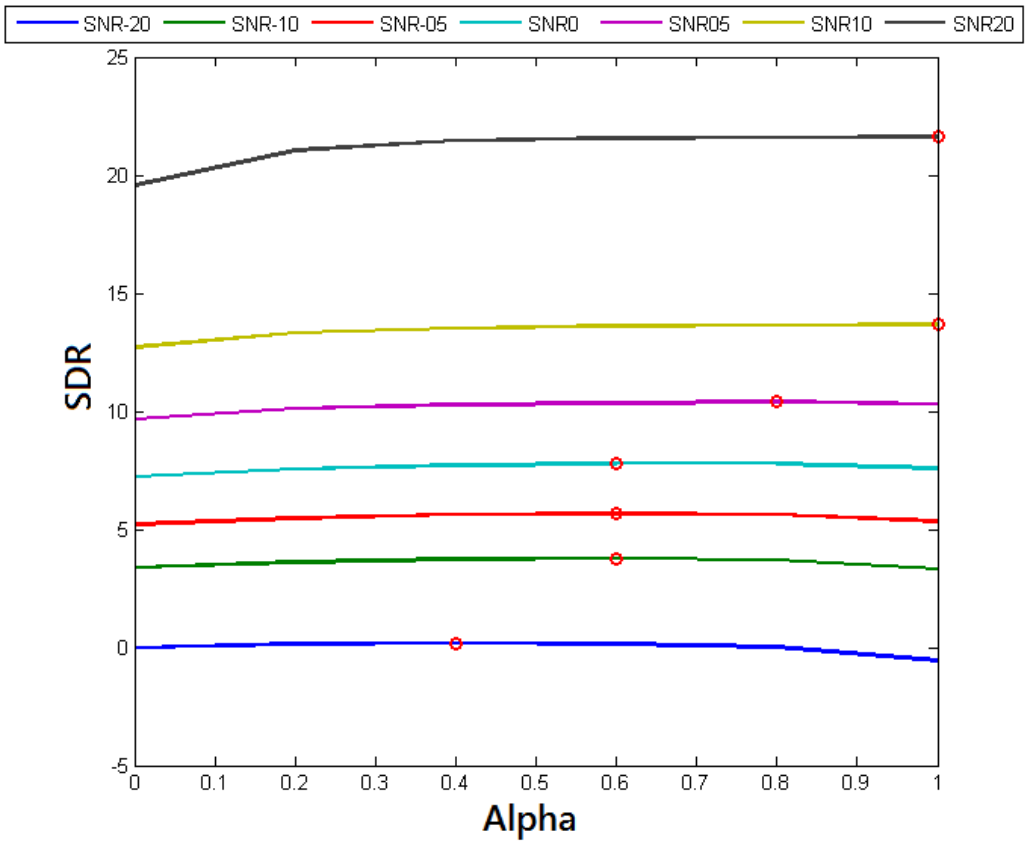


Figure 6.2: *SDR when varying α from 0 to 1 in Equation 6.20. The small red circles are showing the peak values.*

Hence it is concluded that as the target becomes more dominant, the information contained in the audio speech component becomes more useful than the the visual speech component.

Figure 6.3, shows the SAR variations when varying α from 0 to 1. The SAR variations are very flat and are not significant. This suggests that for the various values of α, the introduced algorithmic distortions are almost of the same level.

The SIR and SDR gains are a trade-off, i.e. larger SIR gains are obtained at the cost of reduction in SDR gains. Therefore, the selected optimal value of α

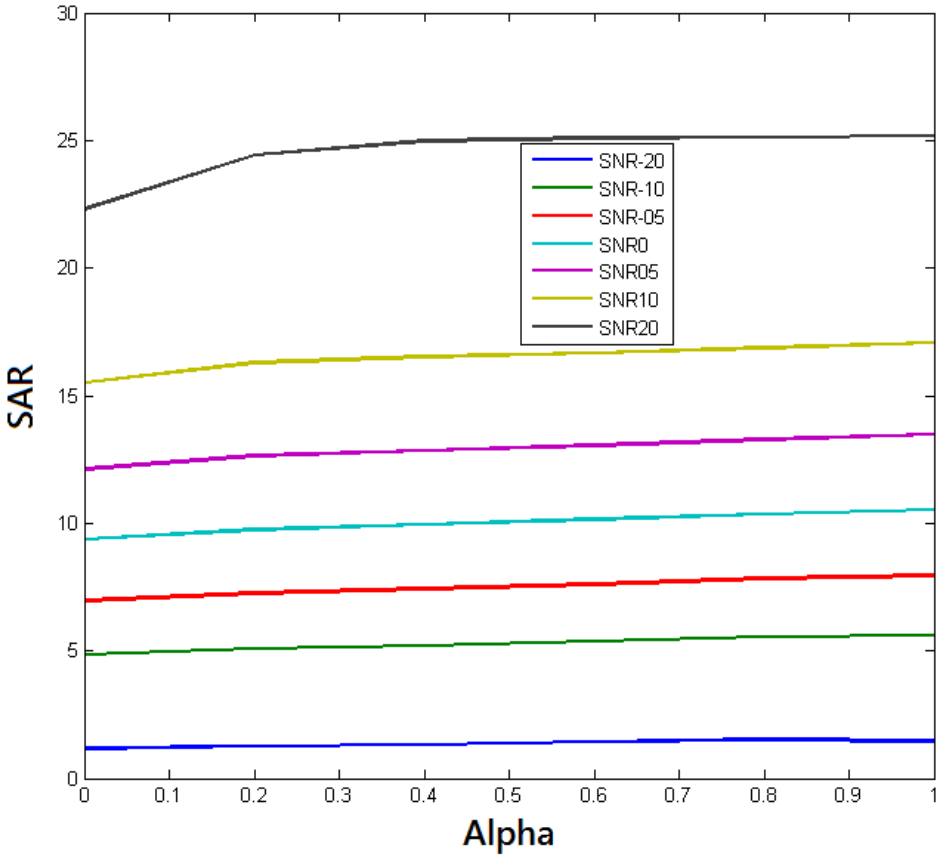


Figure 6.3: *SAR when varying α from 0 to 1 in Equation 6.20.*

should keep a balance between SIR and SDR gains.

6.5.3 AV-Beta

This section examines the results for the AV-Beta method that introduces visual information in both the conditional parts of the A-only soft mask method as described in Equation 6.21. The effect of varying the visual contributions was investigated at an SNR of 0dB. In these tests the optimal value of α was used and

α was not varied when varying β. Figure 6.4 shows the SIR, SDR and recognition accuracy when varying the visual contribution, β, from 0 to 1.

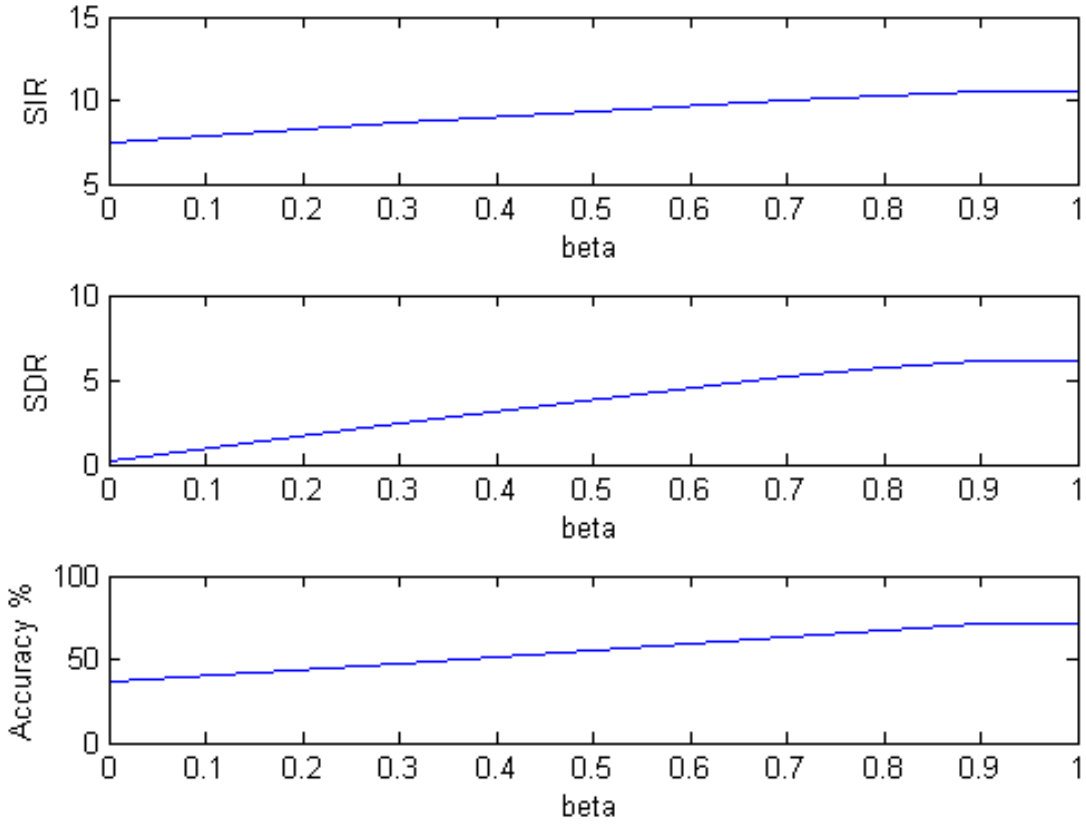


Figure 6.4: *SIR, SDR and recognition accuracy when* $\alpha = 0.35$ *and varying* β *from 0 to 1 in Equation 6.21.*

The results suggest that in situations when the target mean is less than the competing mean, the value α=0.35 is used that was determined to be the optimal in the previous experiments for the AV-Alpha method. For reference, the performance at $\beta = 1$ corresponds to the situation when no visual information is included in the estimate when target mean is greater than the competing mean and the target spectral estimate is made from audio only which is the original soft mask. At this point ($\beta = 1$ and $\alpha = 0.35$), performance is equal to the best

obtained for the AV-Alpha method. As β reduces, the visual information makes more contribution to the estimate. For SIR, SDR and recognition accuracy as more visual information is included, and thereby audio information reduced, performance falls. All three metrics reach minimum levels when the target estimate is based only on visual information, i.e. $\beta = 0$. Therefore an optimal value of β is one. Hence it is concluded that the introduction of β does not give any improvements in terms of quality or intelligibility and is dropped from any further investigation in this work. This suggests that in times when the target speaker is dominant then the audio information is more useful than the visual information.

6.5.4 AV-VW

This section examines the results for AV-VW method that introduces visual information only in the conditional part when the target mean is less than the competing mean as described in Equation 6.22. The introduction of visual information in the form of β, in the conditional part when the target mean is greater than the competing mean, did not give any improvements in terms of quality and/or intelligibility, as was shown in the results for AV-Beta method. Therefore, in AV-VW method, no visual information is introduced in the conditional part when the target mean is greater than the competing mean. The previously discussed methods: A-only, AV-Alpha and AV-Beta, none of these uses/filters directly the observed mixed audio speech y_d, to estimate the target speaker, $\hat{x}_{1d}$, in the conditional part when target mean is less than the competing mean. The AV-VW method, filters directly the mixed audio speech using the visually derived Wiener filter as shown in Equation 6.22.

