# TOWARDS AN INTELLIGENT FUZZY BASED MULTIMODAL TWO STAGE SPEECH ENHANCEMENT SYSTEM

A thesis submitted in accordance with the requirements of the University of Stirling for the degree of Doctor of Philosophy

by

ANDREW KARL ABEL

Computing Science and Mathematics

University of Stirling

April, 2013

# DECLARATION

I hereby declare that this work has not been submitted for any other degree at this University or any other institution and that, except where reference is made to the work of other authors, the material presented is original.

*Stirling, Scotland, April, 2013*

Andrew Karl Abel

# ABSTRACT

This thesis presents a novel two stage multimodal speech enhancement system, making use of both visual and audio information to filter speech, and explores the extension of this system with the use of fuzzy logic to demonstrate proof of concept for an envisaged autonomous, adaptive, and context aware multimodal system. The design of the proposed cognitively inspired framework is scalable, meaning that it is possible for the techniques used in individual parts of the system to be upgraded and there is scope for the initial framework presented here to be expanded.

In the proposed system, the concept of single modality two stage filtering is extended to include the visual modality. Noisy speech information received by a microphone array is first pre-processed by visually derived Wiener filtering employing the novel use of the Gaussian Mixture Regression (GMR) technique, making use of associated visual speech information, extracted using a state of the art Semi Adaptive Appearance Models (SAAM) based lip tracking approach. This pre-processed speech is then enhanced further by audio only beamforming using a state of the art Transfer Function Generalised Sidelobe Canceller (TFGSC) approach. This results in a system which is designed to function in challenging noisy speech environments (using speech sentences with different speakers from the GRID corpus and a range of noise recordings), and both objective and subjective test results (employing the widely used Perceptual Evaluation of Speech Quality (PESQ) measure, a composite objective measure, and subjective listening tests), showing that this initial system is capable of delivering very encouraging results with regard to filtering speech mixtures in difficult reverberant speech environments.

Some limitations of this initial framework are identified, and the extension of this multimodal system is explored, with the development of a fuzzy logic based framework and a proof of concept demonstration implemented. Results show that this proposed autonomous, adaptive, and context aware multimodal framework is capable of delivering very positive results in difficult noisy speech environments, with cognitively inspired use of audio and visual information, depending on environmental conditions. Finally some concluding remarks are made along with proposals for future work.

# ACKNOWLEDGMENTS

I always figured that I'd be all manly and brief about this. However, with the deadline looming, it turns out that the clocks have just gone back, giving me a free hour which I didn't expect to have. Therefore, I am going to use this free hour to get all emotional and dramatic. There are so many people I want to name, but I can't possibly name you all, so if I've missed you out, then I'm sorry!

First of all, I have to thank my supervisor Amir Hussain, who gave me this opportunity in the first place, and who has continued to provide me with so many other opportunities! I really couldn't imagine doing a PhD with anyone else, and he has shown so much faith in me and given me so much freedom, despite all the evidence suggesting this was a mistake.

I want to thank Leslie, my second supervisor as well, whose tips for surviving a conference I faithfully (and occasionally nearly fatally) followed for 3 years. In fact, I'm grateful to so many of the staff, especially Savi and Carron, who provided me with good advice and unleashed me on undergrads, and Grace and Linda, who showed more patience and understanding than any human beings would rightly be expected to. Oh and Kate too, who was always ready to help whenever I turned up at her office with a spreadsheet full of numbers and a look of abject terror.

The shared discussions with all my fellow COSIPRA group members were valuable, and not just the ones that involved muttering about nets, harpoon guns, and phone calls. Oh, and of course, the unofficial COSIPRA group. You know who you are, and I'm sure you'll continue to have those closed door discussions, clandestine email exchanges, and highly secret conversations.

I have to thank my friends who I made at all those Cost 2102 events, who reminded me that we really were all in it together, David, Marco, Erik, and Maria (and her lovely neck!), to name just a few, and who made the (ahem) ordeal of having to fly to a European city twice a

year and spend a week working very hard, going to bed early, and doing a tiny, tiny bit of moderate socialising, into some of the highlights of my PhD!

Oh, and those who I started my PhD with all those years back, Ros (who is probably using the word "anyway" a good deal less these days), Soufiene (who may get let out of the dungeon one day), and especially Jesse, who was a brilliant office mate, and who hates Microsoft Word, although he doesn't like to talk about it. The clique too, Claire and Jamie, with whom things were never at any point weird, and who have made the last few years so much fun, thank you both!

My flatmates too were sometimes the only thing that kept me going, Marco, Rob, Jodie, Euan, the Spices, Chrystie, and Jamie. I was contemplating dedicating my PhD to Ninians, but I thought that would be excessive.

I have to thank my family too, and not just because they'd cut me out of their wills otherwise. Their sarcasm, occasional warm support, frequent highly personal comments, and exasperated commands for me to get on with it were much appreciated, and are exactly why my family are awesome!

Probably most of all though, if it wasn't for the support of Laura, who has been there for me for all these years, I wouldn't still be here. I really don't know where I'd be without our late night text conversations and her patience, sympathy, and good advice.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACRONYMS

| | |
|---|---|
| 2D-DCT | Two Dimensional Discrete Cosine Transform |
| AAM | Adaptive Appearance Models |
| ACM | Active Contour Models |
| AcAMs | Active Appearance Models |
| ANC | Adaptive Noise Canceller |
| ANNs | Artificial Neural Networks |
| ASA | Auditory Scene Analysis |
| ASR | Automatic Speech Recognition |
| ASMs | Active Shape Models |

| | |
|---|---|
| AVCDCN | Audio-Visual Codebook Dependent Cepstral Normalisation |
| AVICAR | Audio-Visual Speech Recognition in a Car |
| BANCA | Biometric Access Control for Networked and E-Commerce Applications |
| BM | Blocking Matrix |
| BSS | Blind Source Separation |
| CCA | Canonical Correlation Analysis |
| CDCN | Codebook Dependent Cepstral Normalisation |
| CRM | Coordinate Response Measure |
| CUAVE | Clemson University Audio Visual Experiments |
| DCT | Discrete Cosine Transform |
| DFT | Discrete Fourier Transform |
| EM | Expectation Maximisation |
| FBF | Fixed Beamformer |
| FIR | Finite Impulse Response |
| FPGAs | Field-Programmable Gate Arrays |
| GMM | Gaussian Mixture Model |
| GMMs | Gaussian Mixture Models |
| GMM-GMR | Gaussian Mixture Model - Gaussian Mixture Regression |
| GMR | Gaussian Mixture Regression |
| GSC | Generalised Sidelobe Canceller |
| HMM | Hidden Markov Model |
| HMMs | Hidden Markov Models |

| | |
|---|---|
| ICA | Independent Component Analysis |
| IIFastICA | Intelligently Initialised Fast Independent Component Analysis |
| IS | Itakura-Saito Distance |
| ITU-T | International Telecoms Union |
| LIF | Leaky Integrate-and-Fire |
| LLR | Log-Likelihood Ratio |
| LPC | Linear Predictive Coding |
| LSP | Line-Spectral Pairs |
| M2VTS | Multi Modal Verification for Teleservices and Security Applications |
| MAP | Maximum a Priori |
| MCMC-PF | Markov Chain Monte Carlo Particle Filter |
| MFCC | Mel Frequency Cepstral Coefficients |
| MLP | Multi Layer Perceptrons |
| MLR | Multiple Linear Regression |
| MOS | Mean Opinion Scores |
| PCA | Principal Component Analysis |
| PDF | Probability Density Function |
| PESQ | Perceptual Evaluation of Speech Quality |
| PS | Power Spectrum |
| RGB | Red Green Blue |
| ROI | Region of Interest |
| SAAM | Semi Adaptive Appearance Models |

| | |
|---|---|
| SegSNR | Segmental SNR |
| SNR | Signal to Noise Ratio |
| STFT | Short-Time Fourier Transform |
| SVM | Support Vector Machine |
| TFGSC | Transfer Function Generalised Sidelobe Canceller |
| VAD | Voice Activity Detector |
| WSS | Weighted-Slope Spectral Distance |
| XM2VTSDB | The Extended Multi Modal Verification for Teleservices and Security Applications Database |

# 1

# INTRODUCTION

Previous research developments in the field of speech enhancement (such as multi microphone arrays and speech enhancement algorithms) have been implemented into commercial hearing aids for the benefit of the hearing impaired community. In recent years, electronic hardware has advanced to such a level that very sophisticated audio only hearing aids have been developed. It is expected that in the future, conventional hearing aids will be transformed to also make use of visual information with the aid of camera input, combining audio and visual information to improve the quality and intelligibility of speech in real-world noisy environments.

The multimodal nature of both human speech production and perception is well established. The relationship between audio and visual speech has been investigated in the literature, demonstrating that speech acoustics can be estimated using visual information. Amongst others, Almajai *et al.* [13] investigated correlation between audio and visual features using Multiple Linear Regression (MLR), and expanded upon this to develop a visually derived Wiener filter for speech enhancement. Sargin *et al.* [160] also performed correlation analysis of multimodal speech, but used Canonical Correlation Analysis (CCA) (Hotelling [84]) as part of a speaker identification task.

The ultimate long term goal of the research presented in this thesis is to improve the lives of those who suffer from deafness. Even state-of-the-art modern hearing aids can fail to cope with rapid changes in environmental conditions such as transient noise or reverberation, and there is much scope for improvement. The work presented in this thesis aims to develop an initial flexible audiovisual speech filtering system, which can then be developed further to become a cognitively inspired, autonomous, adaptive, and context aware multimodal speech enhancement framework, using fuzzy logic as part of the speech filtering process. The research presented in this thesis takes the field of speech enhancement and utilises visual information to extend single modality audio concepts. In recent decades, much research has been conducted into speech enhancement in a range of different environments, and more recently, the use of visual information to filter noisy speech has been studied in more detail.

However, there are limitations with both visual and audio speech filtering approaches, and there are a number of requirements for plausible, real world speech enhancement. A speech enhancement system has to be able to automatically track and extract visual information, it should consist of a scalable framework that can be upgraded in the future, and it should take advantage of both audio and visual information to maximise performance. It should also make use of both speech modalities and filter speech in different ways depending on environmental conditions.

## 1.1 THESIS MOTIVATION

### 1.1.1 *Multimodal Speech Enhancement*

Speech enhancement, the process of improving speech quality by filtering noise, is a field with a long history of development stretching back decades (Zelinski [189], Hussain & Campbell [90]). Today, this field continues to be extremely active, with many recent examples of modern speech processing algorithms in the literature such as Hussain *et al.* [91], Van den Bogaert *et al.* [173], Li *et al.* [113]. There are several techniques that are commonly used, such as beamforming, a multi microphone approach that aims to exploit the spatial diversity between speech and noise sources to filter the speech signal (Gannot *et al.* [67], Griffiths & Jim [73]). Another technique is to make use of Wiener filtering (Wiener [180]) to compare a noisy speech signal to an estimate of the equivalent noise free signal to produce an enhanced signal (Van den Bogaert *et al.* [173]). There are many different other single modality noise cancelling techniques in the literature.

Of growing interest in recent years has been the extension of these single modality algorithms to take account of developments pertaining to multimodality. This takes account of behavioural and physiological properties relevant to speech production and perception. Speech is produced by vibration of the vocal cords and the configuration of the vocal tract (which is composed of articulatory organs), and due to the visibility of some of these articulators such as tongue, teeth, and lips, there is an inherent relationship between the acoustic and visible properties of speech production. The relationship between audio and visual aspects of communication have been established since pioneering works in 1954 by Sumby & Pollack [168], and subsequent developments such as the McGurk effect (McGurk & MacDonald [126]). Audiovisual speech correlation has been deeply investigated in the literature (Barker & Berthommier [22], Almajai & Milner [10], Sargin *et al.* [160]), including in publications by this author (Cifani *et al.* [45], Abel *et al.* [2]), showing the connection between lip movement and acoustic speech.

Multimodal correlation is of interest because this relationship can be exploited to filter noisy speech. To the knowledge of the author, the first example of a functioning audiovisual speech filtering system was proposed in 2001 by Girin *et al.* [69], and this was then followed by other related work by Goecke *et al.* [71], Sodoyer *et al.* [166, 167]. The increased processing power of computers and improved capability of relevant technical components such as video cameras has made the concept of utilising camera input as part of a hearing aid system feasible. There are strengths and weaknesses with using visual information for speech enhancement, but it has proved to have potential for further development. Following on from pioneering work by Girin *et al.* [69], recent work has focused on the use of visual information for use as part of a source separation based system (Rivet *et al.* [147]). Other state-of-the-art work, particularly by Almajai *et al.* [13], has made use of visual information as part of a Wiener filtering speech processing system.

The research presented in this thesis shows the correlation between visual speech features and acoustic speech data in noisy environments, and makes use of visual information to develop a speech enhancement system, extending an audio only two stage approach to become multimodal. This utilises both audio only and visually derived speech filtering techniques.

### 1.1.2 *Multimodal Speech Filtering Framework*

As part of the desire to create a multimodal speech enhancement system, it is important to take the complete filtering framework into account. The speech filtering algorithm is just one individual component of an overall speech processing system, especially with regard to complex multimodal systems. There are many potential components to consider, such as the extraction of audio and visual features from the raw speech input data of both modalities, the individual speech processing algorithms used, audiovisual speech modelling components, and the tracking of visual data in order to identify the ROI. Examples of a multimodal speech filtering front end were reported by Potamianos *et al.* [142], Deligne *et al.* [52]. These authors developed a multimodal system consisting of ROI extraction from visual frames, and audio and visual feature extraction. This framework served as a multimodal front end for speech processing algorithms, which in the work presented by Potamianos *et al.* [142], was either speech recognition or speech enhancement.

This motivation is of interest with regard to this thesis because the work presented will utilise multiple speech filtering techniques. This research is multi disciplinary, and proposes to combine a number of state-of-the-art techniques from a number of different research fields. For example, the visual tracking aspect of this system that is proposed for use in this thesis (Nguyen & Milgram [135]) was developed originally as a standalone visual tracking system, not specifically designed to apply to speech enhancement. Therefore, a framework that enables all of the individual components to be successfully integrated to form a comprehensive speech filtering system is required. Additionally, as there are a number of different techniques involved in such a framework, a well designed framework should be both scalable and component based to allow for the upgrading of individual aspects of the system without requiring a complete rewrite and redesign. So in the case of the visual tracking example mentioned above, it should ideally be possible to replace the tracking technique with an alternative version without difficulty. The same should be true of all proposed components, all are the subject of research in their specific fields, and so it is desirable to design a scalable framework that is easy to upgrade.

### 1.1.3 *Plausible Noisy Speech Environmental Condition Testing*

A crucial aspect of speech enhancement research is the evaluation process. This concerns more than simply the approach used for testing (i.e. objective machine based scoring or subjective listening tests), but also considers the environment and speech/noise data that filtering algorithms are designed for. This includes factors such as the level of background noise in comparison to the speech signal (SNR), the type of noise used (white noise, other speech, automobile noise etc.), the approach used for mixing speech and noise sources (additive or convolved), the composition of the test sentences used, whether these sentences consist of simple vowel-consonant-vowel mixtures or more complex sentences, and many other factors. Many speech enhancement approaches are optimised to work with very specific data, for example, Milner & Almajai [130] have designed audiovisual filtering systems that are trained and tested with data from a single speaker. Historically, early audiovisual systems were tested using simple sentences corrupted with white noise (Goecke *et al.* [71]), or using test sentences consisting of vowel-consonant-vowel combinations (Girin *et al.* [69]).

The evaluation approach is of interest in this thesis because in real world conditions, speech enhancement is required to be flexible with regard to environmental conditions. As a practical example, a hearing aid wearer is not expected to limit themselves to interacting only in specific environments. A hearing aid would be expected to function adequately with a wide range of input data, noise mixtures, environments and speakers. Since early pioneering audiovisual research, evaluation of speech enhancement systems has become much more comprehensive. Some examples of more advanced testing configurations include detailed speech sentences (Almajai & Milner [11]), and much more realistic convolved noise mixtures from reverberant room environments (Rivet & Chambers [152]). However, there is still a focus on evaluating speech enhancement in very specific and consistent environmental conditions rather than taking account of a more plausible, changeable range of scenarios.

This thesis evaluates the performance of the system presented in this work in a range of challenging environments, using different speakers and noises to establish the strengths and limitations of this multimodal approach. The initial system is then extended with the ultimate goal of becoming capable of adapting to changeable audiovisual noisy speech environments.

### 1.1.4 *Cognitively Inspired Intelligent Flexibility*

There have been many different speech enhancement systems developed, both audio only (Van den Bogaert *et al.* [173], Li *et al.* [113]), and in recent years, multimodal (Goecke *et al.* [71], Almajai & Milner [11]). These systems filter noisy speech mixtures in a variety of different ways, with various limitations. Often, these filtering systems are designed to perform best in very specific scenarios. One example of this is that visually derived filtering requires a consistent source of good quality visual information. In a practical, more realistic scenario that a speech enhancement system might be expected to deal with, a degree of flexibility when it comes to speech processing is desirable. An example of how speech can be filtered in a more cognitively inspired manner is seen with neurofuzzy systems (Esposito *et al.* [57]), which process sound in different ways depending on the noisy speech input, and commercial hearing aids, that may use decision rules to determine processing (Chung [44]).

The potential flexibility of a multimodal speech enhancement system is of interest because in a real world environment, a number of conditions may vary. One real world commercial example of this is with the Danalogic 6 6080 DVI hearing aid. This is a modern hearing aid that contains some very advanced features such as adaptive directional sound filtering and multiple band signal processing. However, despite this, limitations still remain. For example, informal investigation by the author, confirming general limitations with directional filtering(Chung [44]), has found that with this hearing aid, directional speech filtering performs well in some situations, but in others can cause the quality of audio signal received by the listener to vary dramatically with very small changes in background noise. Custom programmes can be created for different environments such as for listening to music, or for improving intelligibility in busy environments, but in order to use them, the listener has to manually push a button on their hearing aid. Different speech enhancement algorithms are suited to different conditions, and no single algorithm is ideal for use in all noise environments. It is also important to take account of potential quality issues when it comes to visual input. Although in laboratory conditions it can be assumed that input visual frames are of good quality, in a less restricted environment this may not be the case. For example, the speaker may turn their head, place their hand in front of their mouth, or another person may temporarily obstruct the view of the speaker. Because of this, it is essential that a state-of-the-art multimodal speech filtering system is intelligent and sophisticated enough to take account of both acoustic and visual criteria to optimise speech filtering output.

This thesis develops an initial audiovisual system further by presenting a novel framework that utilises fuzzy logic as part of a two-stage audio and visual speech processing framework. The addition of fuzzy logic allows for the development of an autonomous, adaptive, and context aware system that takes account of different audio and visual environmental conditions to filter the noisy speech in a more cognitive manner. This demonstrates that a multimodal framework can be developed that takes account of volatile noisy speech environments.

## 1.2 RESEARCH AIMS AND OBJECTIVES

The overall aim of the work presented in this thesis is to develop a multimodal speech enhancement system, making use of both audio only and visually derived speech filtering techniques. The resulting system should combine these techniques in an integrated framework, along with state-of-the-art lip tracking and feature extraction techniques. This thesis also aims to investigate the extension of this system to become more autonomous, adaptive, and context aware, by developing a cognitively inspired speech filtering framework that utilises fuzzy logic to determine the most suitable processing techniques to use on a frame by frame basis.

There are a number of detailed objectives of this thesis. Firstly, the relationship between audio and visual aspects of speech will be discussed, followed by an investigation into multimodal speech correlation, in order to demonstrate the viability of developing a multimodal speech filtering system. Secondly, state-of-the-art techniques in a variety of research domains, such as visual tracking (Nguyen & Milgram [135]), multimodal speech modelling (Calinon *et al.* [33]), and Wiener filtering (Almajai *et al.* [13]) will be examined to provide the basis for development of a two-stage multimodal speech enhancement system. Thirdly, this system will be scalable, with the ability to be upgraded to take account of future research developments, and will provide a solid basis for future development. Fourthly, the two-stage speech filtering system presented in this thesis will then be evaluated in depth to identify its strengths and limitations. Finally, the work presented in this thesis aims to extend this initial two-stage system with the use of fuzzy logic to develop a preliminary, autonomous, adaptive, context aware, multimodal speech enhancement framework that takes account of audio and visual environmental conditions to filter speech in a cognitive manner.

## 1.3 original contributions of this thesis

The key contributions of this thesis are listed in this section, and are:

### 1.3.1 *Investigation into Audiovisual Speech Correlation in Noisy Acoustic Environments*

One original contribution of this work is an investigation into audiovisual speech correlation in noisy environments. Given the established relationship between audio and visual aspects of speech, it is of no surprise that speech correlation has been studied extensively in the literature (Almajai & Milner [10], Barker & Berthommier [22], Sargin *et al.* [160]). However, there has been less research into multimodal speech correlation in noisy environments, and this thesis presents new results of an investigation into multimodal correlation in noisy speech environments, published by the author in Cifani *et al.* [45].

This investigation examines the effectiveness of using a beamformer to filter speech, by comparing the correlation of noisy and enhanced speech. The results represent an original contribution of this thesis. These results also successfully demonstrated the preliminary use of several techniques used later in this thesis such as 2D-DCT and audio only beamforming. They also validated the work of others in the literature, particularly research by Almajai & Milner [10], by using a range of similar techniques such as 2D-DCT and Multiple Linear Regression ( MLR), but with a different corpus and the use of a beamformer, and finding a similar pattern of results despite these differences.

### 1.3.2 *Development of a Two-Stage Multimodal Speech Enhancement System*

One significant original contribution of this thesis is the development of a two-stage multimodal speech enhancement system. This system builds on the idea of audio only two-stage filtering systems such as proposed by Li *et al.* [114], Zelinski [189], and extends this concept to utilise multimodal information. This combination of techniques and modalities has not, to the knowledge of the author, been applied previously. This system contains a number of state-of-the-art techniques. However, it is also loosely integrated and scalable, which means that it is relatively simple for any future development to take advantage of state-of-the-art research and upgrade individual components without any particular difficulty.

This multimodal system has also been exhaustively tested in a variety of challenging environments, simulating a range of conditions that a real world speech enhancement system may be expected to encounter. This includes environments such as those with an extremely high level of background noise, or with data that the multimodal system would be expected to perform poorly with, such as with a speaker that the visually derived filtering aspect of the system has not been trained with. This testing identifies the strengths and weaknesses of this novel system and highlights potential areas for improvement. The results present here represent benchmark results, as the author is not aware of any other multimodal two-stage speech enhancement systems in the literature.

### 1.3.3 *Novel Application of State-Of-The-Art Components to New Research Domains*

In order to create the system developed in this thesis, a number of disparate components were integrated. Some of these individual components have precedence for use as part of speech processing research, but others were applied in this thesis to the domain of audiovisual speech enhancement for the first time. One original contribution resulting from the work presented in this thesis is therefore the investigation and application of state-of-the-art techniques to the domain of multimodal speech enhancement. One such example of this is with the Semi Adaptive Appearance Models (SAAM) image processing approach (Nguyen & Milgram [135]). This is a state-of-the-art visual tracking approach that had never been previously applied to the speech enhancement domain. The work in this thesis is built on a collaboration with the developers of the SAAM image processing approach in order to utilise this state-of-the-art technique for lip tracking, and subsequently lip feature extraction, as part of the overall audiovisual speech enhancement system.

Another prominent example of applying a technique to a novel research domain is the audiovisual speech model used. In this work, audio and visual speech features are modelled in order to then estimate audio speech features given visual information. There are many ways to model this, but in this work, a technique new to this domain is tested, Gaussian Mixture Model - Gaussian Mixture Regression (GMM-GMR). This is a technique that was originally developed and utilised by Calinon *et al.* [33] to train robot arm movement. In this research, this is experimented with as part of an audiovisual speech system. This work also makes use of fuzzy logic as part of the overall system. Although fuzzy logic processing has been utilised as part of speech processing systems previously, to the knowledge of the author, this work represents a novel use of fuzzy logic as part of a two-stage audiovisual speech enhancement system.

### 1.3.4 *Towards Autonomous, Adaptive, Context Aware, Multimodal Speech Enhancement.*

The application of intelligence to an audiovisual speech processing system using fuzzy logic is an original contribution of this thesis. This thesis presents an initial audiovisual speech enhancement system in chapter 4, but then expands upon it in chapter 7 to become more cognitively inspired and make more intelligent use of audio and visual information. There are examples of varying the processing method for noisy speech depending on the background noise applied to audio only, single modality speech filtering systems (Esposito *et al.* [57]), but to the knowledge of the author, there are no examples of the application of fuzzy logic to multimodal two-stage speech enhancement systems. This novel contribution transforms the speech processing system originally presented into a much more intelligent speech filtering framework that takes account of audio, visual, and historical inputs, such as the quality of input information available to both modalities and previous processing decisions, to make an automatic decision of the speech processing operation to utilise on a frame by frame basis. This original contribution represents a proof of concept multimodal framework, and promising example results are shown.

## 1.4 PUBLICATIONS ARISING

From the research carried out during the course of this project, the following publications have emerged:

- A. Abel, A. Hussain. Novel Two-Stage Audiovisual Speech Filtering in Noisy Environments. In Cognitive Computation. Conditional Acceptance.
- A. Abel, A. Hussain, Q.D. Nguyen, F. Ringeval, M. Chetouani, and M. Milgram. Maximising audiovisual correlation with automatic lip tracking and vowel based segmentation. In Biometric ID Management and Multimodal Communication: Joint COST 2101 and

2102 International Conference, BioID_MultiComm 2009, Madrid, Spain, September 16-18, 2009, Proceedings, volume 5707, pages 65–72. Springer-Verlag, 2009.

- S. Cifani, A. Abel, A. Hussain, S. Squartini, and F. Piazza. An investigation into audio-visual speech correlation in reverberant noisy environments. In Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions: COST Action 2102 International Conference Prague, Czech Republic, October 15-18, 2008 Revised Selected and Invited Papers, volume 5641, pages 331–343. Springer-Verlag, 2009.
- A. Abel, A. Hussain. Multi-modal Speech Processing Methods: An Overview and Future Research Directions Using a MATLAB Based Audio-Visual Toolbox. Multimodal Signals: Cognitive and Algorithmic Issues: International School Vietri sul Mare, Italy, April 21-26, 2008 Revised Selected and Invited Papers, volume 5398, pages 121–129. Springer-Verlag, 2009.
- M. Faundez-Zanuy, A. Hussain, J. Mekyska, E. Sesa-Nogueras, E. Monte-Moreno, A. Esposito, M. Chetouani, J. Garre-Olmo, A. Abel, Z. Smekal, K. Lopez-de-Ipiña. Biometric Applications Related to Human Beings: There Is Life beyond Security. Cognitive Computation, volume 4, pages 1-16. Springer-Verlag, 2012.

## 1.5 THESIS STRUCTURE

In this section, the structure of the remainder of this thesis is provided.

Chapter two describes background research relevant to this thesis, examining the relationship between audio and visual aspects of speech production and perception. Audiovisual speech correlation is also discussed and investigated. In chapter three, a detailed literature review is presented. A number of prominent state-of-the-art audiovisual speech enhancement techniques are reviewed, such as visually derived Wiener filtering, audiovisual fragment decoding, and multimodal beamforming. Speech processing techniques currently used in commercially available hearing aids are also reviewed. Finally, a number of audiovisual speech corpora are also reviewed.

A novel audiovisual two-stage speech enhancement system is presented in chapter four. This chapter describes the individual components of this novel multimodal system in detail, including audio and visual feature extraction, the SAAM lip tracking algorithm, the simulated room environment and multi microphone array, the audiovisual GMM-GMR technique used for speech estimation, the visually derived Wiener filtering process, and audio only beamforming. The focus of this chapter is on the technical composition of this two-stage system. Chapter five presents a detailed evaluation of the performance of the speech enhancement system described in chapter four. After briefly describing some preliminary investigations to identify the configuration of this system that delivers the best results, the two-stage multimodal system is tested in a range of noisy environments at various SNR levels to evaluate performance in a range of noisy speech scenarios, and the results are reported. This multimodal system is also tested in an environment where audio-only speech filtering algorithms can struggle. This examines a possible limitation with the system. To evaluate the system with untrained data, experiments with data that features a novel (previously not trained with the system) speaker are presented. Finally this chapter summarises the strengths and weaknesses with this initial system and proposes improvements.

To overcome some of the limitations discussed in the previous chapter, the extension of the multimodal system to become an autonomous, adaptive, context aware audiovisual speech enhancement framework is described in chapter six. This chapter introduces and summarises the concept of fuzzy logic, and describes the relevance of this technique to speech enhancement. The design and implementation of this novel fuzzy based system is presented in this chapter, integrating the components discussed in chapter four within the new fuzzy logic framework, illustrating the inputs, rules, and processing. The way in which this novel framework overcomes limitations identified in the previous chapter is discussed Chapter seven then presents some evaluation of the functionality of this proof of concept framework.

Finally, chapter eight summarises this thesis, provides some concluding remarks, and recommends a number of directions for future work.

# 2

# AUDIO AND VISUAL SPEECH RELATIONSHIP

## 2.1 INTRODUCTION

Before the overall aim of this thesis, the development of a fuzzy logic based multimodal speech enhancement framework, can be presented, it is important that the background to this work is discussed. This chapter presents a summary of the general research domain, and also discusses some new experiments, performed by the author, into the relationship between audio and visual aspects of speech. The background to human speech production is briefly discussed, along with a definition of several speech phenomena relevant to this thesis, namely the Cocktail Party Problem, McGurk Effect, and Lombard Effect.

After a summary of this background information, the remainder of this chapter presents an investigation into audiovisual speech correlation. A number of speech sentences are used from the audiovisual VidTIMIT speech database (Sanderson [157]) , and these are used to carry out a number of experiments. The relationship between audio and visual speech features has been deeply investigated in the literature, and a summary of this has been provided, with a particular focus on relatively recent work by Almajai & Milner [10], however, very little work (to the knowledge of the author) has been carried out into audiovisual speech correlation with noisy speech. This chapter presents an investigation into the correlation between audio and visual speech features when noise is mixed with the original speech sentences. This work represents one original contribution of this thesis, and the majority of the results presented here were also published in a paper by the author in Cifani *et al.* [45]. Several experiments are described in this chapter, using Mel Frequency Cepstral Coefficients (MFCC) audio features, 2D-DCT and Cross-DCT for visual feature vectors, and a Transfer Function Generalised Sidelobe

Canceller (TFGSC) beamformer to enhance noisy speech, with correlation carried out using MLR (which is defined later in this chapter).

These experiments firstly examine the difference in correlation when different visual features (2D-DCT or Cross-DCT) are used for correlation. The dimensionalities of the input vectors (both audio and visual) that maximise speech correlation are also investigated, and finally, a comparison of correlation in a number of noisy scenarios is carried out. This is performed by adding a variety of noises to sentences from the VidTIMIT corpus. These noisy mixtures are then enhanced with the use of a beamformer, and a comparison of the correlation in each noisy scenario is performed. This work represents an original contribution of this thesis, in that it performs a new investigation into audiovisual correlation. It also attempts to validate other work in the literature, and finally, serves as a test of some of the techniques such as 2D-DCT and the TFGSC beamformer that are proposed for use as part of the two-stage speech enhancement system presented later in this thesis.

The remainder of this chapter is divided as follows. Section 2.2 presents a brief background of human speech production, and this is followed by a summary of several speech phenomena relevant to this research domain in section 2.3. A review of audiovisual correlation research is then described in section 2.4, and experiments by the author are reported and discussed in section 2.5.

## 2.2 AUDIO AND VISUAL SPEECH PRODUCTION

### 2.2.1 *Speech Production*

The multimodal nature of human speech is established. This use of multiple modalities in speech involves production as well as perception; indeed, speech is produced by the vibration of the vocal cords and the configuration of the vocal tract that contains articulatory organs. Since some of these articulators are visible, there is an inherent relationship between the acoustic and visual aspects of speech. Moreover, the well-known McGurk effect (summarised

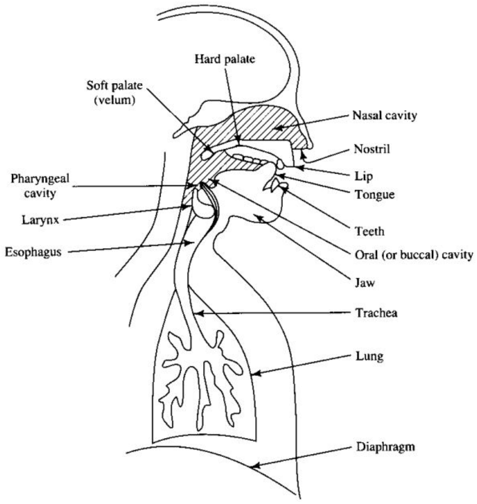


Figure 1: Diagram of the components used in human speech production. Taken from (Rosistem [155]).

later in this chapter) empirically demonstrates that speech perception also makes use of multiple modalities. The high degree of correlation between audio and visual speech has been deeply investigated in literature, with work by Yehia *et al.* [184], Barker & Berthommier [21], and Almajai & Milner [10], showing that facial measures provide enough information to reasonably estimate related speech acoustics. This section provides a summary of human speech production. This is a subject area which has been researched in depth, and there are many detailed summaries in the literature, including by Almajai [9], and Owens & Lynn [137], and this section contains a brief summary.

A diagram of the components used in speech production is shown in figure 1, with the shape of the vocal tract decided by the articulators shown in figure 1, such as the lips, tongue and teeth. Depending on the vibration of the vocal cords, speech can be defined as either voiced or unvoiced. Voiced sounds include all vowels, and some consonants, caused by the vibration of the vocal cords. Unvoiced sounds do not involve vibration, and involve airflow passing through an opening in the vocal cords, with a noise being produced by constriction in the vocal tract. There are also plosive sounds, which are made by closing the lips, allowing air pressure to build, and then opening them again.

### 2.2.2 *Phonemes and Visemes*

The speech units described previously can be split by artificial notation, both audio (phonemes) and visual (visemes). The work presented in this thesis does not directly utilise these concepts, but a brief description is provided here for reference purposes. There are many descriptions of this work in the literature, and one example of a full detailed description can be found in work by Almajai [9]. Phonemes are very basic audio speech units (Rabiner & Schafer [143]). These are broken down from individual words, and there are many different notations used to represent phonemes, depending on the language and source. With regard to British English, phonemes can be broken down into four main categories, vowels, consonants, diphthongs, and semivowels. More detail can be found in many sources, including Almajai [9].

To describe individual visual speech units, visemes are used. However, these do not map exactly to phonemes due to the nature of the articulators used in speech production (i.e. the lips are always visible, whereas others like the tongue are only visible intermittently). There are a variety of ways in which these have been mapped. This is considered to be outside of the scope of this thesis, and so has not been covered here.

## 2.3 multimodal speech phenomena

There are several speech phenomena referred to in this thesis. This section provides a brief summary of these phenomena for reference purposes.

### 2.3.1 *Cocktail Party Problem*

The Cocktail Party problem was first defined by Cherry [38] in 1953, and describes the ability of human listeners to be able to listen to a single speech source while unconsciously filtering out irrelevant background information such as music or competing speech sources. This phenomenon was named after the scenario of two people being able to maintain a conversation while ignoring the sound of a lively party with a myriad of competing speakers and other background noise. Despite being defined in 1953, there is no definitive explanation of this effect, and it represents a very active field of research. A detailed review of this effect and its technological application has been carried out by Haykin & Chen [79].

This effect is of relevance to this field of research because it represents a major challenge for those that suffer from hearing loss. The Cocktail Party effect seems to be binaural, with implications for sound localisation, and those with hearing aids or with 'unbalanced hearing' can particularly struggle to cope with noisy environments. There has also been much research into solving this problem using speech filtering technology. An example of this is Blind Source Separation (BSS), with many recent examples of research aiming to solve this problem such as work by Rivet [151]. Current research along these lines will be discussed in depth in the next chapter.

### 2.3.2 *McGurk Effect*

One phenomenon relevant to multimodal speech perception is called the McGurk Effect. This was first reported in 1976 by McGurk & MacDonald [126]. The significance of this effect with regard to the work carried out in this thesis is that it serves as a physical demonstration of the relationship between hearing and vision in terms of speech perception. Essentially, when a video is played of the syllable 'ga', the viewer recognises it correctly. However, when the video is dubbed with the audio for 'ba', the viewer hears a third syllable, 'da'. The conflict between the sound and the visual effect of the mouthing of a syllable results in the viewer hearing something different.