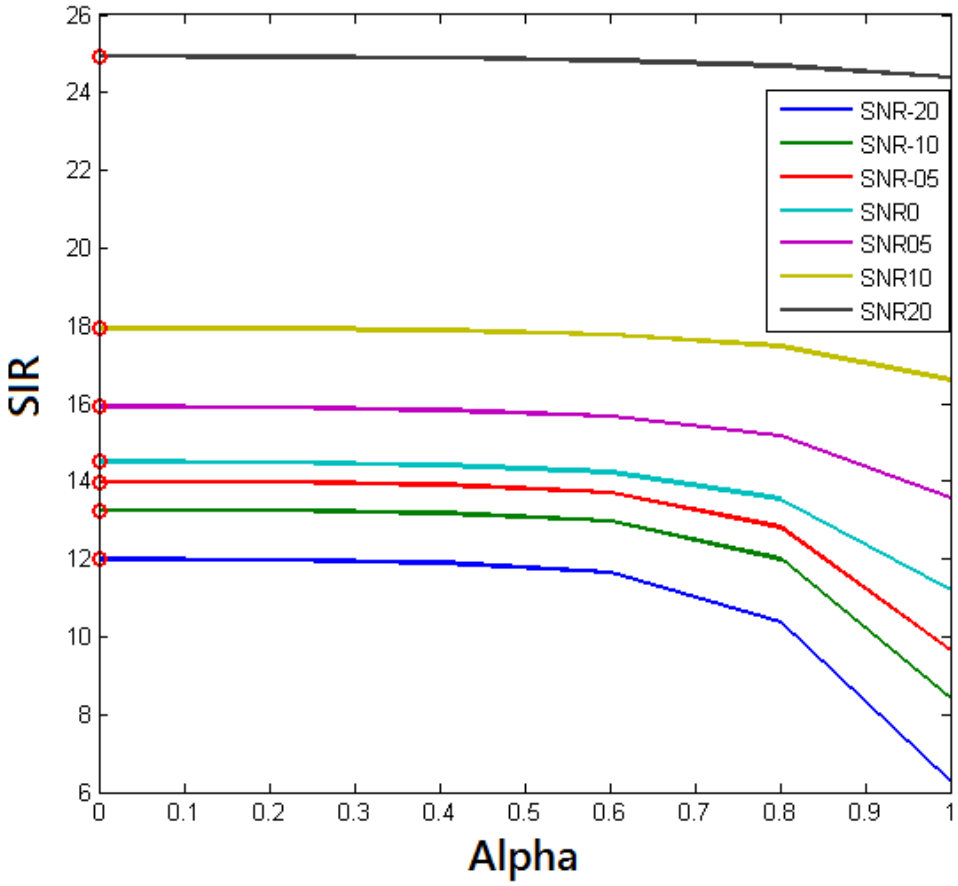


Figure 6.5: *SIR when varying α from 0 to 1 in Equation 6.22. The small red circles are showing the peak values.*

Figure 6.5, shows the SIR variations when varying α from 0 to 1 in Equation 6.22, for different SNRs of -20dB to +20dB. For all the SNRs (-20dB to +20dB), SIR peaks when all the contributions are made by the directly filtered observed mixed audio using the visually derived Wiener filter, W_{1d}, without taking any contributions from the target mean. The SIR gains keep on decreasing slightly till $\alpha = 0.6$, i.e. when the contributions of the visually derived Wiener filter are from 100% till 40%, but beyond this point, the decrease in SIR gains is rapid and drops

to the lowest when $\alpha = 1$, i.e. when the contributions of the visual Wiener filter are zero. These results show that in terms of the SIR gains, the direct filtering of the mixed audio speech using the visually derived Wiener filter, is more useful than using the target mean of the most likely subsource. The figure is also showing that AV-VW method is giving significant gains in SIR over the A-only method i.e. when $\alpha = 1$.

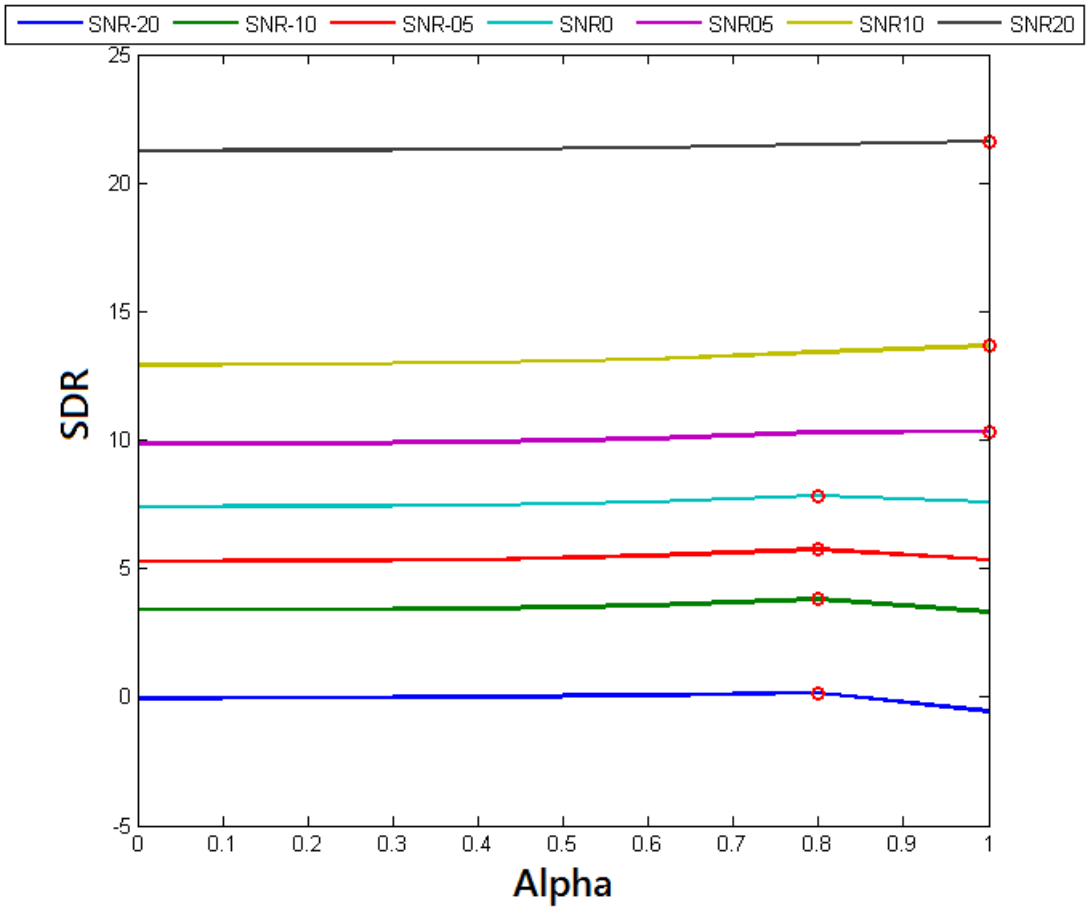


Figure 6.6: *SDR when varying α from 0 to 1 in Equation 6.22. The small red circles are showing the peak values.*

Figure 6.6, shows the SDR variations when varying α from 0 to 1. The SDR

variations are very flat and are not significant although the SDR very slightly peaks when most of the contributions (80%) are made by the target mean rather than the mixed audio filtered through the visually derived Wiener filter. These results show that the visually derived Wiener filter does not give any significant improvements in SDR for the different values of α.

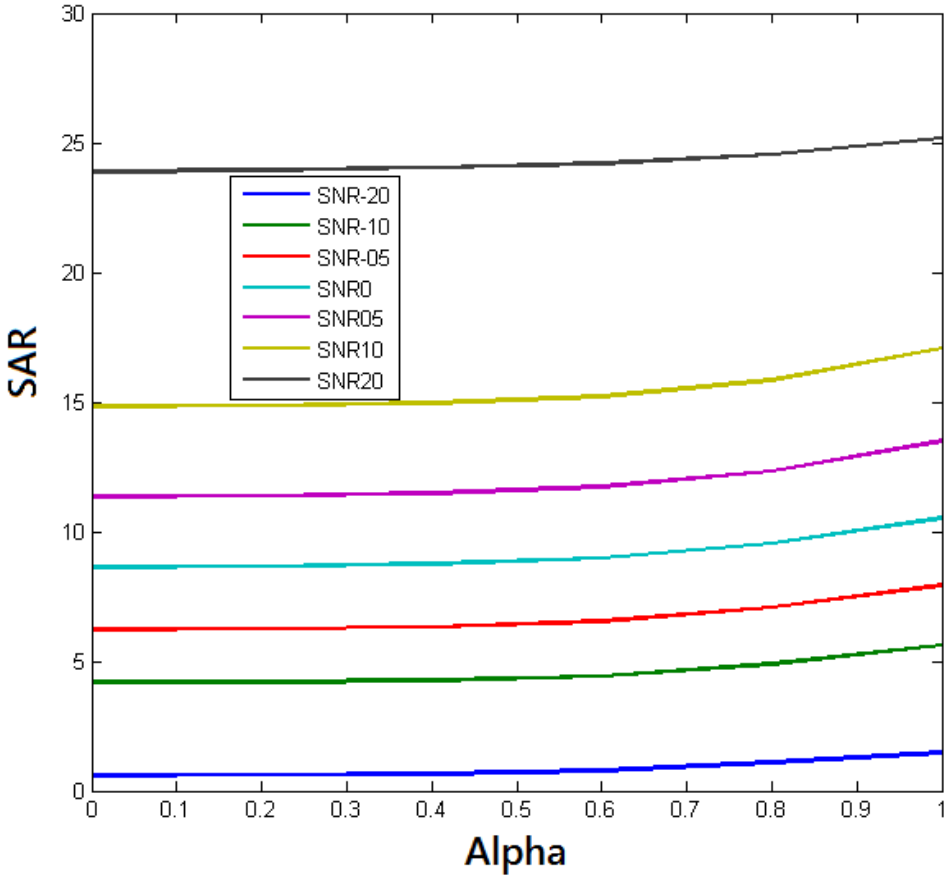


Figure 6.7: *SAR when varying α from 0 to 1 in Equation 6.22.*

Figure 6.7, shows the SAR variations when varying α from 0 to 1. The SAR variations are also not significant although the SAR slightly peaks when no visual

information is used i.e. when $\alpha = 1$. This shows that the introduction of visual information introduces a smaller amount of algorithmic distortion as compared to the A-only processing.

6.5.5 Comparison of A-only, AV-Alpha and AV-VW methods

In this section, a brief comparison is made of the three mentioned methods that were described just above, in terms of estimated speech quality and intelligibility. SIR, SDR and SAR are used as the measures of estimated quality and words recognition accuracy is used as the measure of estimated intelligibility. A detailed comparison of the different methods is to follow in Chapter 7.

Figure 6.8 shows the SIR gains for the three mentioned methods. The AV-VW method gives significant gains in terms of SIR over the A-only soft mask method and AV-Alpha method, at all SNRs.

Figure 6.9 shows the SDR gains for the three mentioned methods. Although the SIR gains of AV-VW are very significant, but in terms of SDR gains, the three methods are performing at almost the same level. In terms of SAR gains, the A-only and AV-Alpha are performing slightly better than the AV-VW method as shown in Figure 6.10

Figure 6.11 shows the word recognition accuracy results for the three mentioned methods. The AV-Alpha method is giving significant gains in recognition accuracy over the A-only and AV-VW methods, in particular, in the lower SNR regions (-20dB to 0dB). The decrease in recognition accuracy of AV-VW method can be compensated by using higher values of α, as in these comparison plots, $\alpha = 0$, for

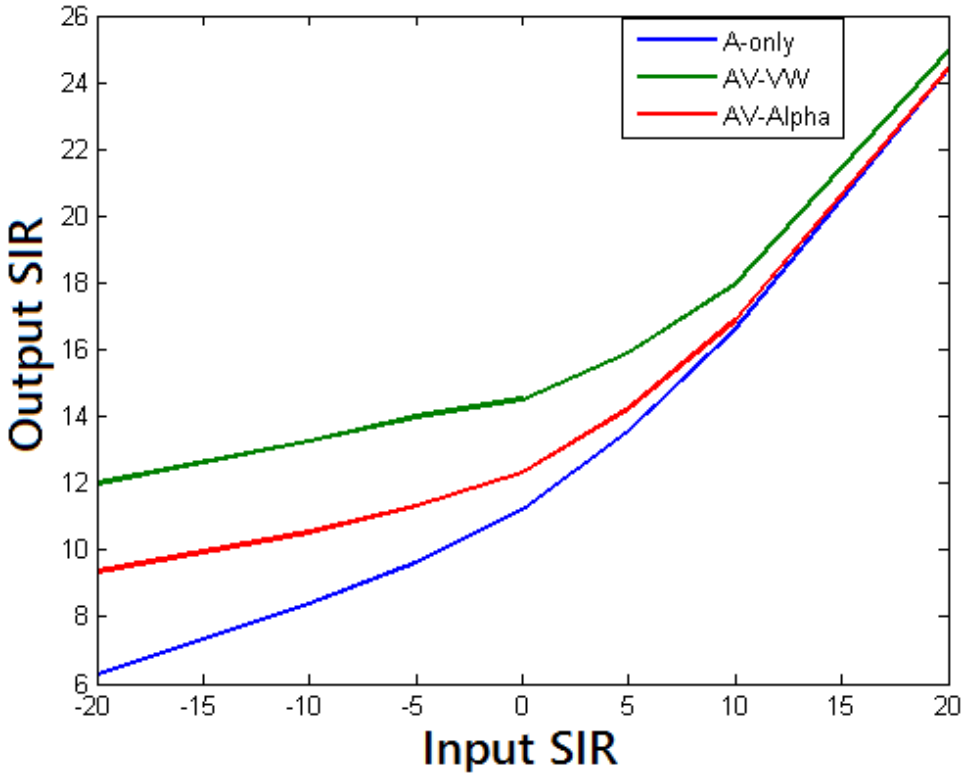


Figure 6.8: *Comparisons of SIR gains for A-only, AV-Alpha and AV-VW methods, for optimal values of α, for SNRs of -20dB to +20dB.*

the AV-VW method, but this will cause a decrease in SIR gains of the method. So depending on the application in hand, whether quality enhancement is more desirable or intelligibility improvements, different values of α can be used.

6.6 Summary

This chapter provided an overview of the A-only soft mask method. This method estimates the target speaker in two different ways for the two conditions. When the target mean is greater than the competing mean, the soft mask method is

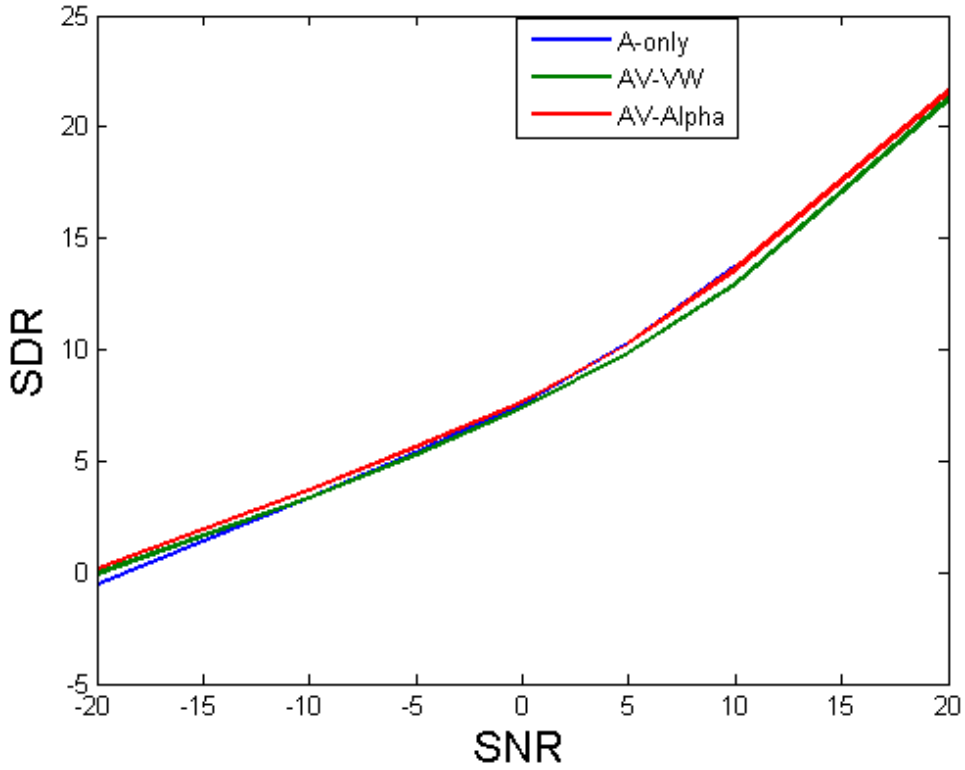


Figure 6.9: *Comparisons of SDR gains for A-only, AV-Alpha and AV-VW methods, for optimal values of α, for SNRs of -20dB to +20dB.*

confident to extract the target speaker from the observed mixture using a Wiener type filter of the mixed speech and the most likely subsource. When the target mean is less than the competing mean, the soft mask method, assumes that no useful information can be extracted from the mixed speech and the estimate of the target speaker is made from the most likely subsource instead.

To improve the performance of this A-only method, three different methods were introduced based upon the work carried out in the previous chapters. The AV-Alpha method, introduced visual information in the condition when the target mean is less than the competing mean. Significant gains are obtained in SIR

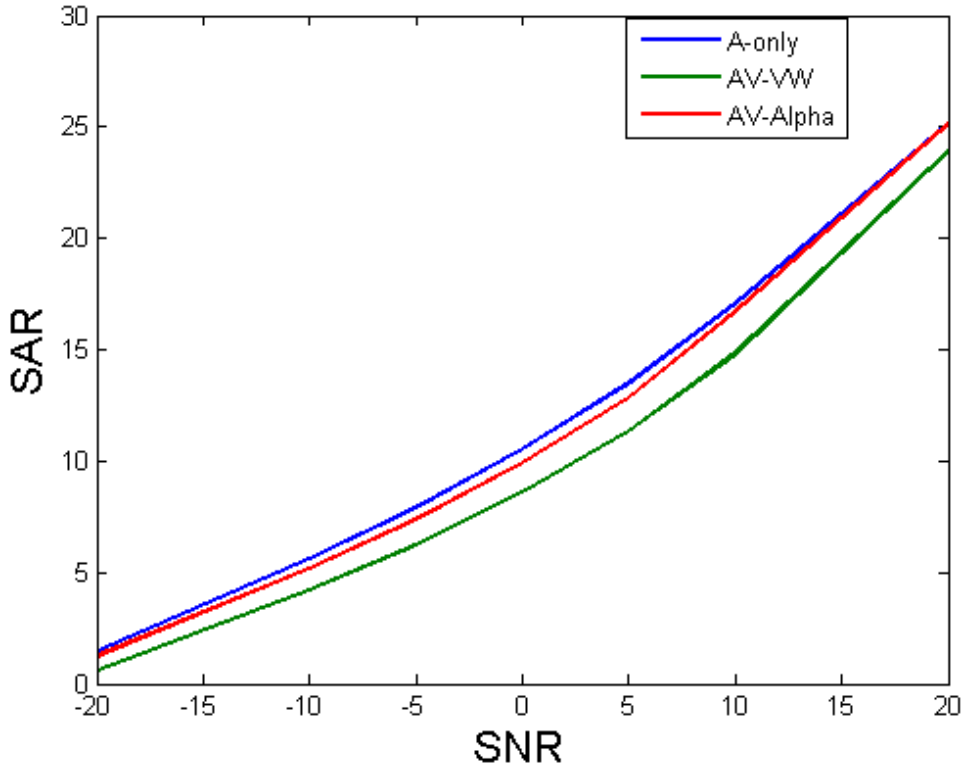


Figure 6.10: *Comparisons of SAR gains for A-only, AV-Alpha and AV-VW methods, for optimal values of α, for SNRs of -20dB to +20dB.*

and SDR over the A-only method. While the SAR variations are flat and not significant.

The AV-Beta method introduced the visual information in the condition when the target mean is greater that the competing mean as well. But this caused the quality and intelligibility to drop. This proves the assumption that when the target becomes more dominant, the information contained in the audio speech component becomes more useful than the the visual speech component.

The AV-VW method, introduced the visually derived Wiener filter to filter the observed mixed speech in the condition when the target mean is less than

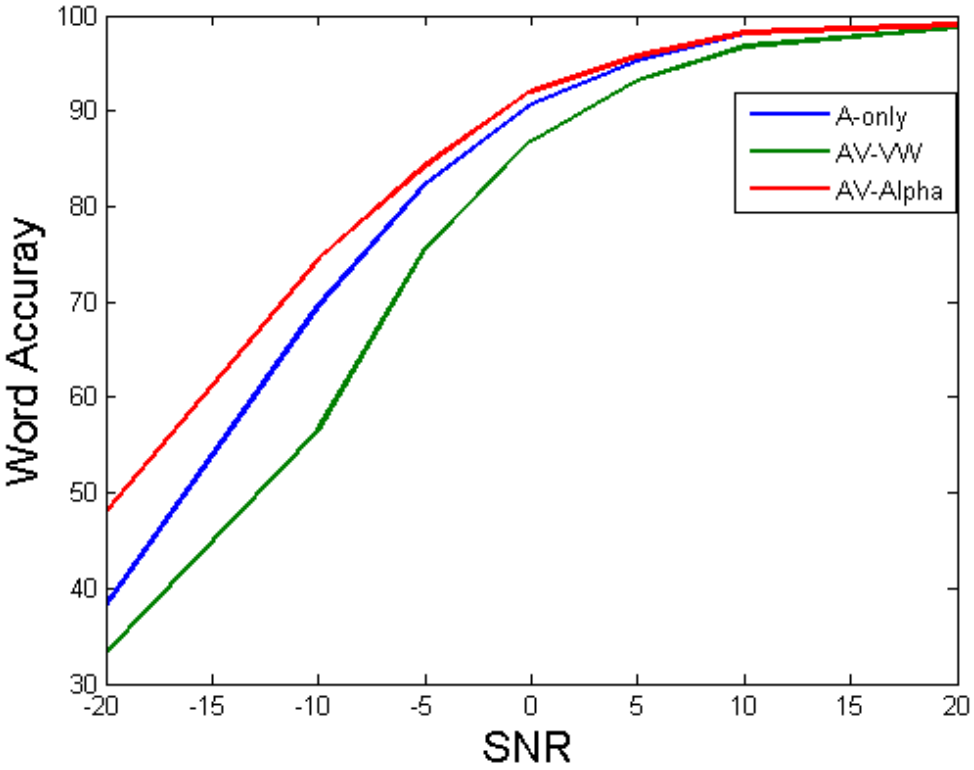


Figure 6.11: *Comparisons of word recognition accuracy (%) for A-only, AV-Alpha and AV-VW methods, for optimal values of α, for SNRs of -20dB to +20dB.*

the competing mean. AV-VW gives significant gains in SIR for higher visual contributions than audio contributions, but the gains in SDR are not significant. Intelligibility decreases at the cost of increase in SIR.

The SIR and SDR gains and intelligibility are trade-offs i.e. larger SIR gains are obtained at the cost of reduction in SDR gains and intelligibility. Therefore, the selected optimal value of α should keep a balance between SIR and SDR gains and intelligibility. So depending on the application in hand, whether quality enhancement is more desirable or intelligibility improvements, different values of α

can be used. Also it is noted that at higher SNRs, the visual component's contributions towards improvement in quality and intelligibility decreases. It is concluded that as the target becomes more dominant, the information contained in the audio speech component becomes more useful than the visual speech component.

Chapter 7

Comparsions of proposed and existing methods

Preface

This chapter presents a comparison of the objective and subjective evaluations of the proposed and existing methods. The objective evaluation methods use the previously used criteria of SIR, SDR and SAR to measure the quality of the extracted speech and *word recognition accuracy* is used as the measure of estimated intelligibility. For subjective evaluation, listening tests are conducted and the subjects are asked to rate the quality of the extracted speech. The three proposed methods and two existing methods were compared along with the reference and unprocessed speech.

7.1 Introduction

The aim of speech processing applications like speech enhancement and speaker separation is to enhance the quality and intelligibility of the processed speech [56]. The perceived quality of speech is the perception of a listener about the speech that how "good" its quality is. The definition of "good" is dependant on the listener [43]. However the natural clean speech in daily life provides a reference point and the listeners rate the quality of any speech in relation to this reference. According to [56], enhancement in speech quality results in reduction of listener fatigue. The accuracy with which listeners hear what is being said to them is called the speech intelligibility and is measured in terms of correctly identified responses [43].

Speech quality and intelligibility are measured using objective and subjective measures. In subjective evaluations, listeners rate the quality of speech according to some reference. While in objective evaluations, a particular physical measure is computed from a reference and a processed speech [43].