### 2.3.3 *Lombard Effect*

The Lombard Effect, discovered in 1909 by Lombard [120], describes the tendency of speakers to attempt to improve the audibility of their voice in loud environments by involuntarily increasing their vocal effort. This applies to more than just volume, as research has shown that in very noisy environments, speakers can change their pitch, frequency, duration, and their method of vocalisation in order to be heard more clearly. This is not a voluntary phenomenon, in that speakers have no control of it, and an example of this effect is found in deaf people, who often speak loudly due to their hearing loss, in order to hear their own voice more clearly. Examples of the Lombard Effect can be found in work by Lee *et al.* [110], who presented the AVICAR corpus (which will be discussed in more detail in chapter 3), with speech recorded at different noise levels. Although the work presented in this thesis does not directly consider this problem, it has implications for speech processing research. One example is that many visual speech estimation or recognition models are trained on clean speech, and so performance may be negatively affected when the input visual signal does not match the model exactly.

## 2.4 AUDIOVISUAL SPEECH CORRELATION BACKGROUND

There is an established relationship between audio and visual aspects of both speech perception and production. This relationship between the two speech modalities can be expressed and calculated as the correlation between audio and visual speech features. There are various methods of calculating correlation such as MLR (Almajai & Milner [10], Cifani *et al.* [45]) and CCA (Hotelling [84], Sargın *et al.* [159]), and a significant quantity of research has been published investigating audiovisual speech correlation. This section presents a brief review of recent research into audiovisual correlation, and then section 2.5 presents the results of further investigation by the author into multimodal correlation.

Relatively early examples of audiovisual correlation can be found in research by Yehia *et al.* [184], Jiang *et al.* [96], and Barker & Berthommier [21]. Yehia *et al.* [184] used Line-Spectral Pairs (LSP) audio features and examined the correlation between these and 3D marker points (calculated using infrared LEDs physically placed on the speaker). This early research found an average correlation of 0.73. Barker & Berthommier [21] investigated correlation between face movement (not confined to lip features alone) and LSP audio features, and found a slightly lower correlation (0.55).

More recent work by Sargın *et al.* [159] makes use of CCA to investigate audiovisual correlation. CCA was first defined in 1936 by Hotelling [84] and was used by Sargın *et al.* [159] to analyse the linear relationships between multidimensional audio and visual speech variables by attempting to identify a basis vector for each variable, that then produces a diagonal correlation matrix. CCA maximises the diagonal elements of the correlation matrix. The main difference between CCA and other forms of correlation analysis is the independence of analysis from the coordinate system describing the variables. Ordinary correlation analysis can produce different results depending on the coordinate system used, whereas CCA finds the optimal coordinate system.

Sargın *et al.* [159] made use of the commonly used MFCC technique for audio features, and then compared the correlation when using three different types of visual features. They utilised 2D-DCT features taken from image frames, 2D-DCT features taken from optical flow features, and also predefined lip contour co-ordinates, and compared the correlation of each to audio features. This work concluded by confirming that individual correlation was greatest when performing 2D-DCT on optical flow vectors. This was then extended by Sargin *et al.* [160] in later work to solve a different (speaker identification) research challenge.

Recent work by Almajai & Milner [10] also investigated the degree of audiovisual correlation between multiple audio and visual features. Almajai & Milner [10] used filterbank vectors (described later in chapter 4 of this thesis) and the first four formant frequencies (the most significant distinguishing frequency components of human speech) as audio features, and three different visual features, 2D-DCT, Cross-DCT, and Active Appearance Models (AcAMs). AcAMs are a commonly used approach for feature extraction, first developed by Cootes *et al.* [47], and operate by building statistical models of shape and appearance, based on a training set. More details can be found in (Cootes *et al.* [47]).

Of the research by other authors reported in this section, it is work by Almajai & Milner [10] that is of most interest in the context of this thesis. The correlation work performed by Sargın *et al.* [159] served as a background to work by the author in Abel *et al.* [2] (not reported in this thesis). The research by Almajai & Milner [10] is expanded upon in the next section by the author and used as a basis for new experiments into audiovisual speech correlation. While there has been much research into audiovisual speech correlation, less investigation into the effect of noise on audiovisual correlation has taken place, and this will be discussed in the next section. Almajai & Milner [10] make use of MLR (described in more detail in the next section) to calculate correlation, and primarily investigate the correlation between filterbank vector sizes and visual feature dimensionality.

The following conclusions were drawn by Almajai & Milner [10]. As the size of the visual vector increased, correlation increased. However, this initially increased rapidly until a dimensionality of 24, and then stabilised. Increasing the size of the vector further was not found to make a particularly significant difference. On the other hand, decreasing the size of the filterbank vector increased correlation. This was argued to come at the cost of potentially useful speech information. It was also concluded that Cross-DCT resulted in a lower correlation than using 2D-DCT or AcAMs. Due to the simplicity of using 2D-DCT visual features rather than AcAMs, it was recommended that the 2D-DCT visual feature technique be used for further research. Additionally, Almajai & Milner [10] found after an investigation of phoneme specific versus global speech correlation that phoneme specific correlation analysis also resulted in an increased correlation. This work by Almajai & Milner [10] was used as the basis for subsequent audiovisual speech enhancement research (Milner & Almajai [130]), that will be described in more detail in chapter 3.

## 2.5 MULTIMODAL CORRELATION ANALYSIS

In addition to the audiovisual correlation work presented in the literature, additional correlation research has been published by the author investigating the relationship between audio and visual elements of speech. This research has been published by the authors (Abel *et al.* [2], Cifani *et al.* [45]), and the work presented by the author in Cifani *et al.* [45] is described in this section. There were several reasons why this particular research was of interest. Firstly, as described in the previous section, much research into audiovisual correlation has been carried out by a variety of researchers; however investigation into multimodal correlation with noisy speech mixtures is much scarcer. However, this is of deep interest with regard to the research presented in this thesis as it concerns the effect of speech filtering on multimodal correlation. This work also served as the preliminary to further speech enhancement research, in that possible techniques for feature extraction were investigated and evaluated with a view to the application of more sophisticated speech processing. Finally, this work also serves as a comparison and validation of similar preliminary work by Almajai & Milner [10], as it uses a different corpus but with some overlap in the techniques used.

The remainder of this section describes experiments that show audiovisual correlation (using the MLR technique) between audio and visual features in reverberant noisy environments. All experiments made use of the well-known MFCC technique as the audio feature extraction method, and 2D-DCT and Cross-DCT were both tested as potential visual feature extraction approaches. The use of AcAMs was considered, but it was ultimately decided that based on existing work in the literature, AcAMs were not found to deliver a performance difference of a degree to justify the complexity involved in the configuration and training of suitable AcAMs. The first set of experiments compares individual correlation of sentences from the multimodal VidTIMIT (Sanderson [157]) corpus to assess the difference between using 2D-DCT or Cross-DCT as the visual feature input. The second set of experiments examines the effect of noisy speech on audiovisual correlation, and also aims to identify the dimensionality of audio and visual feature vectors that maximise correlation. Finally, noise was also added to sentences from the corpus in order to test the difference in correlation between noisy speech and filtered speech enhanced with a beamformer. The beamformer exploits the spatial diversity of noise and speech sources to enhance speech, and makes up a part of the proposed speech enhancement system discussed in chapter 4. The next section describes the MLR correlation technique, and after this, section 2.5.2 presents the results of these experiments, which are then discussed and summarised in section 2.5.3.

### 2.5.1 *Correlation Measurement*

In order to carry out audiovisual correlation, Multiple Linear Regression is used. This is a multivariate approach that assesses the relationship between audio and visual vectors (Chatterjee & Hadi [36]). In the analysis discussed in this chapter, experiments have been carried out by using an audio frame of 25ms and a video frame of 100ms. These were chosen based on the parameters also used by existing work in the literature (Almajai & Milner [10]). This implies that the same visual features are used for four consecutive audio frames. For a speech sentence, each component $F_a(l, j)$ of the audio feature vector is predicted by means of MLR, using the entire visual feature vector $F_v(l, q)$, where $l$ is the time-frame index. This approach mirrors that taken by Almajai & Milner [10], and means that using $Q + 1$ regression coefficients $\left\{b_{j,0}, ..., b_{j,q}, ..., b_{j,Q}\right\}$ the $jth$ component of the audio feature vector can be represented by the visual feature vector $F_v(q) = [F_v\left(l, 0\right), ..., F_v\left(q\right), ..., F_v\left(Q - 1\right)]$,

$$\hat{F}_a\left(l, j\right) = b_{j,0} + b_{j,1}F_v\left(l, 0\right) + ... + b_{j,Q}F_v\left(l, Q - 1\right) + \epsilon_l \tag{2.1}$$

With $\epsilon_l$ representing an error term. The multiple correlation between the $jth$ component of the audio feature vector and the visual vector, calculated over $L$ frames, is given by $R_s$, and is found by calculating the squared value:

$$R_s\left(j\right)^2 = 1 - \frac{\sum_{l=0}^{L}\left(F_a\left(j\right) - \hat{F}_a\left(j\right)\right)^2}{\sum_{l=0}^{L}\left(F_a\left(j\right) - \bar{F}_a\left(j\right)\right)^2} \tag{2.2}$$

$\bar{F}_a\left(j\right)$ represents the mean of the $jth$ component of the audio features vector. In this work, the single correlation value $R$ is found by calculating the mean of each $jth$ component of $R_s$. By this, we mean that $R_s$ returns a vector of correlations, with each value representing the correlation of one audio component to the entire visual vector, and the mean of $R_s$ produces the single correlation value $R$, which represents the correlation of the entire audio vector to the entire visual vector.

### 2.5.2 *Multimodal Correlation Analysis Results*

*Comparison of Visual feature Extraction Techniques*

Twelve sentences were used from the multimodal VidTIMIT corpus, and the audio signal of each sentence was sampled at 8 kHz, and processed at 100fps. This was converted into MFCC features with 6 components used for correlation analysis. The matching visual signal for each sentence was recorded at 25 fps, and was interpolated to 100fps to match the input audio signal, before having Cross-DCT and 2D-DCT transforms performed, with 30 components used for each. The Discrete Cosine Transform (DCT) was originally developed in 1974 by Ahmed *et al.* [6], and is a close relative of the Discrete Fourier Transform (DFT). This was extended for application with image compression by Chen & Pratt [37]. The one-dimensional DCT is capable of processing one-dimensional signals such as speech waveforms. However, for analysis of two dimensional signals such as images, a 2D-DCT version is required. This will be described in more detail in chapter 4. For a matrix of pixel intensities, the 2D-DCT is computed in a simple way: the 1D-DCT is applied to each row of the matrix, and then to each column of the result. Thus, the transform is given by the DCT matrix. Since the 2D-DCT can be computed by applying 1D transforms separately to the rows and columns, the 2D-DCT is separable in the two dimensions. To be used as a feature vector, the 2D-DCT matrix is then vectorised in a zigzag order. Cross-DCT is an alternative approach to consider and consists of taking only the central horizontal row and vertical column of the matrix of pixel intensities and then applying 1D-DCT to each vector. This is a much simpler method, and may contain adequate information for lip reading because the former captures the width of the mouth while the latter captures the height of the mouth. The two vectors are truncated and concatenated to get the visual feature vector.

The correlation of the MFCC component for each selected sentence to the appropriate matching 2D-DCT vector is compared to the MFCC to Cross-DCT correlation of the same sentence. This is shown in figure 2. This graph plots matching sentence pairs of audiovisual correlation when using the two different methods of visual feature DCT. The left side of figure

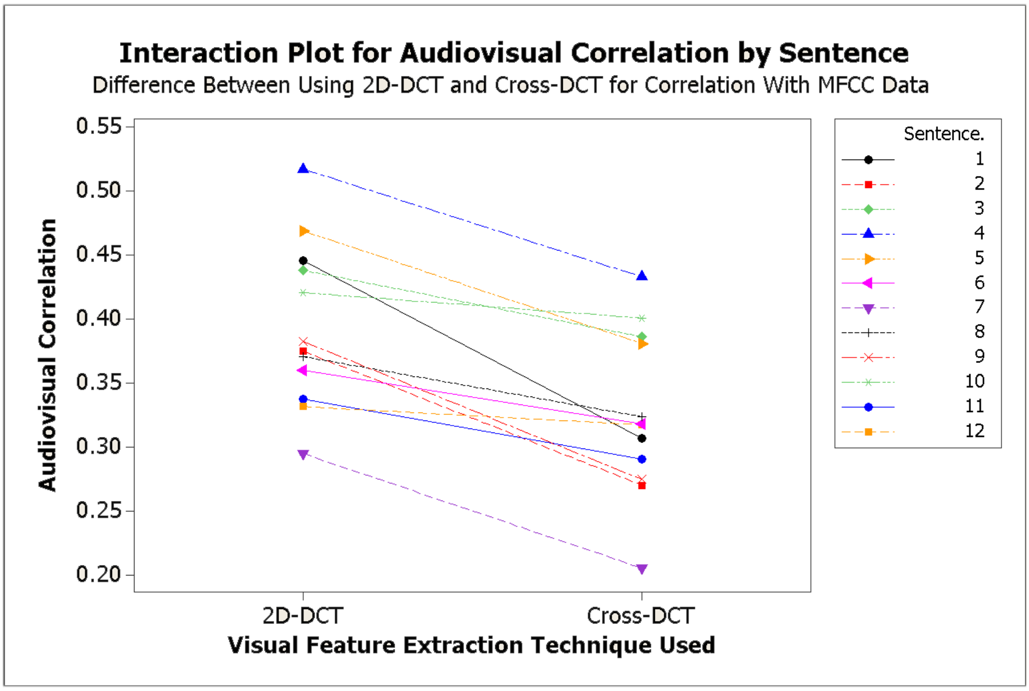


Figure 2: Interaction plot of audiovisual correlation for twelve sentences from the VidTIMIT corpus, showing the difference in correlation when using 2D-DCT or Cross-DCT as the visual feature.

2 shows the correlation found for each sentence when using 2D-DCT. The right side shows the correlation found for each sentence when using Cross-DCT. Matching sentences are linked by a line showing the difference in correlation.

Figure 2 shows that the audiovisual correlation of a sentence from the VidTIMIT corpus found when comparing MFCC correlation to 2D-DCT visual features was greater in all tested cases than for the equivalent correlation when using Cross-DCT. As the interaction plot shows, this result was found with all chosen sentences. This was an expected result because 2D-DCT makes use of all the visual information present in the mouth region of a speaker, whereas Cross-DCT only takes a limited sample of the available visual information.

### *Maximising Audiovisual Correlation*

This section describes experiments carried out to investigate the ideal audio and visual feature vector dimensionalities to use for performing multimodal correlation analysis. The performance of the beamformer was also assessed. This was done by adding white noise to 16 sentences from the VidTIMIT corpus in order to produce noisy speech with a SNR of -3dB, and making use of the beamformer to remove the added noise and produce enhanced

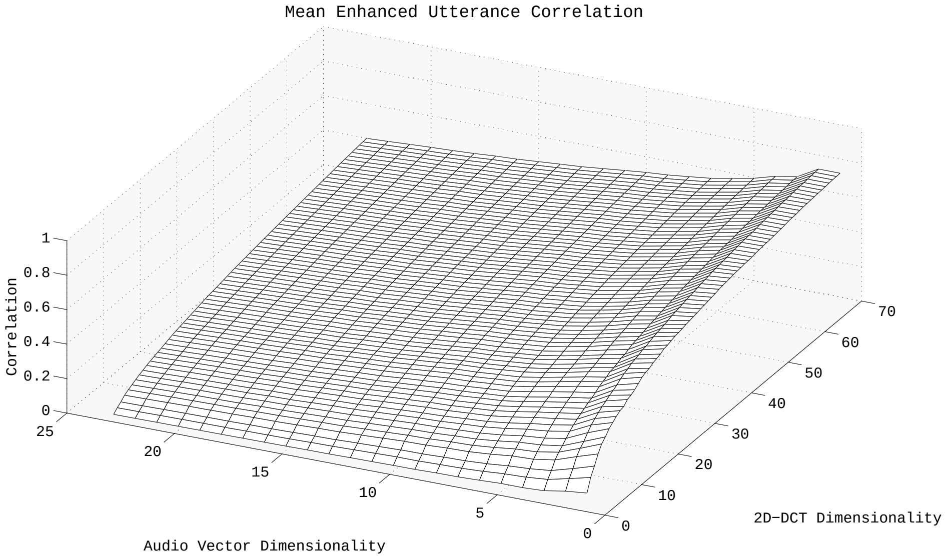


Figure 3: Plot of enhanced speech correlation when varying audio and visual vector dimensionality. This figure shows the results for a mean of 16 sentences from the VidTIMIT corpus.

speech. For the experiments described in this chapter, four microphones were used. The configuration used in this work was subsequently used to produce the initial two-stage speech enhancement system presented later in this thesis, and is described in chapter 4. To find the right combination of audio and visual vectors that maximise correlation, the white noise was removed with the beamformer to produce enhanced speech. The correlation of this enhanced speech, found by performing MLR as described in the previous section, when varying the MFCC and 2D-DCT vector sizes used is shown in figure 3. The correlation of the mean of 16 sentences is shown in this figure, plotted against an audio vector varying between a dimensionality of 1 and 23, and a visual vector that varies in size between 1 and 70. As can be seen in figure 3, a very clear pattern can be seen. Increasing the size of the visual vector increases the correlation, and reducing the size of the audio vector produces a similar effect, peaking at very high (70) visual and very low (3) audio vector dimensionalities, which initially seems to differ from that found in the literature (Almajai & Milner [10]). Almajai & Milner [10] found that the increase in correlation levelled off beyond a visual vector size of 30.

However, this is a misleading result. Figure 4 shows the same mean of 16 sentences, but with the correlation of the visual vector against noisy speech. This obviously results in a lower level of correlation being found. However, a visual comparison of the noisy and enhanced figures in

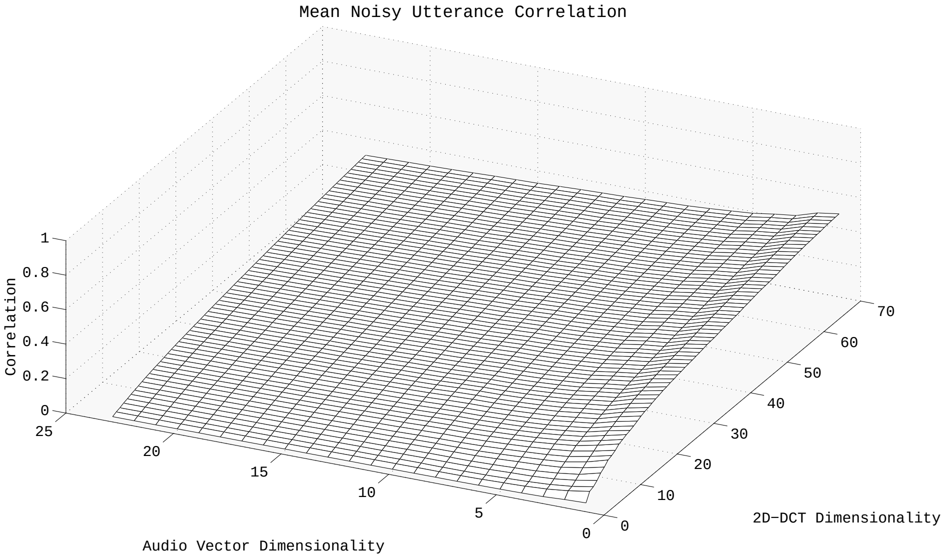


Figure 4: Plot of noisy speech correlation when varying audio and visual vector dimensionality. This figure shows the results for a mean of 16 sentences from the VidTIMIT corpus.

(figure 4 and figure 3) show that they have a similar shape, and where the enhanced correlation is very large; the noisy correlation is also very large. Therefore, it becomes important to find the audio and visual vector sizes that maximise the difference between noisy and enhanced speech. Figure 5 shows the mean of these values.

Figure 5 shows the difference in audiovisual between noisy and enhanced speech correlation plotted against varying audio and visual vector sizes for a mean of 16 sentences respectively. Figure 5 shows that with a very small visual vector, the difference in audiovisual correlation is very small, and that an initial increase results in an increased difference between noisy and enhanced correlation. However, this increase tails off when the visual vector is increased above thirty, with only a very small rate of increase in difference found, and that there is no significant gain to be achieved from increasing this above thirty, matching and validating results found in the literature by Almajai & Milner [10]. Additionally, figure 5 shows that increasing the size of the MFCC dimensionality results in a lower difference in audiovisual correlation and that the highest difference is found with an audio vector size of less than five. However, this is not a practical value to use. A very low MFCC dimensionality contains less spectral information about the input speech, and a compromise between maximising correlation and feasibility for complex speech processing use has to be found.

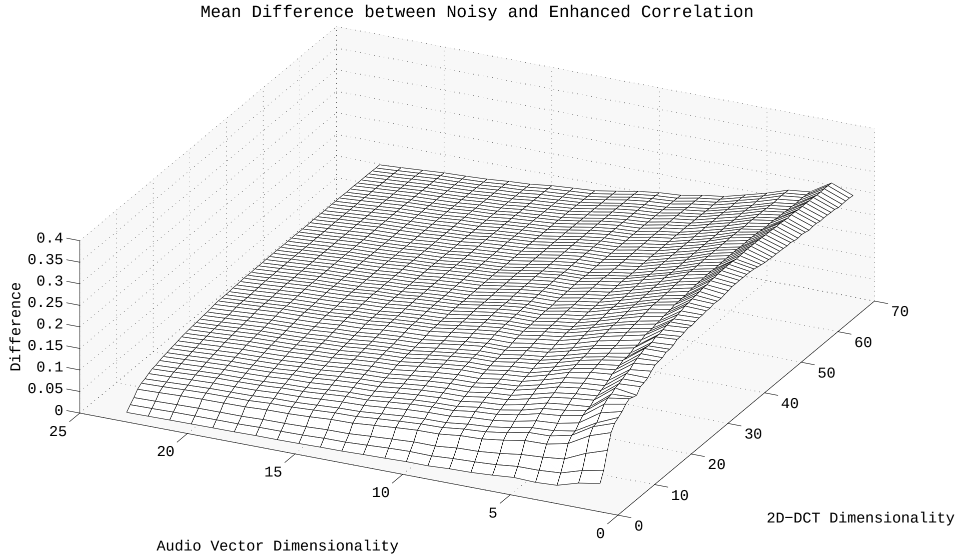


Figure 5: Plot of difference between noisy and enhanced correlation values when varying audio and visual vector dimensionality. This figure shows the results for a mean of 16 sentences from the VidTIMIT corpus.

*Investigation of Noisy Environments*

The identified audio and visual vector dimensionalities that produced the greatest difference in correlation between noisy and enhanced speech signals were used to investigate audiovisual correlation in a range of reverberant noisy environments. As with the previous experiments, MLR was used to measure audiovisual correlation. Based on the results of the experiments in the previous section, it was decided to make use of an MFCC vector dimensionality of 6, and a visual vector size of the first 30 values. To investigate, four types of noise were added to a selection of sentences from the VidTIMIT corpus. Three mechanical noises were chosen, white machine noise, filtered pink noise, and recorded aircraft cockpit noise. An incoherent human babble mixture was also chosen to simulate a busy social environment. The results of these experiments are shown in table 1. Each row shows the mean and the variance of the audiovisual correlation for eight sentences from the VidTIMIT corpus, mixed with different noises, in the noisy and enhanced cases.

Table 1 shows a consistent and statistically significant difference between noisy and enhanced speech correlation in all four types of noise. In every case, the enhanced audio vector for a sentence produces a higher correlation with visual information than the noisy signal

Table 1: Selected results of Bonferroni Multiple Comparison, showing P-Value results for comparison of audiovisual correlation in noisy and enhanced sentences from the VidTIMIT corpus.

| Noise | Mean | | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|---|---|
| | Noisy | Enhanced | | | | |
| Babble | 0.334 | 0.417 | -0.083 | 0.005 | -16.40 | 0.000 |
| F-16 | 0.252 | 0.410 | -0.1578 | 0.005 | -31.16 | 0.000 |
| Pink | 0.229 | 0.415 | -0.1866 | 0.005 | -36.86 | 0.000 |
| White | 0.239 | 0.408 | -0.1690 | 0.005 | -33.38 | 0.000 |

does. This was an expected result, and confirms the usefulness of the beamformer for filtering an audio signal before audiovisual correlation analysis is performed. These experiments also showed a much greater difference in correlation between noisy and enhanced machine noisy speech mixtures (white, pink and aircraft noise), than for incoherent babble mixtures. This smaller difference between noisy and enhanced values in environments containing significant levels of background speech may be explained by the relative similarity of sentences from the corpus and the background speech mixture. In the case of babble, while the enhanced correlation is similar to other results, the noisy speech correlation is much higher for all sentences, demonstrating a lower difference between noisy and enhanced audiovisual speech correlation, due to the incoherent babble noise being similar to speech, potentially causing an incorrect correlation result.

### 2.5.3 *Discussion of Results*

Based on the results described above, there are a number of points that can be discussed. Firstly, the research above represented a successful preliminary test of some of the techniques used later in this thesis for speech enhancement (which will be discussed in chapter 4). Of particular interest is the visual feature extraction technique, with Cross-DCT proving to be less effective with regard to correlation than 2D-DCT. It was also shown in the results above that the use of a beamformer to filter speech resulted in improved audiovisual correlation. Overall, the techniques discussed in these experiments are of interest for future development.

This work also validates results published by Almajai & Milner [10]. The research presented here has similarities, but makes use of a different corpus, and also uses a beamformer to enhance noisy speech. However, despite these differences, the results are similar to other published work, and find that the ideal dimensionality of the audio vector is around six to twelve, and the ideal visual vector size is around thirty to forty, which is in line with the findings of Almajai & Milner [10]. The individual correlation results were found to be lower to those reported by others. It is hypothesised that this is because the work described in this chapter makes use of a different corpus (VidTIMIT), which contains a great deal of background noise, with some sentences having a poor audio quality. This has the effect of producing lower levels of audiovisual correlation than would be found when using a cleaner multimodal speech corpus.

As mentioned, the results presented here also demonstrate the effectiveness of using a beamformer to increase multimodal correlation. This, to the knowledge of the author, demonstrated an original contribution of this thesis, because there is very little investigation in the literature into audiovisual correlation with noisy speech. In the results, it could be seen that a range of noises, including white noise, aircraft cockpit noise, and incoherent speech babble, were added to sentences from the audiovisual corpus in order to investigate the performance of a beamformer with regard to multimodal correlation in noisy reverberant environments. It was found that speech enhanced with a beamformer always produced a higher audiovisual correlation, than results with noisy speech. It was also found that the difference in audiovisual correlation was much larger when noise such as white noise or aircraft cockpit noise was added to the corpus, than when human speech babble was added. The addition of incoherent speech led to a much higher correlation in noisy speech, suggesting that audiovisual correlation is much more accurate in some environments than others. This is a justification for the stated aim of this thesis in moving towards the development of a speech enhancement system that is able to work well in a range of different environments.

## 2.6 SUMMARY

As part of the process of the development of an audiovisual speech enhancement system, this chapter provided some information into the relationship between audio and visual aspects of speech, briefly summarising some of the basic concepts involved in the production of human speech. Some speech phenomena relevant to this thesis were also discussed, namely the Cocktail Party Problem, the McGurk Effect, and the Lombard Effect. This initial part of the chapter provides the fundamental background to the work discussed in the rest of the thesis, covering the concepts that the contributions presented in subsequent chapters are built on.

The relationship between audio and visual aspects of speech was then discussed, describing the extraction of audio and visual features, and the calculation of multimodal correlation. Multimodal correlation has been investigated in the literature, and a summary of existing research was given in this chapter, most significantly, recent work by Almajai & Milner [10]. Some novel experiments were then described. These were first reported in work carried out by the author in Cifani *et al.* [45], and represent an original contribution of this thesis. The experiments reported in this thesis served to validate existing correlation research by Almajai & Milner [10], and also served as a preliminary evaluation of some of the techniques such as 2D-DCT visual feature extraction, and the effect of beamforming. Several experiments were carried out. Firstly, two different visual extraction techniques were compared. A number of sentences were used for comparison, and correlation analysis was carried out (calculated with MLR) on sentences using MFCC audio features and both 2D-DCT and Cross-DCT as the visual feature vectors. It was shown that using 2D-DCT vectors as the visual feature produces a consistently higher correlation than using Cross-DCT. An investigation into maximising correlation was carried out, and the effect of noise and subsequent speech filtering with beamforming was examined. These results demonstrated the positive effect of beamforming on multimodal correlation, and represented an original contribution of this thesis.

This chapter provided a general introduction to the principles behind this work, namely the relationship between the audio and visual speech modalities, and chapter 3 presents a more specific focus on the research context. Chapter 3 builds on the background presented here to review the technical features of current hearing aids, investigate a number of multimodal speech databases to evaluate suitability for use in further research, and crucially, provides a detailed review of the state-of-the-art audiovisual speech enhancement work in the literature.

# 3

# THE RESEARCH CONTEXT

## 3.1 INTRODUCTION

As stated in chapter 1, the ultimate aim of this research is to develop a multimodal speech enhancement framework, with a focus on long term application to future hearing aid development. Chapter 1 introduced the thesis and presented the motivation and goals of this research. Chapter 2 presented the justification of this work by describing the relationship between audio and visual aspects of speech, and also investigated and demonstrated the correlation between the two speech modalities.

This chapter presents a literature review that places the research proposed in this thesis in context, building on the background presented in the previous chapters. Firstly, the overall speech processing domain is briefly discussed. To place the proposed framework in the context of commercially available hearing aid technology, a review of the methods used in modern hearing aids is provided. This review presents examples of listening devices using directional microphones, array microphones, noise reduction algorithms, and rule based automatic decision making, demonstrating that the multimodal two stage framework presented later in this thesis has established precedent in the context of real world hearing aid devices. The other significant aspect vital to the research context of this work is the field of audiovisual speech filtering. This chapter presents a review of multimodal speech enhancement, with a discussion of the initial early stage audiovisual speech filtering systems in the literature, and the subsequent development and diversification of this field. A number of different state of the art speech filtering systems are examined and reviewed in depth, particularly multimodal beamforming and Wiener filtering. Finally, in order to perform the experiments required in this research, a suitable audiovisual speech database has to be chosen. In recent years, there has been a large increase in the range of available corpora, and several audiovisual speech databases are evaluated to assess their suitability for use in this project.

The remainder of this chapter covers a number of different areas. Firstly, section 3.2 presents a general overview of speech processing. This is followed by a review of contemporary hearing aid technology in the next section, demonstrating that the two stage system proposed in this thesis, and the extension of this to use fuzzy logic, has an established single modality precedent. Section 3.4 then presents a summary of audiovisual speech enhancement research and a detailed review of three different state of the art speech filtering techniques. A review of lip detection and tracking is presented in section 3.5.The next section reviews a number of audiovisual speech corpora, and finally, this chapter is summarised in section 3.7.

## 3.2 speech processing overview

Speech processing is an established and very active field of research with many different areas of focus, such as recognition, enhancement, and synthesis. The work presented in this thesis is solely focused on speech enhancement with the aim of improving speech quality by filtering a noisy speech input signal to remove noise. In recent decades, many different audio-only speech enhancement solutions have been proposed, such as those by Zelinski [189] and Hussain & Campbell [90]. There are many examples of modern single modality speech processing algorithms in the literature (Hussain *et al.* [91], Van den Bogaert *et al.* [173], Li *et al.* [113]). A common technique is to use multiple-microphone techniques such as beamforming that can improve speech quality and intelligibility by exploiting the spatial diversity of speech and noise sources (Gannot *et al.* [67], Griffiths & Jim [73]). Another possible technique is to make use of Wiener filtering (Wiener [180]) to compare a noisy speech signal to an estimate of the equivalent noise-free signal to produce an enhanced signal (Van den Bogaert *et al.* [173]). There are also approaches such as that proposed by Zelinski [189] and refined by others, including Li *et al.* [113, 114], Van den Bogaert *et al.* [173], that propose a two stage audio-only speech enhancement solution that makes use of both adaptive beamforming and Wiener filtering in a single system. There are many different other audio-only single modality noise cancelling techniques in the literature.

This thesis focuses on the use of multimodal information for the extraction of an enhanced speech signal from a noisy speech mixture, and so a detailed review of audio-only speech enhancement techniques is considered to be outside the scope of this thesis. However, because the context of this research is the development of a framework for the aid of those who use hearing aids, the next section focuses on reviewing the technology used in current commercial hearing aids. This places the proposed research into context, and this is followed by a detailed description of multimodal speech filtering algorithms in section 3.4.

## 3.3 APPLICATION OF SPEECH PROCESSING TECHNIQUES TO HEARING AIDS

### *Introduction*

This thesis considers speech filtering from the point of view of potential application to hearing aids for the benefit of users with deafness. However, this is very much a long term focus, with the system presented in this thesis focusing exclusively on early stage software development. However, it is considered appropriate to provide an overview of some state of the art features of modern hearing aids. Much of the content in this section is adapted from a detailed review by Chung [44], and this section provides a summary of current technology.

### *Directional Microphones*

In the detailed review by Chung [44], current hearing aid technology is divided into two categories, directional microphones and noise reduction algorithms. With regard to directional microphones, the most relevant topics which are summarised in this thesis are first order directional microphones, adaptive directional microphones, second order directional microphones, and assistive array microphone devices. Overall, directional microphones are an established technology that has been around since the 1970s (Chung [44]). They operate on the premise that speakers are more likely to be located to the front of the listener, and so directional microphones are designed to be more sensitive to sounds arriving from the front of the speaker. This technique is of relevance to this thesis because it demonstrates a practical use of speech filtering techniques, and shows the use of multiple microphones for directional filtering, as utilised in chapter 4 of this thesis.

FIRST ORDER DIRECTIONAL MICROPHONES The most common type of directional microphone processing was identified by Chung [44] as being first order directional microphones. These are designed to direct the focus of microphone sensitivity to sounds coming from the front of the listener and reduce sensitivity to sounds arriving from the side or rear of the listener. Hearing aids equipped with this technology can be designed to use single or dual microphones. In the single microphone configuration, a hearing aid has a microphone with two ports, anterior and posterior. Sound entering the posterior port is delayed and subtracted from the input to the anterior port. This delay is determined by physical factors such as the distance between the two microphone ports. These have been superseded in recent years by dual microphone hearing aids, which work in a similar fashion. These have two omnidirectional microphones, with an anterior and posterior microphone. The microphone inputs are combined with 'delay and subtract' processing, similar to that discussed for single microphone hearing aids. The difference is that the delay and therefore the directional focus is software based and can be adjusted and programmed with signal processing algorithms (Ricketts & Mueller [145]).

With regard to performance, it is consistently found in laboratory testing that the use of directional microphones can lead to improved results, as discussed in Valente [172]. The more focused the directionality is, the better the results. It was also found that optimum performance was produced when there were less discrete noise sources, and less reverberation. This is because a large room produces a lot of reverberation and these echoes reduce the effectiveness of directional focus. However, in actual practical usage, many users do not perceive the benefit of directional microphones. There are several reasons for this. Firstly, the desired signal is not always located directly to the front of the listener, many times; the speaker will be at an angle to the listener (Kuk *et al.* [106]). There are also a wide variety of reverberation environments in real life, with many unsuited to directional microphones, and so the large improvements seen in laboratory experiments do not always translate well into practical use.

Another issue is that research found that the majority of hearing aid users simply do not use the directional setting. Modern hearing aids can come with a variety of different settings, two of which can be omnidirectional and directional microphone modes. Studies by Cord *et al.* [48, 49] found that there was little active switching by users between modes, with the majority preferring to use the omnidirectional mode at all times. Many users did not notice the benefit of directional microphones, or even found the performance to be worse than using the standard omnidirectional settings. There are significant limitations to using directional microphones in many environments. For example, in quiet environments, when there is wind noise present, or when reverberation is at a significant level, omnidirectional microphones are suggested, and many users prefer to use this setting at all times. These recommendations are due to a number of technical limitations with this type of hearing aid. Firstly, at low frequencies, wind noise can dominate and affect directional processing, meaning that omnidirectional microphones are recommended. Also, the internal noise from dual microphone hearing aids is louder than for the equivalent single microphone systems. Often, this internal noise is masked by environmental noise, but louder dual microphone noise is more noticeable. There is also the issue of low frequency roll off. This is when the delay and subtract processing is carried out with low frequencies, and sounds are subtracted at a similar phase. This can result in under amplification of low frequency sounds and a 'tinny' sound quality output. So although very commonly used, directional microphones have significant limitations.

ADAPTIVE DIRECTIONAL MICROPHONES To improve upon the limitations of directional microphones, adaptive directional microphone techniques have been developed. While first order directional microphones assume a fixed source location at the front of the listener, adaptive microphones do not make this assumption, and are designed with the aim of maximising sensitivity in the direction of the dominant source location and minimising sounds originating from the opposite direction to this source. This makes adaptive dual microphone hearing aids theoretically more suited to real world environments.