This chapter measures the speech quality and intelligibility using both objective and subjective measures. Section 7.1.1 and Section 7.1.2 describes the proposed and existing methods to be compared using objective and subjective evaluation. This is followed by Section 7.2 that explains the audio and visual data used in this chapter, experimental set up for the listening tests and then the actual results of speech quality and intelligibility using objective and subjective measures.

7.1.1 Proposed methods

In this work, three methods for speaker separation were proposed. These were presented in detail in Chapters 4, 5 and 6. Each of these methods have different variants. One variant of each of these methods was selected for the comparison testing based on its performance measured previously. These methods along with the selected variants are described below.

1. **Speaker separation using Wiener filtering and perceptual gain functions:** This method was discussed in detail in Chapter 4 along with its variants in the form of perceptual gain functions $H1$, $H2$, $H3$ and $H4$. $H1$ as shown in Equation 4.12 was selected for final testing because of its higher SDR and SAR gains and higher recognition accuracy.

2. **Speaker separation using visually-derived binary masks:** This method was discussed in detail in Chapter 5 along with the effects of different numbers of filterbank channels on speech quality and intelligibility. This method with variant when number of filterbank channels $D = 23$, was selected for the final testing because it gives a good balance between quality and intelligibility improvements.

3. **Audio-visual method of speaker separation:** This method was discussed in detail in Chapter 6 along with its variants AV-Alpha, AV-Beta and AV-VW. These variants were further studied using different contributions of audio and visual information through α and β. AV-VW method as shown in Equation 6.22 with $\alpha = 0$ was selected for the final testing because of its huge SIR gains and comparable intelligibility scores.

7.1.2 Existing methods

The three above mentioned proposed methods were compared with the following two existing methods.

1. **A-only method:** This audio-only soft-mask method [70] was discussed in detail in Chapter 6. This method gives the final audio estimate of the target speaker as the weighted combination of the Wiener type filtering of the observed mixed speech and the target mean as shown in Equation 6.16 in Chapter 6.

2. **CASA method:** This CASA method [36], uses the traditional CASA approach of segmentation and grouping. The segmentation and grouping for voiced speech is based on fundamental frequencies. While the unvoiced speech segregation is based on onset/offset analysis.

7.2 Experimental Results

In this section, the performance of the proposed and existing methods of speaker separation, is compared. First, the experimental set up is described along with the audio and visual features. Second, the speaker separation results in terms of quality and intelligibility for the different methods are compared and discussed using objective and subjective measures.

7.2.1 Audio and visual data

The GRID audio-visual speech database is used in these experiments [15]. A male speaker (speaker 6) is used as the target and a female speaker (speaker 4) as the

competing speaker. Of the 1000 utterances spoken by each speaker, 800 are used for training and the remaining 200 for testing. The audio for both the speakers was down-sampled to a sampling frequency of 8KHz and log spectral vectors extracted at 10ms intervals. The video was up-sampled to 100 frames per second to match the audio frame rate. For both speakers, 2D-DCT visual features were captured from the mouth region centred on the speaker's mouth. The extraction process of the audio and visual features, and the GRID database were described in detail in chapter 2.

The test scenario assumes that the two speakers are talking simultaneously and are located close together. Video is captured from each speaker with a separate camera. The mixed audio is created by taking speech from the target speaker and mixing it with the speech from the competing speaker that is scaled to create the desired SNR levels of -20dB, -10dB, -5dB, 0dB, 5dB, 10dB and 20dB. For the tests reported, the male speaker is the target and the female the competing speaker. The 200 test utterances from the male speaker were mixed with the 200 utterances from the female speaker with the restriction that no mixture used the same two sentences. Similar results were obtained when considering the female as the target and the male as the competing speaker.

7.2.2 Experimental set up for subjective tests

In the subjective quality assessment of speech quality, 20 human listeners participated. The listening test were carried out in a sound proof room. The listeners used headphones and only the computer screen and mouse were inside the room while rest of the computer was outside to avoid any noise coming from the com-

puter fans. Out of the 200 hundred test utterance, 10 different utterance were chosen for each subject at random. The tests were carried out for SNR levels of -10dB, -5dB, 0dB, 5dB and 10dB. The following guidelines and instructions as in [23], were provided and explained to the subjects. These are copied here as they are in [23] with slight variations and added figures.

"GUIDELINES FOR LISTENING TEST

This listening test aims to rate the quality of a set of signals produced by source separation systems. Source separation aims to extract the signal of a target source from a mixture of several sound sources as shown in Figure 7.1. The resulting signals may include several types of degradations compared to the clean target source, including distortions of the target source and remaining sounds from other sources. The test is in three parts:

Test 1

To rate the quality in terms of the amount of suppression of the interfering (female) speaker.

Test 2

To rate the quality in terms of the preservation of the target (male) speaker.

Test 3

To rate the overall quality of the speech compared to the reference signal.

Each test has a training part and the training GUI is shown in Figure 7.2,

where you can listen to examples of the speech signals. Once comfortable, the actual evaluation takes place.

Each test involves 10 experiments and the evaluation GUI is shown in Figure 7.3. For each experiment you will need to rate the quality of seven test sounds compared to the reference sound and mixture on a scale of 0 to 100. Larger numbers indicate higher quality. You can listen to the sounds as many times as you wish. You should make sure that

- The ratings between pairs of sounds are consistent, i.e. if one sound has better quality than another, it should be rated better.
- The ratings between different experiments are consistent, i.e. if two sounds from different experiments have the same quality, they should be rated equally.
- The whole rating scale is used, i.e. sounds with perfect quality should be rated 100 and the worst test sound over all experiments (but not necessarily the worst test sound in each experiment) should be rated close to 0.

  The expected total duration of the test is 30 minutes that is 10 minutes per test".

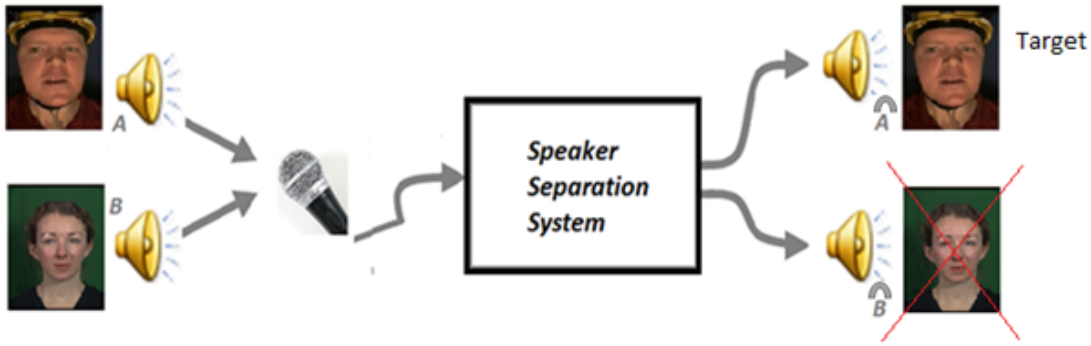


Figure 7.1: Speaker separation task: To extract the target male speaker and suppress the competing female speaker.

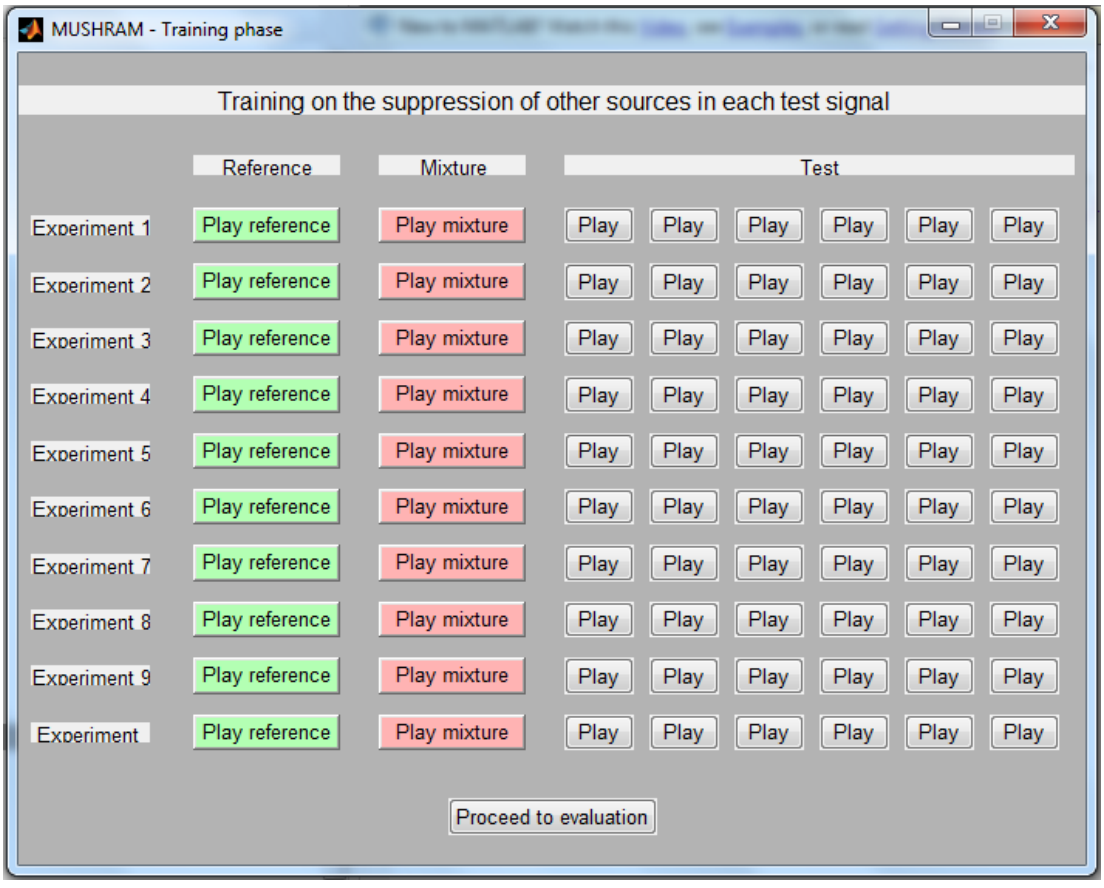


Figure 7.2: GUI for the training phase of the listening test.

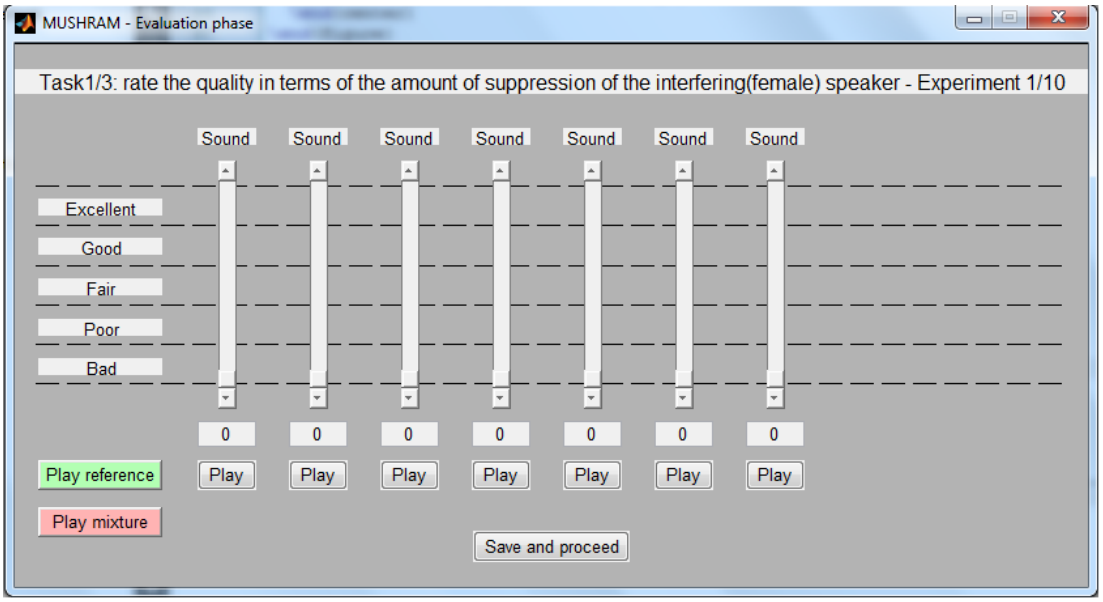


Figure 7.3: GUI for the evaluation phase of the listening test.

7.2.3 Objective measures

In this section the results for the evaluation of the objective measures of speech quality and intelligibility are presented.

Speech Quality

SIR, SDR and SAR measures are used as the objective measures for the evaluation of speech quality. These measures were discussed in detail in Chapter 4 and are computed using the 'BSS evaluation' toolbox [29].

The *SIR* gains comparisons for the different methods at input SIRs of -20dB, -10dB, -5dB, 0dB, 5dB, 10dB and 20dB are shown in Figure 7.4.

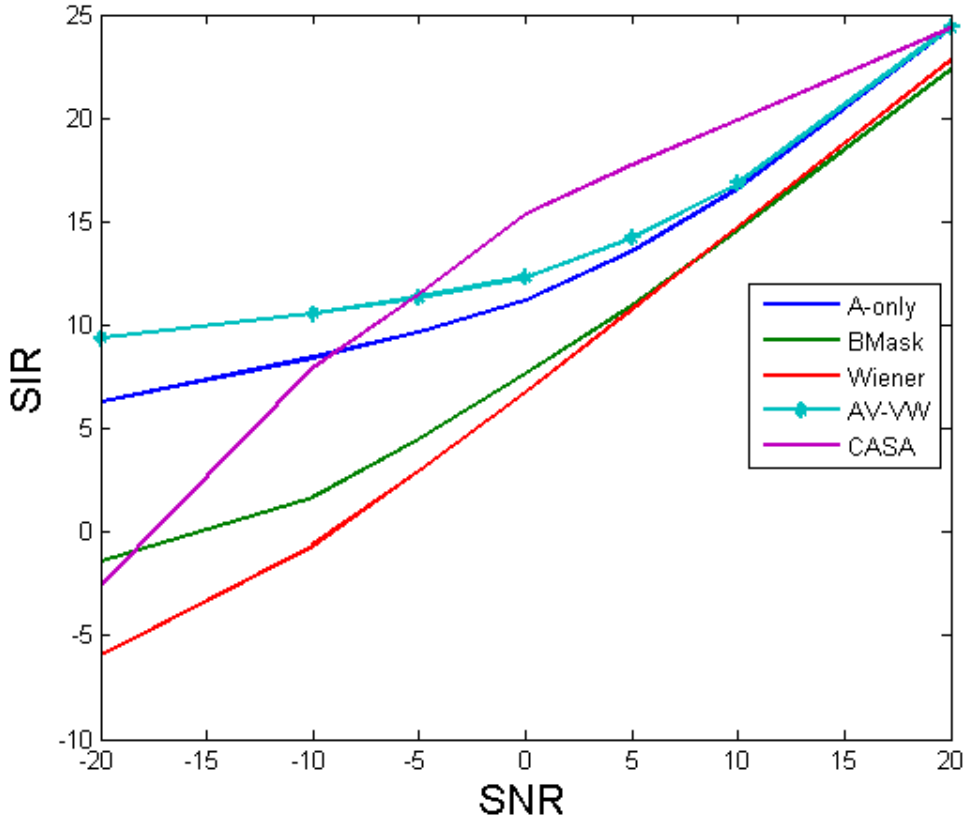


Figure 7.4: *SIR gains comparisons for the various methods at the shown input SNRs for the target speaker.*

At the lower input SNRs of -20dB to -5dB, AV-VW method is giving the highest SIR gains while in the higher input SNRs region of 0dB to +10dB, the CASA method gives higher SIR gains and at +20dB the three methods i.e. CASA, A-only and AV-VW method gives the same SIR gains. The SIR gains of the visually-derived binary masking method and perceptual Wiener method are very low at the lower input SIRs but at higher input SIRs, their SIR gains are comparable to the other three methods. The AV-VW and the A-only methods, give consistently good SIR gains at all input SIRs although the AV-VW method is leading all the

time and at huge margin in the lower input SIRs region.

The reason for the consistently good SIR gains of AV-VW over the A-only method especially in the lower SNR regions is the use of visual information as was shown in Chapter 6 that when the competing speaker dominates the target speaker then the visual stream plays more important role. The CASA method is also based on audio information only and its performance is poor in the lower SNR region. The SIR gains of the visually-derived binary masking method and perceptual Wiener method are very low as compared to the other three methods. As these methods rely totally on the visual information for separation, it becomes obvious here that the audio information is also required for a better separation.

The *SDR* comparisons for the different methods at input SIRs of -20dB, -10dB, -5dB, 0dB, 5dB, 10dB and 20dB are shown in Figure 7.5. AV-VW and A-only methods perform consistently better at all input SIRs in terms of SDR gains and the AV-VW method leading by a small margin in the lower input SIRs region of -20dB to 0dB. The CASA method is performing slightly better than the AV-VW and A-only methods at 0dB and +5dB. The SDR results almost follow the same pattern as the SIR gains.

The *SAR* comparisons for the different methods at input SIRs of -20dB, -10dB, -5dB, 0dB, 5dB, 10dB and 20dB are shown in Figure 7.6. AV-VW and A-only methods perform consistently better at all input SIRs in terms of SAR gains and the A-only method is leading by a very small margin almost all the time. The reason for the slight gains of the A-only method in SAR is the not spectrally detailed coarse estimates of the audio features from the visual features. The visually-derived binary mask is performing the worst because it removes any segment that is identified as the masker dominated causing the target to lose its

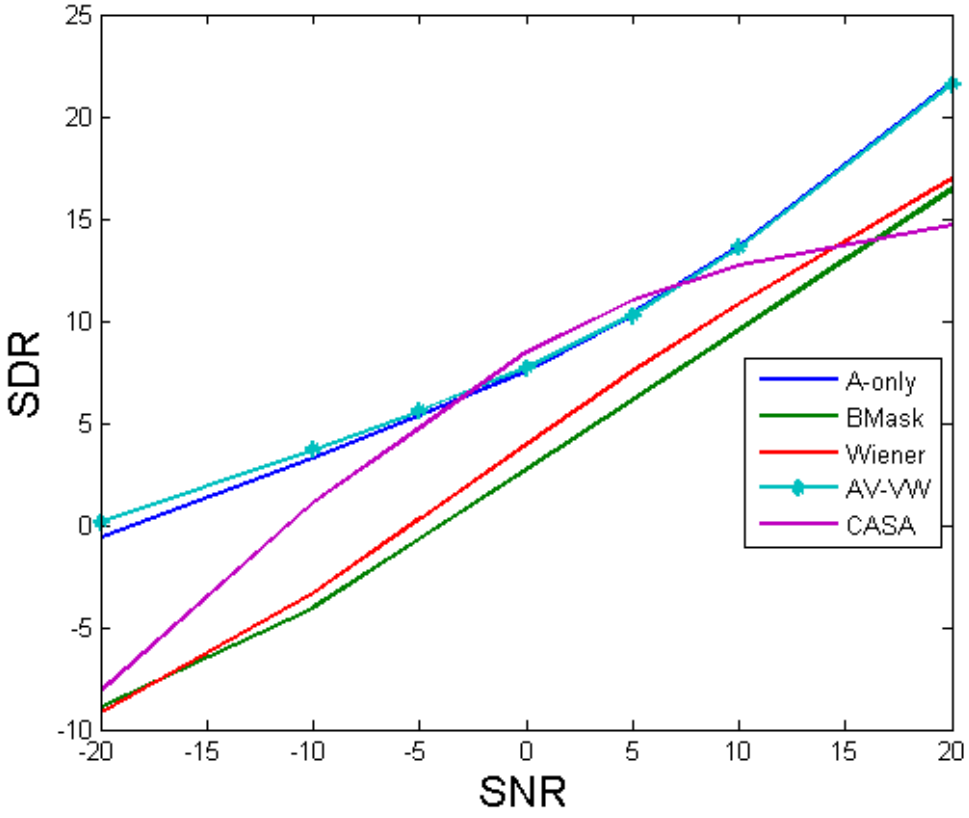


Figure 7.5: *SDR comparisons for the various methods at the shown input SNRs for the target speaker.*

segments and it leads to larger algorithmic distortions in the target in binary masking.

Looking at the SIR, SDR and SAR results, it is concluded that in terms of overall quality, AV-VW is proving to be the best followed by the A-only and CASA methods.

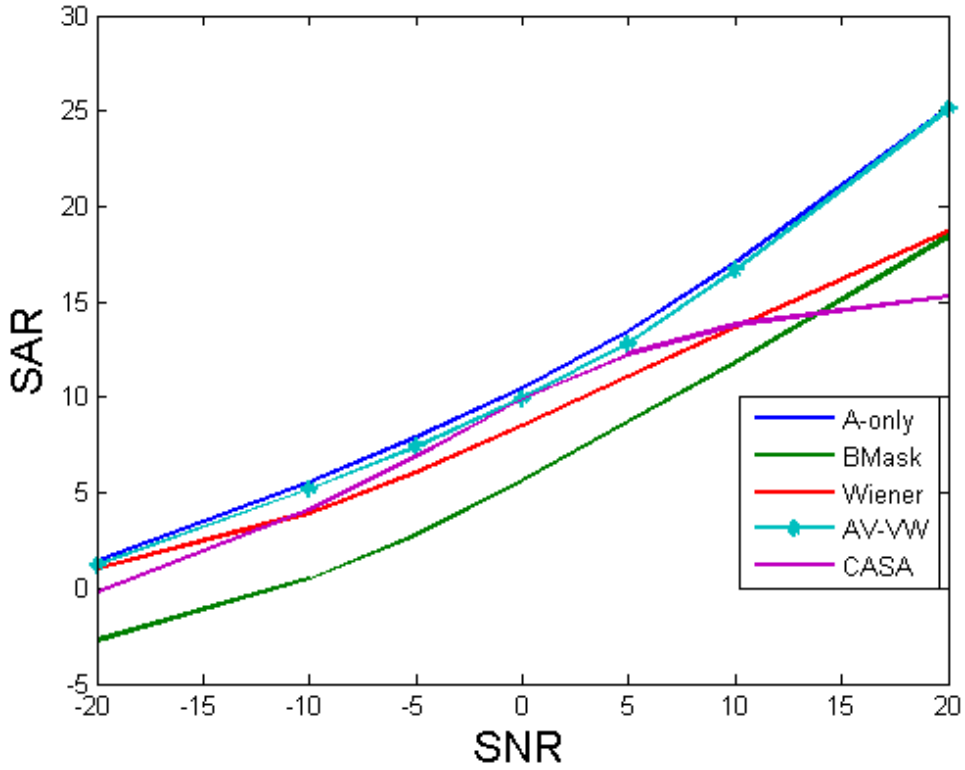


Figure 7.6: *SAR comparisons for the various methods at the shown input SNRs.*

Intelligibility

To measure the intelligibility, a whole word speech recogniser was used. In GRID database, Of the 1000 utterances spoken by each speaker, 800 were used for training and the remaining 200 for testing. Each utterance follows a grammar containing six words of the structure

command→*colour*→*preposition*→*letter*→*digit*→*adverb*.

From the estimates of the target speaker's speech, filterbank vectors were extracted and the resulting word accuracy was used as an estimate of intelligibility.

The intelligibility comparisons for the different methods at input SIRs of -20dB, -10dB, -5dB, 0dB, 5dB, 10dB and 20dB are shown in Figure 7.7.