In terms of functionality, adaptive microphone hearing aids function in a similar way to the first order directional microphones described above, but the direction and microphone settings are not fixed. These settings (such as the posterior delay and omni or uni directional mode) can be automatically changed with sophisticated signal processing. Different manufacturers use a range of proprietary methods, which are not generally publicly disclosed, but Chung [44] provides a general overview of the functionality of adaptive microphones, specifying four steps. Firstly, signal detection and analysis is carried out. Secondly, the appropriate microphone mode is determined. In the third stage, the polar pattern to be used is decided, and finally, the chosen configuration is executed. The signal detection stage involves an analysis of the inputs. Chung [44] mentions a number of sensors that various manufacturers use in hearing aids. A level detector is often used. This measures the level of the incoming sound, with the principle that directional microphones should only be used if the input level is above a set threshold. A modulation detector may also be used. This can be used as a Voice Activity Detector (VAD) to detect the presence of speech in an incoming signal. Speech generally has a modulation between 2 and 50 Hz (Rosen [154]), and a centre modulation between 4 and 6Hz (Houtgast & Steeneken [85]), with noise generally outside this range. Therefore, the modulation detector can be used to estimate the presence of speech, and also for SNR estimation. A hearing aid may also be equipped with a wind detector, and also a front/back level detector. This detects the difference in level between front and rear microphone inputs, which can aid with identification of the location of the dominant source.

The second step is the determination of the operational mode. This is simply the decision whether to use the omnidirectional or directional microphone setting. Although switching can be manual in some hearing aids, it is often automatic, and takes account of a variety of factors, as mentioned above. Chung [44] gives a prominent example of this in the Oticon Syncro hearing aid, which has three different operational modes. These are; full directional mode; directional mode at certain frequencies; and omnidirectional mode. To identify the most suitable mode, the level and modulation detectors are used, along with two alarm detectors in the form of a wind detector and a front/back detector. The alarm detectors are given priority, with the microphone mode selection aiming to maximise the SNR input. More detail is given in Flynn & Lunner [61], Flynn [60]. If the directional microphone setting is chosen, then the most suitable polar pattern is then selected. By this, it is meant that the focus of the directional microphone configuration, i.e. the source location identified as being the dominant speech source, is decided. Finally, this configuration is executed.

In terms of results, Chung [44] concludes that adaptive directional microphones have not been found to deliver a worse performance than fixed first order directional microphones in many scenarios, but in scenarios with a very narrow angle of speech, can produce improved results. More detailed results are given in Ricketts & Henry [144].

SECOND ORDER DIRECTIONAL MICROPHONES An alternative to first order directional microphones is to use second order directional microphone hearing aids. First order microphones are normally found to result in a 3 to 5 dB SNR improvement. However, for those with a significant hearing loss, this improvement may not be noticeable. Second order directional microphones utilise more than two microphones, but otherwise use similar delay and subtract processing techniques. The downside to this approach is that there is a lot of low frequency roll off, which is difficult to amplify without having an impact on the amplification of internal noise. Bentler *et al.* [24] tested one such system, the Siemens Triano, and found only small benefits. This design of hearing aid is reported by Chung [44] as being relatively uncommon.

ASSISTIVE LISTENING ARRAY MICROPHONES The previous sections discussed hearing aids that delivered several decibels of SNR improvement. For those with more than 15dB of hearing loss though, the gains are still not sufficient to adequately increase the SNR of a desired speech source in difficult environments. The traditional solution for this is to use a FM radio microphone system. This consists of a microphone placed close to the mouth of a speaker (such as a wearable clip on microphone), which is then transmitted via radio to a receiver worn by a listener. This is then transmitted to the listener's hearing aids directly. This delivers very good performance, but comes at the cost of removing almost all background noise. It is

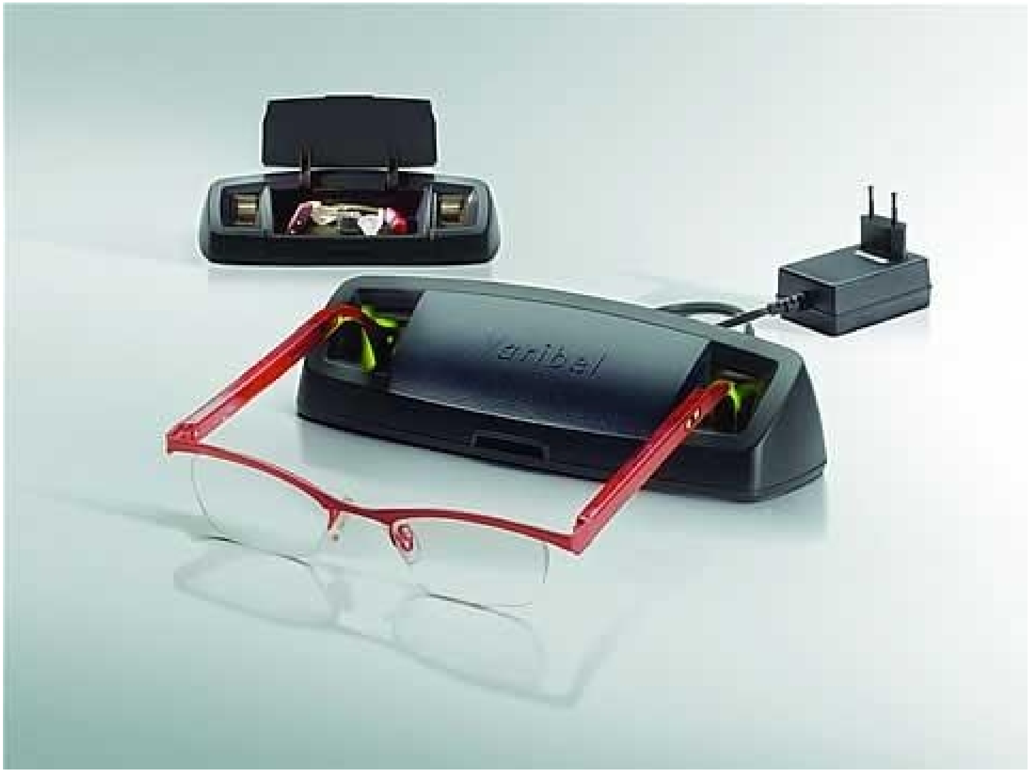


Figure 6: State of the art glasses mounted microphone array, the Varibel Hearing Glasses. Image shows glasses and charging station. Image taken from Varibel website (Varibel-Innovations [174]).

also limited to a single speaker, which is of limited usefulness in a scenario where multiple speakers are involved. One alternative to this traditional approach is to use a microphone array.

A multiple microphone array consists of a series of linked microphones, with the inputs combined to provide directional focus. Delay and sum processing is used with each microphone to increase the directional effect to a greater extent than for conventional hearing aids. The array input then overrides the user's hearing aid input. These arrays can be hand held (i.e. a device that the listener points at the desired target), or head worn (Mens [128]). An example of a state of the art head worn microphone array can be found with the Hearing Glasses, manufactured by Varibel, shown in figure 6. This device has four small microphones positioned in each arm of the glasses to provide directional sound filtering.

There are a number of advantages to microphone array devices such as these. Firstly, because they are directed by the listener, multiple speakers can be listened to, with the listener moving the microphone to suit. There is also the advantage of being extremely precise with directionality. In terms of results, Christensen *et al.* [43] reported a 7 to 10dB improvement, which was backed up by Laugesen & Schmidtke [109].

However, there are limitations with these directional microphone arrays. Firstly, the extreme directional focus results in a very significant loss in noise from other directions. This can make it difficult to use in an environment where multiple speakers are present. Although array microphones are theoretically designed to deal with this by being mobile, it is possible for the listener to miss the start of utterances due to the time taken for the array to be moved and directed correctly. Also, although a 7 to 10dB gain is reported, a traditional FM transmitter system still has a much greater level of improvement, and is still the recommended solution for noisy environments with a single speaker.

### *Noise Cancelling Algorithms*

While directional microphones take advantage of the spatial differences between speech and noise sources, noise reduction algorithms aim to exploit the spectral and temporal differences between speech and noise. A detailed review is provided by Chung [44], and this section presents a summary. As stated, noise reduction differs from the use of directional microphones, and relies on speech filtering software algorithms with decision rules used to decide the appropriate level of filtering. These rules rely on the input from various detectors such as wind noise, signal level, and modulation detectors, as described previously. Again, the exact configuration and use of detectors tends to be proprietary, although the use of a modulation detector is considered to be standard practice. This is because speech tends to have a modulation centred around 4 to 6 Hz, whereas in most cases, noise tends to have a modulation range outside of this, and so the specific modulation of speech can often be detected. Another type of modulation that can be detected is co-modulation. This is generated by the opening and closing of vocal folds during voicing of vowels and voiced consonants (Rosen [154]). The rate can be visualised as spikes on a spectrogram, with spikes showing instances when vocal folds open and darker regions showing instances with no speech energy. These rapid spikes are a good indicator of speech as noise generally does not display this pattern of rapid co-modulation. This rate of co-modulation is known as the fundamental frequency. Chung [44] distinguishes between algorithms that detect the specific modulation of speech (multichannel adaptive noise reduction algorithms), and those that detect co-modulation (synchrony detection noise reduction algorithms).

MULTICHANNEL ADAPTIVE NOISE REDUCTION ALGORITHMS Multichannel adaptive noise reduction algorithms are described by Chung [44] as being the most commonly commercially implemented algorithms. The principle behind them is that they aim to reduce noise interference at frequency channels where noise is dominant. These algorithms are most effective when there is a significant spectral difference between speech and noise, but suffer if the noise source is speech from a competing speaker. In general, these algorithms are described by Chung [44] as having three stages. First, signal detection and analysis is carried out. This is then followed by the application of decision rules, and finally, appropriate gain reduction (adjustment of the ratio of output to input) is carried out.

The first stage, signal detection and analysis, is similar to that for directional microphones, in that a variety of detectors such as wind and modulation detectors are used to analyse the input signal, as described previously in this chapter. Two particular additional features that are specifically of interest with regard to noise reduction algorithms are intensity-modulation-temporal changes in each frequency channel (Tellier *et al.* [171]), and spectral-intensity-temporal patterns across frequency channels (Kuk *et al.* [105]). Intensity-modulation-temporal change detection operates on the basis that one single property is not fully adequate for correctly classifying speech. Of the individual dimensions, intensity can be used, i.e. the amplitude of the signal. The input is divided into events, and it is assumed that high amplitude represents speech and low amplitude noise. This does not take account of changeable conditions though. Another dimension, modulation, is as described previously, and the final dimension considers time varying properties. Tellier *et al.* [171] examines the Conversa from Unitron Hearing, which analyses the input signal in 16 bands and classifies the input as being either stationary noise (e.g. automobile noise), pseudo stationary noise (e.g. incoherent speech babble), transient noise (e.g. a door slam), or the desired input speech or music signal, using the three dimensions

Table 2: A continuum exists between three specific characteristics found in signals (intensity changes, modulation, and duration). and four types of noise. Taken from (Tellier *et al.* [171]).

| | Stationary | Pseudo Stationary | Desirable | Transient |
|---|---|---|---|---|
| Intensity Changes | smallest | < .......................................... > | | largest |
| Modulation Frequency | lowest | < .......................................... > | | highest |
| Time Duration | longest | < .......................................... > | | shortest |

described above, as seen in table 2. Spectral-intensity-temporal pattern analysis works on a similar basis, but considers spectral information rather than modulation.

A key aspect for many hearing aids equipped with multichannel adaptive noise reduction algorithms is to estimate the SNR in each frequency channel. The exact methods used to do this are proprietary and vary between manufacturers, but generally, this is done by calculating the modulation depth of the signal identified as speech. If the depth is calculated to be high at an individual frequency channel, then it is assumed that speech is dominant at that channel. After the initial input, the second phase is the application of decision rules. Again, this varies between individual hearing aids, and Chung [44] reports that a range of factors such as the level of input signal, the SNR at individual channels, and the type of noise reduction programmed for the individual user during the hearing aid fitting process are considered. The outcome of the decision rules is to apply gain reduction at the appropriate frequency channels. Chung [44] states that generally, the amount of gain reduction applied at each frequency channel is inversely proportional to the SNR of the input signal of that channel. The general justification that forms the basis of the decision rules is that if the SNR at a frequency channel is estimated to be high, then it is assumed that speech is the dominant signal and the channel is not filtered. If a moderate to low SNR is estimated, then it is assumed that in that channel, speech and noise co-exist, or noise is dominant. In this case, the gain is reduced at this channel to decrease interference. If no speech is found, then the channel has the maximum level of gain reduction applied.

The actual gain reduction varies between hearing aids, and the application of this is complex and relies on both calculations and individual judgement (Tellier *et al.* [171]). The application of gain reduction considers a number of temporal constraints. The engaging/adaption/attack time, which is the time between the detection of noise at a frequency channel and the application of gain reduction, is considered, as well as the speed of this reduction. The time between the detection of an absence of noise and the ending of gain reduction is also taken into account, as well as the speed of this process. As stated, this is not an exact science, and different balances between listening comfort and performance are found on different hearing aids.

Overall results of these algorithms are reported as being mixed. Although positive results have been reported in the literature by Levitt [112] and Boymans *et al.* [28], others have reported no significant improvements (Alcántara *et al.* [7]). There is also the issue reported by Ricketts & Dahr [146] and Alcántara *et al.* [7] that in broadband noise such as in an automobile, noise reduction of this type can lead to no benefits because the algorithm reduces gain in all frequency channels with noise domination, meaning that speech and noise is reduced at the same rate, leading to no benefits. Generally, the bigger the difference between speech and noise, the greater the benefits of this form of noise reduction.

synchrony detection noise reduction algorithms Synchrony detection noise reduction algorithms take advantage of the detection of co-modulation in speech to distinguish between speech and noise (Elberling [56]). These algorithms essentially detect the fast modulation of speech across frequency channels. The signal detector monitors high frequency parts of the incoming signal and looks for the high frequency spectral spikes that indicate co-modulation. If these rapid bursts of speech energy are detected then it is assumed that speech is present and gain levels are kept at the default level. However, if co-modulation is not detected, then it is assumed that speech is not present and the overall gain of the hearing aid is gradually reduced (Schum [164]).

Chung [44] reports that this algorithm is less useful when the noise source is competing speech because this speech also has co-modulation, and that this algorithm can be combined as part of an overall noise reduction package combining both multichannel and synchrony detection algorithms, rather than being used individually.

*Summary*

This section described some current developments in commercially available hearing aids. It can be seen that there exist some very sophisticated developments in modern hearing aids that utilise the audio modality. Modern hearing aids can include multiple microphones (either as part of an array, or as part of a directional microphone system), and may also contain noise cancelling algorithms to emphasise speech and reduce levels of background noise. There also exist hearing aids that include rule based decision-making to automatically adjust the degree of speech filtering, depending on environmental conditions, as well as those that combine both directional microphones and noise reduction algorithms. This combination of techniques is similar to the multimodal system proposed in this thesis. The research discussed here is entirely focused on utilising the audio modality; with no commercial audiovisual hearing aid released as yet (to the best knowledge of the author) that combines visual information with these audio techniques. With recent technological advances, hearing aid technology is at a level where extremely sophisticated processing can be carried out. The concept of extending the techniques outlined above to create a multimodal hearing aid that combines rule based decision making, noise cancellation algorithms, multiple microphones, and visual information, is much more feasible than was previously considered.

## 3.4 AUDIOVISUAL SPEECH ENHANCEMENT TECHNIQUES

### 3.4.1 *Introduction*

This section presents a summary of several state of the art multimodal speech enhancement techniques. Given the audio and visual speech relationship described previously, and recent correlation research, it was obvious that the concept of audio-only speech enhancement systems would be extended to become multimodal. A pioneering multimodal speech enhancement technique was proposed by Girin *et al.* [69]. This was, to the knowledge of the author, the first example of a functioning multimodal speech enhancement system. This approach made use of the height, width, and area of the lips, recorded using blue lipstick and chroma key technology to record this information and exclude other details, and made use of simple linear regression to estimate a Wiener filter to enhance speech contaminated with white noise. Results performed on selected data from vowel/consonant/vowel sequences showed an improvement over similar audio-only approaches, but were limited due to the linear nature of the filter. A non-linear artificial neural network utilising Multi Layer Perceptrons (MLP) was proposed, and this more complex filter was found to produce improved results. This early work demonstrated the potential of multimodal enhancement, and was developed further by Goecke *et al.* [71], who proposed a linear mean square error estimation method. The research focus of these authors was on the development of Automatic Speech Recognition (ASR) rather than purely enhancement, but this work identified that this simple filtering approach resulted in better ASR performance than noisy unfiltered speech. However, the authors also found that the results of this enhanced speech were inferior to the early integration fusion of visual and noisy audio information. Deligne *et al.* [52] demonstrated a non-linear approach for speech enhancement, Audio-Visual Codebook Dependent Cepstral Normalisation (AVCDCN), which was an extension of an audio-only Codebook Dependent Cepstral Normalisation (CDCN) (Acero & Stern [3], Deng *et al.* [53]) approach. The authors found that limited experiments using the same noise for testing as was used in training showed improved results when the audiovisual approach was used rather than audio-only. This AVCDCN approach was also tested by Potamianos *et al.* [142] as part of a speech recognition system, who found that the non-linear AVCDCN approach outperformed audio-only processing and a linear audiovisual model that was also examined.

Since this early work, there have been many recent advances in this research field. Three of the most relevant state of the art developments that particularly build on this early work are discussed in depth in this section. The concept of non-linear speech enhancement has been built on and expanded into a more sophisticated speech filtering system by Almajai & Milner [12]. Work by a range of authors, including Rivet & Chambers [152] (also Rivet *et al.* [147, 148, 149, 150]) has expanded on the initial speech enhancement work of Girin *et al.* [69] and focused on the development of algorithms for audiovisual source separation. In a manner similar to the AVCDCN approach mentioned above, multimodal speech fragment decoding, developed in 2007 by Barker & Shao [18], aims to improve on existing audio-only fragment decoding techniques. This is not exclusively a speech enhancement system as it combines both recognition and enhancement together. Each of these three strands of state of the art research developments will be described in detail in the remainder of this section.

### 3.4.2 *Audiovisual Blind Source Separation*

*Summary of Work and Previous Papers*

A range of authors including Rivet & Chambers [152] (also Rivet *et al.* [147, 148, 149, 150]) have been involved in the development of multimodal BSS systems, which aim to filter a speech source from a noisy convolved speech mixture. This expands upon previous related work by these authors, and is an extension of audio-only BSS solutions. These authors have proposed several different ways to use visual information to overcome limitations with the single modality BSS approach.

BSS, first proposed by Jutten & Herault [98] (also Herault *et al.* [81]) was designed with the aim of separating individual speech sources from a mixture of competing speakers. The problem of separating speech mixtures and recovering individual speech sources is one that is of great interest to researchers due to its relevance to the Cocktail Party Problem (Cherry [38]). It is similar to speech enhancement in that a single source can be extracted from a mixture, but the goal is to separate sources rather than noise cancellation. It is a difficult problem to tackle as real world speech mixtures are convolved. By that, it is meant that the sources are mixed, reflected off of different surfaces in the speech environment and weakened before they are picked up by the microphones. The 'blind' aspect of the name refers to a lack of knowledge regarding information about the number of sources and the mixing matrix.

There are a number of requirements in order for BSS to be successfully performed. Generally, the sources must be statistically independent, i.e. each source must be independent and uncorrelated from the others, and the speech mixture must be a linear combination of speech sources, with no additional background noise present. Also, in order to adequately separate sources, the number of observations (e.g. microphones) must be at least equal to the number of sources. If the number of microphones is less than the number of sources, then the resulting reconstructed sources produced after BSS will not be single speech sources, but will continue to be mixtures of speech sources. The assumption of independence between the sources is known as Independent Component Analysis (ICA). This approach works by estimating a demixing matrix, which is the inverse of the mixing matrix, to reconstruct a number of independent sources.

Like alternative approaches (Almajai *et al.* [13], Barker & Shao [18]) to multimodal speech processing covered in this section, this problem is tackled in the frequency domain. This means that power spectrum density matrices are used, and for all frequency bins, the power spectrum of a source is a diagonal matrix. Therefore, for efficient BSS, the demixing matrix should be adjusted so that the recovered source power spectrum density is also a diagonal.

There are a number of limitations with BSS. Firstly, as mentioned above, the number of microphones must be the same (or greater) as the number of speech sources. Secondly, the scale of a reconstructed source cannot be determined. As a scalar multiplier could be extracted from a source and multiplied by the mixing matrix, the amplitude assigned to a reconstructed source is just arbitrary. The separated output could be inverted, or of greater or lesser amplitude than the original source. Finally, BSS does not have any prior knowledge of the sources, and so the order cannot be determined. It is not known which source is of interest, and so while BSS may produce a separated source, it may not be the desired output.

When applying the above limitations, this means that the demixing matrix is limited by a diagonal matrix to represent the scale, and a Permutation matrix to represent the order indeterminacy, i.e. the source of interest. In the frequency domain, in order to ensure a good quality reconstruction of sources, the scale and permutation need to be the same for each frequency bin. If each permutation is the same, this means that a single reconstructed source comes from only one original source, meaning there is no interference from other sources. If the scale is the same, this ensures that the amplitude and reconstruction is correct. Various audio-only approaches have been attempted with varying degrees of success.

The research presented in this section attempts to use visual information to tackle these limitations. A detailed review is presented in a recent paper by Rivet & Chambers [152], and is summarised here. In all of the examples presented, visual information is used to assist with the estimation of the permutation and diagonal matrices. This work builds upon previous related work by Girin *et al.* [70], Jutten & Herault [98], Sodoyer *et al.* [167, 166], which has experimented with the fusion of audio and visual information. Initial multimodal source separation work (Sodoyer *et al.* [167, 166]), focused on maximising an audiovisual statistical model in order to extract the correct signal. This was found to be computationally expensive, especially when convolved speech mixtures were considered. One approach considered was to maximise the relationship between audio source information and lip movement with a statistical model, as detailed further in work by Rivet *et al.* [148, 147]. Another approach to solve the permutation indeterminacy problem is arguably more computationally efficient and makes use of a VAD to identify silent periods in speaker utterances (Rivet *et al.* [149, 150]), and so extract the correct source at all frequency bins.

*Key Output*

An investigation of the literature identified that a state of the art example of this research technique for multimodal speech processing was developed by Naqvi *et al.* [133]. This research builds on the concept of using visual information to solve permutation indeterminacies, and develops an audiovisual beamforming approach to solve the problem of source separation in an environment consisting of a mixture of overlapping moving speech sources. In effect, this is a technique for solving the Cocktail Party Problem (Cherry [38]). As described earlier in this section, previous research work of this nature originally made use of the audiovisual coherence between audio and lip information. However, this approach was found to be computationally intensive with regard to the use of visual information, and so Naqvi *et al.* [133] utilise a simpler approach that uses speaker tracking to identify source locations, and then uses these coordinates for beamforming. A diagram of the system is shown in figure 7.

One significant aspect of this system is the 3D visual tracking approach utilised. Speaker tracking is used to identify visual data and makes use of state of the art techniques. Firstly,

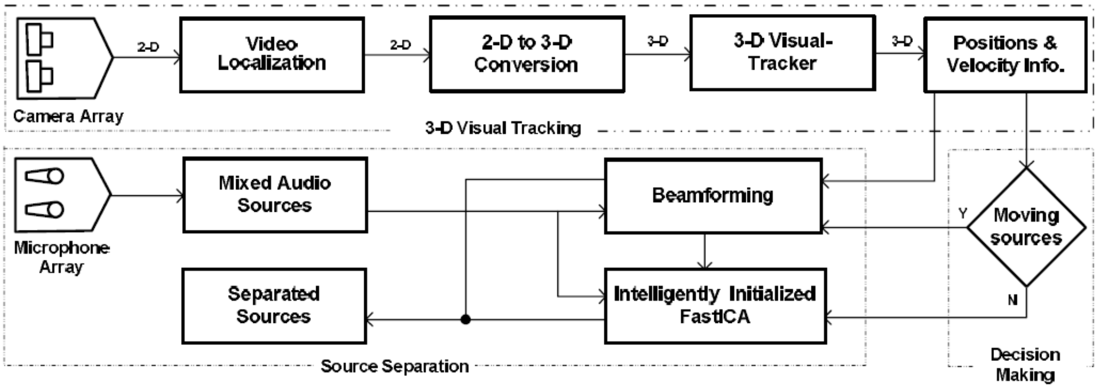


Figure 7: Diagram of system proposed by Naqvi *et al.* [133] . Utilises 3D visual tracking and source separation . Taken from (Naqvi *et al.* [133]).

some assumptions are made. It is assumed by Naqvi *et al.* [133] for the purposes of their paper that a full face image of each speaker is visible at all times, and that a geometric cue (i.e. the centre of the face) is available. The experiments are performed in a simulated office environment (a small room), using two high quality cameras, mounted above head height and synchronised using an external hardware trigger module. This provides a high vantage point, and the input from the two cameras is used to convert a two dimensional view of the room to 3D. To carry out the tracking, a Viola-Jones face detector is used (Viola & Jones [175]). This is a face detector that operates with a cascade of boosted classifiers. In each input image, parts of the image are sub sampled at a variety of scales and locations within the frame. There are three stages in the face detection process. Firstly, all sub windows are normalised in order to take illumination into account. Secondly, the cascade of classifiers, with each classifier being more complex than the last, is then applied to each sub window in order to identify whether a candidate for a face is present or not. The final stage of the face detection process is the subsequent merging of overlapping sub windows that identified a candidate for a possible face, in order to output the final identification of all faces in each visual frame. To then track the face movement for multiple speakers, a Markov Chain Monte Carlo Particle Filter (MCMC-PF) is used. The use of this is described in detail by Naqvi *et al.* [133]. The tracking stage ultimately calculates the 3D position and velocity of each speaker, which is used in the source separation stage.

After receiving visual information, the second stage is to perform source separation. In this work, the authors assume that noise is either non-existent, or considered to be a separate source, and that the number of input sources is equal to the number of desired outputs. The visual information is first used to determine whether the sources are moving or stationary. This is done by considering the visual information of the sources, calculated in the tracking stage. If the sources are considered to have been stationary for at least two seconds, then Intelligently Initialised Fast Independent Component Analysis (IIFastICA) is used to separate the sources. If however, the sources are determined to be moving, beamforming is used, with the aid of visual information.

When the sources are moving, beamforming is used for source separation. Working in the frequency domain, the mixing process is defined by Naqvi *et al.* [133] for $M$ statistically independent real speech sources $s(\omega) = [s_1(\omega), \ldots, s_M(\omega)]^H$, with $(.)^H$ denoting Hermitean transpose, and $\omega$ denoting discrete normalised frequency. This work states that a multichannel Finite Impulse Response (FIR) filter producing $N$ mixed signals $u(\omega)$ can be defined as,

$u(\omega) = H(\omega)\, s(\omega)$

with $H(\omega)$ representing the filter. The source separation process of extracting these mixed signals can then be described as,

$y(\omega) = W(\omega)\, u(\omega)$

with $W(\omega)$ representing the unmixing matrix, and $y(\omega)$the estimated sources to be output. To calculate this unmixing matrix, the authors follow an approach defined in Parra & Alvino [140] that utilises geometric information as one factor in the determination of the unmixing matrix. For this, the angle of arrival, i.e. the position of the speech sources in relation to the microphone array of each source, is calculated using the visual information. This helps to solve the permutation problem, and focuses the direction of the beamformer to eliminate non relevant noise for each source.

If the sources are considered to be stationary, then IIFastICA is used. This uses the estimated FIR filter and whitening (Hyvarinen *et al.* [92]) to initialise the FastICA algorithm (Bingham & Hyvarinen [25]). At each frequency bin, the FIR filter with visual information used to calculate it and the whitening matrix are used to initialise the FastICA algorithm,

$w_i(\omega) = Q(\omega)\, h_i(\omega)$

where Q is the whitening matrix and $h_i(\omega)$ represents the $i - th$ column of $H(\omega)$. Fast ICA is then used to separate the sources, as described in more detail in Bingham & Hyvarinen [25]).

In the experiments reported by Naqvi *et al.* [133], the system is evaluated with a room containing two speakers, with audio recorded at 8 kHz and video at 25 Hz, and audio and visual data manually synchronised. Initially, the 3D tracking was evaluated, and was found to be accurate and effective. The angle-of-arrival data (using visual positioning) was also found to be correct with regard to the experimental data. With regard to the source separation problem, various system configurations (both audio-only and audiovisual) were tested by Naqvi *et al.* [133], and they found that the use of visual information improved overall results and algorithm performance.

*Strengths*

There are a number of strengths with this work. Firstly, this research demonstrates the value of using visual information as part of a speech filtering system. It builds on prior work and shows an improvement on existing audio-only techniques (for example, using audio-only geometric information for beamforming initialisation). The results demonstrate that the proposed multimodal system is effective. In particular, this work uses visual data to solve the permutation problem, i.e. it manages to successfully identify the correct source of interest. The technique used, tracking speakers and returning 3D coordinates for further processing, does not use detailed lip information or attempt to analyse speech content, but uses a much less computationally intensive approach that delivers effective results. The recent state of the art work discussed above makes used of modern 3D tracking technology and displays a nuanced approach to speech processing. The type of source separation performed varies depending on environmental conditions (the movement of the speech sources), showing an intelligent use of multimodal information.

*Limitations*

There are a number of practical limitations to this work. Firstly, there are a number of assumptions made. It is assumed that the room being used is small enough to keep the reverberation level low, and it is also assumed that good visual information is visible, with a good quality full facial image available for each speaker at all times. This is adequate for the limited experiments discussed in this paper, but in a practical real world environment, these conditions are unlikely to be met. The cameras and microphones are also fixed in position, with the microphone array in the centre of the room at all times, and the cameras mounted above head height. This is adequate for simulations, but may produce poor results if experiments are extended to a more realistic environment with regard to hearing aid wearers. A hearing aid user would not be expected to remain stationary, and this means that the cameras and microphone would be mobile, making calculations such as the angle of arrival of different speech sources and accurate 3D tracking much more difficult. The system is also aimed specifically at solving the source separation problem, with noise that doesn't originate directly from a competing speech source not considered.

*Summary of Work*

Overall, there are positives to this research, in that it demonstrates a nuanced use of audio and visual information, and shows that visual information can be effectively and efficiently used as part of a speech processing system. It also demonstrates an alternative speech filtering approach, source separation rather than noise reduction, and uses state of the art techniques to extend an audio-only approach to become multimodal. However, the system relies on a number of environmental assumptions, and ultimately aims to solve the source separation problem rather than concentrating on speech enhancement. So while the research is of interest because of the use of visual information, it is not entirely relevant to the problem discussed in this thesis.

### 3.4.3 *Multimodal Fragment Decoding*

#### *Summary of Work and Previous Papers*

Multimodal speech fragment decoding, developed by Barker & Shao [18], was designed to improve on existing audio-only fragment decoding techniques (Barker *et al.* [23]). This is not primarily a speech enhancement approach as the ultimate aim of the research described here is to develop a multimodal speech recognition system. However, in order to achieve this, an approach is used that combines both filtering and recognition, and so the source separation technique used is of interest with regard to this thesis. The primary problem that this system attempts to tackle is that of speech recognition in environments where the speech source is obscured by a competing simultaneous speech source, and this research is particularly focused on the problem of masking.

There are two types of masking that represent a problem for speech recognition systems. The first is energetic masking, which occurs when the speech energy of the masker is greater than the energy of the speech source. An example of this is when a vowel from a competing masking speaker obscures an unvoiced part of the target speech source. In general, this is a problem that can be tolerated by many speech recognition systems as it is normally clear which parts of an utterance are masked. The second type of masking, informational masking, is more challenging. This is when it is unclear which part of the noisy speech input signal is dominated by the target speaker and which by a masking speaker. This research aims to tackle this by using visual information to identify the target speaker with greater accuracy. As stated previously, this is primarily a speech recognition system, but the use of visual information as part of the speech filtering process makes this research of relevance.

The idea behind audio-only fragment decoding (Barker *et al.* [23, 20]) is to combine source recognition and separation in a single framework. Source separation research is inspired by Auditory Scene Analysis (ASA) (Bregman [30, 31]), which is the process by which humans organise sounds into meaningful elements. The concept behind speech fragment decoding is that in a noisy speech mixture, there exist elements within this mixture (in the spectral-temporal domain) where speech energy is concentrated sufficiently to ensure that noise source energy has a negligible effect. There are two elements to the fragment decoding process. The first is the generation and identification of spectral temporal fragments, i.e. those fragments which are dominated by one single source, either target or masking source. These fragments are automatically labelled and are then used to create a segregation (represented by a binary mask) of these labelled fragments. A segregation hypothesis is then searched for, where the noisy input is then matched to statistical models of clean speech, with missing data speech recognition performed by matching fragments to Hidden Markov Models (HMMs) (trained on clean speech), with the type of processing dependent on the labelling, in order to produce the best matching word sequence from the noisy input mixture.

There are practical limitations to an audio-only approach though. While it is possible to determine whether a fragment is dominated by a single source, it can be difficult to determine whether that source is the target speaker or background noise. In environments containing competing background speech, this is particularly difficult. Visual information can be used to extend and improve this approach. This system deals with both recognition and enhancement, and visual information has a role to play in both aspects. Firstly, word recognition is improved by the addition of a visual vector in addition to the audio. In the audio-only approach, the recognition is done by comparing the input audio vector and a mask labelling the input, to trained audio models. In the multimodal approach, the visual information is concatenated with the audio and is simply treated as additional spectral information for use with trained audiovisual models. This was developed by Barker & Shao [19, 18].

*Key Output*

State-of-the-art multimodal work carried out using this technique (Barker & Shao [19, 18]) utilises visual information to assist with the identity of fragments dominated by the target speaker, as shown in an example in figure 8, where visual information is used alongside audio to jointly identify correct fragments. In the case of labelling appropriate fragments as dominated by target or noise, visual information helps to increase the accuracy of this. In the simplest case, visual features can determine the likelihood of the target speaking or being silent at a given point in time. Trained audiovisual models can identify audio fragments that match the equivalent visual information well, increasing the accuracy of fragment labelling.

In order to extract the audio features, the input signal is passed through a 64-channel filterbank, and temporal difference features are computed with 5-frame linear regression and added to the 64-channel filterbank output to create an audio vector with a dimensionality of 128. The visual features take the form of 2D-DCT features extracted from the lip region of the target speaker, with 36 (6 by 6) low-order coefficients extracted. Temporal difference features are added to these to create a vector with a dimension of 72. As the video is recorded at 25 fps, this is up-sampled to 100 fps to match the audio vector.

The research presented in this work takes the audio and visual inputs and combines them in a variety of ways using HMMs, so both early stage feature fusion (concatenating audio and visual vectors before processing), and decision fusion (visual and audio computed separately and then merged) approaches are considered. The audiovisual information is used for both fragment labelling and word recognition, but as speech recognition is outside the scope of this thesis, the recognition process is not discussed here. In terms of spectral temporal fragment identification, the authors describe visual features as a form of 'scaffolding' that supports the fragment identification process, as shown in figure 8. So while there may be masked speech fragments that could be identified as a good match for the target speaker model on the audio-only level, the addition of visual information reduces the likelihood of this, because the masked speaker fragments are unlikely to be a good match when visual information is also used in the speech models. Even if no phonetic information is provided, the visual information at a fundamental level can identify the presence or likelihood of the target speaking, making the informational masking problem less of an issue. This research required the training of HMMs for each speaker tested.

The authors compared this approach to a similar audio-only fragment decoding approach using sentences from the GRID Corpus (Cooke *et al.* [46]), and found that while the audio-only approach produced similar results as the audiovisual system with an SNR of +6dB, as the SNR decreased to -9dB the audio-only approach produced a much steeper drop off in performance,

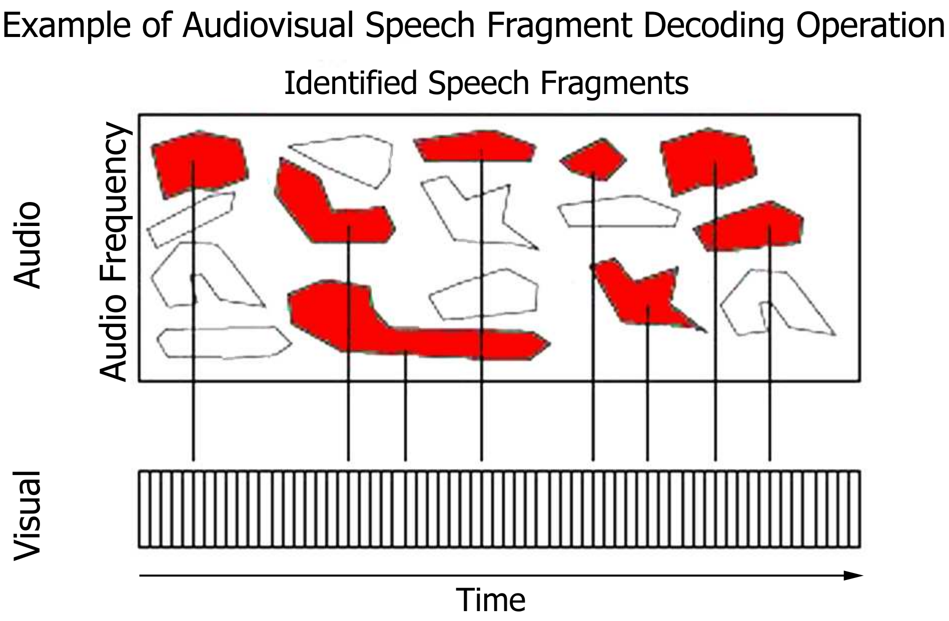

Example of Audiovisual Speech Fragment Decoding Operation
Identified Speech Fragments
Audio
Audio Frequency
Visual
Time

Figure 8: Example of fragment decoding system operation, using visual information to identify speech fragments from noise within frequency groupings. Taken and adapted from Barker & Shao [18].

with the audiovisual approach performing significantly better. This shows the benefits of a multimodal approach.

Although this is primarily a speech recognition system, it does deal with enhancement to an extent. It extends an existing audio-only approach, and it fuses audio and visual information. This means that both audio and visual information is used together. It does this while making use of commonly used feature extraction techniques, so visual information is tracked automatically and the DCT of the lip region is extracted. As it makes use of an ICA approach, it deals with speech mixtures containing two similar speakers, and makes use of sentences from the GRID corpus, which is a high quality audiovisual speech database. The system is trained for multiple speakers from this corpus, and this means it can work with a range of subjects. In order to test performance, the authors tested their system against equivalent audio-only systems, and it was consistently found that in adverse conditions, word recognition of the target speaker was stronger when using the audiovisual approach. The authors also tested the system using a variety of low resolution visual features, finding that even with a simple 1 x 1 pixel DCT vector; an improvement was found over audio-only results.

*Strengths*

The most important thing to take from this work is that visual information can be used effectively as part of a speech filtering system. The research presented here successfully used visual information to deal with a very challenging speech environment. Recognition tests performed found that the use of visual information enabled this system to outperform audio-only approaches. The use of poor quality visual information was also tested, with even low quality visual data found to improve results. Overall, this is a very interesting approach with scope for further development.

*Limitations*

While this research demonstrates the benefit of a multimodal approach, it has some limitations with regard to specific application to this research project. In these experiments, trained HMMs are used, and the assumption is made that the target speaker is also part of the training set. This means that experiments have not been attempted with completely novel data, only with a limited selection of sentences from a single corpus, and this limits the possible practical application of this work. Individual HMMs are used for each speaker used in the testing and training process. Also, this approach deals with enhancement and recognition in parallel, rather than having an enhancement stage followed by a separate recognition process. This makes the two components difficult to separate, and limits its potential relevance in terms of speech enhancement alone. It is also tested with speech recognition rather than enhancement. Therefore the true impact of noise cancellation is not thoroughly tested. The system as it is makes use of visual information for both the separation and recognition, with the two linked, limiting the relevance of this work with regard to this thesis.

*Summary of Work*

The fragment decoding work described here is an example of the extension of audio-only algorithms to the multimodal domain by utilising visual information. This research considers a challenging speech problem, that of filtering speech when a competing speech source is masking the target speech. This research demonstrates that in a challenging environment, visual information (even in a crude form) can help to extend the usability of speech filtering algorithms, with further scope for extension of this work. However, the key limitation from the perspective of the research proposed in this thesis is that this is primarily a speech recognition system rather than a speech enhancement system. So while visual information is utilised for speech source separation, this is linked with the recognition aspect. Therefore, while the concept is of interest, this system is not of particular relevance for further development in this thesis.

### 3.4.4 *Visually Derived Wiener Filtering*

*Summary of Work and Previous Papers*

Almajai & Milner [12] have developed a multimodal speech enhancement system that makes use of visually derived Wiener filtering (Wiener [180]). This approach builds on previous published work by the same authors. Firstly, Almajai & Milner [10] demonstrated a high degree of audiovisual correlation between the spectral output of speech and the shape of the mouth, and then built on this to filter speech by making use of visual features to estimate a corresponding noiseless audio signal, and then filtering a noisy audio signal (Almajai *et al.* [13], Almajai & Milner [11, 12], Milner & Almajai [130]). Wiener filtering works by comparing a noisy input signal to an estimation of an equivalent noiseless signal. This approach is commonly used in other research fields such as the restoration of damaged or distorted photographic images (Hiller & Chin [83]), but has also been applied to audio-only speech enhancement (Zelinski [189]). The key problem with this approach with regard to speech processing is the difficulty in obtaining an accurate estimation of the noise-free speech signal. Almajai & Milner [12] first created a basic Wiener filtering approach that initially made use of a simple joint audiovisual model and basic competing white noise, and then expanded upon it to produce a more sophisticated and comprehensive audiovisual speech enhancement system (Almajai & Milner [11]).

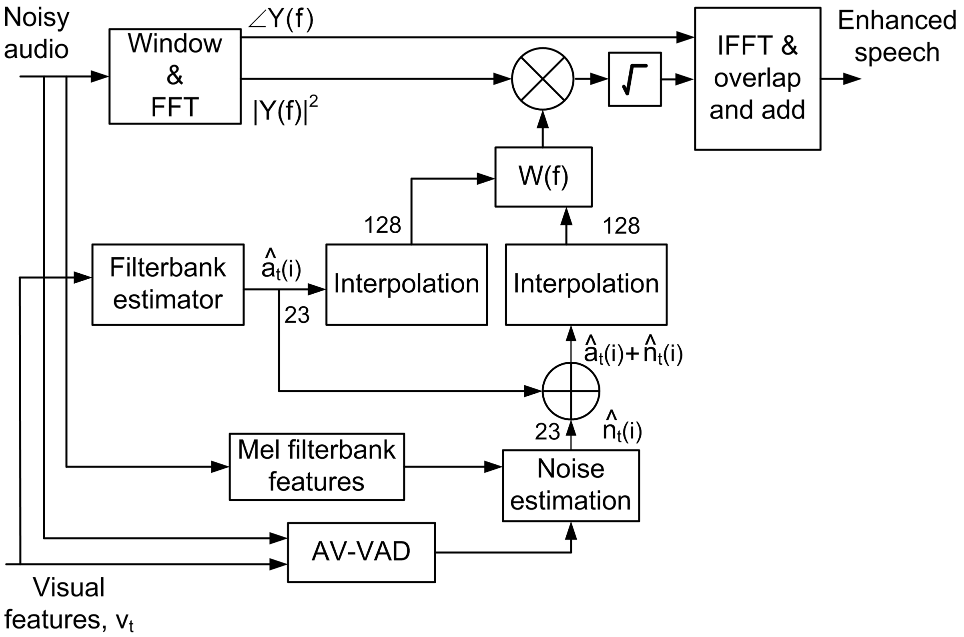


Figure 9: Diagram of an audiovisual speech filtering system for speech enhancement, utilising visually derived filtering, and an audiovisual VAD. Taken from Almajai & Milner [11].

*Key Output*

Recent work from Almajai & Milner [11] makes use of a set of 277 UK English sentences from an audiovisual speech database, spoken by a single male speaker. Of these sentences, 200 of them have been used as training data for the audiovisual noiseless speech estimation models, and the remaining 77 used for testing. Visual information was recorded using a head mounted camera, and 2D-DCT was used to extract relevant lip information. This was then upsampled to match the equivalent audio information. The system design used in this paper is shown in figure 9.

In a system containing a noisy time domain audio signal $y(n)$ (with $n$ representing sample number) and visual information taken from the facial region $v(i)$, with $i$ representing frame number, $v(i)$ is used to produce an estimate of the log filterbank vector of the noiseless audio signal $\hat{x}(i)$. This is transformed into a linear filterbank estimate $L_{\hat{x}}(m)$, with $m$ being a filterbank channel. This is then compared to an estimation of the noiseless speech plus noise. A noise only estimate, $L_y(m)$, is calculated from noise only periods of the utterance, identified with the aid of an audiovisual VAD. The combination of the speech and noise estimates is used to calculate the filterbank Wiener filter $L_w(m)$, which compares the noise-free estimate to the speech plus noise estimate:

$$L_w(m) = \frac{L_{\hat{x}}(m)}{L_{\hat{x}}(m)+L_y(m)}$$

The authors then interpolate $L_w(m)$ in order to match the dimensionality of the power spectrum of the audio signal to produce the frequency domain Wiener filter. This is used to calculate the enhanced speech power spectrum, which is then combined with the phase of the audio input and an inverse Fourier transform is used to return the enhanced speech to the time domain.

With Wiener filtering, the most complex aspect is the method of estimating the noise-free signal. A Maximum a Priori (MAP) estimate of the noise-free speech can be found with the use of visual information. The authors make use of phoneme-specific estimation. 36 monophone HMMs, plus an additional one for silence are trained using the training dataset. The training dataset makes use of forced Viterbi alignment to split each training utterance into phoneme sequences, and these labelled utterances are split into vector pools, with Expectation Maximisation (EM) clustering used to train a Gaussian Mixture Model (GMM) for each phoneme. The resulting HMMs are then used for speech estimation.

In the test sentences, an audiovisual speech recogniser is used to identify the phoneme. It is assumed in this work that the first few frames of the utterance are noise only, and an estimate of the SNR is taken from these. The speech recogniser then combines audio and visual recognition to identify the phoneme. In conditions with a low SNR, more emphasis is put on visual information than on audio. This attempts to identify the phoneme being spoken in each frame, which identifies the most suitable GMM to use for each frame, and returns the log filterbank speech-only noise-free estimate, $\hat{x}(i)$.

To estimate the noise-alone signal, an average of the non-speech vectors preceding the speech frame is taken. To correctly identify non-speech frames, an audiovisual VAD is used. This uses a pair of Gaussian Mixture Models (GMMs) trained using the manually labelled audiovisual vectors from the training dataset. It is established that in noisy speech, with a low SNR, it becomes more difficult for an audio-only VAD to correctly label frames, so here, visual information is used. The SNR estimate, taken from the first few frames of the utterance (assumed to be noise only) is used to define how much weight to apply to the audio information. In environments with a low SNR detected, less audio information is used. The frames identified as non speech are averaged, and this produces the noise alone estimate, $L_y(m)$.

The results presented are generally positive, with both objective and subjective scores displaying the potential of this work, as shall be discussed in the next sections.

*Strengths*

Fundamentally, this work demonstrates the potential for using visual information purely as part of a multimodal speech enhancement system, with the visual information used to remove noise from speech. Recent work builds on and extends preliminary work by Almajai & Milner [12], with refinements throughout development, such as the addition of a VAD and the increased sophistication of the speech filtering model. The system combines audio and visual feature extraction, a multimodal VAD, and a visually derived Wiener filtering approach. This is a sophisticated system that takes account of the level of noise when it comes to phoneme decoding, and filters the signal differently depending on the phoneme identified.

The authors also report good results. It can be seen that objective speech evaluation using three commonly used measures, PESQ, Log-Likelihood Ratio (LLR), and Itakura-Saito Distance (IS) all showed a significant improvement in speech enhancement performance when compared to the original noisy speech and a standard audio-only spectral subtraction approach. Subjective human listening tests also showed that visually derived Wiener filtering was effective at removing noise from speech, with a significant improvement being found at all reported SNR test levels from 20dB to 5dB. This improvement was not reflected in the speech distortion level (with visually derived filtering considered worse than unfiltered speech), but the overall speech quality was considered to be an improvement over an audio-only spectral subtraction technique at all SNR levels.

One strength of this work is that it focuses exclusively on delivering enhanced speech. Much of the audiovisual speech work in the literature has focused on the fusion of audio and visual modalities for different purposes, such as biometric authentication or automatic speech recognition, and so this speech enhancement work is significant and pioneering, with potential for further development. The use of 2D-DCT as a feature extraction technique is validated, as well as the potential of using a GMM for speech estimation and Wiener filtering.

*Limitations*

There are a number of limitations. Firstly, when the results are analysed, the most significant results to consider are the subjective listening tests. While the work of the authors has reduced the noise intrusion score, this is at the cost of increasing the speech distortion score, meaning that even at relatively low SNR levels; listeners still have a slight preference for unfiltered speech over visually derived speech. Additionally, the results are limited to relatively high SNR levels (+5dB to +20dB), meaning that the system has not been tested in noisy environments. The model also makes use of a complex phoneme dependent model, and when forced alignment is used (i.e. manual labelling of phonemes), slightly improved results are obtained over standard automatic phoneme recognition. This is because in the presence of noise, phoneme decoding accuracy falls to 30% at 0dB, meaning that the accuracy of this system in noisy conditions is poor.

Another limitation with this work is the relatively constrained range of the database used for training. An audiovisual database containing 277 utterances for a single speaker is used. This means that although good results have been found, it is effectively only trained and tested with input from a single speaker. The speech estimation model is also trained with training data from the same speaker, meaning that there is a potential lack of robustness in the GMMs used. Another issue is that the visual filtering approach makes use of visual information tracked by a camera, with the relevant ROI acquired with the aid of AcAMs; however, the system proposed in this research does not take account of situations where a poor visual feature-extraction result is returned. There are many ways in which a poor result could be returned, for example, if the AcAMs do not correctly identify the ROI due to the visual signal being corrupted by noise or the target speaker moving unexpectedly. More fundamentally, the camera is head mounted and aimed by the listener, and while this is adequate in a laboratory environment, in a real situation, the listener may not always be looking at the speaker, or the lip region may not be seen because of room lighting conditions or obstacles in the way (e.g. a hand over the mouth, or another person coming between listener and speaker). So there are limitations with this system due to being dependent on visual information.

When the Wiener filtering approach makes use of a VAD to determine speech and non speech, it then calculates a noise-only estimate. However, this estimate makes the assumption that the first few frames of the noisy speech signal are non-speech. The noisy speech itself is also described as "corrupted", but it is not specified whether the noise is simple additive noise, or whether a more complex mixing matrix is used to provide a more realistic speech filtering challenge. There is also the issue that speech is articulated differently in the presence of noise, as described by the well-known Lombard Effect (Lee *et al.* [110], Lane & Tranel [107]), which is not accounted for in the training of this system. It is clear that although the results are promising, there is still a lack of flexibility in the application of this speech filtering approach.

#### *Summary of Work*

The visually derived filtering work carried out by Almajai & Milner [11] is of great interest for further research. The use of visual information as part of a Wiener filtering system has been shown to be effective and positive results have been found. Almajai & Milner [11] also added complexity to the system by utilising an audiovisual VAD to provide more nuanced filtering. There is much potential for further research development. However, there are some limitations to the work such as being trained with a limited training set, and the system as presented being strongly reliant on visual information, and not taking account of situations without suitable camera input.

## 3.5 VISUAL TRACKING AND DETECTION

### 3.5.1 *Introduction*

One aspect of relevance with regard to the system proposed in this thesis is the extraction of relevant ROI information. This research focuses on developing an audiovisual framework, and part of that framework is the use of a suitable lip detection and tracking algorithm. This is an area of active research and development (Wakasugi *et al.* [176], Liew *et al.* [117]), and while it is relevant to this research, the development of a novel lip tracking algorithm was considered to be outside the scope of this thesis. However, an automated detection and tracking approach is required in order to accurately extract visual information for further processing, and for the work presented later in this thesis, an existing approach is adapted using a ROI detector (Viola & Jones [175]) and a lip tracking technique (Nguyen & Milgram [135]). It is therefore felt to be relevant to summarise some recent developments in this research domain.

The remainder of this section provides a brief summary of recent developments in this field. Firstly, some developments specifically to lip tracking are briefly covered, giving a number of recent developments in a range of categories. The development of more general ROI detectors, particularly the seminal Viola-Jones detector (Viola & Jones [175]), is also discussed, as some lip tracking approaches still depend on a manual initialisation in the first frame, and utilising a ROI detector can automate this process.

### 3.5.2 *Lip Tracking*

Lip tracking is an active field of research, with many different examples in the literature, such as Shape Models Nguyen & Milgram [135], and Active Contour Models (ACM) (Kass *et al.* [99]). Lip tracking represents a challenging research area, as it can be difficult to track lip images due to issues such as a weak colour contrast between skin and lip areas (Das & Ghoshal [51]), and also the elastic shape and non rigid movement of the lips during speech (Cheung *et al.* [39]). There is also the issue of environmental conditions, with issues such as variable lighting conditions to be taken into account, as well as the issue of overall face movement (i.e. not just lip movement). A brief summary of a number of recent developments is discussed here.

Cheung *et al.* [39] divide lip tracking approaches into two main categories, edge based approaches, and region based approaches. Edge based approaches, as suggested by the name, rely on colour and edge information to track movement. This can rely on identifying colour contrasts (Zhang & Mersereau [192], Eveno *et al.* [58]), key points (Eveno *et al.* [59]), or points considered to be 'corners', as proposed by Das & Ghoshal [51]. These approaches work well under desired conditions (i.e. a clean background and distinct features), but will produce poor results if the image is not ideal (for example, in poor lighting conditions or when the subject is wearing cosmetics that may interfere with contrast detection (Cheung *et al.* [39]). Some other approaches include the use of ACM (Kass *et al.* [99]), also known as 'Snakes', to detect edges.

ACM were first proposed in 1988 by Kass *et al.* [99] and are designed to fit lines (hence the reason they are known as 'snakes') to specific shapes for feature extraction. In the context of lip tracking, this means fitting a contour around the edge of the lips in order to identify and extract the lip shape. Snakes are based on minimisation of energy and operate by identifying edges. The contour is shaped by the idea of external and internal energy. Ideally, internal energy is minimised when the snake has a shape relevant to the desired object, and external energy is minimised when the snake has correctly identified the boundary of the desired object. There are many implementations of this technique for edge based lip tracking, and some examples of this approach include work by Kass *et al.* [99] and Freedman & Brandstein [62]. However, these can again deliver poor results in sub-optimal conditions. There is also the limitation that models of the lip region may not accurately fit the precise edges.

Another example of an edge based approach is to use corner detection Das & Ghoshal [51]. This technique converts images to binary images, identifies the lower half of a face, and then uses horizontal profile projection (Ji *et al.* [95]) (defined as the sum of pixel intensities in each row of an image) to identify the rows of an image corresponding to the lip region. This is done by identifying the two maximum points of the horizontal projection vector for the lower part of the face (assumed by Das & Ghoshal [51] to correspond to the upper and lower boundaries of the lip region. Of this region, corners are identified, using the Harris Corner Detector proposed by Harris & Stephens [77]. Corners are defined as points for which there are two dominant and different edge directions in the local region, and the Harris Corner Detector identifies these corners based on the autocorrelation of image gradient or intensity values. Although this implementation is interesting, it appears to be untested in anything other than extremely basic environments, and like other edge based methods, appears vulnerable to noise within images.

Region Based approaches primarily make use of region based information. Cheung *et al.* [39] identify four main categories of interest, Deformable Templates, Region Based ACM approaches, AcAMs and Active Shape Models (ASMs). Firstly, ASMs use a set of landmark points, derived manually from a training set of images to create a sample template to apply to the area of interest. When presented with a new image, the template is applied, and then the points are iteratively moved to match the face (fitting). This is a relatively simplistic approach, although is relatively fast. It does require intensive and time consuming training however, and suffers from a lack of robustness with images not similar to those found in the training set. Some examples of this include research by Luettin *et al.* [122], who applied manual points to train lip images, and also Nguyen *et al.* [136], who used ASMs to learn lip shapes by applying multi-features of lip regions. This algorithm was also extended by the development of the SAAM approach, which is discussed in more depth in chapter 4, and performs lip tracking by online training of models, and also AcAMs which were first proposed by Cootes *et al.* [47]. These techniques again have the limitation of being dependent on initial parameter selection, and the initial detection phase is separate from the tracking process.

AcAMs were originally developed by Cootes *et al.* [47], for face recognition and operate by creating statistical models of visual features, making use of shape (as described above) and texture information. The shape is defined as geometrical information that remains when location, scale and rotational effects are filtered out from an object, and the texture refers to the pixel intensities across the object. The initial shape models are combined with grey-level variation in a single statistical appearance model. Models are trained with manually labelled test-sets, and can then be applied to unseen images, the points are warped to make the model fit the image, and this produces output parameters. Using AcAMs has the advantage that detailed models can be produced, while still being relatively computationally efficient. The disadvantage with this technique is that the models require time consuming and intensive training before use, and can struggle to generalise accurately when presented with novel data.

Deformable Templates were introduced by Yuille *et al.* [186], and operate in a similar way to the ACM approach described above. An initial template is specified, with parameters matching a lip shape. The minimisation of energy approach is then utilised to alter these parameters, and then as the parameters are adjusted, the template is altered to gradually match the boundary of the desired lip shape. This initial approach has been extended by others (Liew *et al.* [117], Chiou & Hwang [41]). The limitation of this approach is that if there is a very irregular lip shape, or the image has a very widely open mouth, then poor performance has been found.

Region Based ACM approaches, are similar to the ACM approaches defined above, but rather than searching for an edge, looks specific regions within an image are inspected in order to minimise energy by dividing images into lip and non lip regions. This approach is described by Cheung as generally being highly dependent on initial parameter initialisation, and while it can outperform traditional edge based ACM approaches in some conditions, can also perform poorly with complex lip shapes. Examples of this work include Wakasugi *et al.* [176], Cheung *et al.* [39].

Cheung *et al.* [39] have proposed to extend a Localised Active Contour Model (Lankton & Tannenbaum [108]), by using colour information, to create a Localised Colour Active Colour Model (LCACM). Briefly, this is first initialised and the parameters set in the first frame by using a 16 point deformable model, as proposed by Wang *et al.* [178], using image intensity based techniques to identify the contrast between lip and non lip regions and therefore the initial points. These points were then utilised to create the deformable model. The tracking process then used these initial parameters to minimise the energy in colour based difference between areas defined as lip and non lip region. This results in a system that is adaptive to lip movement and produces accurate results in optimal conditions. This approach is an example of a promising technique still in the relatively early stages of development. In terms of tracking, it appears to produce promising results, but only on very limited test data (pre-processed faces that are presented as being straight on with no image noise). It also relies, like other techniques, on separate parameter initialisation.

While these represent some lip tracking options, as can be seen from the examples above, the initial lip detection and the lip tracking are often initialised in two distinct steps (Cheung *et al.* [39], Nguyen & Milgram [135]), and are dependent on initial parameters being accurately defined. Therefore it is also of interest to consider overall ROI detection algorithms that can be used and adapted to automatically detect regions of interest and generate initial parameters. The system chosen for use in this thesis, as described makes use of a lip tracking algorithm proposed by Nguyen & Milgram [135] and tested in collaboration work with the author (Abel *et al.* [2]). However, the initial parameters are required to be generated, and so looking at ROI detection is of interest.

### 3.5.3 *Region of Interest Detection*

ROI detection is a key aspect of the image processing aspect of this research. In order to be able to automatically track images, the ROI has to be successfully identified. In the system discussed later in this thesis, the mouth region is utilised to process visual information, and this is required to be automatically identified by the system. As part of this ROI detection process, the role of face detection is of relevance for discussion.

There are a number of factors that make the detection of faces in an image a challenge, as defined by Yang *et al.* [183]. These include pose, occlusion, image orientation, image conditions, facial expression, and the presence of structural components. Pose refers to the position of the face in relation to the camera, so the face may be frontal, at an angle, and significant features may not be visible to the camera. Occlusion is related to this, in that the pose may result in features being occluded, but it also refers to features being blocked by other objects. There is also the issue of image orientation and conditions, concerning different rotations around the optical axis of the camera, and also factors such as lighting conditions. Finally, structural components refer to facial furniture such as beards and glasses, which may affect the detection process.

Yang *et al.* [183] in a paper from 2002, and Wang & Abdel-Dayem [177] classify face detection approaches into four key categories. Firstly, Knowledge based approaches such as proposed in work by Kotropoulos & Pitas [103], Yang & Huang [182] are rule based systems. These make use of rules, often defined by human experts with facial features represented by differences and positions from each other. From these rules, features can be extracted, and candidate features can be identified based on these rules. However, the wide range of potential facial features can make it difficult to translate facial features to good rules that are applicable to more general cases.

The second category outlined by Yang *et al.* [183] covers Feature Invariant Approaches. These work by extracting structural features of the face, specifically focusing on identifying features that will be identifiable, even in environments when conditions like lighting and pose vary greatly. The theory behind this is that humans are capable of identifying faces, even in extremely degraded environmental conditions. Some examples of this approach include Kjeldsen & Kender [100], Yow & Cipolla [185].

The third category identified covers template matching methods. These use standard patterns of a face which have been trained and stored, either for identifying entire faces, or for individual features. To detect ROI information, the correlations between the input image and stored patterns are computed for detection. This can include shape models Luettin *et al.* [122] using deformable templates, and also using 'snakes' for contour information (Kass *et al.* [99]). Although these approaches are widely used in state of the art tracking and identification applications, with regard to detection, many template tracking approaches still require a separate initialisation of the ROI (Cootes *et al.* [47], Nguyen & Milgram [135]). Appearance based approaches are more commonly used in state of the art ROI detection approaches.

Appearance based methods are the final category of approaches identified by Yang *et al.* [183]. This category describes models that are learned from training images. Rather than templates, which are manually trained and configured, appearance based approaches rely purely on trained true or false results from the training data. This approach requires a considerable quantity of training data in order to be effective. Yang et al. identified a number of approaches used, such as Eigenfaces (Kohonen [102]), Distribution-Based Methods (Sung

[169]), and Neural Networks (Agui *et al.* [5]). This category of approaches is the most common approach used in state of the art research in this field, and specifically, the development of the Viola-Jones approach, pioneered by Viola & Jones [175], is one of the most influential developments in recent years. Much research is focused on exploring, utilised, and extending this approach.

*The Viola-Jones Detector*

The Viola-Jones detector (Viola & Jones [175]) is arguably one of the most important developments in the field of face detection. This is an appearance based method, with three main components, the integral image, classifier learning with adaboost, and an attentional cascade structure. The first aspect of the Viola-Jones detector to consider is the integral image, also known as the summed area table. This is a technique used for quickly computing the sum of values in a rectangular subset of a grid. This was first introduced to the field by Crow [50], and is used for rapid calculation of Haar-Like features (which will be explained below). Essentially, the value at any point of an integral image is the sum of all pixels above and to the left of that point. It can be computed efficiently, and then any rectangle in that integral image can be accomplished quickly. This is used in the Viola-Jones detector for the calculation of Haar-Like features.

Haar-Like features represent an improvement on calculating all image intensities. They were adapted from Haar wavelets, and use the integral image technique to calculate the sum of intensities in specific rectangular regions within an image of interest. The sum of intensities can then be compared for neighbouring regions, and then the difference between each of these regions can be calculated, so for example, a lip boundary can be identified by finding a difference between lip and non-lip pixel intensities. The Viola-Jones detector uses comparisons of 2, 3, and 4 rectangles as part of the detection process. The detector is trained using a very intensive process of training, requiring many hours of training images. Trained Haar-Like features are available as part of the OpenCV library (Bradski [29]), limiting the requirement for further training.

There have been refinements to the original Haar-Like features. For example, Lienhart & Maydt [116] introduced rotated features (rectangles at a 45 degree angle to the overall image), and rectangles with flexible sizes and overlap distances were introduced by (Li *et al.* [115]). Mita *et al.* [131] proposed the use of joint features, based on the principle that human faces had incidences of co-occurrence of multiple features. There are many such developments, and a more comprehensive description of these can be found in a detailed review by Zhang & Zhang [191].

In order to identify the optimal features to use to reduce errors, an approach known a boosting (Meir & Rätsch [127]) is utilised by Viola & Jones [175]. This is an approach that aims to produce a very accurate hypothesis of a classification result by combining many weak classifiers. The initial approach utilised by Viola & Jones [175] is a modified version of the Adaboost (Freund & Schapire [63]) algorithm. The theory behind this approach is that the number of Haar-Like features in any image sub-window will naturally be very large, and in order to produce a usable and quick classification, the vast majority of features must be excluded, with focus given to a very small number of critical features. At each stage of the boosting process, a weak learning algorithm is designed to select one single Haar-Like feature that best separates two distinct regions with the minimum number of errors. Each weak classifier only depends on a single feature, and at each stage of the boosting process, the strongest weak classifiers are weighted accordingly to produce the overall classification with the least errors. With regard to face detection, Viola & Jones [175] found that the first feature to be selected was a large feature demonstrating a strong difference between the eye region and the upper cheek, and the second feature was contrast in image intensity between the two eye regions and the bridge of the nose. There have been a number of refinements to the original Adaboost algorithm, such as Gentleboost (Friedman *et al.* [64], Brubaker *et al.* [32]), Realboost (Li *et al.* [115], Bishop & Viola [26], Schapire & Singer [161]), and JS-Boost (Huang *et al.* [88]). Again, a very detailed summary is given by Zhang & Zhang [191].

The final component used in the Viola-Jones detector makes use of an attentional cascade structure. As has been stated previously, most of the many sub-windows produce a negative result, and so are not of relevance for classification. The cascade structure proposed by Viola

& Jones [175] aims to exploit this to reduce computation time, by using a tree (cascade) of trained classifiers. Simpler trained classifiers are used in the early stages to reject the majority of sub-windows, with more complex classifiers used in later stages. In practice, this means that computationally efficient classifiers are used to discard the vast majority of sub-windows immediately, with a second stage being called only if a positive result was found in the first stage. If a positive result is found here, then a third stage is then used. A negative result at any stage results in the sub-window being rejected. This was found to be an efficient approach with regard to face detection. However, the training of classifiers at each stage was found to be extremely time consuming, with timescales of months talked about for early versions of face detectors.

### 3.5.4 *Summary*

Overall, based on the discussions above, the research presented in this thesis makes use of shape models for lip tracking, following the work of Nguyen & Milgram [135], as originally tested in the collaborative work presented in Abel *et al.* [2]. This is an approach which was felt to be suitable for the initial testing of the system described in this research due to its robust nature and accurate results in tests of sentence from the relevant speech corpora used for overall system testing. As described above, this approach requires the initial location of the ROI to be specified. To accomplish this, the commonly used Viola-Jones detector is used, using features trained specifically for lip detection. This follows the basic principles outlined above of training using a cascade of classifiers. As the training process can be extremely time consuming (Zhang & Zhang [191]), pre-trained HAAR-like cascades available from the OpenCV library (Bradski [29]) are used. This allows for automatic selection of the lip region and along with the lip tracking approach used, completely automates the visual feature detection process.

## 3.6 audiovisual speech corpora review

There are a number of audiovisual corpora now available for use as part of an audiovisual system. While there are many audio-only speech corpora available for use, good quality and large-scale audiovisual speech databases are less widely available. Many such as Clemson University Audio Visual Experiments (cuave) (Patterson *et al.* [141]) are small scale corpora designed for specific tasks, with often only the collection of isolated words or digits. For more general speech processing use, larger databases with a range of speakers and sentences are more useful. This section presents a review of selected audiovisual speech databases. There are many potential databases, but this thesis only considers those that meet selected criteria. Databases that are single modality are not considered to be within the scope of this work, and neither are invasive multimodal databases, such as those that record subjects with physical markers on their face. Corpora such as these are considered to be outside the scope of this research, as are the multitude of older, specialist audiovisual speech databases. Of the databases that meet these criteria, a selection of relevant databases is described in this section.

### 3.6.1 *The BANCA Speech Database*

The banca audiovisual speech database (Bailly-Bailliere *et al.* [16]) was primarily designed for the purpose of biometric authentication. The banca project was a European wide project that aimed to develop and implement multimodal security for applications such as remote banking. This was done by developing verification schemes using audio and visual information. One output of this project was the banca database. This database was proposed at a time when the number of publicly available multimodal speech databases were very limited (Chibelushi *et al.* [40]). It consists of a wide range of speech sentences (208) recorded from across Europe, with data recorded in four languages, and in a range of different scenarios. This last condition was especially relevant for authentication testing, for example, in controlled environments with a good quality camera and microphone, and also in busier scenarios with poorer equipment (as

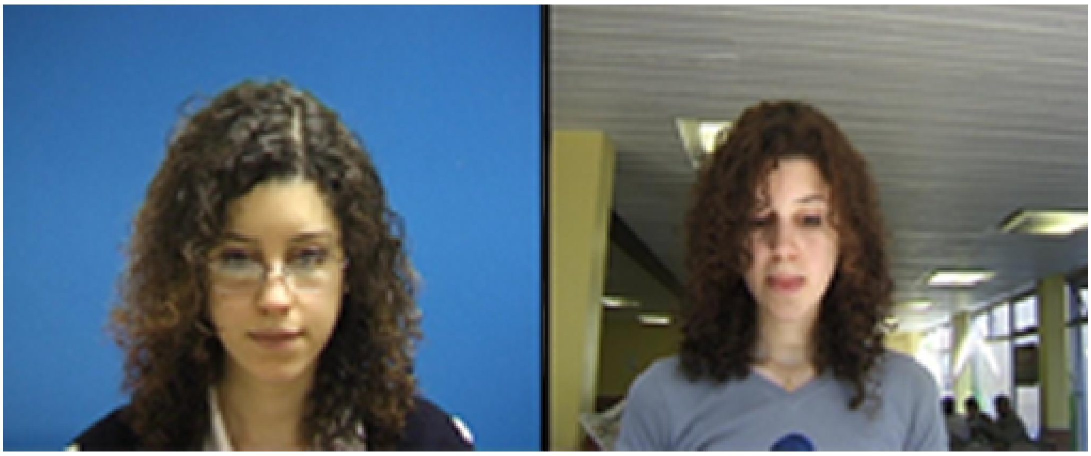


Figure 10: Example of a speaker from the BANCA audiovisual database (Bailly-Bailliere *et al.* [16]) in different recording environments.

shown in figure 10). Two types of camera, good and poor quality were used, two different quality audio recordings were used, and data was recorded in three different environmental scenarios (controlled, degraded, adverse). 52 subjects were used in different scenarios. In each recording, the speaker was expected to provide two items of speech information. Both of these consisted of data associated with biometric authentication, i.e. a series of numbers, and a name, address, and date of birth.

One of the strengths of this dataset is that there is a wide variety of speech in a range of environments. In the context of the audiovisual research area covered in this thesis, the difference between adverse and controlled environments is of relevance to this work. The corpus is available to purchase from the University of Surrey for a fee, although at the time of writing this, currently only English language data is available. However, a crucial limitation of this corpus is in the lack of natural speech data. Although there are very few truly natural speech corpora due to the artificial nature of recording, this database is very limited because the recording is designed for authentication research, with speakers mainly limited to reading out sequences of digits or lines from a postal address, limiting its potential for speech filtering work.

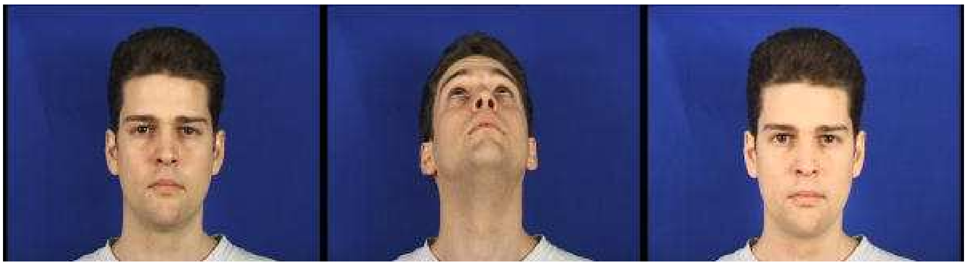


Figure 11: Example of a speaker from XM2VTSDB Messer *et al.* [129] showing frames from a head rotation shot. Image sequence taken from the official XM2VTSDB website (XM2VTS [181]).

### 3.6.2 *The Extended M2VTS Database*

XM2VTSDB (Messer *et al.* [129]) is another audiovisual speech database available from the University of Surrey. This database contains data from 295 British-English speakers, with each speaker reading three sentences. There are also head-rotation recordings for each speaker. This corpus is designed for authentication and biometric purposes, with a large quantity of data to enable security-focused multimodal recognition system training. This is similar to the aim of the BANCA corpus described above. The data was recorded on a video camera and then transferred to computer. Each speaker recorded the same three sentences "0 1 2 3 4 5 6 7 8 9", "5 0 6 9 2 8 1 3 7 4", and "Joe took fathers green shoe bench out".

This dataset has significant limitations with regard to speech enhancement research. Although it contains a large number of speakers, the recordings make use of only a very small range of sentences in limited conditions, and like the BANCA database, are more suitable for dedicated biometric authentication, than for speech filtering research.