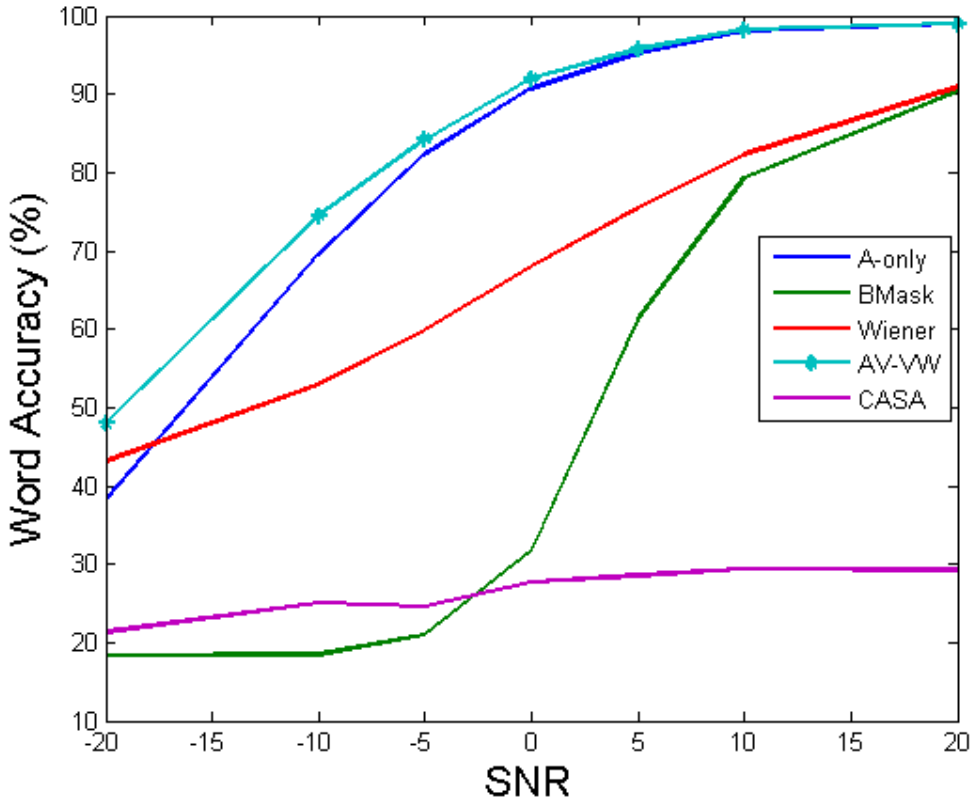


Figure 7.7: *Word recognition accuracy for the various methods at the shown input SNRs for the target speaker.*

The AV-VW method is giving the best intelligibility scores followed by the A-only method. The reason for the consistently high accuracy of AV-VW over the A-only method especially in the lower SNR regions is the use of visual information as was shown in Chapter 6 that when the competing speaker dominates the target speaker then the visual stream plays more important role. The visually-derived binary mask is performing the worst at the lower input SNRs because of the

largely masked segments at the lower input SIRs. The poor performance of CASA is because of mismatch between the training and test data.

7.2.4 Subjective measures

In this section the subjective evaluation results of speech quality are presented in terms of the suppression of the competing speaker, preservation of the target and the overall quality of the extracted speech.

The quality ratings in terms of the suppression of competing speakers for the different methods at input SIRs of -10dB, -5dB, 0dB, 5dB and 10dB are shown in Figure 7.8. The results show that the CASA methods is performing the best in terms of the suppression of the competing speaker followed by the AV-VW method and the results produced by these two methods are comparable at the lower SIRs. The results also show that all the methods are giving huge suppression of the competing speaker over the unprocessed mixed speech.

The quality ratings in terms of the preservation of the target speaker for the different methods at input SIRs of -10dB, -5dB, 0dB, 5dB and 10dB are shown in Figure 7.9. The preservation results show that the unprocessed mixed speech is performing the best. One reason for this is that suppression and preservation are trade-offs i.e. improvements in suppression introduces degradation in preservation and vice versa. As the mixed speech is not processed for suppression therefore there is no degradation introduced in it and the target speaker is preserved in the mixture. The second reason is the layout of the listening test where the listeners have access to the reference target speech as well. Thus any degradation introduced in the target speech by the competing speaker is compensated by the

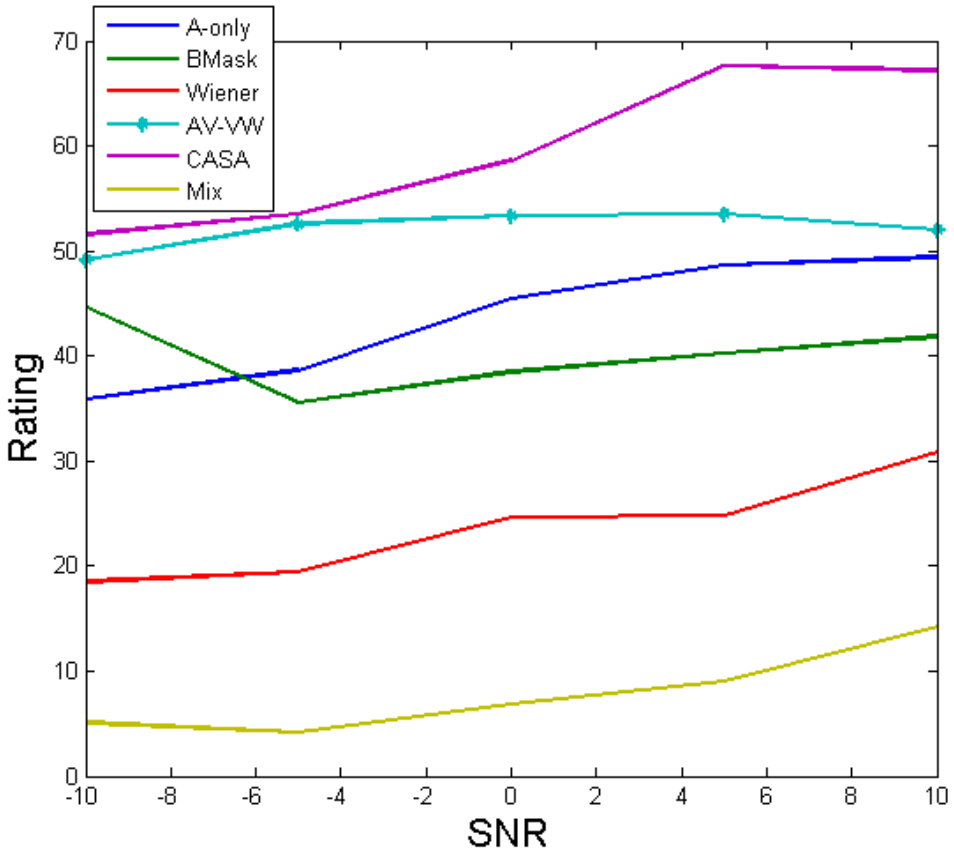


Figure 7.8: *Speech quality rating in terms of the suppression of the competing speaker for the different methods at various input SNRs.*

listeners by listening to the reference speech. The performance of perceptual Wiener, AV-VW and A-only method are comparable. While the visually-derived binary mask and CASA are performing the worst because of the masking effects in these methods.

The quality ratings in terms of the overall quality of the target speaker for the different methods at input SIRs of -10dB, -5dB, 0dB, 5dB and 10dB are shown in Figure 7.10. The overall quality rating results show that the AV-VW method is performing the best most of the time except at +5dB and +10dB where the

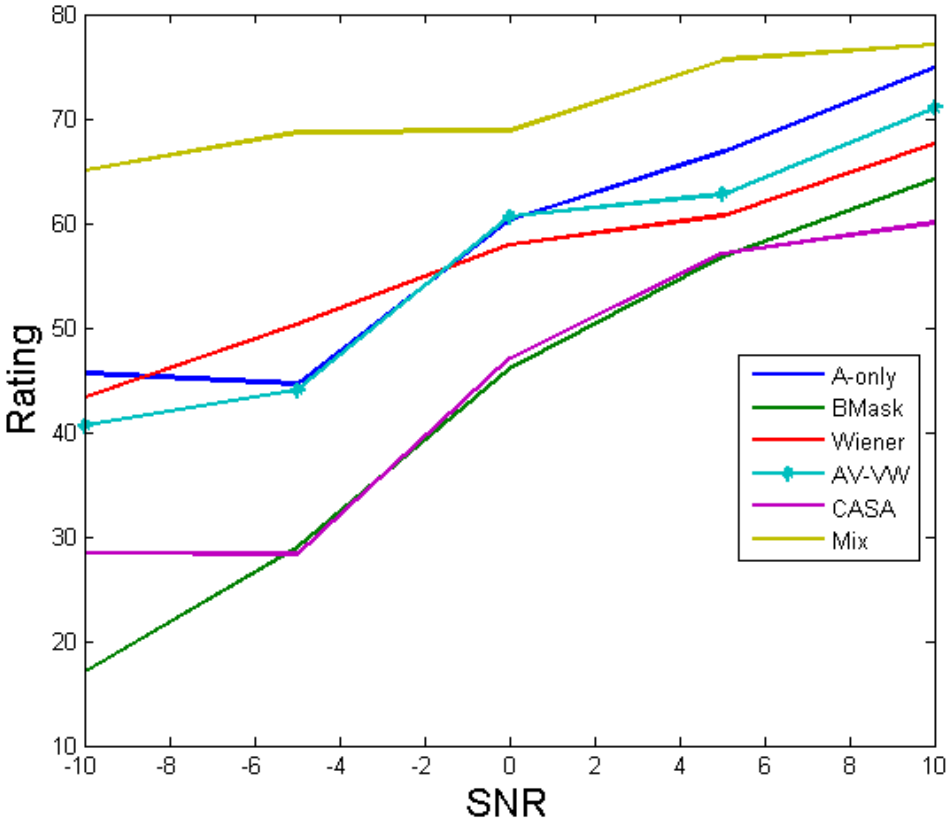


Figure 7.9: *Speech quality rating in terms of the preservation of the target speaker for the different methods at various input SNRs.*

CASA method is performing slightly better. The AV-VW method is followed in performance by the A-only and CASA methods. While the perceptual Wiener and the visually-derived binary masks are performing the worst.

7.3 Summary

This chapter briefly described the proposed and existing methods that were used in the comparison testing. The objective and subjective measures used for the

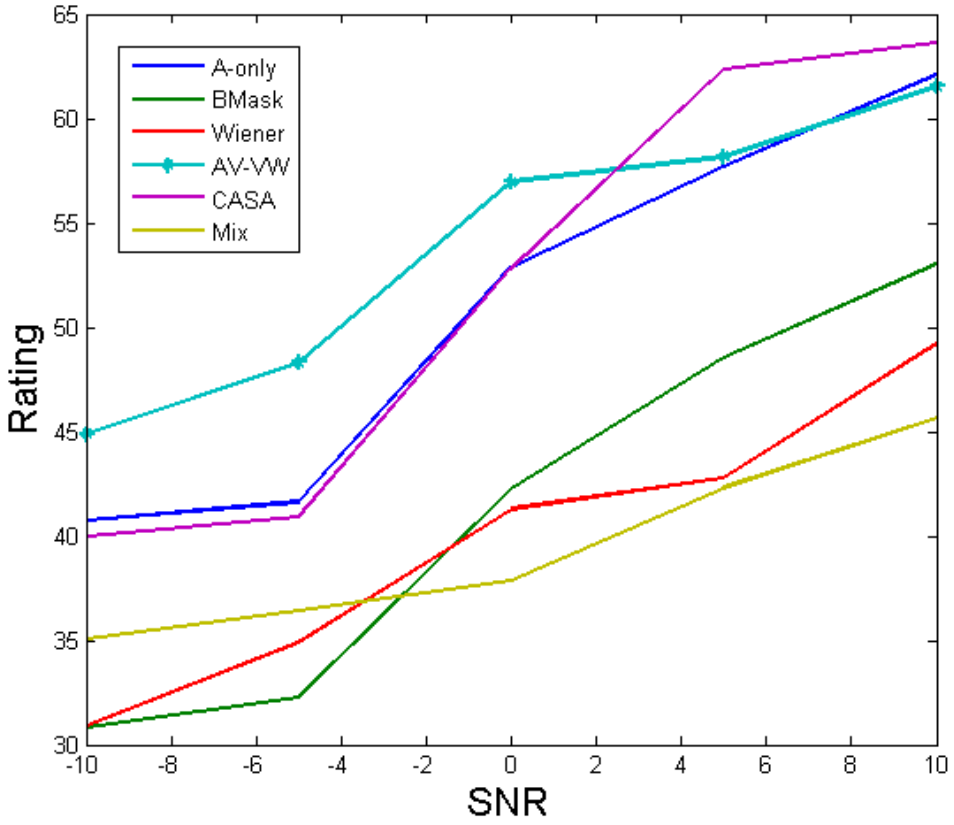


Figure 7.10: *Overall quality ratings of the target speaker for the different methods at various input SNRs.*

quality and intelligibility assessment of the extracted target speech were introduced followed by the detailed experimental set up for the listening tests. SIR, SDR and SAR were used as objective measure to assess the quality of the extracted target speech while word recognition accuracy was used as an objective measure of the intelligibility of the extracted target speech. For subjective quality assessment, listeners were recruited and they rated the quality in terms of the suppression of the competing speaker, preservation of the target and the overall quality of the extracted speech.