### 3.6.3 *The AVICAR Speech Database*

AVICAR (Lee *et al.* [110]) is a multimodal speech database recorded and released by the University of Illinois at Urbana-Champaign. This was recorded as part of the AVICAR project, and was designed for speech recognition. It contains video files and the associated audio files for 100 speakers from a range of backgrounds, and is recorded in a noisy environment. As people speak differently in the presence of significant levels of background noise by

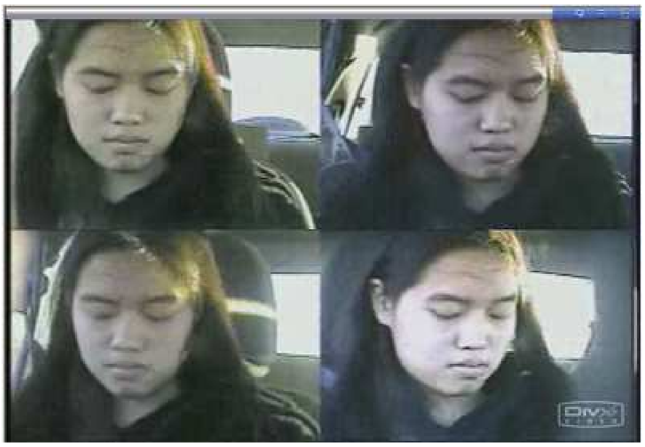


Figure 12: Example Screenshot taken from sample AVICAR video file.

involuntary raising their voice (the Lombard effect (Lee *et al.* [110], Lane & Tranel [107])), this database takes account of this effect and accommodates a range of conditions by recording speakers in an automobile. This provides different levels of background noise (i.e. different car speeds, with associated varying levels of engine noise and wind effects), and as shown by figure 12, there are four cameras recording the speaker from different angles, providing significant levels of raw data to work with. Each speaker recites a range of data, including single digits, single letters, phone numbers, and full speech sentences. There are audio recordings of 87 native English language subjects, and 86 matching video sets available.

This corpus has the advantage of containing a wide variety of speech recordings; all recorded at a variety of noise levels, and with a range of phonetically balanced Timit sentences used. The dataset also takes account of the Lombard Effect, which can have an effect on vocalising of words, and overall, this is a good corpus to consider for more advanced speech filtering research.

### 3.6.4 *The VidTIMIT Multimodal Database*

VidTIMIT is a non-invasive Australian-English audiovisual speech corpus, recorded by Sanderson & Paliwal [158], Sanderson [157]. It contains videos split into image sequences (see figure 13) and matching audio files for 43 speakers reciting a number of phonetically balanced Timit sentences. There are ten sentences recorded for each speaker, with each subject speaking a

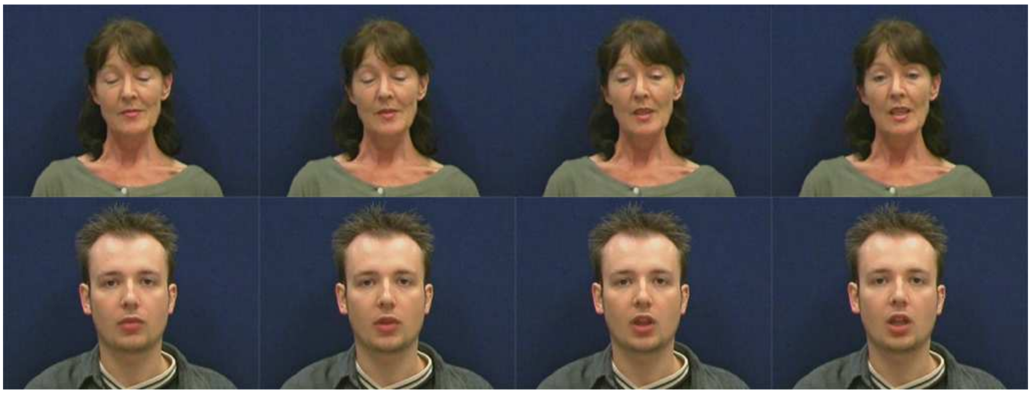


Figure 13: Sample frames from the VidTIMIT database.

variety of sentences. Each subject also recorded a head-rotation sequence. This corpus was recorded using a broadcast-quality video camera, with image sequences stored as jpeg files, and matching audio files provided. The corpus is available for use free of charge. VidTIMIT was used by its author for biometric authentication research, but its range of sentences and the size of corpus means that it does not suffer from the same limitations as many other authentication corpora. Like the other non-invasive corpora reviewed in this section, the speech data is recorded with no invasive markers or other distractions, and the speakers have been given permission to move their head naturally.

Because whole sentences were spoken, rather than isolated words or digit sequences, this provides a better simulation of spontaneous speech than some of the other corpora reviewed in this section. The corpus is freely available and is very easy to download and use. This makes the corpus a useful one, with a suitable range of speakers and sentences. However, the main disadvantage of this corpus is the presence of continuous background noise, as it is recorded in an office environment, resulting in a lack of truly clean speech. This means that there is no completely clean speech data to use for model training purposes, making this corpus more suitable for speech recognition or biometric authentication.

### 3.6.5 *The GRID Corpus*

The GRID corpus (Cooke *et al.* [46]) (example frame shown in figure 14), is an English language multimodal speech corpus developed in the University of Sheffield. It contains 1000 sentences

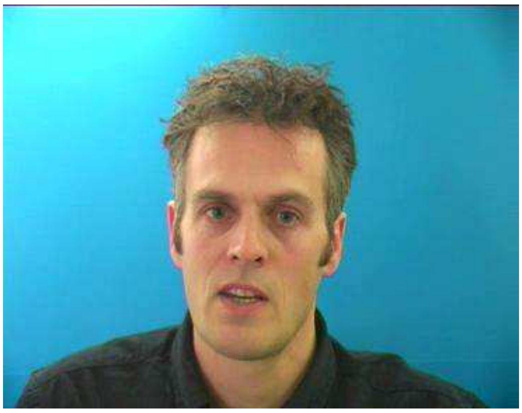


Figure 14: Example of a frame from the GRID corpus.

from each of 34 speakers, 18 male and 16 female. This corpus is free to download, and comes in the form of a number of videos of each speaker, with one video per sentence. The audio data is also provided in a separate file, making the corpus easy to use. Like VidTIMIT, there are no physical restrictions placed on the speaker, and there are a very large range of speakers and sentences available. Additionally, this corpus does not have the problem of background noise, as the data is recorded in a visually and acoustically clean environment. The sentence structure is influenced by the Coordinate Response Measure (CRM) task (Moore [132], Bolia *et al.* [27]). CRM sentences were recorded in the format of "READY callsign GO TO color digit NOW", with the lower case words in each recording of a sentence being replaced with a variety of data. In the GRID corpus, this was extended to use sentences in the form of "COMMAND COLOUR PREPOSITION LETTER DIGIT ADVERB", with each word being changeable, producing example sentences such as "Put red at G 9 now." This variation results in a very large corpus of 34000 sentences.

This corpus has some extremely useful benefits. It is very large, with a large selection of words, speakers, and sentences. It is also recorded in a visually and acoustically clean environment, without any significant background noise. It is also very easy to download and use.

### 3.6.6 *Audiovisual Speech Corpora Summary*

The number of available audiovisual speech corpora has grown significantly in recent years, with many large databases now available. Databases like CUAVE (Patterson *et al.* [141]) were not considered in this work due to the limited speech content, and single modality databases were also not considered. Of the corpora reviewed, the three most suitable multimodal databases were found to be the GRID corpus, VidTIMIT, and AVICAR. Of these, AVICAR was considered the least convenient to use, and GRID and VidTIMIT were chosen for use in this thesis. VidTIMIT was used for early correlation research, but the main speech enhancement experiments (as described in chapter 5), utilised the GRID corpus due to the superior clean audio speech source quality.

## 3.7 SUMMARY

The established relationship between audio and visual aspects of speech production and perception and the correlation between these speech modalities was examined by the author in chapter 2. This chapter presented a detailed review of the literature relevant to this work. Firstly, as the aim of this research is to develop a framework with consideration of potential long term practical applications to future hearing aid design, a summary of current hearing aid technology was provided, describing directional microphones, noise reduction algorithms, and microphone arrays, which are proposed to be used as part of the system presented in this thesis. Chung [44] also gave examples of systems where several of these elements are combined, so for example, combining directional microphones with noise reduction algorithms, which provides precedent for the two-stage speech enhancement system proposed in the next chapter. This chapter also showed that another feature of modern hearing aids is that some make use of decision rules, where a range of detectors are applied to the input signal, with subsequent speech filtering decided by application of decision rules based on different inputs. Although these are single modality approaches, this serves as the basis for further development later in this thesis, when the initial multimodal system is enhanced to utilise fuzzy logic as part of a more intelligent fuzzy-logic based speech enhancement system.

As well as placing this work in the context of audio-only hearing aid research, another relevant research field summarised in this chapter is audiovisual speech filtering. This chapter presented a review of current audiovisual speech enhancement research, describing the initial speech enhancement algorithms published in the literature, as well as subsequent state-of-the-art developments in this field. To build up a picture of state of the art multimodal speech enhancement, three particular strands of research were focused on in detail. Audiovisual fragment decoding (Barker & Shao [18]) was examined, as well as visually derived Wiener filtering (Almajai & Milner [11]), and multimodal blind source separation (Rivet [151]). The system proposed in this thesis aims to utilise techniques similar to those pioneered by Milner & Almajai [130], as part of a novel multimodal speech enhancement system, utilising both audio and audiovisual speech filtering techniques. In addition, a brief summary was provided of state of the art lip detection approaches, most prominently, the widely used Viola-Jones detector.

Finally, the development of the system described in the next chapter has to be tested with an audiovisual speech database. There are a range of research corpora available for this purpose, and a selection of these were evaluated to assess their suitability for use as part of the speech filtering research discussed in this thesis. It was found that the most suitable corpora for use were the GRID Corpus (Cooke *et al.* [46]), and VidTIMIT (Sanderson [157]).

With the problem defined, and the research context established in this chapter, chapter 4 presents a proposed two-stage audiovisual speech enhancement system inspired by the audiovisual speech filtering research discussed in this chapter. The system proposed in chapter 4 follows the principle described in this chapter of using directional microphones and noise reduction algorithms together, as found in currently available audio-only commercial listening devices.

# 4

# A TWO STAGE MULTIMODAL SPEECH ENHANCEMENT SYSTEM

## 4.1 INTRODUCTION

The overall aim of this thesis is to utilise the relationship between audio and visual aspects of speech in order to develop a speech enhancement system. This proposed multimodal speech enhancement system aims to make use of both audio-only and visually derived filtering techniques as part of a multi-stage filtering approach. Chapter 2 described the general background behind this research, showing the established relationship between audio and visual aspects of both speech perception and production, and also presented research into multimodal speech correlation. Chapter 3 built on this background work by presenting a detailed literature review summarising current relevant audiovisual speech enhancement work in this research domain.

This chapter builds on the previous two by providing a detailed description of the initial two-stage multimodal speech enhancement system presented in this thesis. This represents a combination of a variety of state-of-the-art techniques, all integrated into one novel system. Each individual component is described in detail, covering feature extraction, audiovisual Wiener filtering, the audiovisual model required by this filtering approach, and audio-only beamforming. This system is described in this chapter and an evaluation of the strengths and weaknesses of this approach is presented in the following chapter. Chapter 4 presents the technical description of this multimodal speech enhancement system.

The remainder of this chapter is divided into a number of sections. In section 4.2, an overall summary of the system is presented, with the motivation and the framework summarised. The simulated reverberant room used for production of noise and speech mixtures is then described in section 4.3. Each component of the proposed system is then described in depth, starting with the microphone array (4.4), with sections also dedicated to audio feature extraction, lip tracking using the SAAM approach, and then visual feature extraction. The visually derived Wiener filtering approach is then discussed in section 4.7, followed by a description of the GMM-GMR technique used for audiovisual speech modelling. Section 4.9 presents the audio-only noise cancelling directional microphone beamforming approach (which exploits the spatial diversity of different speech and noise source locations to suppress signals from unwanted directions) used in this work, and finally, section 4.10 summarises this chapter.

## 4.2 OVERALL DESIGN FRAMEWORK OF THE TWO-STAGE MULTIMODAL SYSTEM

The speech filtering system presented in this chapter is an extension of existing audio-only concepts, in that it extends the concept of an audio-only two-stage filtering system that combines multiple audio-only filtering techniques into one integrated system, as demonstrated in examples by Zelinski [189], Van den Bogaert *et al.* [173]. These systems are theoretically more powerful than those using only a single technique due to the additional filtering offered by utilising a combination of techniques and the addition of visual information may add more potential still. The system presented in this chapter extends this idea by combining audio beamforming with visually derived Wiener filtering to produce a novel integrated two-stage speech enhancement system, theoretically capable of functioning in extremely noisy environments. The overall diagram of this multimodal system is shown in figure 15.

The overall system diagram, as seen in figure 15, shows that the system is presented with two inputs. Firstly, there is the audio input, which consists of a mixed speech and noise source, and there is also a visual input, in the form of a video recording of a matching speech source. The system receives the mixed audio signal input in the form of a microphone array and then processes the signal with first visually derived Wiener filtering and then beamforming. Although this is a state-of-the-art filtering system, the feasibility of a hardware implementation of these algorithms in the near future is not beyond the realms of possibility. An example of exciting new research developments being implemented in hardware is with hearing aid

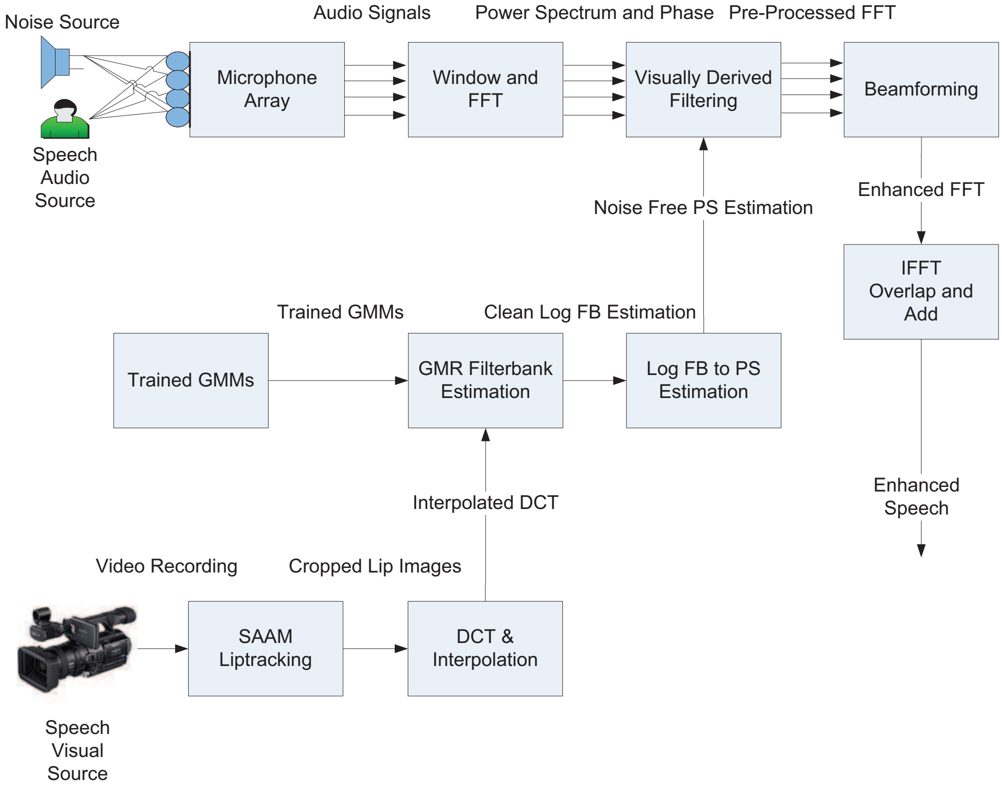


Figure 15: Block diagram of multimodal two-stage filtering system components.

products that have a multiple microphone array and processing hardware built discreetly into a pair of glasses (Mens [128] Varibel-Innovations [174]). This demonstrates that miniaturisation of speech processing hardware is reaching a stage where the concept of making use of a camera in addition to microphones to augment such a system is becoming more feasible for the end user.

To summarise the components utilised in this system and shown in figure 15, the audio signal is received by the microphone array, and this signal is then windowed and transformed into the frequency domain. This audio signal is then used as the noisy input into a visually derived Wiener filtering process. In order to carry out this pre-processing operation, associated visual features have to be extracted. This is carried out by utilising SAAM lip tracking to extract mouth region information from the input image sequence, and 2D-DCT features are then extracted and interpolated to match the audio input. This DCT information is then used along with GMMs (which are trained offline) to produce an estimate of the noise-free audio signal in the filterbank domain. This is then used to perform Wiener filtering. After this, the pre-processed signals are then filtered using audio-only beamforming to produce an enhanced frequency-domain signal. Finally, this is then transformed back to the time domain and output. The individual components of this system are each described in full detail in the remainder of this chapter.

As can be seen in figure 15, there are a number of different components in this novel integrated system. One strength of this system is that these individual components are loosely coupled. For example, with regard to visual feature extraction, the GMM-GMR filterbank estimation approach requires visual DCT information, but this does not have to be provided specifically by the SAAM lip-tracking approach. The system is designed so that the lip tracking technique can be easily upgraded without having to redesign the whole system. The same is true for the GMM-GMR speech estimation stage, and also the two individual speech-filtering stages. The system is explicitly designed to be scalable in order to take account of future state-of-the-art developments in each of the specific associated research domains.

## 4.3 REVERBERANT ROOM ENVIRONMENT

In order for speech filtering to be performed in an experimental environment, the speech and noise sources have to be mixed. There are two main alternatives, additive or convolved mixtures. Additive mixtures are the most simple method of combining speech and noise, and simply consist of combining speech plus noise "speech+noise" to create a noisy speech mixture. Although this is the simplest type of mixture to calculate and filter, and is used in the literature (Almajai *et al.* [13]), it is not always necessarily a realistic mixture. A simple additive mixture does not take into account factors such as the difference in location of source and noise, atmospheric conditions such as temperature and humidity, or reverberation (a natural consequence of broadcasting sound in a room). Reverberation, in the context of this research, refers to the situation where large numbers of echoes are built up during the transmission of a sound due to environmental factors such as a small room. These echoes take time to dissipate, and so have an effect on the input received by a microphone or human listener. Convolved mixtures of speech and noise can provide a more realistic noisy speech

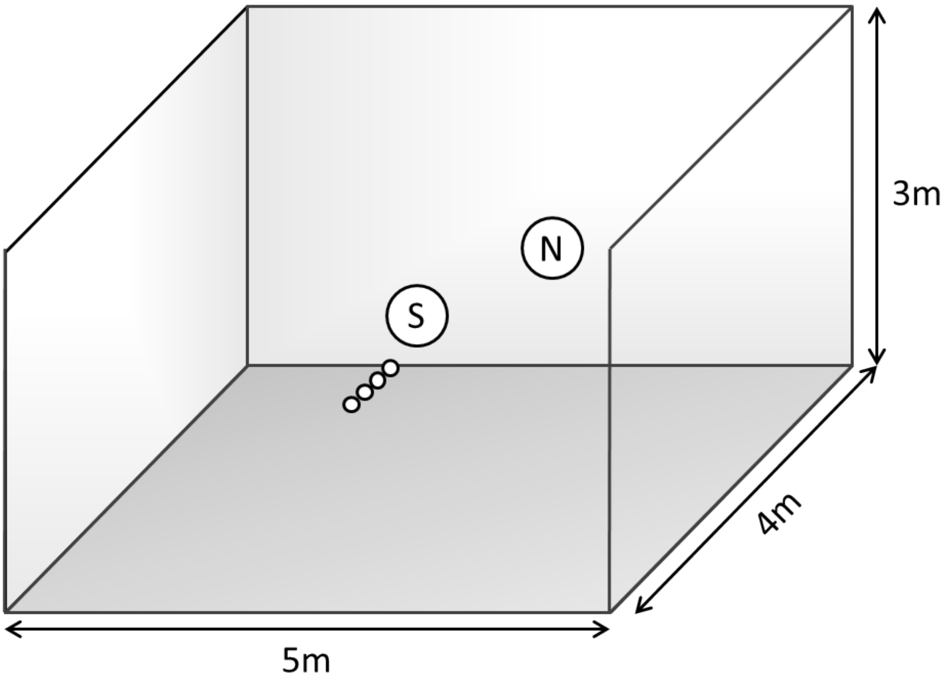


Figure 16: Diagram of simulated reverberant room used in this work to create noise and speech mixtures. S represents the speech source, and N the noise source. The four circles represent the microphone array used to receive the speech mixtures.

mixture. These convolved mixtures do not simply add the sounds together, but necessitate the construction of a mixing matrix. Convolved mixtures are used in source-separation based speech-filtering problems Rivet *et al.* [147], Hussain *et al.* [91], and represent a better and more detailed example of a realistic noisy speech mixture in the context of this research.

In the work presented in this thesis, the noisy speech mixtures used are mixed in a convolved manner. To do this, a simulated room environment is used, with the speech and noise sources transformed with the matching impulse responses. Impulse responses represent the characteristics of a room when presented with a brief audio sample, and these are then applied to the speech and noise signals in the context of their location within the simulated room. This gives them the characteristics of being affected by environmental conditions with regard to microphone input. These sources are then convolved. This process is described in more detail in section 4.4 of this chapter.

In order to create this speech mixture, the simulated room used in this work has a number of parameters that have to be defined. Firstly, it is assumed that the speech and noise sources originate at different locations within this simulated room, and a diagram of this room is shown in figure 16. It can be seen in this figure that this room has been designed with dimensions of 5 by 4 by 3 metres and this room is considered to be a closed room (i.e. completely walled with a ceiling). It is assumed for the purposes of calculating speed of sound that the air temperature is 20 degrees Celsius, with humidity of 40%. In this diagram, figure 16 also shows the positions of the speech and noise sources, with the speech source (marked in the diagram with 'S') located at an xyz position of 2.50 metres, 2.00 metres, 1.40 metres, and the noise source (marked by 'N') having xyz coordinates of 3.00 metres, 3.00 metres, 1.40 metres. Finally, the position within the room of the simulated microphone input array is represented in the diagram by four small circles, and is located at an x coordinate of 2.20 metres, y coordinates ranging from 1.88 to 2.12 metres, and at a height (z coordinate) of 1 metre. The exact convolution calculations are discussed in section 4.4. The specific speech and noise sources vary depending on the experiments performed and the different SNR desired, and these are discussed in more detail in chapter 5.

## 4.4 multiple microphone array

The previous section described the reverberant room environment used in this thesis. This environment consisted of a closed room with speech and noise sources originating at different locations, as shown in figure 16 (with 'S' representing the speech source, and 'N' the noise source). This figure also shows that a multiple microphone array is used to receive the noisy speech mixtures. The reason for these multiple microphones is to allow the directional beamforming aspect of the integrated two-stage system to function. Directional microphone filtering is commonly used in state-of-the-art hearing aids, as summarised in chapter 3. As shown in figure 15 and the room diagram in figure 16, in the work presented in this chapter, the noise and speech mixtures are received by an array of four microphones, with each microphone at the same x coordinate, the same z coordinate, but slightly different y coordinates. The first microphone is positioned 2.20 metres along the length of the room (x coordinate), at a height of 1 metre (z coordinate), and with a y coordinate of 1.88 metres. The second microphone is positioned at the same x and y coordinates, but with a y coordinate of 1.96 metres. The third and fourth microphones also have the same x and z coordinates, but have y coordinates of 2.04 metres and 2.12 metres respectively.

In the simulated room environment, it is assumed in this work that the noise and signal come from different sources, as shown in figure 16. In this research, the noisy speech mixtures are received by an array of four microphones. To summarise this concept, assume $M$ microphone signals, $z_1(t), \dots, z_M(t)$ record a source $x(t)$ and $M$ uncorrelated noise interfering signals $\hat{x}_1(t), \dots, \hat{x}_M(t)$. Thus, the $m$-th microphone signal is given by,

$$Z_m(t) = a_m(t) * x(t) + \hat{x}_m(t), \quad 1 \leqslant m \leqslant M \tag{4.1}$$

where $a_m(t)$ is the impulse response of the $m$-th sensor to the desired source, and $*$ denotes convolution. In the frequency-domain convolutions become multiplications. Furthermore, since in this thesis there is no interest in balancing the channels, the source is redefined so that the first channel becomes unity. Hence, applying the Short-Time Fourier Transform (STFT) to (4.1), this results in,

$$Z_m(k,l) = A_m(k,l)X(k,l) + \hat{X}_m(k,l), \quad 1 \leqslant m \leqslant M \tag{4.2}$$

where $k$ is the frequency bin index, and $l$ the time-frame index. Thus, this produces a set of $M$ equations that can be written in a compact matrix form as,

$$\mathbf{Z}(k,l) = \mathbf{A}(k,l)X(k,l) + \hat{\mathbf{X}}(k,l) \tag{4.3}$$

with,

$$\begin{aligned} \mathbf{Z}(k,l) &= [Z_1(k,l)Z_2(k,l)...Z_M(k,l)]^T \\ \mathbf{A}(k,l) &= [A_1(k,l)A_2(k,l)...A_M(k,l)]^T \\ \hat{\mathbf{X}}(k,l) &= [\hat{X}_1(k,l)\hat{X}_2(k,l)...\hat{X}_M(k,l)]^T \end{aligned} \tag{4.4}$$

This produces the convolved Fourier transformed microphone mixtures of speech and noise, represented by $Z_1(k,l), \dots, Z_4(k,l)$. Each of these represents the transformed noisy input of an individual microphone, sampled at 8 kHz. These input signals are then used to extract audio log filterbank features, described in section 4.5, to produce the audio log filterbank values $F_a(1), ..., F_a(4)$ for each microphone input, using an audio frame of 25ms with a 50 percent overlap. At the same time, matching visual DCT features $F_v$ are extracted from the video recordings, which will be described in section 4.6.

## 4.5 AUDIO FEATURE EXTRACTION

The previous section described the process of receiving noisy speech signals from the microphone inputs and transforming them to the frequency domain. These transformed signals are used in the second stage of this integrated multimodal system to carry out audio-only beamforming. Before the second filtering stage, the initial stage of the filtering process makes use of visually derived Wiener filtering. For this, the audio input has to be transformed to make it possible for audio to be estimated given visual data (as is described in more detail in section 4.7). The initial Fourier transformed noise mixtures are further transformed to produce the magnitude spectrum, and subsequently log filterbank values. These are used as part of the visually derived filtering process.

Therefore, in the system presented in this chapter, a filterbank dimension of $M = 23$ filters is used. This is motivated by other work in the literature such as Almajai *et al.* [13], which uses the same size of filterbank. The relationship between linear and Mel frequency is given by,

$$\hat{f}_{mel} = 2595 \cdot \log 10 \left(1 + \frac{f_{lin}}{700}\right) \tag{4.5}$$

In this implementation, the limits of the frequency range are the parameters that define the basis for the filter bank design. The unit interval $\Delta\hat{f}$ is determined by the lower and the higher boundaries of the frequency range of the entire filter bank, $\hat{f}_{high}$ and $\hat{f}_{low}$, as follows,

$$\Delta\hat{f} = \frac{\hat{f}_{high} - \hat{f}_{low}}{M + 1} \tag{4.6}$$

the centre frequency $\hat{f}_{cm}$ of the $m - th$ filter is given by,

$$\hat{f}_{cm} = \hat{f}_{low} + m\Delta\hat{f}, \quad m = 1, ..., M - 1 \tag{4.7}$$

where $M$ represents the total number of filters in the filterbank. The conversion of the centre frequencies of the filters to linear frequency (Hz) is given by,

$$\hat{f}_{cm} = 700 \cdot \left(10^{\hat{f}_{cm/2595}} - 1\right) \tag{4.8}$$

and the shape of the $m$-th triangular filter is defined by,

$$H_m(k) = \begin{cases} 0 & k < f_{b_{m-1}} \\ \frac{k-f_{b_{m-1}}}{f_{b_m}-f_{b_{m-1}}} & f_{b_{m-1}} \leqslant k \leqslant f_{b_m} \\ & \qquad\qquad m = 1, ..., M \\ \frac{f_{b_{m+1}}-k}{f_{b_{m+1}}-f_{b_m}} & f_{b_m} \leqslant k \leqslant f_{b_{m+1}} \\ 0 & k > f_{b_{m+1}} \end{cases} \tag{4.9}$$

where $f_{b_m}$ are the boundary points of the filters and $k = 1, ..., K$ corresponds to the $k - th$ coefficient of the K-points DFT. The boundary points $f_{b_m}$ are expressed in terms of position, which depends on the sampling frequency $F_{samp}$ and the number of points $K$ in the DFT,

$$f_{b_m} = \left(\frac{K}{F_s} \cdot f_{cm}\right) \tag{4.10}$$

For computing the log filterbank parameters, the magnitude spectrum $|Z(k)|$ of each mixed noisy audio signal acts as the input for the filterbank $H_m(k)$. Next, the filterbank output is logarithmically compressed to produce the log filterbank audio signal for further processing,

$$F_a = \ln\left(\sum_{k=0}^{K-1} |Z(k)| \cdot H_m(k)\right) \tag{4.11}$$

This produces the log filterbank signal $F_a$ of each microphone input, which is used, along with matching visual information (the extraction of which is discussed in the next section) to carry out visually derived speech filtering.

## 4.6 VISUAL FEATURE EXTRACTION

As has been stated previously, the system presented in this thesis is multimodal. This means that in addition to the audio features discussed in the previous sections, in order to perform visually derived filtering, there is an obvious requirement for visual feature extraction. The visual filtering algorithm relies upon DCT vectors taken from lip information, and this section discusses the extraction of these features.

In this thesis, visual lip features are extracted by using the state-of-the-art SAAM approach pioneered by Nguyen & Milgram [135] and also used in work carried out by the author in Abel *et al.* [2]. There are many different approaches to lip tracking and visual tracking in general, with much active research and many state of art commercial solutions such as the Microsoft Kinect, that makes use of infrared and Red Green Blue (RGB) cameras to track skeletal frames. There are many such techniques, and the detailed exploration of these is considered to be outside the scope of this thesis, and will not be discussed in this work. It was decided to make use of the SAAM technique for a variety of reasons. Firstly, although it is possible to extract lip information by manually cropping each frame, this technique is obviously extremely time consuming. Although the approach of manually extracting each ROI from a video sequence by cropping was used for some of the preliminary correlation work presented in the background chapter (chapter 2), it was decided that this process had to be automated when research was extended to a larger scale. The chosen approach had one key advantage in that it was considered state-of-the-art, and collaborative work took place with the developers of this technique (Nguyen & Milgram [135]) who also reported good results with this technique, to adapt and test this system in the context of multimodal speech processing, resulting in the publication of the first utilisation of the SAAM approach in the context of speech correlation in Abel *et al.* [2]. Another benefit of using this approach was that the chosen tracking technique was a standalone component that could be integrated into the system without difficulty. Its loose integration with the other components of this system also means that it is feasible for this approach to be replaced in future work with a different front end without any difficulty. The results of informally testing this SAAM approach with a range of different speech sentences from the GRID corpus found that the system was able to reliably and successfully track data from this corpus; making it suitable for use in this work. Finally, there was also the 'convenience' aspect of using this approach. As will be explained later in this section, the system uses online training, and learns each sentence automatically, frame by frame, allowing the focus of this research to be mainly on the speech filtering aspect.

There are three main components needed for the successful utilisation of visual information as part of a speech processing system. The first is accurate ROI detection. The second is the ability to automatically track and extract the ROI in each frame of a speech sentence, taking into consideration that the speaker may not remain completely still throughout. Finally, each cropped ROI frame has to be extracted and transformed into a format suitable for further processing.

While there are many lip tracking approaches that have been proposed in the literature (Cheung *et al.* [39], Nguyen & Milgram [135]), some of these require the initial identification of the ROI, with approaches such as AcAMs (Cootes *et al.* [47]) being highly dependent on the setting of initial parameters. Therefore, while the SAAM lip tracking approach is used in this thesis, the extraction of the ROI is a separate issue. As stated, there are many examples of both lip tracking and ROI detection in the literature (Iyengar *et al.* [93], Wark & Sridharan [179]), and a summary of these was presented in chapter 3. One key point is that widely used appearance based techniques require considerable time consuming training, due to the requirement for manual training on images. Therefore, it was decided to make use of an implementation of the Viola-Jones (Viola & Jones [175]) detector by (Kroon [104]). The Viola-Jones detector, and variations of it, are amongst the most widely used ROI detection techniques for facial features, with the implementation discussed in more detail in chapter 3.

Essentially, the Viola-Jones has three main components, the integral image, classifier learning with adaboost, and an attentional cascade structure. The integral image (Crow [50]), also known as the summed area table, is a technique used for quickly computing the sum of values in a rectangular subset of a grid, and is used for rapid calculation of Haar-Like features. Haar-Like features use the integral image technique to calculate the sum of intensities in specific rectangular regions within an image of interest. The sum of intensities can then be compared for neighbouring regions, and then the difference between each of these regions can be calculated, so for example, a lip boundary can be identified by finding a difference

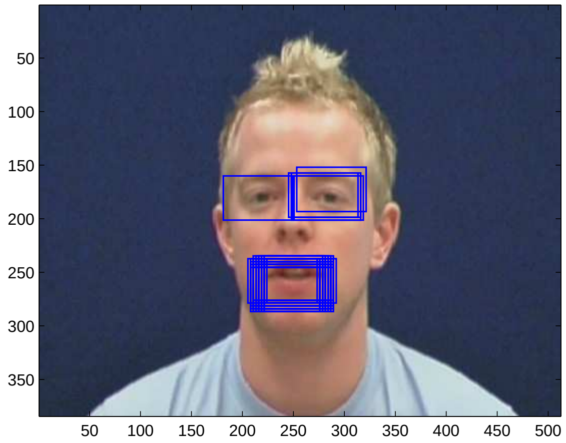


Figure 17: Example of initial results from Viola-Jones Lip Detection process. Rectangles indicate a possible candidate ROI.

between lip and non-lip pixel intensities. The Viola-Jones (Viola & Jones [175]) detector uses comparisons of 2, 3, and 4 rectangles as part of the detection process.

The training of an appearance based lip detector can be a very intensive and time consuming process, and so in this work, it was decided to make use of Haar-Like features that have been specifically trained for lip detection and are available as part of the OpenCV Bradski [29] library, limiting the need for further training. The chosen lip detection implementation runs the algorithm on a single image frame, and then returns a number of potential candidate ROI areas. An example of this initial output is shown in figure 17.

It can be seen in the example given in figure 17 that a number of these candidates are correct, with the majority centred around the mouth region as would be expected. However, it can also be seen that a number of other potential matches are found, in the example of figure 17 a number of candidates can be seen around the left eye region of the image, as well as another candidate around the right eye. This is a common occurrence, and the final decision was made by customising the initial code in a similar manner to other work, including by Viola & Jones [175]. The candidate areas were divided into a number of different subsets by firstly identifying overlapping candidates. If one candidate had an overlap with another of 65% or greater (value chosen manually after investigation of preliminary test data), then the candidates were considered to belong to that subset. This had the effect of both dividing the candidates into subsets, and also removing values with only a small overlap. It was

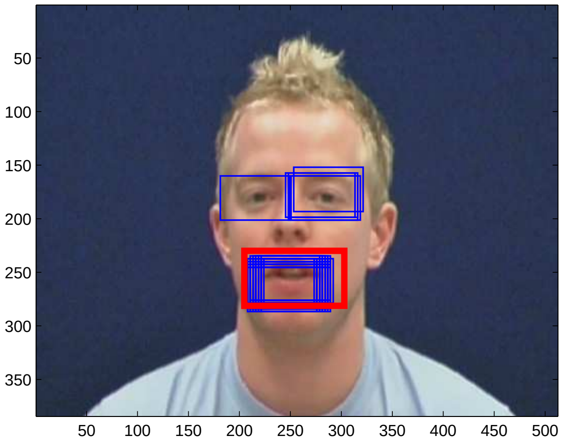


Figure 18: Example of initial results from Viola-Jones Lip Detection process. Thinner rectangles indicate a possible candidate ROI, and the thicker rectangle represents the final selected mouth region to be used for tracking.

then assumed for the purposes of this research that the subset with the greatest number of candidate members was the relevant mouth area. This is because the data used for this research concentrates on single face scenarios.

To calculate the final ROI area, the mean of all the values within the most relevant subset was calculated, producing a single rectangular ROI. Testing (which will be discussed in more depth in chapter 5) of the detector identified a number of issues. Firstly, to produce a more relevant ROI, it was found that the detected areas were smaller than ideal. This was solved by adding 20 pixels to the width and 10 to the height in order to contain more useful data. An example of this refined output can be seen in figure 18. This is the same image as figure 17, but the thicker rectangle represents the final output.

The second issue was that in a very limited number of cases, the detector did not successfully identify the lip region. The problem lies with the trained Haar-Like features, rather than in the code, and this was solved by running a check on the initial number of candidates. If an image was found to produce less than 7 candidate ROI areas, then a second scan of the image was run, using different Haar-Like features. These were trained to identify the whole face, and were found to work reliably in all tested incidences. The face region was then cropped to produce an estimate of the lip region. This was found to be sufficient to produce a valid result, and was deemed to be suitable for use as part of the work presented in this thesis. Lip detection testing is discussed in more detail in section 5.4 of chapter 5.

This lip detection was used in the first frame of the image to identify the ROI. As previously stated, the second aspect, lip tracking, is handled by using the SAAM technique. Finally, the extraction of visual information into a usable format 2D-DCT) is covered later in this section.

Lip tracking essentially deals with non-stationary data, as the appearance of a target object may alter drastically over time due to factors like pose variation and illumination changes. The lip tracking framework discussed is based on the Adaptive Appearance Models (AAM) approach by Levey & Lindenbaum [111], which allows for the updating of the mean and Eigen vectors of g-dimensional observation vectors $v_o \in R^g$. First, the AAM technique is extended by inserting a supervisor model (Golub & Van Loan [72]) that verifies the AAM performance at each frame in the sequence, by using a Support Vector Machine (SVM) to filter the AAM result for an individual frame, as shown in (4.12) ,

$$y\left(v_o\right) = sgn\left(\sum_i \alpha_i \hat{v}_i \omega\left(v_o i, v_o\right) + b\right) \tag{4.12}$$

Where $y(v_o) \in \{-1, 1\}$ signifies whether $v_o$ represents a good or bad result. $\alpha, b$ are trained offline with the SVM (Cauwenberghs & Poggio [34]), $\omega\left(.\right)$ is the Gaussian kernel function and $v_{oi}$,$v$ are trained and observation vectors respectively. Each $\hat{v}_i$ represents the desired output of each example $v_{Ti}$ from the offline training dataset.

Secondly, shape models are constructed to allow the SAAM technique to track feature points in video sequences. To model deformation, a shape model is formed,

$$S^\circ = S^\circ + P_s b \tag{4.13}$$

where $S^\circ = \left(v^\circ_{o1}, \hat{v}^\circ_1, \ldots, v^\circ_{oj}, \hat{v}^\circ_j\right)$ is a normalised shape and j represents a number of feature points. To track these, it is sufficient to find the parameters,

$$p = [b_{o1}, \ldots, \phi_v, \phi_q, \theta, s] \tag{4.14}$$

where $b_i$ is the coefficient to deform $S^{\circ}$, and $\phi_v, \phi_q, \theta, s$ represent translations, rotation and scale parameters respectively. To track a target object, the aim is to maximize the cost function given in (4.15) as follows,

$$p^* = \arg\max_p (y_e) \tag{4.15}$$

Where $y_e$ is a negative exponential of projection error between $v_{ot}$ and the Principal Component Analysis (PCA) subspace created by earlier observations, defined by equation 4.16 as follows,

$$y_e = exp\left(-\left\|(v_{ot} - \bar{v}_o) - UU^T(v_{ot} - \bar{v}_o)\right\|^2\right) \tag{4.16}$$

Note that the distance $y_e$ is a Gaussian distribution, with Eigen vectors $U$ and mean $\bar{v}_o$, $y_e = p(v_{ot}|p) \propto J\left(v_{ot}; \bar{v}_o, UU^T + \epsilon I\right)$ as $\epsilon \to 0$, and the inverse matrix can be solved by applying the Woodbury formula (Golub & Van Loan [72]), given in equation 4.17 as follows,

$$(UU + \epsilon I)^{-1} = \epsilon^{-1}\left(I - (1+\epsilon)^{-1} UU^T\right) \tag{4.17}$$

The optimal parameter $p^*$ is found with a number of iterations. Here, empirical gradient is used, since the cost function is evaluated in the neighbourhood of the current parameter vector value. The tracking algorithm used in this thesis works as follows,

1. Manually locate target object in the first frame (t=1). Eigen vectors $U$ are initialized as empty. The tracker initially works as a template based tracker.

2. At the next frame, find the optimal parameters $p^* = argmax\left(\left\{y_e\left(p_i^{\omega}*\right)\right\}\right)$ over a number of iterations:

   - For each parameter $p_i$.
   - For each $\Delta p$ and $\omega \in \{-1, 1\}$, compute $p_i^{\omega}(p_1, \ldots, p_i + c\Delta p, \ldots, p_{c+4})$
   - Compute $i^* = max\left\{y_e\left(p_i^{\omega}\right)\right\}$
   - Do $p \leftarrow p_i*$, store $y_e\left(p_i^{\omega}*\right)$

3. Check the observation vector: $v_o = v_o\left(\Xi\left(S_e^{'}, p^*\right)^{-1}\right)$ where $\Xi$ is a transformation matrix, with result estimation phase as shown in equation 4.12

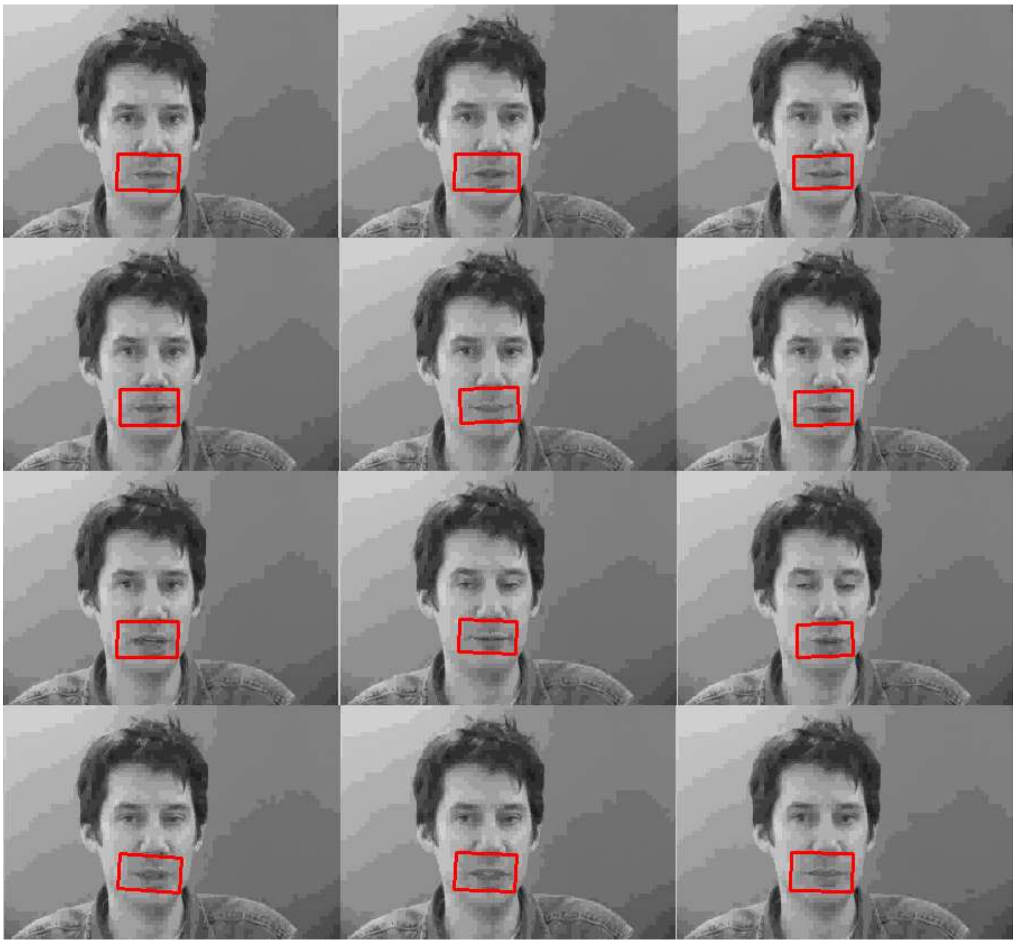


Figure 19: Demonstration of tracker running on an image sequence from the GRID audiovisual corpus. Selected frames from one sequence are shown. This figure demonstrates the automatic movement of the ROI to maintain focus on the lip region.

4. If $y(v_o) = 1$, this signifies a good result to add to the model. When the desired number of new images has been accumulated, perform an incremental update.

5. Return to step 2.

The performance of this tracker is shown in figure 19, which shows selected frames from one image sequence. The rectangle around the lip region represents the ROI, which was manually identified in the first frame, and subsequent frames then automatically tracked this region to maintain focus on the desired area.

After tracking a sequence of lip images with this technique, the 2D-DCT vector $F_v = 2D - DCT(v)$ of each image in the sequence is found. A number of different visual feature extraction techniques have been used in the literature, but as shown in chapter 2, 2D-DCT is extremely common and has been used by others, such as Sargın *et al.* [159], Almajai & Milner [10]).

DCT was originally developed in 1974 by Ahmed *et al.* [6], and is a close relative of the DFT. This was extended for application with image compression by Chen & Pratt [37]. The one-dimensional DCT is capable of processing one-dimensional signals such as speech waveforms. However, for analysis of two dimensional signals such as images, a 2D-DCT version is required. For a $V_U$ x $V_V$ matrix $V_P$ of pixel intensities, the 2D-DCT is computed in a simple way: the 1D-DCT is applied to each row of $V_P$ and then to each column of the result. Thus, the transform of $V_P$ is given by the DCT matrix $V^{DCT}$ ,

$$V^{DCT}_{m,n} = V_{Wm} V_{Wn} \sum_{Vu=0}^{V_U-1} \sum_{Vv=0}^{V_V-1} V_{P vu,vv} \cos\left( m(2V_u+1) \cdot \frac{\pi}{2V_u} \right) \cos\left( n(2V_v+1) \cdot \frac{\pi}{2V_v} \right) \quad (4.18)$$

with $0 \leqslant m \leqslant V_U - 1, 0 \leqslant n \leqslant V_V - 1$,

$$V_{Wn} = \begin{cases} \sqrt{1/V_v} & \text{if n=0} \\ \sqrt{2/V_v} & \text{otherwise} \end{cases} \quad \text{and } V_{Wm} = \begin{cases} \sqrt{1/V_U} & \text{if m=0} \\ \sqrt{2/V_U} & \text{otherwise} \end{cases} \quad (4.19)$$

Since the 2D-DCT can be computed by applying 1D transforms separately to the rows and columns, this means that the 2D-DCT is separable in the two dimensions. The first 30 2D-DCT components of each image are vectorised in a zigzag order to produce the vector for a single frame in an image sequence. The resulting 2D-DCT sequence of frames is then interpolated to match the equivalent audio log filterbank matrices for the matching speech sentence. As the video used in this work is recorded at 25fps, this means that the 2D-DCT sequence is upsampled to match the audio features by using the same visual feature frame for four consecutive audio frames, a technique commonly used in the literature.

## 4.7 VISUALLY DERIVED WIENER FILTERING

Wiener filtering (Wiener [180]) is a signal processing technique that aims to clean up a noisy signal by comparing a noisy input signal with an estimation of a noiseless signal. This technique is very commonly used in image processing for image reconstruction (Hiller & Chin [83]) by removing noise from degraded images, and has also been experimented with in speech enhancement to filter an audio speech signal (Almajai *et al.* [13], Zelinski [189]). One challenging aspect of Wiener filtering is the acquisition of an estimation of the noiseless signal. Unlike some other speech filtering approaches, some knowledge of the original signal is required.

In this thesis, visual information is used to filter speech by making use of the relationship between acoustic and visual components of speech production as the means of producing the estimate of the original audio signal, and comparing this estimate to the noisy signal. This represents the first stage of filtering in this two-stage approach and acts as the pre-processing step before the audio-only beamforming described later in this chapter. The Fourier transformed audio signal is used as an input, in tandem with associated visual information. In this system, the Wiener filter, $W(\gamma)$, is calculated in the frequency domain from the Power Spectrum (PS) estimate of clean speech ($\Psi_{\hat{a}}(\gamma)$) and the noisy speech mixture PS ($\Psi_{a}(\gamma)$) as,

$$W(\gamma) = \frac{\Psi_{\hat{a}}(\gamma)}{\Psi_{a}(\gamma)} \tag{4.20}$$

This produces the Wiener filter to be applied to the input signal. However, as stated, this is challenging to implement in this form, due to the difficulty of accurately calculating the clean PS, $\Psi_{\hat{a}}(\gamma)$, from visual information. Firstly, there are a variety of ways to calculate the PS of the noisy signal. One example is utilised by Almajai & Milner [11], where a VAD is used to identify non speech frames, and a noise alone PS is calculated. This is then added to the estimated speech alone PS to produce an estimated noisy speech PS. However, in this work, it was decided to use the power spectrum of the noisy speech mixture as a whole. This is a parameter that can be varied without difficulty, and using the noisy PS on a frame by frame basis also allows for a wide frame by frame variation in potential volume or type (aircraft, white noise etc.) of noise source. So therefore, $\Psi_{a}(\gamma)$ is simple to calculate from the noisy audio signal, but it is less straight forward to estimate the noise free PS.

This is where the input data from visual feature information can be utilised. Although it is very hard to estimate PS information directly from visual information, it is possible to estimate log filterbank values. Therefore, in the system, it is proposed to make use of the log filterbank vectors $F_{a}(1), ..., F_{a}(4)$, as described in section 4.5, and the 2D-DCT vector $F_{v}$, which is calculated as outlined in section 4.6, as inputs into the filter, with each audio channel being processed separately. Previous work by others (Sargin *et al.* [160], Almajai *et al.* [13]) showed that it is possible to estimate audio features from visual features, and produce the estimated noise free log filterbank vectors $F_{\hat{a}}(1), ..., F_{\hat{a}}(4)$. The production of these estimates is described in more depth in section 4.8. For each channel, this is then transformed into a linear filterbank estimate of the clean audio signal, which is then interpolated to match the dimensionality of the audio 2D-DCT $\Psi_a(\gamma)$ with pchip interpolation (Fritsch & Carlson [65]). This produces an estimate of the noise free power spectrum, $\Psi_{\hat{a}}(\gamma)$, which can be used to find $W(\gamma)$ as shown in equation 4.20. To find the enhanced power spectrum value, $\Psi_{\bar{a}}(\gamma)$, the noisy power spectrum $\Psi_a(\gamma)$ and the Wiener filter $W(\gamma)$ can be used as given in equation 4.21,

$$\Psi_{\bar{a}}(\gamma) = \Psi_a(\gamma)W(\gamma) \tag{4.21}$$

The key aspect of equation 4.20 is producing an estimate of the clean audio filterbank signal $F_{\hat{a}}$ to use as part of the filter. In this work, it is proposed to make use of Gaussian Mixture Regression (GMR), as described by Calinon *et al.* [33], and outlined in section 4.8. Following this filtering, the phase, $\varpi(\gamma)$, of each $F_a$ is calculated and combined with $\Psi_{\bar{a}}(\gamma)$, to update the frequency domain Fourier transform $\mathbf{Z}(k, l)$ (see equation 4.3), for further processing.

## 4.8 GAUSSIAN MIXTURE MODEL FOR AUDIOVISUAL CLEAN SPEECH ESTIMATION

As mentioned in the previous section, a crucial aspect of this system is the production of an estimate of the noise free signal for use by the Wiener filter. In order to provide such an estimate, the joint audio and visual speech relationship has to be modelled. There are a variety of different approaches, for example, it is possible to make use of an approach utilising GMMs as demonstrated by Almajai & Milner [11]. There are also a range of other modelling alternatives available, such as using a number of different GMMs for each speech phoneme, requiring significant speech segmentation both in training and in the actual system. One alternative to this approach though, is one that was developed by Calinon *et al.* [33], GMM-GMR. This is a technique that was originally developed for robot arm training, and in the work presented in this thesis, it has been adapted to be applied to the estimation of log filterbank audio vectors given a training set for offline training and valid visual input data. Although the

Maximum a Priori (MAP) approach has been used in this field previously, to the knowledge of the author, this research represents the first example of applying this GMM-GMR technique to audiovisual data for speech filtering. The performance of this approach is discussed in chapter 5.

To implement this GMM-GMR approach, this work makes use a method first outlined by Calinon *et al.* [33] to encode the audiovisual signals in a mixture of GMMs, by considering each visual DCT vector $F_v$ as an input in order to find an estimation of the equivalent noiseless audio signal by using GMR.

A mixture model of Q components of the joint audiovisual vector $F_{av}$ is defined by a Probability Density Function (PDF),

$$e(F_{av}) = \sum_{q=1}^{Q} e(q)e(F_{av}|q) \tag{4.22}$$

with $e(q)$ representing the prior, and $e(F_{av} \mid q)$ representing the conditional PDF.

To model the joint audiovisual data $F_{av}$ of dimension $C$, an offline training set is needed to train a mixture of Q Gaussians of dimensionality $C$. The performance of this aspect of the system is dependent on the training data provided, and so a training set using the GRID Corpus was used for this purpose, combining audio and visual data into a single training set. The detailed composition of the training set is discussed in chapter 5. Returning to the PDF described in (4.22), the parameters in equation 4.22 become,

$$\begin{aligned} e(q) &= \pi_q \\ e(F_{av}|q) &= N(F_{av}; \mu_q, \Sigma_q) \\ &= \frac{1}{\sqrt{(2\pi)^C|\Sigma_q|}} e^{\frac{1}{2}((F_{av}-\mu_q)^T \Sigma_q^{-1}(F_{av}-\mu_q))} \end{aligned} \tag{4.23}$$

with $\pi_q$ representing the prior, $\mu_q$ the mean, and $\Sigma_q$ the covariance matrix of Gaussian component $q$. K-means clustering is applied to the joint vector training set to produce an initial estimate of GMM parameters, and Maximum Likelihood Estimation is performed on the model using Expectation Maximisation. The trained GMMs can then be used to perform GMR and with the aid of the input visual vector $F_v$, return an estimated value of the noiseless filterbank audio vector. For each frame of the speech signal, the mean and the covariance matrix of the Gaussian component q are divided into their visual and audio components, as defined by,

$$\mu_q = \{\mu_{v,q}, \mu_{a,q}\} \quad , \quad \Sigma_q = \begin{pmatrix} \Sigma_{v,q} & \Sigma_{va,q} \\ \Sigma_{av,q} & \Sigma_{a,q} \end{pmatrix} \tag{4.24}$$

For each Gaussian component q, $\hat{\Sigma}_{a,q}$, the expected conditional covariance, of $F_{a,q}$ given $F_v$ is defined as,

$$\hat{\Sigma}_{a,q} \quad = \Sigma_{a,q} - \Sigma_{av,q}(\Sigma_{v,q})^{-1}\Sigma_{va,q} \tag{4.25}$$

and $F_{\hat{a},q}$, the conditional expectation of $F_{a,q}$ given $F_v$ is defined as,

$$F_{\hat{a},q} \quad = \mu_{a,q} + \Sigma_{av,q}(\Sigma_{v,q})^{-1}(F_v - \mu_{v,q}) \tag{4.26}$$

$F_{\hat{a},q}$ and $\hat{\Sigma}_{a,q}$ are mixed depending on the probability that $q \in \{1, \ldots, Q\}$ has of being responsible for $F_v$, as shown by,

$$\beta_q = \frac{e(F_v|q)}{\Sigma_{i=1}^{Q} e(F_v|q)} \tag{4.27}$$

For a mixture of Q components, $\hat{\Sigma}_a$, the conditional covariance, and $F_{\hat{a}}$, the conditional expectation of $F_a$ given $F_v$ are defined as,

$$\hat{\Sigma}_a = \sum_{q=1}^{Q} \beta_q^2 \hat{\Sigma}_{a,q}, \quad F_{\hat{a}} = \sum_{q=1}^{Q} \beta_q F_{\hat{a},q} \tag{4.28}$$

Where $F_{\hat{a}}$ represents the estimated log filterbank signal to be processed further as described in section 4.7. This audio log filterbank estimation is then used as part of the visually derived filtering approach, and enables the first stage of the two-stage speech filtering process to be performed. The resulting filtered signals are then used for audio-only beamforming.

## 4.9 BEAMFORMING

Multiple microphone techniques such as beamforming can improve the quality and intelligibility of speech by exploiting the spatial diversity of speech and noise sources to filter speech. This is an active research field, with many different techniques developed. Within these techniques, one can differentiate between fixed and adaptive beamformers. The former combines the noisy signals by a time-invariant filter-and-sum operation, the latter combine the spatial focusing of fixed beamformers with adaptive noise suppression, such that they are able adapt to changing acoustic environments and generally exhibit a better noise reduction performance than fixed beamformers. The Generalised Sidelobe Canceller (GSC) is a very widely used structure for adaptive beamformers and a number of algorithms have been developed based on it. Among them, the general Transfer Function Generalised Sidelobe Canceller (TFGSC) suggested by Gannot *et al.* [67], has shown impressive noise reduction abilities in a directional noise field, while maintaining low speech distortion.

In this work, the TFGSC beamformer is used on the pre-processed speech and noise mixtures as extracted in section 4.4 and pre-processed in section 4.7. This single modality technique receives multiple microphone signals, and then utilises them to output a single filtered signal. This follows examples of directional microphones utilised in commercial hearing aids and multi microphone array listening aids, as summarised in chapter 3. In the system presented in this chapter, the input signals have been pre-processed by the visually derived Wiener filtering before being processed by audio-only beamforming. The beamforming approach used here is loosely integrated into the system, and so can be replaced by a different filtering mechanism to take account of further state-of-the-art research developments.

The general GSC structure is composed of three main parts: a Fixed Beamformer (FBF) $\mathbf{G}(k)$, a Blocking Matrix (BM) $\bar{\mathbf{G}}(k)$, and a multichannel Adaptive Noise Canceller (ANC) $\mathbf{H}(k, l)$. The FBF is an array of weighting filters that suppresses signals arriving from unwanted directions. The column of the BM can be regarded as a set of spatial filters suppressing any component impinging from the direction of the signal of interest, thus yielding $M - 1$ reference noise signals $\tilde{}(k, l)$. These signals are used by the ANC to construct a noise signal to be subtracted from the FBF output. This technique attempts to eliminate stationary noise that passes through the fixed beamformer, yielding an enhanced output signal $\bar{X}(k, l)$. Thus, the enhanced beamformer output $\bar{X}(k, l)$ can be written as,

$$\bar{X}(k, l) = \bar{X}_{FBF}(k, l) - \bar{X}_{NC}(k, l) \tag{4.29}$$

where $\bar{X}_{FBF}(k, l)$ represents the output of the FBF, and $\bar{X}_{NC}(k, l)$ the noise signal to be subtracted from the FBF. The FBF output can be described as,

$$\bar{X}_{FBF}(k, l) \quad = \mathbf{G}^{H}(k, l)\mathbf{Z}(k, l) \tag{4.30}$$

with $\mathbf{G}^{H}(k, l)$ , representing the FBF, and $\mathbf{Z}(k, l)$, the Fourier transformed microphone inputs, as described in section 4.4. The noise signal to be subtracted from this, $\bar{X}_{NC}(k, l)$ , is defined as thus,

$$\bar{X}_{NC}(k, l) \quad = \mathbf{H}^{H}(k, l)\tilde{}(k, l) \tag{4.31}$$

where $\mathbf{H}^{H}(k, l)$ , represents the ANC, value and $\tilde{}(k, l)$ the reference noise signals, defined as,

$$\tilde{}(k, l) \quad = \bar{\mathbf{G}}^{H}(k, l)\mathbf{Z}(k, l) \tag{4.32}$$

with $\bar{\mathbf{G}}(k)$ being the BM and $\mathbf{Z}(k, l)$, as already mentioned, being the Fourier transformed microphone inputs. The FBF and BM matrices are constructed using the ATF ratios as follows,

$$\mathbf{G}(k, l) = \frac{\mathbf{A}(k, l)}{\|\mathbf{A}(k, l)\|^{2}} \tag{4.33}$$

$$\bar{\mathbf{G}}(k, l) = \begin{bmatrix} -\frac{A_2^*(k,l)}{A_1^*(k,l)} & -\frac{A_3^*(k,l)}{A_1^*(k,l)} & \cdots & -\frac{A_M^*(k,l)}{A_1^*(k,l)} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & 0 \\ 0 & 0 & \dots & 1 \end{bmatrix} \tag{4.34}$$

Note that the computation of both $\mathbf{G}(k)$ and $\bar{\mathbf{G}}(k)$ requires the knowledge of the ATF ratios. In this work, for simplicity, the true impulse responses $a_m(t)$, are directly transformed as defined in section 4.4, into the frequency domain. This filtering operation produces the output frequency domain filtered output signal $\bar{X}(k, l)$ . To transform this signal back to the time domain, an inverse Fourier transform is carried out, resulting in $\bar{x}(t)$, the production of the final output signal of the two-stage filtered speech system. This final output is then used for performance analysis of the system, as described in chapter 5.

## 4.10 SUMMARY

In recent years, the established relationship between audio and visual aspects of speech production and perception has been conclusively demonstrated, as summarised in chapter 2. This relationship has been exploited by various recent developments in audiovisual speech filtering, where audio-only algorithms are extended to become multimodal with the addition of visual information. This has been carried out in a variety of ways in the literature, for example, Rivet [151] developed a system where the visual information is used purely as part of a source separation system. Another example summarised in chapter 3 was to use visual information as part of a single stage visually derived speech filtering system (Almajai & Milner [12]).

The multimodal speech filtering system designed by the author and presented in this chapter was inspired by both the relationship between audio and visual speech information, and by single modality work in the literature that combines multiple speech techniques as part of one integrated speech enhancement system. The system described in this chapter uses visually derived Wiener filtering in addition to audio-only beamforming as part of a novel integrated, multimodal speech enhancement system, designed to function in adverse environments. This chapter describes each component of this system in detail. The general configuration of the reverberant room environment was discussed in Section 4.3, followed by a description of the multi microphone array used for audio input. The next section, section 4.5, described the filterbank based audio feature extraction process. In section 4.6, visual lip tracking using the state-of-the-art SAAM tracking approach, and the subsequent visual feature extraction process using 2D-DCT was discussed. This was then followed by a description of the visually derived filtering process in section 4.7, and then the GMM-GMR technique used to create the audiovisual speech estimation model. Finally, the second stage of the speech filtering process used in this system, audio-only beamforming, was described in section 4.9. These components were combined to create a loose, scalable, framework that is possible to upgrade with state-of-the-art developments in the future.

However, the system presented in this chapter is in need of thorough evaluation to assess its strengths and weaknesses. This chapter only presented a detailed system description of the individual components, not a comprehensive performance analysis. Chapter 5 carries out a thorough review of the strengths and limitations of this initial system.

# 5

# EXPERIMENTS, RESULTS, AND ANALYSIS

## 5.1 INTRODUCTION

As discussed in chapter 2, the multimodal nature of both human speech production and perception is well established. The relationship between audio and visual aspects of speech has been investigated in literature, with decades of historical work since pioneering work by Sumby & Pollack [168] in 1954, and is further demonstrated by the well-known McGurk Effect (McGurk & MacDonald [126]). Almajai & Milner [10] demonstrated correlation between audio and visual features using MLR , and expanded upon this to devise a visually derived Wiener filter for speech enhancement (Almajai & Milner [11, 12], Almajai *et al.* [13]). The work presented in this thesis has utilised this multimodal speech relationship to design a two-stage multimodal speech filtering system, making use of both audio and visual information. The detailed system design was presented in chapter 4.

The two-stage audiovisual speech enhancement system described in chapter 4 utilised both audio-only beamforming and visually derived Wiener filtering, and this system is evaluated in this chapter. This approach combines the individual elements described in chapter 4 including: (a) lip detection and SAAM lip tracking, (b) visual feature extraction using the DCT technique, (c) audio filterbank extraction, (d) visually derived Wiener filtering, and (e) audio-only beamforming. The performance of this integrated multimodal system is evaluated in this chapter.

This chapter presents a detailed investigation of system performance in several different audiovisual scenarios. Initially, the results of a preliminary investigation to identify the ideal system configuration are discussed. The findings of this preliminary investigation are used to carry out a number of different experiments. The first experiment evaluates the speech filtering performance of the system in a range of very noisy environments. This represents an example of a challenging real world situation that a potential user of a speech enhancement system may encounter, for example, on an aircraft, where there is consistent intrusive background noise that can make communication very challenging. With this type of difficult environment in mind, this chapter presents results of experiments with speech sentences from the GRID corpus mixed with aircraft cockpit noise at a variety of different SNR levels, ranging from -40dB, to +10dB.

One potential limitation with audiovisual speech enhancement systems is the adaptability of the visually derived filtering. The results in section 5.5 make use of a test-set derived from the same set of speakers as were used for training the system, but this raises the question of performance with an unknown speaker. In a real world environment, listeners cannot be expected to interact only with known speakers. Work such as by Almajai & Milner [12] focuses on very limited (e.g. single speaker) datasets, but to do this is to ignore a significant potential limitation of visually derived speech filtering systems. This work investigates the performance impact on this system when an unknown speaker outside of those used for training is tested, and when a different corpus is used for training and for testing. Another limitation of speech enhancement systems is dealing with inconsistent noisy environments. Rather than a consistent noise, as modelled in the initially presented set of experiments, there are many real world environments where the background noise rapidly changes in volume and source. In this chapter, an example of an inconsistent noise source has been created by mixing speech sentences with the sound of clapping. This noise contains silences between claps, and the noise is inconsistent and transient.

Overall, this chapter will evaluate the performance and flexibility of the two-stage multimodal system presented in this work, using both objective listening tests and subjective evaluation by human listeners to identify the strengths and limitations of the system. The discussion will also identify areas in which these limitations can be overcome.

This chapter is organised as follows. Firstly, the speech evaluation approaches used in this work are discussed, with both subjective listening tests and objective measures described. Section 5.3 presents a preliminary investigation into system performance, carried out to ascertain the optimum configuration with regard to training set size and the number of GMM components to use in the audiovisual noise free speech estimation model used in the system. This is followed by a description of the testing of the visual lip detector in section 5.4. In section 5.5, a comprehensive evaluation of the proposed audiovisual speech enhancement system is performed. The performance of the two-stage system is evaluated using both objective and subjective speech testing, and these results are discussed. An evaluation of potential limitations with the system is conducted in section 5.6, an evaluation of testing the system with a novel corpus that it has not been trained on is presented in section 5.7, and this limitation is also explored in section 5.8. Finally, this chapter is summarised in section 5.9.

## 5.2 speech enhancement evaluation approaches

To evaluate the performance of the two-stage multimodal speech enhancement system presented in this work, a variety of measures are used. These come in two forms, objective and subjective. Objective measures are those that are calculated automatically by machine. These have the benefit of being faster to perform than subjective listening tests. There are many different approaches with regard to objective tests such as the PESQ (Rix *et al.* [153]) measure, the IS (Hu & Loizou [86], Hansen & Pellom [76]), or SNR level gain. Subjective testing is also used for the evaluation performed in this chapter, which is carried out with the use of human volunteers, and is widely regarded to be the most accurate way of evaluating the performance of speech enhancement algorithms Hu & Loizou [87]. In these tests, listeners hear filtered speech sentences and score each sentence. Both objective and subjective measures are described in more detail in the remainder of this section.

### 5.2.1 *Subjective Speech Quality Evaluation Measures*

One approach for evaluating speech is to use subjective listening tests. These are tests which make use of human volunteers to evaluate speech. Many different approaches have been used, such as word identification in a sentence (Hussain & Campbell [90]), and scoring the overall quality of speech according to the opinion of the listener. However, the number of different approaches makes comparison of results between different publications more difficult, and a simple measure of overall speech quality does not always provide a comprehensive picture of listener opinion. Many speech enhancement algorithms introduce distortion as well as removing noise, and so this should be taken into account in listening tests. In recent years, a standardised approach has been developed by the International Telecoms Union (ITU-T), and released as ITU-T recommendation P.835 (P.835 [138]).

The aim of the ITU-T approach is to provide clearer guidance to listeners with regard to the evaluation of speech sentences, and this approach has been utilised in reviews of objective test measures (Hu & Loizou [87]) and recent work by Almajai & Milner [12]. This approach requires the listener to listen to each sentence, and then score it from one to five based on three criteria. Firstly, a score for speech distortion level is recorded, with 1 indicating the most, and 5 indicating the least distortion. Secondly, the listener gives a score (again between 1 and 5, with 5 indicating the least noise intrusiveness) for the level of noise intrusiveness, before finally giving a score for the overall speech quality. These three scores were used to produce Mean Opinion Scores (MOS) for each evaluation measure.

In the listening tests reported in the remainder of this chapter, nine volunteers participated, and each volunteer heard sentences from the test-set at six different SNR levels (-40dB to +10dB). Six of the nine volunteers were male and three were female, all with a good level of hearing and all spoke English fluently (six of the volunteers spoke English as a first language, three did not). All volunteers were postgraduate research students, although none were speech processing specialists. These tests took place in a soundproofed room using headphones, and each sentence was played randomly to listeners, who then assigned a score from 1 to 5, with 1 being worst and 5 best, using the three criteria discussed above (speech signal distortion, noise intrusiveness level, and overall speech quality). For purposes of comparison, three versions of each sentence were played. The noisy sentence with no speech processing, the sentence processed with an audio-only spectral subtraction approach (Fritsch & Carlson [65]), and thirdly, the sentence processed with the audiovisual approach proposed in this thesis.

### 5.2.2 *Objective Speech Quality Evaluation Measures*

Hu & Loizou [87] state in their research into various speech measurement algorithms (Hu & Loizou [86, 87]) that the most accurate method for speech evaluation is to make use of subjective listening tests. However, there are a number of issues with the use of subjective testing. Firstly, the availability of listeners can be problematic. Comprehensive listening tests can be time consuming and dull, leading to listener fatigue, and it can be difficult to find an adequate number of suitable volunteers (Loizou [119]). Suitable listening test volunteers should have a good mastery of the language the speech system is tested with, and must also have a good level of hearing. Because of these limitations, it can often be useful to run objective tests in addition to listening tests. Objective tests are carried out automatically by machine rather than by using the subjective opinion of human volunteers. These have the advantage of being much quicker to conduct, and to varying extents, can correlate and confirm the results of subjective tests. Many different measures have been devised to objectively assess the performance of speech enhancement algorithms. Work by Hu & Loizou [87, 86], Loizou [119] has focused on the evaluation of many of these measures, some of which were not originally designed for assessing the performance of speech enhancement. The development and testing of composite measures is also covered in work by Hu & Loizou [86], combining a number of objective measures into a single, theoretically more accurate measure. This section focuses on the relevant objective approaches used in the thesis that contribute to the creation of the composite measures. Other measures such as the IS (Hu & Loizou [86], Hansen & Pellom [76]) are not used in this thesis, and so are not described here.

One objective measure that is very widely used is the PESQ (Rix *et al.* [153]) algorithm. This has been recommended by ITU-T recommendation P.862 P.862 [139] for measurement of narrow band telephony related speech enhancement. It is the successor to ITU-T recommendation P.861, and has time-alignment built into the algorithm. Being a full reference approach, it compares a clean reference signal to the speech signal to be evaluated and returns a score ranging from

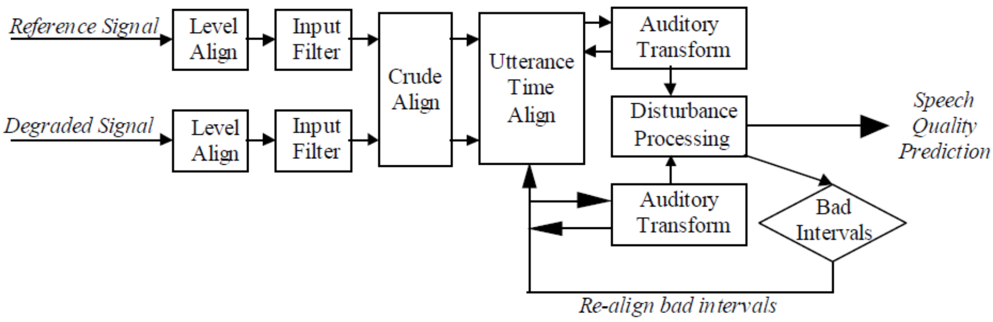


Figure 20: Block diagram of PESQ Algorithm operation. Taken from Malden Electronics Ltd. [125].

-0.5 to 4.5, which means it can be compared to subjective MOS results. A description of PESQ functionality is shown in figure 20, taken from Malden Electronics Ltd. [125].

It can be seen from figure 20 that there are a number of stages involved in this algorithm. These are described in full detail in (Malden Electronics Ltd. [125]). Firstly there is an alignment performed to attempt to ensure that both reference and test signal are at the same audio level by applying a gain to both. The signals are filtered to simulate the effect of transmission through a telephone handset. These are then time aligned in a multi stage process. The overall signal is aligned, and then overlapping frames are aligned. The third stage then attempts to correct any errors in the initial alignment process, and is carried out after a transformation process. After the initial two temporal alignment stages, both signals are transformed in a manner that simulates human hearing. Finally, disturbance processing is carried out to look for errors in the signal being tested. This approach is widely used and is used in this thesis both as a standalone objective measure, and as part of composite objective measures.

Another objective measure that is used in this work is Weighted-Slope Spectral Distance (WSS). This is an established technique, devised by Klatt [101], and adjusted by Hu & Loizou [87], and it is another full reference measure. WSS functions by comparing the spectral slope distance in each spectral band. The spectral slope refers to the distance between adjacent spectral magnitudes, and is measured in decibels. This measure is less complex to calculate than PESQ, and is used in this work as part of the composite measures.