The AV-VW and the A-only methods, perform consistently better for SIR, SDR, SAR and the word accuracy measures as compared to the other methods. The AV-VW method gave gains over the A-only method especially in the lower SNR regions for all the measures except SAR. The reason for these gains of AV-VW over the A-only method is that when the competing speaker dominates the target speaker then the visual stream plays more important role and provides useful information for the separation process. The SIR gains of the visually-derived binary masking method and perceptual Wiener method are very low as compared to the other three methods. As these methods rely totally on the visual information for separation, it becomes obvious here that the audio information is also required for a better separation.

The subjective test results correlate heavily with the objective results. The AV-VW and the A-only methods perform consistently better in subjective tests as well and the AV-VW getting gains over the A-only method.

Chapter 8

Conclusions and future work

Preface

This chapter gives a review of the work carried out in this thesis in Section 8.1. The important conclusions drawn from the work are presented in Section 8.2. Finally, in Section 8.3, some suggestions and directions for future work are discussed.

8.1 Review

This section gives a review of the work in this thesis. The introductory chapter explained the aims and motivations for this thesis. The bimodal nature of speech and its importance in human speech perception and in speech processing applications was highlighted. Then an overview of speaker separation methods both in audio only and audio-visual domains was presented. An overview of the thesis structure was also presented in this chapter.

Chapter 2 started with a description of the human speech production process

along with the functioning of the main organs involved. The relation between audio and visual speech units was described in terms of phonemes to visemes mapping. The three AV speech databases used in this work were described. Then methods for the extraction of the audio and visual speech features from these databases were discussed. Then correlation between these audio and visual features was discussed. And finally the correlation results for these three databases used were presented. Maximum average correlation for filterbank features were found to be 0.84 for speaker 6 of GRID database, 0.77 for speaker 4 of GRID database, 0.77 for Messiah database and 0.58 for LIPS2008 database. In the same way the maximum average correlation for the log power spectral features were found to be 0.80 speaker 6 of GRID database and 0.73 for speaker 4 of GRID database.

Chapter 3 discussed the estimation of clean audio speech features from visual speech features. 2D-DCT features were used as the visual features and log filterbank and log power spectral features were used as the audio speech features. The joint density of the audio-visual vectors of each speaker was modelled using a GMM with various number of clusters. Then using these trained models, a MAP estimate of the acoustic speech features from the visual speech features was made. The accuracy of the estimation was measured in terms of mean percentage filterbank estimation errors and mean percentage log power spectral estimation errors. The results showed that the number of dimensions in the visual vector and the number of clusters in the GMM affect the accuracy of estimation. The lowest mean percentage filterbank estimation errors were found to be 9.44 for speaker 6 of GRID database, 10.41 for speaker 4 of GRID database, 8.12 for Messiah database and 12.85 for LIPS2008 database. In the same way the lowest mean percentage log power spectral estimation errors were found to be 15.91 for speaker 6 of GRID

database and 17.18 for speaker 4 of GRID database.

Chapter 4 proposed a method of single-channel audio speaker separation that used visual speech information to extract a target speaker's speech from a mixture of speakers. The method required a single audio input and visual inputs from each speaker in the mixture. The visual information from speakers was used to create a visually-derived Wiener filter. The Wiener filter gains were then non-linearly adjusted by a perceptual gain transform to improve the quality and intelligibility of the target speech. Experimental results were presented that measured the quality and intelligibility of the extracted target speaker and a comparison was made of the different perceptual gain transforms. These showed that significant gains are achieved with the visually-derived Wiener filtering over the original mixture and the gains are further improved by the application of the perceptual gain functions.

Chapter 5 proposed another solution for the problem of single-channel speaker separation and exploited the visual speech information to aid the separation process. The visual features were used to create a time-frequency binary mask that identifies regions where the target speaker dominates. These target dominant regions were retained and formed the estimate of the target speaker's speech. While the regions where the competing speaker was dominant, are masked and discarded. Experimental results compared the visually-derived binary masks with ideal binary masks which showed a useful level of accuracy. The effect of the number of filterbank channels on mask accuracy was also studied. The accuracies of binary mask estimation were found to be 73.76% at -10dB and 83.23% at +20dB. The effectiveness of the proposed method of speaker separation using visually-derived binary masks was then evaluated through estimates of speech quality and speech intelligibility. These results showed substantial gains in quality and intelligibility

for the processed speech over the original mixture.

Chapter 6 proposed another method to exploit both audio and visual speech information to extract a target speaker from a mixture of competing speakers. The chapter began by taking an effective audio-only method of speaker separation, namely the soft mask method, and modified its operation to allow visual speech information to improve the separation process. Experimental results were presented that compared the proposed audio-visual speaker separation method with the audio-only soft mask method using both speech quality and intelligibility metrics.

Chapter 7 presented a comparison of the objective and subjective evaluations of the developed and existing methods. The objective evaluation methods used SIR, SDR and SAR to measure the quality of the extracted speech and word recognition accuracy was used as the measure of estimated intelligibility. For subjective evaluation, listening tests were conducted and the subjects were asked to rate the quality of the extracted speech. The three developed methods and two existing methods were compared along with the reference and unprocessed speech.

8.2 Conclusions

The following important conclusions can be drawn from the work carried out in this thesis:

- High levels of correlation exist between audio and visual speech features that confirms that communication using speech (both audio and video) is an audio-visual experience. This bimodal (audio-visual) nature of speech can be exploited for speaker separation.

- The correlation between audio and visual speech features for a particular speaker is dependant on how well the speech articulators are visible and how well and clearly the speech is articulated.

- The correlation between audio and visual speech features is of significant levels which make the estimation of less spectrally detailed audio features vectors possible from the visual features vectors using GMMs.

- Visual speech features have several limitation. For example the mapping from phonemes to visemes is not unique as several phonemes have the same viseme representation. Also visual speech features for the open mouth in the silence region can mislead the estimation process and the non-speech regions are taken as speech regions because of the mouth opening.

- The estimated audio features vectors from the visual vectors can be used in the construction of visually derived Wiener filter for speaker separation and in the computation of visually derived binary masks that can identify the regions dominated by the target speaker and hence these regions can be used in the final estimate of the target speaker from the mixed speech.

- The visual speech features are more important when the SNR level drops because the audio speech features are much susceptible to acoustic noise but the visual features are not.

- Audio only speaker separation is more efficient in the regions where the target speaker dominates while visual information is more efficient in the regions where the competing speaker dominates and no useful information can be derived from the mixed audio.

- The SIR and SDR gains and intelligibility are trade-offs i.e. larger SIR gains are obtained at the cost of reduction in SDR gains and intelligibility. Therefore, the parameters affecting these measures should be selected in such a way to keep a balance between SIR and SDR gains and intelligibility. Or depending on the application in hand, whether quality enhancement is more desirable or intelligibility improvements, different values of the parameters can be used. For example in sensitive communication like the one used by military, intelligibility is more important than quality.

8.3 Future work

Following from the findings and conclusions of the work carried out some future work directions are suggested here as:

- looking at the correlation levels between audio and visual speech features and the estimation errors of estimated audio features from the visual features, it is obvious that there is a space for improvements. Hence better visual features can be investigated that can increase the correlation levels.

- Better estimation models can be investigated that will reduce the estimation errors like recently deep neural networks (DNN) are becoming more popular [34].

- At present the proposed methods use speaker-dependent models, which is typical of model-based single channel speaker separation methods, it would be desirable to have a speaker-independent system. The high levels of speaker variability in the visual domain make this challenging, but meth-

ods of speaker adaptation and speaker-independent visual features could be investigated.

- At present the number of speakers in the mixture is two which can be increased along with some other acoustic noises.

- The speech mixtures created in this work are instantaneous and reverberation can be introduced in it that could be investigated further.

References

[1] I. Almajai. *Audio Visual Speech Enhancement*. PhD thesis, University of East Anglia, Norwich, UK, January 2009.

[2] I. Almajai and B. Milner. Maximising audio-visual speech correlation. In *Auditory-Visual Speech Processing 2007 (AVSP2007)*, 2007.

[3] I. Almajai and B. Milner. Using audio and visual features for robust voice activity detection in clean and noisy speech. In *EUSIPCO*, Switzerland, August 2008.

[4] I. Almajai and B. Milner. Visually-derived Wiener filters for speech enhancement. *IEEE Trans. Audio, Speech and Language Processing*, 19(6):1642–1651, August 2011.

[5] I. Almajai, B. Milner, and J. Darch. Analysis of correlation between audio and visual speech features for clean audio feature prediction in noise. In *Interspeech*, pages 2470–2473, USA, September 2006.

[6] A. J. Aubrey, Y. A. Hicks, and J. A. Chambers. Visual voice activity detection with optical flow. *IET Image Processing*, 4(6):463–472, December 2010.

[7] J. Barker, A. Coy, N. Ma, and M. Cooke. Recent advances in speech fragment decoding techniques. In *INTERSPEECH 2006 - ICSLP, International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 2006.

[8] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green. The PASCAL CHiME speech separation and recognition challenge. *Computer Speech and Language*, 27(3):621–633, 2013.

[9] L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1):191–199, January 2006.

[10] E. Bozkurt, Q. Erdem, E. Erzin, T. Erdem, and M. Ozkan. Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation. In *3DTV Conference, 2007*, pages 1–4, May 2007.

[11] A. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, 1994.

[12] G. J. Brown and M. Cooke. Computational auditory scene analysis. *Computer Speech & Language*, 8(4):297–336, 1994.

[13] D. A. Bulkin and J. M. Groh. Seeing sounds: visual and auditory interactions in the brain. *Current Opinion in Neurobiology*, 16(4):415 – 419, 2006.

[14] D. Burshtein and S. Gannot. Speech enhancement using a mixture-maximum model. *IEEE Transactions on Speech and Audio Processing*, 10(6):341–351, September 2002.

[15] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *JASA*, 150(5):2421–2424, November 2006.

[16] M. Cooke and D. P. W. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35(3-4):141–177, 2001.

[17] J. Darch. *Robust Acoustic Speech Feature Prediction from Mel-Frequency Cepstral Coefficients*. PhD thesis, Universy of East Anglia, Norwich, UK, February 2008.

[18] J. Darch, B. Milner, and X. Shao. Formant prediction from MFCC vectors. In *COST278 and ISCA Tutorial and Research Workshop (ITRW): Robustness Issuses in Conversational Interaction (Robust2004)*, Norwich, UK, August 2004.

[19] J. Darch, B. Milner, X. Shao, S. Vaseghi, and Q. Yan. Predicting formant frequencies from MFCC vectors. In *ICASSP*, volume 1, pages 941–944, Philadelphia, PA, USA, March 2005. DOI: 10.1109/ICASSP.2005.1415270.

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

[21] S. Ding, J. Huang, D. Wei, and A. Cichocki. A near real-time approach for convolutive blind source separation. *IEEE Trans. on Circuits and Systems*, 53-I(1):114–128, 2006.

[22] D. P. W. Ellis. Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures. *Speech Communication*, 27(3-4):281–298, 1999.

[23] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and objective quality assessment of audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2046–2057, September 2011.

[24] Y. Ephraim. Statistical-model-based speech enhancement systems. *Proceedings of the IEEE*, 80(10):1526–1555, October 1992.

[25] Y. Ephraim, D. Malah, and B. Juang. Speech enhancement based upon hidden Markov modeling. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89*, pages 353–356 vol.1, May 1989.