Segmental SNR (SegSNR) is a time domain objective measure. It is established that the correlation between SNR level improvement and subjective speech quality is very poor (Hansen

& Pellom [76]), which means that on its own, SNR level improvement isn't an adequate objective measure for assessing speech enhancement performance. SegSNR is a technique that aims to improve on this by averaging SNR level estimates from frame to frame. There are also thresholds set up due to very small and very large SNR values not being an accurate reflection of signal quality (Hansen & Pellom [76], Hu & Loizou [87]). In this thesis, the same thresholds (-10dB, +35dB) are used as by Hu & Loizou [87], as part of composite measures.

The final individual objective measure that will be discussed here is the LLR. This is a full measure, requiring both clean and test signal, and is a form of Linear Predictive Coding (LPC) based evaluation. LLR is defined in Hansen & Pellom [76] and Hu & Loizou [87] as,

$$d_{LLR}\left(\vec{a}_p, \vec{a}_c\right) = \log\left(\frac{\vec{a}_p \mathbf{R_c} \vec{a}_p^{\,T}}{\vec{a}_c \mathbf{R_c} \vec{a}_c^{\,T}}\right)$$

with $\vec{a}_c$ representing the LPC vector of a clean reference speech frame, and $\vec{a}_p$ being the equivalent frame of filtered speech. $\mathbf{R_c}$ is defined as the autocorrelation matrix of the clean speech signal. Hu & Loizou [87] limit the segmental LLR values to between 0 and 2 to reduce outliers as well as using only the smallest 95% of frame values. Again, in this work, this measure is used only as part of the overall composite measures.

In addition to making use of individual objective testing approaches, it is also possible to combine these measures to create unified composite speech evaluation techniques. Loizou [119], Hu & Loizou [86, 87] investigated the correlation between subjective listening tests and a wide range of objective measures. As might be expected, there were variations into how strong the relationship was between subjective and objective measures, which is to be expected when it is considered that many commonly used measures were not explicitly designed for the evaluation of speech enhancement algorithms. Furthermore, Hu & Loizou [87] point out that research in this area very rarely assesses the performance of objective measures specifically with regard to speech distortion, overall quality, and noise distortion. The majority of the work generally focuses only on overall quality.

Hu & Loizou [87] carried out such an investigation, and found that the level of correlation between subjective and objective scores for different measures varies depending on the subjective measure considered. They state that an individual objective measure is unlikely to correlate highly in all three aspects (speech/noise distortion, overall quality), and one conclusion from their work was that basic objective measures correlated very poorly with noise distortion. With this in mind, three composite measures were then defined by Hu & Loizou [87], that combined the strongest individual objective measures described above. The first measure deals with signal distortion, and combines LLR, PESQ, and WSS. The second measure is for noise distortion, and utilises PESQ, WSS and SegSNR. Finally, the overall quality measure again combines PESQ, LLR, and WSS. These measures are related to the subjective tests outlined in ITU-T recommendation P.835 (P.835 [138]), and described in section 5.2.1. These composite measures are defined by Hu & Loizou [86] as,

$$C_{SIG} = 3.093 - 1.029 \cdot \mathrm{LLR} + 0.603 \cdot \mathrm{PESQ} - 0.009 \cdot WSS$$

$$C_{BAK} = 1.634 + 0.478 \cdot \mathrm{PESQ} - 0.007 \cdot WSS + 0.063 \cdot segSNR$$

$$C_{OVL} = 1.594 + 0.805 \cdot \mathrm{PESQ} - 0.512 \cdot \mathrm{LLR} - 0.007 \cdot WSS$$

In this work, these composite measures are used along with the PESQ measure to evaluate the performance of the multimodal speech enhancement system, and subjective listening tests are also utilised.

## 5.3 preliminary experimentation

### 5.3.1 *Problem Description*

This chapter presents a detailed investigation of the performance of the multimodal system presented in this work, focusing on system performance in difficult environments, and the particular strengths and weaknesses of this multimodal approach. Before undertaking such a detailed investigation, it was necessary to configure a number of individual parameters in order to optimise the system. To find the ideal system configuration, preliminary experiments have been carried out, which subsequently enabled a full analysis to be performed. In this section, the specific results are of less significance, and the remainder of this section provides an overview of the initial preliminary investigation, including the parameters to be optimised, the training and test data used, a summary of results, and the conclusions that were drawn.

### 5.3.2 *Experiment Setup*

The main aspect of this system to be configured was the GMM-GMR audiovisual model component. This component of the system, described in more detail in chapter 4, is used to calculate a smoothed estimate of the noise-free audio signal, based on matching visual information. There were two main parameters that required configuration. Firstly, the number of GMM components, and secondly, the ideal composition of the training set. The number of components to use was relatively straightforward to investigate. As described in chapter 4, a single GMM with no phoneme-specific speech segmentation is used in the audiovisual speech model, but the number of components used within this mixture model is variable. In these preliminary experiments, models consisting of 8, 10, 12, and 16 components were trained. The other variable is the composition of the training dataset. The detailed investigations presented later in this chapter make use of a relatively large test-set, and so it is desirable to have a model that is as flexible as possible. There are a number of potential training sets that can be used. Sentences from four speakers from the GRID corpus (Cooke *et al.* [46]) are used as the test-set in the work presented in this thesis, and so different combinations of training data using these speakers are tested. These range from using 50 sentences from each speaker and creating a large 200 sentence dataset, to simply using 100 sentences from a single speaker. Different combinations were experimented with, with the hypothesis that using more than one speaker as part of the training set produces a more flexible model.

To assess possible configurations of this system, a dedicated test-set was created specifically for this investigation. This test-set was relatively small and contained three sentences from each of the four speakers selected for use in this work, creating a 12 sentence test-set. This dataset makes use of different sentences from those used in the training sets, and these 12 sentences are also different from the larger test-set used in the next section. In addition to these two variables, a further factor to consider is the performance at different SNR levels. This has to be taken into consideration in order to ensure that system performance is consistent and the best overall configuration for all levels was chosen, rather than selecting a configuration that only functions effectively at a single SNR. Therefore, the combination of different possible system configurations was also tested at three different SNR levels, -50dB, -20dB, and 0dB, to simulate a range of noisy and less noisy environments.

### 5.3.3 *Evaluation Approach*

To assess the various combinations of GMM components and training sets, the PESQ approach described in section 5.2 is used. This approach produces a simple output and is quick to calculate. It was felt that at this preliminary stage, detailed listening tests were not required, and that simple objective results would produce an adequate picture of overall performance. However, in order to obtain a more informed opinion, informal listening was also used to establish the overall best performing configurations. The 12 sentence test-set described in the previous section was used, with PESQ means and peak individual scores calculated.

### 5.3.4 *Results and Discussion*

The first parameter assessed was the number of GMM components. As described above, three different SNR levels were used, -50dB, -20dB, and 0dB, to simulate environments with different levels of noise. A number of different training sets were also used for comparison, particularly a dataset containing sentences from all four chosen speakers. Initially, this four person dataset was used to train audiovisual models containing different numbers of GMM components to be compared with the 12 sentence test-set. To evaluate performance, PESQ scores were used, with both mean and individual best scores considered in tandem with informal listening tests. At each SNR level with the four-speaker training set used to train the model, the test-set was evaluated with the system several times, each time with a different number of GMM components used for training. The number of components was 8, 10, 12, and 16.

The results of PESQ and informal listening showed that for each SNR level considered, models trained with the use of 12 components produced the best results. To verify this, a similar assessment was made using a number of other training sets; GMM models with 10 and 12 components consistently produced the best scores, along with the highest audio quality filtered sentences, with a general preference for 12 components. Therefore, for the remainder of the work described in this thesis, the GMM used as part of the GMM-GMR speech estimation process is trained using 12 components.

The second parameter to be decided was the composition of the training set. The training set is used to train the audiovisual model, and so the exact composition of this dataset is crucial to the success of the model. This training set was decided after the number of GMM components had been confirmed, as described above, and so the optimal number of components, 12, was used for all tests. The same three different SNR levels were also used (-50dB, -20dB, 0dB). The composition of the training set has many potential combinations. As described earlier, the test-set for the main investigation makes use of audiovisual speech data gathered from four speakers. For the training dataset, it was expected that some combination of training sentences from more than one speaker would be used. The list of combinations experimented with, ranging from extended single speaker datasets, to combinations of 50 sentences from all four chosen speakers is shown in table 3.

Each training set listed in table 3 was used to train a different audiovisual GMM-GMR model with 12 components, and each of these models was then used as part of the two-stage speech filtering system, and tested at the three different SNR levels (with the small test-set described previously).

To evaluate the models and establish the most suitable training set, mean PESQ scores were compared, along with individual best scores, and informal listening tests. Of the results, there was considerable variation between the datasets, with some trained models producing filtered sentences that contained too much distortion for the PESQ evaluation measure to function adequately. Of the models that did not produce any failed PESQ results in the test-set, it was

Table 3: List of potential audiovisual training datasets evaluated as part of preliminary experimentation. All examples make use of 50 sentences from each speaker unless specified otherwise.

| Number | Combination of Speakers | Total No. of Sentences | Total Speaker No. |
|---|---|---|---|
| 1 | Speaker 1, 3, 4 | 150 | 3 |
| 2 | Speaker 1, 3 | 100 | 2 |
| 3 | Speaker 1, 2, 3 | 150 | 3 |
| 4 | Speaker 1, 2, 3, 4 | 200 | 4 |
| 5 | Speaker 1, 2 | 100 | 2 |
| 6 | Speaker 1(100 sentences), 2 | 150 | 2 |
| 7 | Speaker 3, 4 | 100 | 2 |
| 8 | Speaker 2, 4 | 100 | 2 |
| 9 | Speaker 2, 3, 4 | 150 | 3 |
| 10 | Speaker 1 (100 sentences) | 100 | 1 |

found that the most consistent and effective training sets of those listed in table 3 at all SNR levels were number 4 (all four speakers) and number 7 (combining two speakers). Therefore, it was decided for the remainder of this work to make use of the four speaker training set.

In summary, the informal trials described in this section, using objective measures and informal listening tests, produced the conclusion that the best configuration of the audiovisual model used in this work is for the GMM to make use of 12 components, and for the training set to make use of sentences from all four speakers.

## 5.4 AUTOMATED LIP DETECTION EVALUATION

### 5.4.1 *Problem Description*

To successfully exploit audiovisual information, it is important that the appropriate visual ROI (in this case lip-region information) is correctly identified and tracked. Manual frame by frame identification of the ROI is time consuming and represents an impractical approach. In the work presented in this thesis, a lip tracker, as described in chapter 4 has been utilised in order to extract the lip-region automatically in each frame of an image sequence. One issue with this approach is the identification of the initial ROI in the first frame of the sequence. Initially, this was successfully carried out by manually selecting the corner points of the relevant region in the first image of a sequence. However, although feasible for small scale demonstrations, this solution does not represent a realistic solution, and does not take account of current developments in the field of audiovisual ROI detection (Viola & Jones [175], Li *et al.* [115]).

In this work, a Viola-Jones (Viola & Jones [175]) detector has been implemented to automatically identify the initial lip-region. This has been described in more depth in chapter 4. This approach is widely used in the literature (Wang & Abdel-Dayem [177], Kroon [104], Scherer *et al.* [162]). The detector used in this work is a standard ROI detector, utilising commonly available parameters (Haar cascades) to specifically detect the lip-region. These parameters are available for use as part of the OpenCV library (Bradski [29]). Some customisation was made to the initial code in order to handle potential poor results (as will be discussed later in this section) and identify the final ROI to use. The remainder of this section describes the testing of this lip detector.

### 5.4.2 *Experiment Setup*

To assess performance of the lip detector, images from both of the main corpora used in this thesis (GRID and VidTIMIT) are used. Six speakers from the GRID corpus are used, along with four speakers from the VidTIMIT database. These speakers are split by gender (six male speakers, four female), as well as representing a number of different ethnicities, in order to test a range of speakers. This data was used to create several test-sets. The first test-set used images from both corpora, and consisted of the first image from a single sentence sequence, producing a small test-set of 10 images. The aim of this first test-set was to test the initial implementation of the detector and refine any problems that were identified. The second

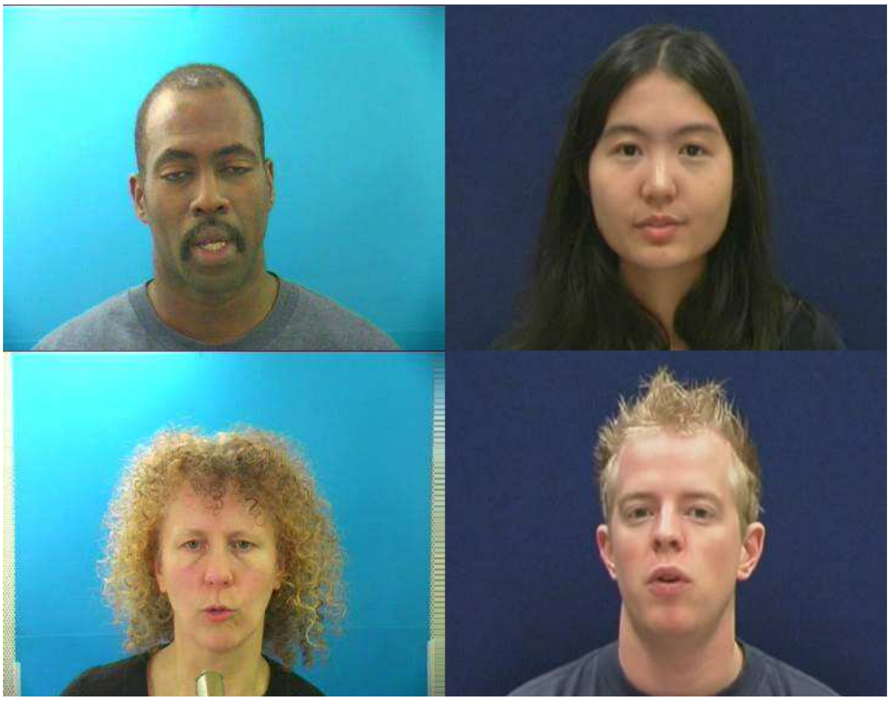


Figure 21: Example image frames from the GRID (top and bottom left) and VidTIMIT (top and bottom right) Corpora.

test-set created was used for the main evaluation of the detection approach. This consisted of sixty images from the ten speakers. Two images from three different sentence sequences from each speaker were used, the first image and then one chosen from the sequence at random (to provide different mouth shapes). A number of example images used in the test-set are shown in figure 21. The final test-set consists of a number of video files from the GRID Corpus. The three sentences were chosen from each of the six speakers, resulting in 18 videos. The aim of this test-set was to inspect whether after identifying the correct ROI, the tracker could correctly use this location to track the correct location in subsequent frames.

### 5.4.3 *Evaluation Approach*

In order to assess the performance of the lip detection approach, a subjective evaluation is used because it is difficult to objectively assess the performance of a lip detection system without visual inspection. Therefore, for this aspect of the speech filtering system, the Viola-Jones detector discussed in section 4.6 of chapter 4 is evaluated by performing an inspection of a number of test-sets. As discussed in chapter 4, the Viola-Jones detector initially produces a number of potential candidates for the object it is tracking. In this work, the primary aim is to identify the lip-region, and so the detector initially produces a number of potential lip-region candidates, before selecting one final ROI to use. To evaluate the effectiveness of this approach (i.e. whether the final lip-region is identified correctly), the initial candidate locations are visually compared to the final chosen ROI.

The testing approach described in this section firstly uses a small test-set to identify initial issues, followed by a larger set to perform a full evaluation. Finally, an inspection of subsequent tracking performance using video files is made.

### 5.4.4 *Results and Discussion*

As described above, the results in this section were produced by visual inspection, as this was felt to be a reliable method to assess the performance of the lip detector. The first dataset was used to identify bugs and test refinements. The lip detector was initially found to successfully identify a range of possible ROI candidates and then select a suitable final ROI (based on the technique outlined in chapter 4) in 9 of the 10 initial images. Examples of successfully detected ROI can be seen in figure 22. It can be seen that a number of possible mouth objects are detected (thinner rectangles), and then the final chosen ROI is correctly identified. One image was found to produce an incorrect result (shown in figure 23). Figure 23 shows that the problem with this image (confirmed by other informal tests on additional images to replicate the result) is the lack of detected candidates, rather than a problem with the final ROI. This is a problem with the standard trained model, and it was felt that to correct this would be outwith the scope of this research, due to the focus of this thesis not being exclusively on lip detection. As discussed in earlier chapters, training a Viola-Jones detector from can be very time consuming and intensive, and so the widely used standard Haar features are felt to be an acceptable compromise.

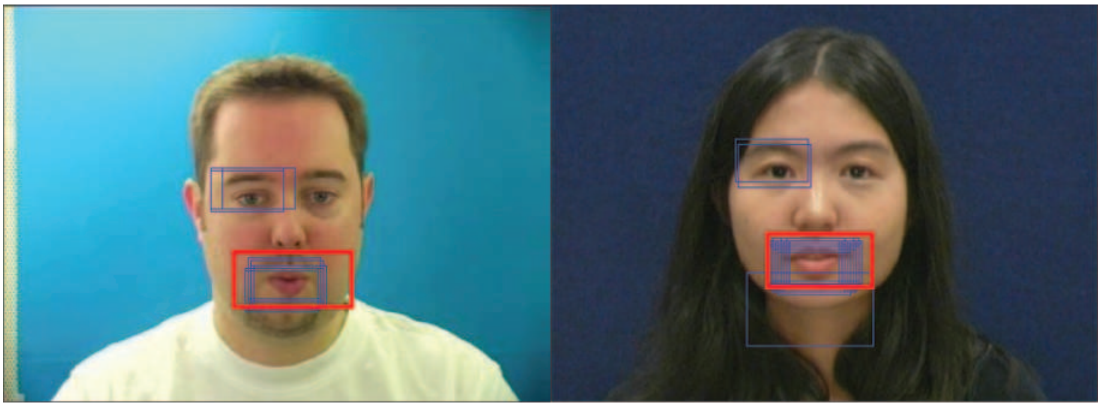


Figure 22: Example of candidate lip-regions, indicated by narrow rectangles. The thicker rectangle shows the final chosen ROI.

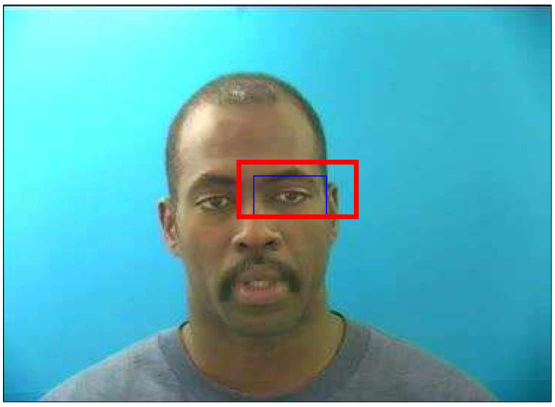


Figure 23: Poor selection performance by the initial lip detection approach. It can be seen that only a single candidate location was identified (resulting in a narrow rectangle around the eye region), resulting in an incorrect final ROI output.

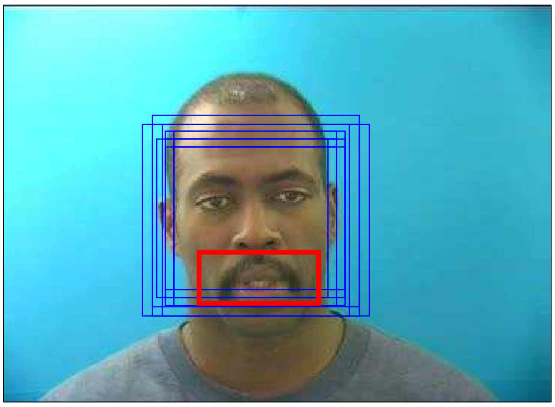


Figure 24: Example of successful face detection.

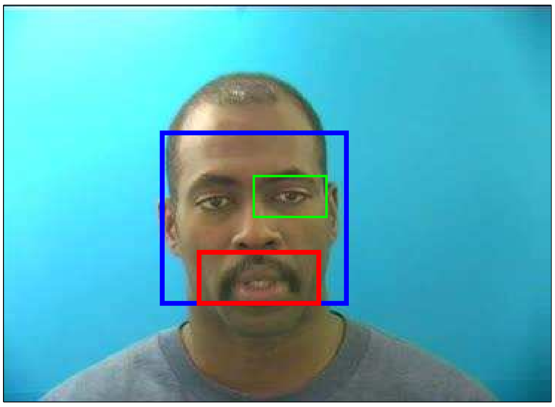


Figure 25: Comparison of mean face detection location, compared to incorrect lip-region identification (narrow rectangle), showing final cropped lip-region output.

This problem was solved by checking the number of resulting potential candidates before outputting a final ROI. The solution was described in depth in section 4.6, but to briefly summarise, the number of output candidate lip-regions was identified. If the number was below a threshold (manually defined in this work as 7), it was decided that this was insufficient to guarantee a correct result. A full face detector is then run on the image (which was found to work without any problems in a variety of tests), and the mean face location is then cropped to only keep the approximate lip area. This was found to be a solution which resulted in error free results with the larger dataset.

After the refinement of the detector, the second set of 60 images was tested. All were visually inspected to check whether the final chosen ROI was appropriate, and also to observe the number of initial candidates. All images were found to identify the lip-region correctly. Where a sufficient number of candidates were identified, the correct ROI was chosen (as can be seen in figure 22), and in cases with a smaller number of candidates, then the face detector was run, and the automatic cropping identified an acceptable ROI, as shown in figure 25. The final tests involved checking if the detected image would allow for the image tracker (as described in chapter 4) to function correctly. The video test-set described previously was tested by combining the lip detector and the tracker. In all of the 18 tested cases, the tracker functioned correctly with no problems and tracked the correct region.

As a result of the successful testing described above, it was felt that this lip detection approach was suitable for use as part of the work discussed in this thesis. The chosen lip detector uses a standard Viola-Jones (Viola & Jones [175]) approach, with some modifications to identify the correct final output region. This is a widely used approach (Li *et al.* [115], Mita *et al.* [131], Scherer *et al.* [162]). However, there are some limitations with this approach. Firstly, as described above, it was found that the lip detector did not work in every situation. A solution was found to this problem, however; this could be further improved on. The other limitation identified with this work is that the detector has only been tested with images using a single speaker in the frame. In order to identify the correct speaker if there is any conflicting visual information, further research would have to be performed, and an appropriate audiovisual VAD would have to be utilised. However, it is felt that this is outside the scope of this thesis. Also, this approach was tested only on data from the GRID and VidTIMIT corpora, meaning that while it functions well with these relatively clean corpora, it has not been fully tested in more difficult situations. Overall, it is felt that this represents an effective solution to the problem of automatically identifying the correct mouth ROI for lip tracking.

## 5.5 noisy audio environments

### 5.5.1 *Problem Description*

One key challenge for speech enhancement algorithms is achieving performance in extremely noisy environments. A real world example of one such environment is on board an aircraft. In this environment, it can become very difficult for conventional hearing aids to function due to the extremely high level of background noise. The use of visually derived filtering in addition to conventional audio-only beamforming adds an extra level of speech enhancement capability and should theoretically allow for successful filtering in very noisy environments where conventional single stage speech filtering may perform badly.

This section focuses on the evaluation of the speech filtering system described in the previous chapter. To do this, we test the system in a very noisy and difficult environment, simulating the real world problem outlined above. Aircraft cockpit noise was added to the simulated room environment as the noise source at a variety of SNR levels, ranging from being relatively quiet (+10dB), to levels in which it is impossible for human listeners to feasibly identify a speech source from noise (SNR levels as low as -40dB). The speech source (sentences from the GRID Corpus (Cooke *et al.* [46])) was mixed with this noise to create a convolved mixture of speech and noise.

The noisy speech mixtures were then processed by the multimodal two-stage speech enhancement system, and the resulting filtered speech sentences were then evaluated by both subjective listening tests and a variety of objective measures.

### 5.5.2 *Experiment Setup*

In order to assess the performance of this system in extremely noisy environments, the multimodal approach described in chapter 4 was tested with speech and noise mixtures that were combined in a simulated room environment. The resulting noisy-speech mixture is received by an array of four microphones. The room environment used is the same as described in chapter 4. These microphone signals are processed with visually derived Wiener filtering and beamforming to produced filtered speech. The parameters required by the audiovisual GMM were defined by the investigation described in section 5.3. The number of components used in the GMM was set at 12, and the training set contained 200 sentences from four speakers from the GRID Corpus.

To provide the speech source data, sentences from the GRID audiovisual corpus were used. For testing, 61 different sentences were used. These sentences were different from those used in the training set, but made use of speakers that the audiovisual model had previously encountered. Each sentence was three seconds in length and when divided up into frames, produced 299 frames per sentence. The noise source was provided by using recorded F16 aircraft cockpit noise. These sources were mixed in the simulated room to produce the noisy speech mixture. Each test sentence was mixed with the aircraft noise at six different SNR levels, ranging from +10dB (a relatively quiet level of noise) to -40dB (a very loud noise source). For evaluation, the commonly used PESQ measure is used, along with listening tests for additional verification. Three versions of each speech sentence were compared at each SNR level. Firstly, the noisy sentence without filtering was used. Secondly, an audio only spectral subtraction approach (Lu & Loizou [121]) was used to produce a filtered signal, and finally, these two sentences were compared to sentences enhanced using the two-stage multimodal system described in chapter 4.

### 5.5.3 *Evaluation Approach*

The PESQ objective measure is used in this investigation to evaluate the filtered sentences. The recently developed composite measure (described in section 5.2) is also used.

It was also felt that a more accurate approach than only using objective speech evaluation measures would be to also carry out subjective listening tests using human volunteers. Nine volunteers participated in these tests, and each volunteer was played sentences from the test-set at six different SNR levels (-40dB to +10dB), as described in section 5.2.1. For purposes of comparison, three versions of each sentence were played to the volunteers. The noisy sentence with no speech processing, the sentence processed with an audio only spectral subtraction approach (Lu & Loizou [121]), and thirdly, the sentence processed with the new audiovisual approach presented in this thesis. This produced three MOS for the three different approaches, one for the overall score, one for speech distortion, and one for background noise intrusiveness. These scores were analysed and compared to the objective scores.

### 5.5.4 *Results and Discussion*

For this experiment, 61 test sentences from the GRID audiovisual corpus were used, as described above. The test sentences use the same speakers as the training set mentioned in section 5.3, but different sentences. Each sentence was mixed in a simulated room environment with aircraft cockpit noise at a variety of different SNR levels to produce convolved noisy speech mixtures. These mixtures were then filtered with two-stage audiovisual speech enhancement to produce enhanced speech signals. Two objective measures, PESQ and composite, were utilised, and listening tests were also used. The performance of this system in very difficult environments is of particular interest, and the focus was on very low SNR levels (-40 to +10 dB).

PESQ was used for initial comparisons, and mean PESQ scores of noisy unfiltered speech sentences were compared to sentences filtered with the audiovisual method. The results of this comparison are shown in table 4.

Table 4 shows that at very low SNR levels, enhanced speech consistently scores higher than noisy speech, with significantly improved results at all levels from -40dB to 0dB ($p<0.05$). This shows that in very noisy environments, this approach is capable of delivering statistically significant improvements, as seen by the p-values in the table, calculated by a repeated measures analysis of variance. In a quieter environment with less noise present (+10dB), the unprocessed noise scores higher, suggesting that distortion introduced by the filtering makes

Table 4: Mean PESQ Scores at Varying SNR Levels.

| SNR Level | Noisy Speech | Audiovisual Filtered Speech | Adjusted P-value |
|---|---|---|---|
| -40dB | 1.232 | 1.697 | 0.000 |
| -30dB | 1.220 | 1.462 | 0.008 |
| -20dB | 1.187 | 1.580 | 0.000 |
| -10dB | 1.412 | 1.958 | 0.000 |
| 0dB | 1.757 | 1.985 | 0.020 |
| +10dB | 2.207 | 1.975 | 0.016 |

the signal less clear. However, there needs to be a degree of caution in interpreting these results. It can be seen in table 4 that there is a clear difference between the noisy and enhanced PESQ scores but it could be argued that this difference is not as large as might be expected. When listening to the unfiltered signal at very low SNR levels, it was often impossible to identify speech, which was not reflected in the PESQ results. Therefore, it was felt that a more comprehensive objective measure had to be used.

Composite measures (which are described in more detail in section 5.3) are used in this work for objective evaluation of test sentences. The results for the Overall score (COvl), Background score (CBak), and Signal score (CSig) are shown in figures 26, 27, 28, with data also displayed in tables 5, 6, and 7. The enhanced audiovisual filtered speech (Avis) results were compared to noisy unfiltered speech (Noi), and speech filtered with audio only spectral subtraction (Spec). Interaction plots for the means of each composite measure (for overall score, noise intrusiveness, and speech signal distortion) are shown in figures 29, 30, and 31 respectively.

Considering speech signal distortion first, it can be seen in figure 27 that at very low SNR levels, both unfiltered speech and spectral subtraction results produced a small negative value at -40dB and -30dB, a very small result at -20dB, and a very low positive result at -10dB. This is because the testing algorithms were unable to identify an adequate level of speech in these results to assign a quality score. However, speech filtered with the audiovisual approach produced much better scores at these SNR levels, returning positive scores at all

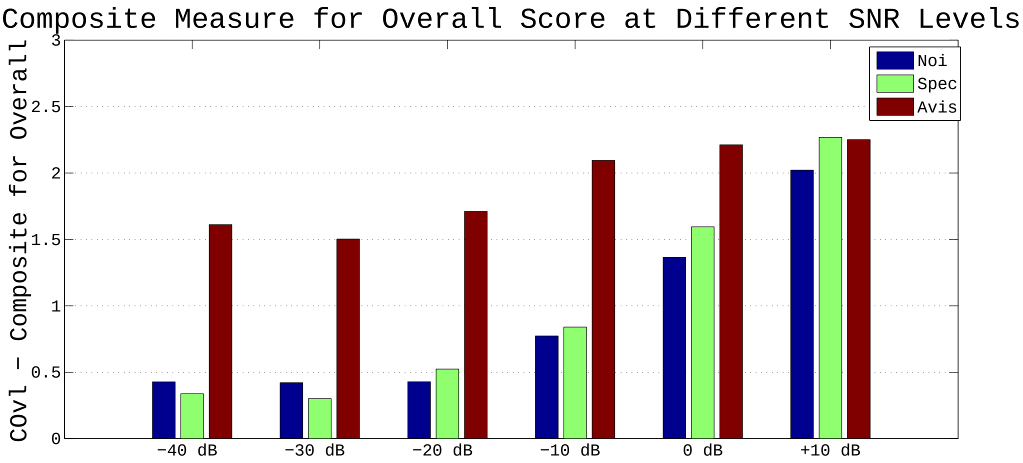


Figure 26: Composite objective mean test scores for overall speech quality, for Unprocessed Noisy Signal (Noi), Spectral Subtraction (Spec), AV Enhancement (Avis).

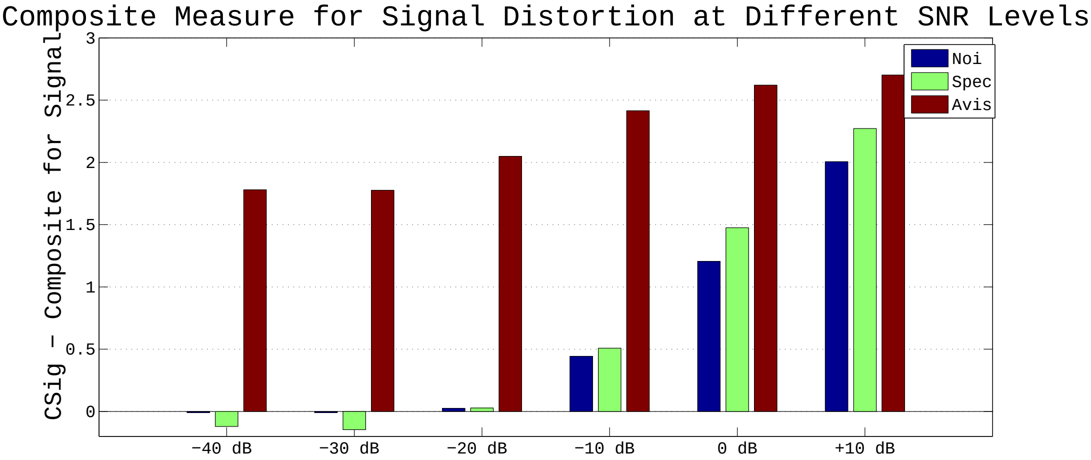


Figure 27: Composite objective mean test scores for noise distortion level for Unprocessed Noisy Signal (Noi), Spectral Subtraction (Spec), AV Enhancement (Avis).

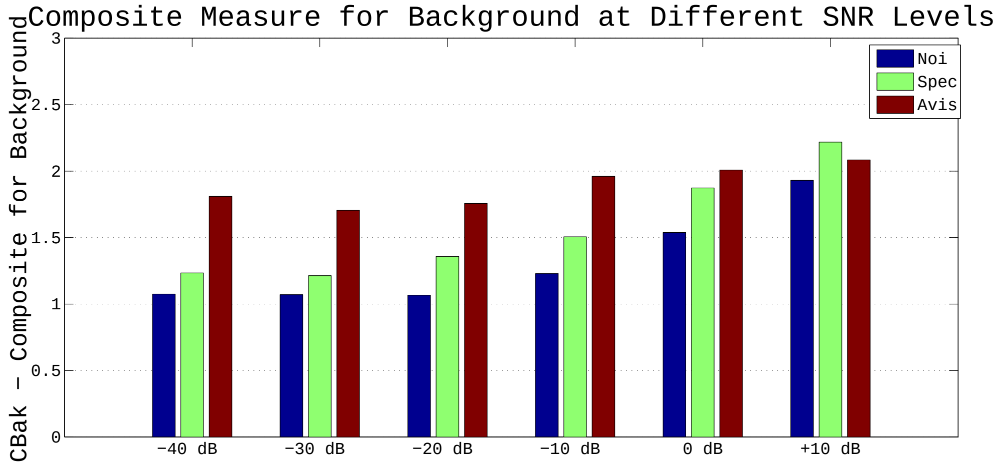


Figure 28: Composite objective mean test scores for speech distortion level for Unprocessed Noisy Signal (Noi), Spectral Subtraction (Spec), AV Enhancement (Avis).

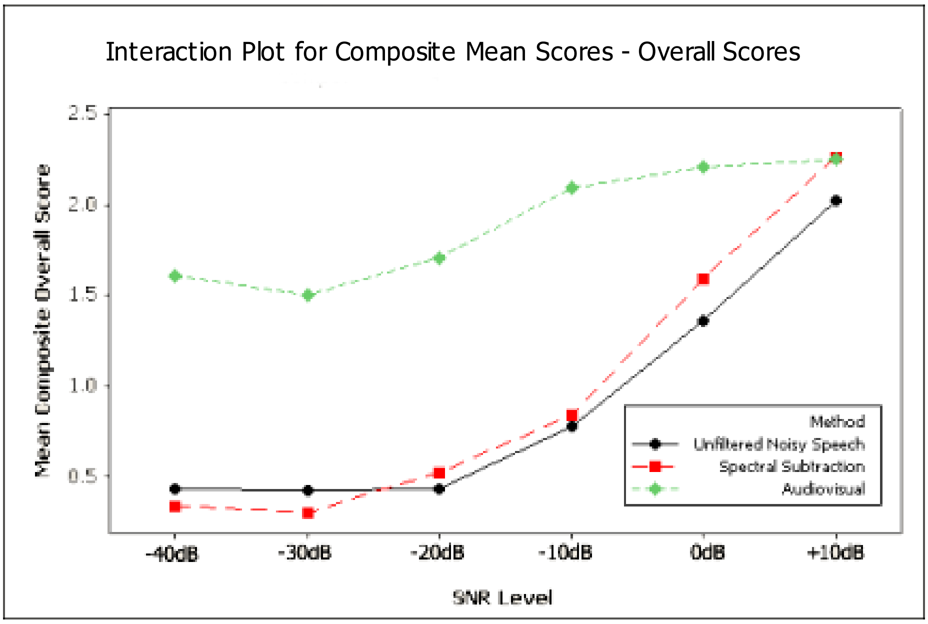


Figure 29: Interaction plot for overall composite objective mean score at varying SNR levels, showing Unfiltered Noisy speech, Spectral Subtraction, and Audiovisual Filtering scores.