[26] Y. Ephraim and N. Merhav. Lower and upper bounds on the minimum mean-square error in composite source signal estimation. *Information Theory, IEEE Transactions on*, 38(6):1709–1724, November 1992.

[27] ETSI. Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms. ES 201 108 version 1.1.2, ETSI STQ-Aurora DSR Working Group, April 2000.

[28] ETSI. Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algo-

rithm. ES 202 212 version 1.1.1, ETSI STQ-Aurora DSR Working Group, November 2003.

[29] C. Fevotte, R. Gribonval, and E. Vincent. BSS EVAL toolbox user guide, 2005. Available from http://www.irisa.fr/metiss/bss eval/.

[30] M. J. R. Gomez, D. P. W. Ellis, and N. Jojic. Multiband audio modeling for single-channel acoustic source separation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004*, pages 641–644, 2004.

[31] L. Gu. *Single-channel speech separation based on instantaneous frequency.* PhD thesis, Carnegie Mellon University, May 2010.

[32] J. A. Hartigan and M. A. Wong. Algorithm as 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[33] J. Hershey and M. Casey. Audio-visual sound separation via hidden Markov models. In *Proc. Neural Information Processing Systems*, 2001.

[34] G. Hinton, L. Deng, D. Yu, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.Dahl, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012.

[35] G. Hu and D. L. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans. Neural Networks*, 15(5):1135–1150, 2004.

[36] K. Hu and D. Wang. An unsupervised approach to cochannel speech separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(1):122–131, January 2013.

[37] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 2000, 5-9 June, 2000, Hilton Hotel and Convention Center, Istanbul, Turkey*, pages 2985–2988, 2000.

[38] J. Junqua, B. Reaves, and B. Mak. A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognizer. In *EUROSPEECH*. ISCA, 1991.

[39] F. Khan and B. Milner. Speaker separation using visual speech features and single-channel audio. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 3264–3268, 2013.

[40] F. Khan and B. Milner. Speaker separation using visually-derived binary masks. In *Auditory-Visual Speech Processing, AVSP 2013, Annecy, France, August 29 - September 1, 2013*, pages 215–220, 2013.

[41] M. S. Khan, S. M. Naqvi, A. Rehman, W. Wang, and J. A. Chambers. Video-aided model-based source separation in real reverberant rooms. *IEEE Trans. Audio, Speech & Language Processing*, 21(9):1900–1912, 2013.

[42] S. Khayam. The discrete cosine transform (DCT): Theory and application. Technical report, Department of Electrical & Computer Engineering, Michigan State University, 2003.

[43] K. Kondo. *Subjective Quality Measurement of Speech: Its Evaluation, Estimation and Applications*. Springer, 2012.

[44] T. T. Kristjansson, H. Attias, and J. R. Hershey. Single microphone source separation using high resolution signal reconstruction. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004*, pages 817–820, 2004.

[45] T. T. Kristjansson, J. R. Hershey, P. A. Olsen, S. J. Rennie, and R. A. Gopinath. Super-human multi-talker speech recognition: the IBM 2006 speech separation challenge system. In *INTERSPEECH 2006 - ICSLP, International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 2006.

[46] Y. Lan, B.-J. Theobald, R. Harvey, and E. Ong. Improving visual features for lip-reading. In *Auditory-Visual Speech Processing, AVSP 2010, Hakone, Kanagawa, Japan, September 30 - October 3, 2010*, pages 7–3, 2010.

[47] P. Li, Y. Guan, B. Xu, and W. Liu. Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech. *IEEE Trans. Audio, Speech & Language Processing*, 14(6):2014–2023, 2006.

[48] Y. Li and D. Wang. On the optimality of ideal binary time-frequency masks. *Speech Communication*, 51(3):230–239, 2009.

[49] P. Liu and Z. Wang. Voice activity detection using visual information. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004*, pages 609–612, 2004.

[50] Q. Liu, W. Wang, and P. Jackson. Audio-visual convolutive blind source separation. In *Sensor Signal Processing for Defence (SSPD 2010)*, 2010.

[51] Q. Liu, W. Wang, and P. Jackson. Bimodal coherence based scale ambiguity cancellation for target speech extraction and enhancement. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 438–441, 2010.

[52] Q. Liu, W. Wang, and P. Jackson. Use of bimodal coherence to resolve spectral indeterminacy in convolutive BSS. In *Latent Variable Analysis and Signal Separation*, volume 6365 of *Lecture Notes in Computer Science*, pages 131–139. Springer Berlin Heidelberg, 2010.

[53] Q. Liu, W. Wang, and P. Jackson. A visual voice activity detection method with adaboosting. In *Sensor Signal Processing for Defence (SSPD 2011)*, pages 1–5, September 2011.

[54] Q. Liu, W. Wang, and P. Jackson. Use of bimodal coherence to resolve the permutation problem in convolutive BSS. *Signal Processing*, 92(8):1916 – 1927, 2012. Latent Variable Analysis and Signal Separation.

[55] Q. Liu, W. Wang, P. J. B. Jackson, M. Barnard, J. Kittler, and J. Chambers. Source separation of convolutive and noisy mixtures using audio-visual

dictionary learning and probabilistic time-frequency masking. *IEEE Transactions on Signal Processing*, 61(22):5520–5535, November 2013.

[56] P.C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, Inc., 2007.

[57] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis. Model-based expectation-maximization source separation and localization. *IEEE Trans. Audio, Speech & Language Processing*, 18(2):382–394, 2010.

[58] R. Martin. Spectral subtraction based on minimum statistics. In *Proc. Euro. Signal Processing Conf. (EUSIPCO*, pages 1182–1185, 1994.

[59] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *Speech and Audio Processing, IEEE Transactions on*, 9(5):504–512, July 2001.

[60] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech and Audio Processing*, 13(5-2):845–856, 2005.

[61] I. Matthews, G. Potamianos, C. Neti, and J. Luettin. A comparison of model and transform-based visual features for audio-visual LVCSR. In *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*, pages 825–828, 2001.

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

[63] T. Nakatani and H. G. Okuno. Harmonic sound stream segregation using localization and its application to speech stream segregation. *Speech Communication*, 27(3-4):209–222, 1999.

[64] S. M. Naqvi, W. Wang, M. S. Khan, M. Barnard, and J. A. Chambers. Multimodal (audiovisual) source separation exploiting multi-speaker tracking, robust beamforming and time-frequency masking. *IET Signal Processing*, 6(5):466–477, July 2012.

[65] S. M. Naqvi, M. Yu, and J. A. Chambers. A multimodal approach to blind source separation of moving sources. *J. Sel. Topics Signal Processing*, 4(5):895–910, 2010.

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

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

[68] M. Pedersen, J. Larsen, U. Kjems, and L. Parra. A survey of convolutive blind source separation methods. In J. Benesty, M. M. Sondhi, and Y. Huang, editors, *Springer Handbook of Speech Processing*. Springer, 2007.

[69] T. F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall, 2001.

[70] M. H. Radfar and R. M. Dansereau. Single-channel speech separation using soft mask filtering. *IEEE Trans. Audio, Speech and Language Processing*, 15(8):2299–2310, November 2007.

[71] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan. A joint probabilistic-deterministic approach using source-filter modeling of speech signal for single channel speech separation. In *16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, pages 47–52, September 2006.

[72] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan. Performance evaluation of three features for model-based single channel speech separation problem. In *INTERSPEECH 2006 - ICSLP, International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 2006.

[73] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan. A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation. *EURASIP J. Audio, Speech and Music Processing*, 2007, 2007.

[74] B. Raj, M. L. Seltzer, and R. M. Stern. Reconstruction of missing features for robust speech recognition. *Speech Communication*, 43(4):275–296, September 2004. DOI: 10.1016/j.specom.2004.03.007.

[75] A. M. Reddy and B. Raj. A minimum mean squared error estimator for single channel speaker separation. In *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*, 2004.

[76] A. M. Reddy and B. Raj. Soft mask estimation for single channel speaker separation. In *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing, ICC, Jeju, Korea, October 3, 2004*, page 158, 2004.

[77] A. M. Reddy and B. Raj. Soft mask methods for single-channel speaker separation. *IEEE Trans. Audio, Speech and Language Processing*, 15(6):1766–1776, August 2007.

[78] 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 Trans. Audio, Speech and Language Processing*, 15(1):96–108, January 2007.

[79] 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.

[80] 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. In *15th International Conference on Digital Signal Processing*, pages 223–226, July 2007.

[81] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers. Audiovisual speech source separation: An overview of key methodologies. *IEEE Signal Processing Magazine*, 31(3):125–134, May 2014.

[82] N. Roman, D. Wang, and G. J. Brown. Speech segregation based on sound localization. *The Journal of the Acoustical Society of America*, 114(4):2236–2252, 2003.

[83] S. T. Roweis. One microphone source separation. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 793–799, 2000.

[84] S. T. Roweis. Factorial models and refiltering for speech separation and denoising. In *Eurospeech*, pages 1009–1012, 2003.

[85] S. Sanei, S. M. Naqvi, J. A. Chambers, and Y. Hicks. A geometrically constrained multimodal approach for convolutive blind source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007, Honolulu, Hawaii, USA, April 15-20, 2007*, pages 969–972, 2007.

[86] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano. Blind source separation combining independent component analysis and beamforming. *EURASIP J. Adv. Sig. Proc.*, 2003(11):1135–1146, 2003.

[87] H. Sawada, S. Araki, and S. Makino. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Trans. Audio, Speech & Language Processing*, 19(3):516–527, 2011.

[88] J.-L. Schwartz, F. Berthommier, and C. Savariaux. Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, 93(2):B69–B78, 2004.

[89] X. Shao and J. Barker. Audio-visual speech recognition in the presence of a competing speaker. In *Proc. INTERSPEECH*, 2006.

[90] X. Shao and J. Barker. Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment. *Speech Communication*, 50(4):337–17353, April 2008.

[91] D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J.-L. Schwartz, and C. Jutten. A study of lip movements during spontaneous dialog and its application to voice activity detection. *The Journal of the Acoustical Society of America*, 125(2):1184–1196, 2009.

[92] S. Srinivasan, J. Samuelsson, and W. B. Kleijn. Codebook driven short-term predictor parameter estimation for speech enhancement. *IEEE Trans. Audio, Speech & Language Processing*, 14(1):163–176, 2006.

[93] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26(2):212–215, 1954.

[94] S. Taylor, B.-J. Theobald, and I. Matthews. The effect of speaking rate on audio and visual speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3037–3041, May 2014.

[95] B.-J. Theobald. *Visual Speech Synthesis using Shape and Appearance Models*. PhD thesis, Universy of East Anglia, Norwich, UK, August 2003.

[96] B.-J. Theobald, S. F. Elisei, and G. Bailly. LIPS2008: Visual speech synthesis challenge. In *Interspeech*, pages 2310–2313, 2008.

[97] C. W. Therrien. *Discrete Random Signals and Statistical Signal Processing*. Prentice-Hall, 1992. ISBN: 0-13-217985-7.

[98] T. Tsalaile, S. M. Naqvi, K. Nazarpour, S. Sanei, and J. A. Chambers. Blind source extraction of heart sound signals from lung sound recordings exploiting periodicity of the heart sound. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, March 30 - April 4, 2008, Caesars Palace, Las Vegas, Nevada, USA*, pages 461–464, 2008.

[99] R. Tucker. Voice activity detection using a periodicity measure. *Communications, Speech and Vision, IEE Proceedings I*, 139(4):377–380, August 1992.

[100] S. V. Vaseghi. *Advanced Digital Signal Processing and Noise Reduction*. John Wiley and Sons Ltd, 2000.

[101] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Trans. Audio, Speech and Language Processing*, 14(4):1462–1469, July 2006.

[102] D. Wang and G. J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[103] W. Wang, S. Sanei, and J. A. Chambers. Penalty function-based joint diagonalization approach for convolutive blind separation of nonstationary sources. *IEEE Trans. Signal Processing*, 53(5):1654–1669, 2005.

[104] R. J. Weiss and D. P. W. Ellis. Estimating single-channel source separation masks: relevance vector machine classifiers vs. pitch-based masking. In *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition, SAPA 2006, Pittsburgh, PA, USA, September 16, 2006*, pages 31–36, 2006.

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