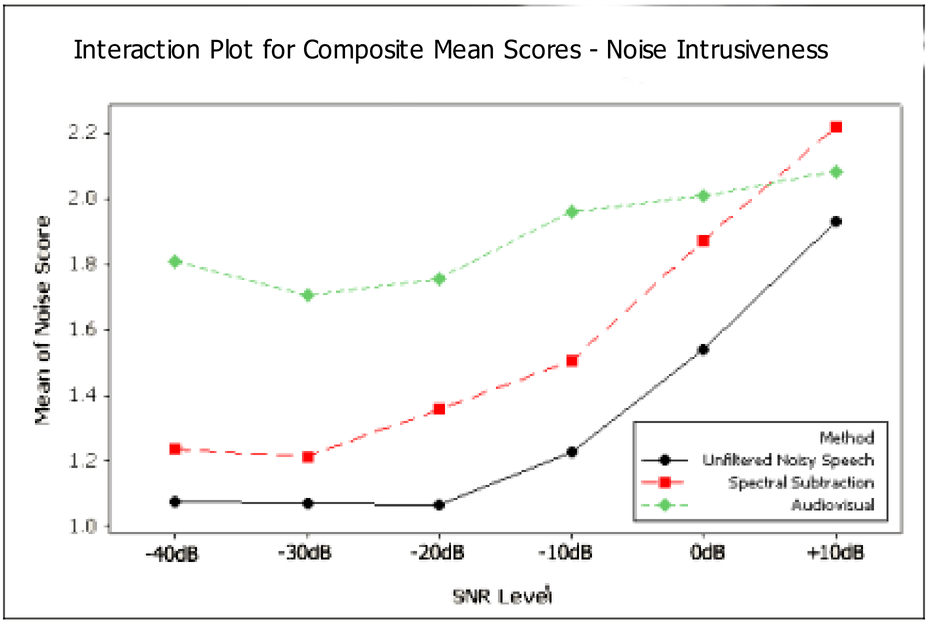


Figure 30: Interaction plot for noise intrusiveness composite objective score at varying SNR levels, showing Unfiltered Noisy Speech, Spectral Subtraction, and Audiovisual Filtering scores.

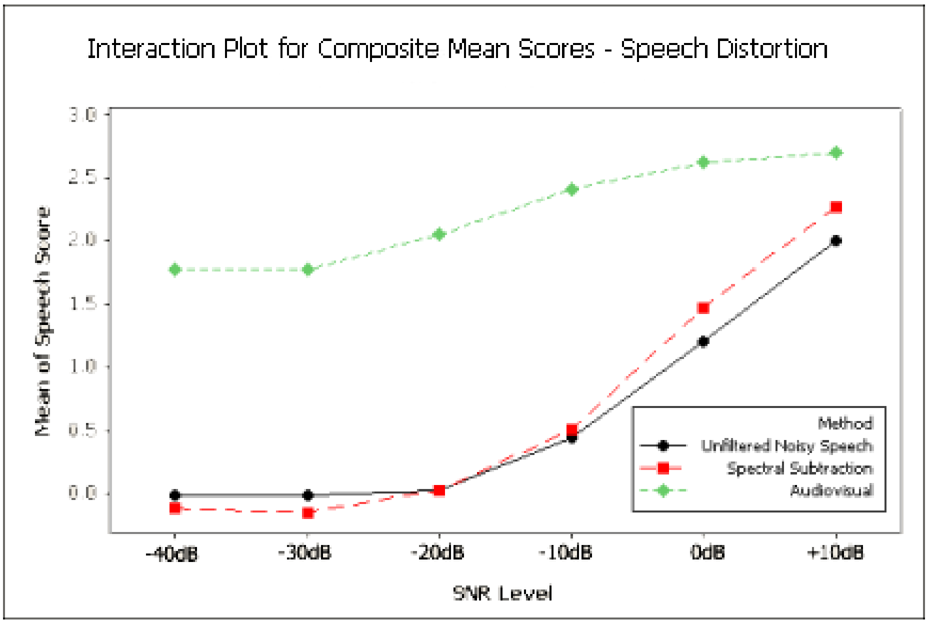


Figure 31: Interaction plot for speech distortion composite objective score at varying SNR levels, showing Unfiltered Noisy speech, Spectral Subtraction, and Audiovisual Filtering scores.

levels, increasing as the SNR level increased. At higher SNR levels (0 and +10 dB), spectral subtraction and unfiltered speech produced much improved results, with only a small, but statistically significant improvement ($p<0.05$) over the unfiltered and audio only options seen when two-stage filtering is used. This can be seen more clearly in the interaction plot in figure 31, and the results of Bonferroni multiple comparison in table 5. As it was found that the results for unfiltered speech and audio only spectral subtraction were very similar, table 5 focuses only on the p-values between unfiltered speech and audiovisual filtering.

The noise intrusiveness scores show slightly different results. The results show that there is significant improvement for noise intrusiveness at low SNR levels when using audiovisual filtering, as shown by the interaction plot in figure 30 and the selected p-values given in the results of Bonferroni multiple comparison in table 7. The difference between the three scores is not as great as might be expected, and this difference tends to be lower than the signal distortion scores. At +10dB, spectral subtraction slightly outperforms the audiovisual method. However, the most important scores to consider are the overall mean scores presented in figure 26, and the associated interaction plot in figure 29. These show that at low SNR levels, the audiovisual approach significantly outperforms conventional spectral subtraction, as confirmed by selected Bonferroni multiple comparison results in table 6. However, when

Table 5: Selected results of Bonferroni Multiple Comparison, showing p-value results for difference between Unfiltered Speech and Audiovisual Filtering for speech distortion composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 1.789 | 0.069 | 26.084 | 0 |
| -30dB | 1.786 | 0.069 | 26.035 | 0 |
| -20dB | 2.202 | 0.069 | 29.495 | 0 |
| -10dB | 1.971 | 0.069 | 28.737 | 0 |
| 0dB | 1.415 | 0.069 | 20.627 | 0 |
| +10dB | 0.695 | 0.069 | 10.135 | 0 |

Table 6: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Unfiltered Speech and Audiovisual Filtering for overall composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 1.184 | 0.066 | 17.974 | 0 |
| -30dB | 1.081 | 0.066 | 16.421 | 0 |
| -20dB | 1.282 | 0.066 | 19.475 | 0 |
| -10dB | 1.321 | 0.066 | 20.064 | 0 |
| 0dB | 0.847 | 0.066 | 12.869 | 0 |
| +10dB | 0.230 | 0.066 | 3.497 | 0.075 |

there is less background noise present, the benefits are less obvious, with very similar overall scores for all three methods at +10dB. This suggests that the audiovisual filtering approach is most effective in extremely noisy environments, with relatively little improvement found in environments containing a lower level of noise.

To confirm the objective composite measure results above, subjective listening tests were used. As described in section 5.2.1, nine volunteers participated in listening tests, and each volunteer was played sentences from the test-set at six different SNR levels (-40dB to +10dB). These tests took place in a soundproof room using noise cancelling headphones, and each

Table 7: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Unfiltered Speech and Audiovisual Filtering for noise intrusiveness composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 0.734 | 0.037 | 19.660 | 0 |
| -30dB | 0.634 | 0.037 | 16.981 | 0 |
| -20dB | 0.690 | 0.037 | 18.457 | 0 |
| -10dB | 0.732 | 0.037 | 19.593 | 0 |
| 0dB | 0.470 | 0.037 | 12.586 | 0 |
| +10dB | 0.153 | 0.037 | 4.107 | 0.007 |

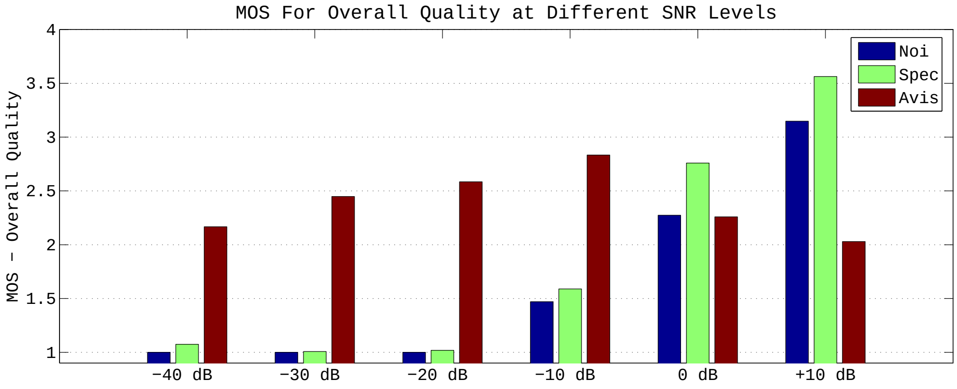


Figure 32: Mean Opinion Score for overall speech quality for unprocessed noisy signal (Noi), spectral subtraction (Spec), audiovisual enhancement (Avis).

sentence was played randomly to listeners, who then assigned a score from 1 to 5, with 1 being worst and 5 best, for three criteria, speech signal distortion, noise intrusiveness level, and overall speech quality. As with the composite measure above, for purposes of comparison, three versions of each sentence were played. The noisy sentence with no speech processing, the sentence processed with an audio only spectral subtraction approach, and thirdly, the sentence processed with the audiovisual approach. The MOS for the three different approaches are shown in figures 32, 33, and 34. In addition, interaction plots of each measure (overall score, speech quality, noise intrusiveness) are shown in figures 35, 36, and 37 respectively.

Firstly, looking at the overall results, it is clear that the listeners were consistently unable to hear unprocessed speech at very low SNR levels, and spectral subtraction was also of little

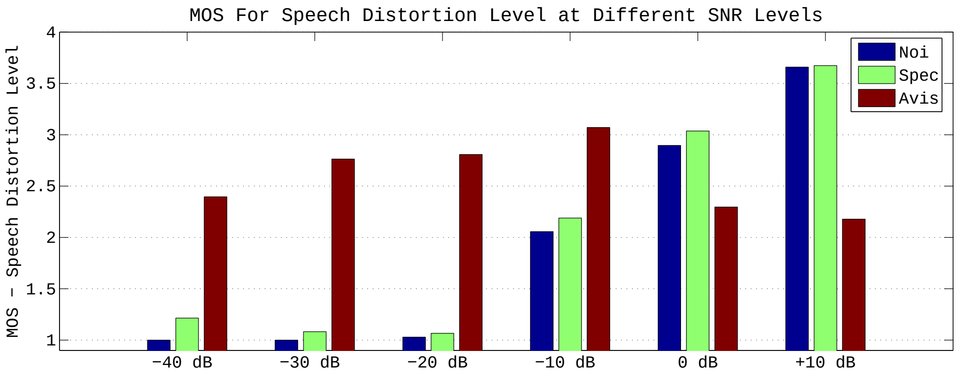


Figure 33: Mean Opinion Score for speech distortion level for unprocessed noisy signal (Noi), spectral subtraction (Spec), audiovisual enhancement (Avis).

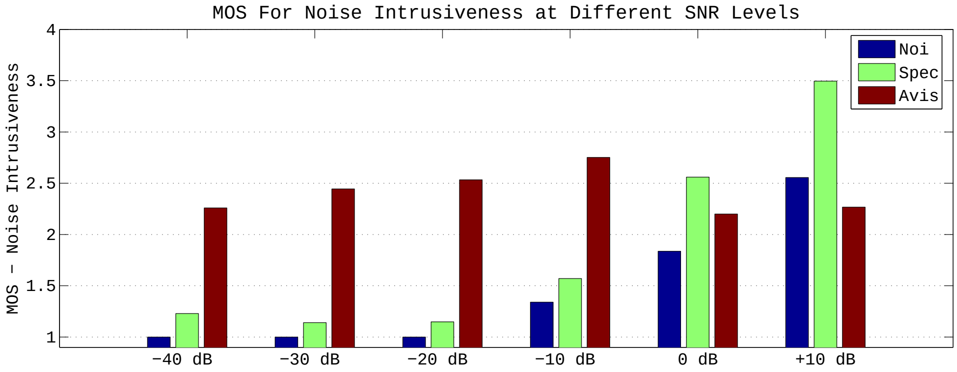


Figure 34: Mean Opinion Score for noise intrusiveness level for unprocessed noisy signal (Noi), spectral subtraction (Spec), audiovisual enhancement (Avis).

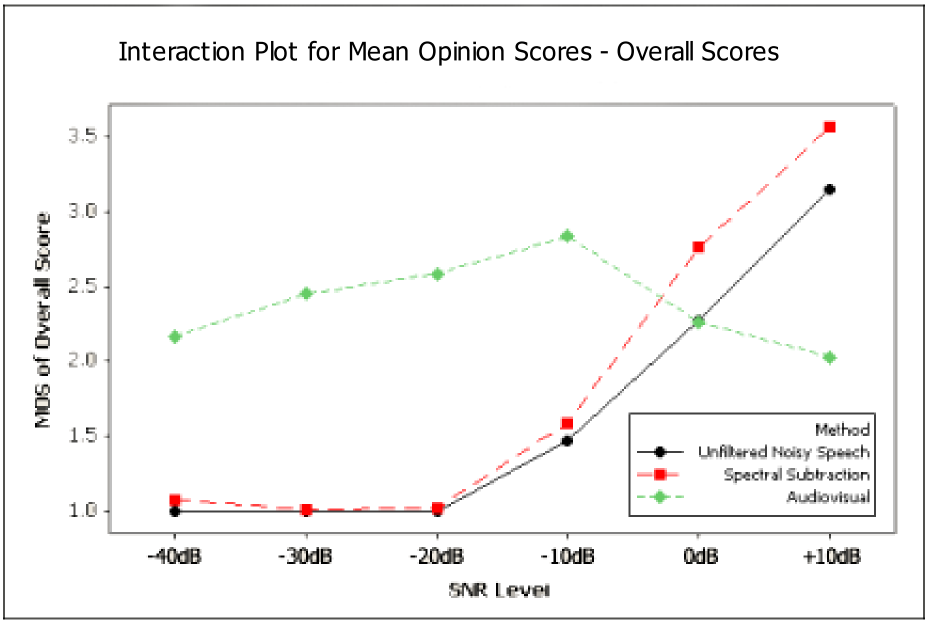


Figure 35: Interaction plot for overall MOS score at varying SNR levels, showing Unfiltered Noisy speech, Spectral Subtraction, and Audiovisual Filtering scores.

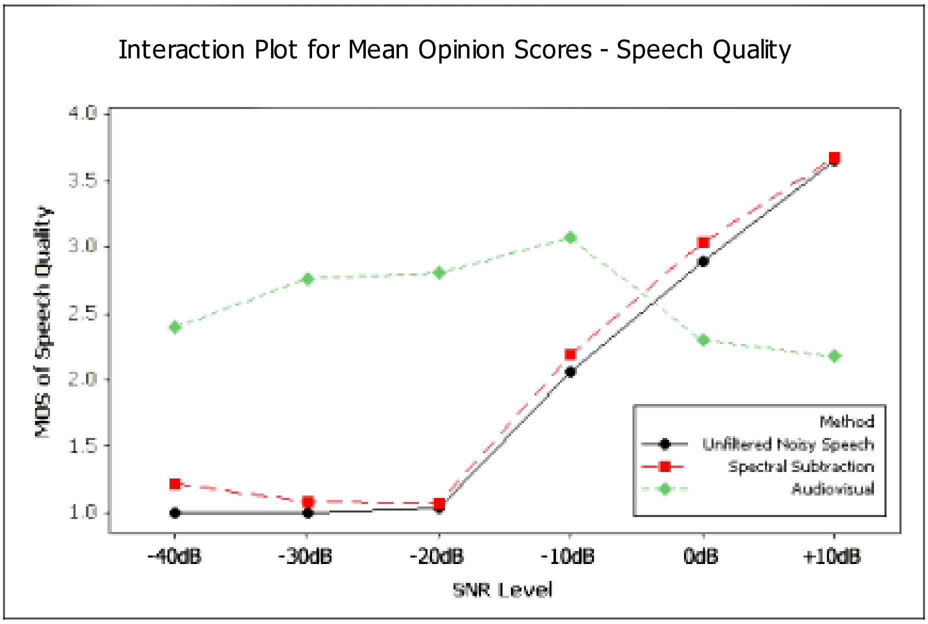


Figure 36: Interaction plot for speech quality MOS score at varying SNR levels, showing Unfiltered Noisy Speech, Spectral Subtraction, and Audiovisual Filtering scores.

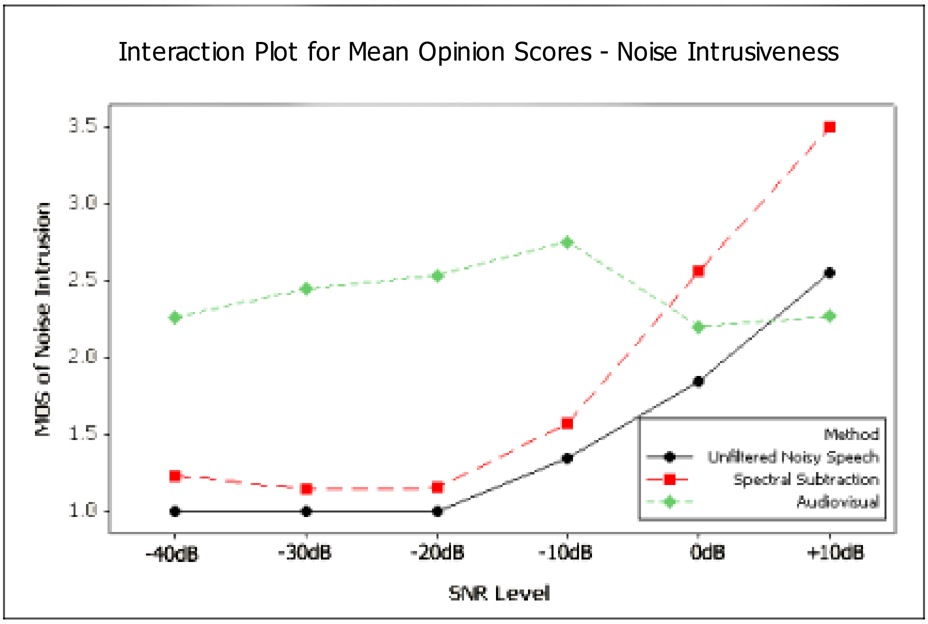


Figure 37: Interaction plot for noise intrusiveness MOS Score at varying SNR levels, showing Unfiltered Noisy speech, Spectral Subtraction, and Audiovisual Filtering scores.

Table 8: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Unfiltered Speech and Audiovisual Filtering for overall MOS scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 1.167 | 0.736 | 15.837 | 0 |
| -30dB | 1.448 | 0.736 | 19.658 | 0 |
| -20dB | 1.585 | 0.736 | 21.519 | 0 |
| -10dB | 1.363 | 0.736 | 18.502 | 0 |
| 0dB | -0.015 | 0.736 | -0.201 | 1 |
| +10dB | -1.119 | 0.736 | -15.180 | 0 |

use. However, in these noisy environments, the audiovisual approach produced higher scores, with the listeners able to identify speech. This pattern is mirrored for speech distortion levels, and also for noise, with the audiovisual approach demonstrating a large improvement at low SNR levels, showing that in very noisy environments, this two-stage approach can produce significantly improved results when it comes to speech quality and the overall score. This closely and accurately matches the results found with the composite measures. With regard to the significance of these results, it can be seen from the interaction plots that the mean scores for unfiltered speech and spectral subtraction are very similar, especially at very low SNR levels and so the focus is on the difference between unfiltered speech and audiovisual filtering. The relevant results of Bonferroni multiple comparison are summarised in tables 8, 9, 10, showing the difference between audiovisual filtering and unfiltered speech for each of the three MOS results, with p-values of $p<0.05$ showing that the difference at low SNR levels is statistically significant for all three measures.

However, at higher SNR levels (0dB and +10dB), it can be seen in figure 32 that the overall score of the two-stage system is lower than at lower SNR levels, and this is especially noticeable in the signal quality level. Analysing the results in detail, it can be seen that spectral subtraction outperforms the multimodal approach at 0dB, with listeners assigning a higher overall MOS to spectral subtraction, and a very similar score to noisy and two-stage filtered speech. At a

Table 9: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Unfiltered Speech and Audiovisual Filtering for speech quality MOS Scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 1.396 | 0.732 | 19.084 | 0 |
| -30dB | 1.763 | 0.732 | 24.095 | 0 |
| -20dB | 1.778 | 0.732 | 24.298 | 0 |
| -10dB | 1.015 | 0.732 | 13.870 | 0 |
| 0dB | -0.600 | 0.732 | -8.201 | 0 |
| +10dB | -1.481 | 0.732 | -20.250 | 0 |

Table 10: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Unfiltered Speech and Audiovisual Filtering for noise intrusiveness MOS Scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 1.259 | 0.083 | 15.152 | 0 |
| -30dB | 1.444 | 0.083 | 17.380 | 0 |
| -20dB | 1.533 | 0.083 | 18.450 | 0 |
| -10dB | 1.411 | 0.083 | 16.980 | 0 |
| 0dB | 0.363 | 0.083 | 4.367 | 0.020 |
| +10dB | -0.289 | 0.083 | -3.476 | 0.079 |

SNR of +10dB, the audiovisual method performed very poorly, with a very low overall MOS in comparison to both noisy and spectral subtraction sentence scores. Looking at these results in detail, it can be seen that the main drop in audiovisual filtered MOS is displayed in the speech distortion score. The noise intrusiveness score for audiovisual filtering increases between -40 and -10dB, but there is a small drop at 0dB, followed by another drop at +10dB. The most relevant result is the speech distortion score. It can be seen that at 0dB, the audiovisual output speech quality is significantly reduced. The p-values in table 8 show that there is no significant difference between the overall scores at 0dB, but individual speech and noise scores are significantly different. At +10dB, the speech quality score is extremely low in comparison to all other approaches, and tables 8, 9, 10 show that the audiovisual approach performs significantly worse. This suggests that the main reason for the low overall score is because of the level of speech distortion introduced, and listening to filtered sentences confirmed this. This represents a more accurate picture than presented by the objective results alone and shows that although objective measures are valuable, they need to be supplemented with listening tests.

One hypothesis is that this distortion is due to problems with the visual filterbank estimation approach. The approach evaluated in this paper, GMR, was originally designed to calculate efficient robot arm movement, and while it is shown in this work that it can be applied to audiovisual data, the resulting signal is not always accurately calculated, and so this technique can introduce distortion into the speech. Filtering at lower SNR levels produces good results, but when there is less background noise to remove, more distortion is introduced into the speech. A key solution to improving results in future work is to consider an improved audiovisual speech estimation model.

State-of-the-art work by Almajai & Milner [12] makes use of a MAP GMM approach and also further enhancements such as a VAD and using a number of different GMMs to represent individual speech phonemes. It was also shown in previous work by the author in Abel *et al.* [2] that a degree of asynchrony between audio and visual frames may also improve results, so there are many ways in which this initial audiovisual two-stage system can be improved in the future. While this two-stage approach produces improved results in very noisy environments, the use of visual information is not always helpful. The potential availability of audio and visual information in a more realistic speech environment also has to be considered. This work makes use of a simulated environment, with pre-recorded visual information and static speech and noise sources, but in a more realistic environment, such as one which a hearing aid might be expected to function in; there may be multiple inconsistent moving noise sources. Furthermore, visual information may not always be usable. Changes in light, pose, and actions by the speaker such as placing a hand over their mouth may render some frames of visual information unusable. Therefore, in addition to improving the visually derived filtering approach as described above, it is also important to consider how a multimodal system can best take advantage of audio and visual information to deliver good results on a frame by frame basis.

Overall, the results show that visual information can be used as part of a speech enhancement system. In very noisy environments, it can be seen that the two-stage speech enhancement system presented in this thesis is capable of successfully filtering speech, as proven by PESQ, composite objective scores and subjective listening tests. While the poor scores at high SNR levels indicate that that the individual components of the system can still be refined further, for example, with a more sophisticated audiovisual speech model, the initial system has been shown to successfully filter speech in challenging environments.

## 5.6 untrained visual environment

### 5.6.1 *Problem Description*

One issue with using a visually derived filtering model is that the system may only be able to process data similar to that which the system was trained with. When presented with entirely new data such as a brand new speaker, the system may struggle to generalise. In the specific case of this system, the limitation is that while the system has been trained using data from four speakers, when presented with visual data from an unknown speaker, it

Figure 38: Screenshots of speakers used for training and test datasets.

can be hypothesised that speech enhancement results will suffer due to a lack of training. This represents a limitation with the current system in that a real world environment is extremely unlikely to only contain known speakers. A potential user of a multimodal speech enhancement system would be expected to interact with a variety of different speakers in various environments. Unknown speakers may have different mannerisms and communication styles to those speakers that the system is trained on.

To evaluate this hypothetical limitation, the results of the multimodal speech filtering experiments described in section 5.5 were compared to audiovisual filtered results of sentences spoken by a novel speaker that the audiovisual model had not encountered in training. This simulated encountering a previously unknown speaker.

### 5.6.2 *Experiment Setup*

The speech enhancement system parameters used in this experiment are the same as used in section 5.5. The preliminary investigation discussed previously identified that a GMM model containing twelve components and trained using sentences from four different speakers produced good results, and these parameters are used in this experiment.

Figure 39: Screenshot of speaker used for testing with previously unseen data.

The training and test-set primarily utilised in this work makes use of data from the GRID audiovisual corpus. The preliminary investigation summarised in section 5.3 identified that a model trained with data from four different speakers delivered positive results, and so the primary test-set made use of the same four speakers. An image of the chosen speakers is displayed in figure 38. To compare the test-set results from this dataset to unknown speakers, 11 sentences from an alternative speaker in the GRID corpus (as shown in figure 39), which was not used in the audiovisual mixture model training process, were selected and used for testing. These were then compared to the results of the test-set from known speakers.

To match the test-set, the same noise source is used. Aircraft cockpit noise is added to a simulated room environment as the noise source at a number of different SNR levels (-40dB, -30dB, -20dB, -10dB, 0dB, +10dB). The same simulated room and audiovisual system configuration was used as has been described in the other experiments in this chapter.

### 5.6.3 *Evaluation Approach*

To evaluate this work, a small test-set was used for comparison. It was considered suitable to run the objective PESQ speech evaluation measure. 11 sentences from a different speaker were compared to the equivalent PESQ scores from speakers the system had encountered in offline

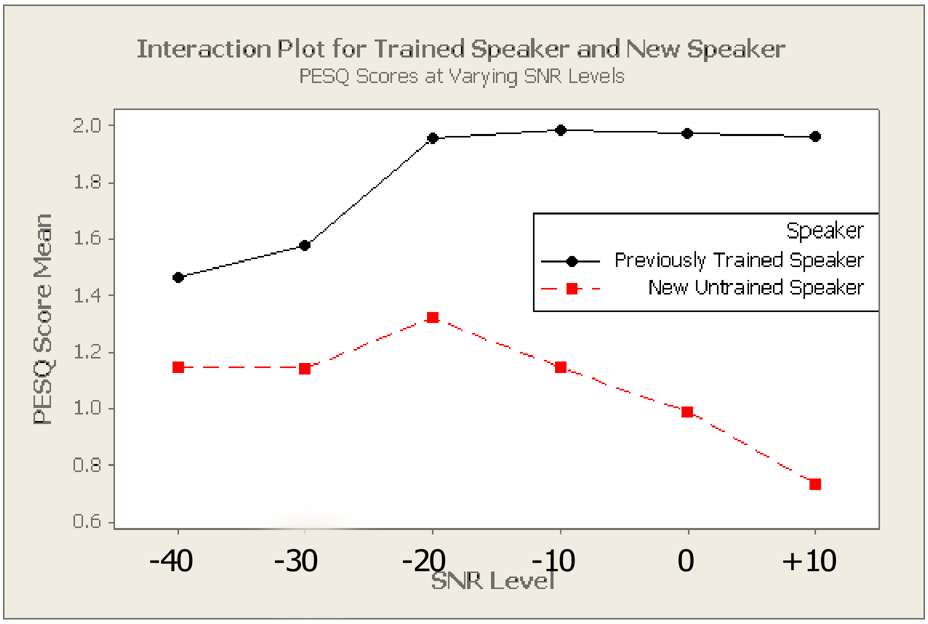


Figure 40: Two way analysis of variance comparison for previously trained versus unseen speaker, comparing PESQ scores at different SNR levels.

training (the results described in section 5.5). The two sets of PESQ scores were evaluated at each chosen SNR and a two way analysis of variance was run on the means of this data, using Bonferroni multiple comparison. This means that the p-values were compared and the means plotted.

### 5.6.4 *Results and Discussion*

After performing speech enhancement on the noisy audio files from the previously unseen speaker, PESQ scores were calculated for each sentence at each SNR level. The individual scores were then combined to produce a mean value at each SNR level (from -40dB, to +10dB). These were then compared to mean values produced from test sentences from speakers that had also been used for training. As explained previously, although sentences from the same speakers had been used for training, the specific sentences used for the test data were new to the audiovisual model. The means of both the new speaker data and the previously trained speaker data at each SNR level are plotted in figure 40.

It can be seen from the interaction plot in figure 40 that the means of unseen versus trained speaker are different at each SNR level plotted. The novel speaker can be consistently seen

Table 11: Results of two way analysis of variance for comparison of trained and untrained visual data, showing adjusted P-Value.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -0.313 | 0.113 | -2.765 | 0.392 |
| -30dB | -0.435 | 0.113 | -3.845 | 0.009 |
| -20dB | -0.636 | 0.113 | -5.61 | 0 |
| -10dB | -0.839 | 0.113 | -7.41 | 0 |
| 0dB | -0.987 | 0.113 | -8.72 | 0 |
| +10dB | -1.226 | 0.113 | -10.83 | 0 |

to produce lower PESQ mean scores at all SNR levels, from -40dB to +10dB. As the SNR level increases, the difference between means also increases. To assess the significance of the result, the results of the two way analysis of variance were examined to assess significance, with particular focus on the adjusted p-value. The results are shown in table 11, and show that the difference is not significant at -40dB, due to the low PESQ mean score for data, but at all other SNR levels, the difference between the previously seen and the unseen speaker data is statistically significant, with the novel speaker producing significantly worse performance.

This experiment identifies a limitation of the system. Positive performance results reported in previous sections of this thesis, and also in similar research (Almajai & Milner [12, 11]) cannot currently be applied universally. There is a lack of generalisation, and when the system is applied to new data from a speaker that it was not directly trained with, despite the data being from the same corpus and consisting of similar sentences, results are poor. This weakness was expected, due to the limitations in the mixture model utilised (as discussed previously), and is in line with expectations, due to previously reported work of a similar nature (such as that by Almajai & Milner [12, 11]) being trained and tested with only a single speaker corpus. Although the visual tracking approach used in this work removes many differences in the raw visual data such as background, eye data, clothing etc. there are still differences in the lips and in the manner of speech (emphasis, tone, speed, accent, etc.). To explore this limitation further, the next section tests this approach with data from a completely different corpus, the VidTIMIT audiovisual speech corpus Sanderson [157].

## 5.7 testing with novel corpus

### 5.7.1 *Problem Description*

As mentioned in section 5.5, one key challenge for speech enhancement algorithms is achieving performance in extremely noisy environments, such as on board an aircraft. In this environment, it can become very difficult for conventional hearing aids to function due to the extremely high level of background noise. However, there is another limitation to consider, the performance of the system with speakers outside the range that it has been trained with. This is a known limitation with the approach, with Wiener Filtering work by Almajai [9] (also Milner & Almajai [130], Almajai & Milner [11, 12]) explicitly being trained with a single speaker corpus, and then tested with the same corpus. Although four speakers were used for training this system rather than one, it can be seen from the results presented in section 5.6 that the system performs poorly with an unknown speaker from the GRID Corpus. This suggests a limitation with this approach that should be explored in more depth.

This section also evaluates the speech filtering system described in the previous chapter. In order to be able to successfully compare the results in this section to the results reported in section 5.5, many of the parameters are similar to the noisy environment evaluation discussed previously. Therefore, the same noise source is used (aircraft cockpit noise), and this is mixed with speech at a variety of SNR levels, ranging from being relatively quiet (+10dB), to levels in which it is impossible for human listeners to feasibly identify a speech source from noise (SNR levels as low as -40dB). The key difference with the results presented in this section is that a completely different corpus is used as the speech source. Sentences from the VidTIMIT audiovisual corpus (Sanderson [157]) are used to create the convolved noisy-speech mixtures. This allows for a direct comparison between the results of using a new corpus (not previously trained with those speakers, different recording conditions, different accent and sentence composition), and the results presented in section 5.5. Although the results presented in section 5.5 use sentences that the system has not been trained with, they are spoken by the same speakers as the system was trained with.

The noisy VidTimit based speech mixtures were then processed by the multimodal two-stage speech enhancement system, and the resulting filtered speech sentences were then evaluated by objective measures for comparison to the results presented previously.

### 5.7.2 *Experiment Setup*

To assess the performance of this system in extremely noisy environments, the multimodal approach described in chapter 4 was tested with speech and noise mixtures that were combined in a simulated room environment, in the same manner as described in section 5.5. Therefore, the room is the same as described in chapter 4, and the parameters required by the audiovisual GMM were defined by the investigation described in section 5.3. The number of components used in the GMM was set at 12, and the training set contained 200 sentences from four speakers in the GRID Corpus.

To provide the speech source data, sentences from the VidTIMIT audiovisual corpus were used. For testing, 66 different sentences were used. As these sentences were from the VidTIMIT corpus, this meant that the system was completely untrained with speakers from this corpus. This corpus uses different speakers, speaking Australian-English, different sentences (TIMIT sentences rather than the simpler commends in GRID, with each sentence around four seconds in length. The noise source was provided by using recorded F16 aircraft cockpit noise. These sources were mixed in the simulated room to produce the noisy-speech mixture. Each test sentence was mixed with aircraft noise at six different SNR levels, ranging from +10dB (a relatively quiet level of noise) to -40dB (a very loud noise source). To evaluate the resulting sentences, the three composite objective measures used in section 5.5 were used.

### 5.7.3 *Evaluation Approach*

The recently composite measure described in section 5.2 is used in this investigation to evaluate the filtered sentences. For purposes of comparison, three versions of each sentence were compared. The noisy sentence with no speech processing, the sentence processed with an audio only spectral subtraction approach (Lu & Loizou [121]), and thirdly, the sentence processed with the audiovisual approach presented in this thesis. The results were also compared to the results in section 5.5.

### 5.7.4 *Results and Discussion*

For this experiment, 66 test sentences from the VidTIMIT audiovisual corpus were used, as described above. As the system is trained with data from the GRID corpus, this is a completely novel corpus. In order to be able to compare the results in this section to the findings presented in section 5.5, the same mixing procedure was used. Each sentence was mixed in a simulated room environment with aircraft cockpit noise at a variety of different SNR levels to produce convolved noisy-speech mixtures. These mixtures were then filtered with two-stage audiovisual speech enhancement to produce enhanced speech signals. The composite objective measures were used to produce the results in this section, with the performance of this system in very difficult environments (very low SNR levels) being of particular interest.

Composite measures (which are described in more detail in section 5.3) are used in this work for objective evaluation of test sentences. The results for the Signal score (CSig), Background score (CBak), and Overall score (COvl), are shown in figures 42, 44, and 46, with data also displayed in tables 12, 13, and 14. As with previous experiments, the enhanced audiovisual filtered speech (Avis) results were compared to noisy unfiltered speech (Noi), and speech filtered with audio only spectral subtraction (Spec). Interaction plots for the means of each

Table 12: Selected results of Bonferroni multiple comparison of VidTIMIT corpus, showing P-Value results for difference between Unfiltered Speech and Audiovisual Filtering for speech distortion composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 0.534 | 0.045 | 11.883 | 0.000 |
| -30dB | 0.494 | 0.045 | 10.997 | 0.000 |
| -20dB | 0.584 | 0.045 | 12.990 | 0.000 |
| -10dB | 1.336 | 0.045 | 29.730 | 0.000 |
| 0dB | 2.190 | 0.045 | 48.730 | 0.000 |
| +10dB | 3.239 | 0.045 | 72.07 | 0.000 |

composite measure (for speech signal distortion, noise intrusiveness, and overall score) are shown in figures 41, 43, and 45 respectively.

Considering speech signal distortion first, it can be seen in figure 42 that at all SNR levels the best performing sound file was the unfiltered noisy signal. The difference between the unfiltered signal and the enhanced signal was statistically significant ($p<0.05$) as shown by table 12, and was also confirmed by listening to the files. The spectral subtraction approach was found to produce files that introduced some noticeable distortion, as confirmed by the difference between the scores, particularly at higher SNR levels. As expected, the audiovisual method performed particularly poorly with regard to speech distortion. This was expected, due to the limitations of the Wiener filtering approach, and confirmed the findings in section 5.6, that the system performed poorly when tested with completely novel data. As the SNR increased, the speech distortion score decreased. This was confirmed by listening to the files.

The noise intrusiveness scores show slightly different results. The results show that there is significant improvement for noise intrusiveness at low SNR levels when using audiovisual filtering, as shown by figure 44 and the selected p-values given in the results of Bonferroni multiple comparison in table 13 (and also the associated interaction plot in figure 43). At SNR levels of -40dB, -30dB, and -20dB, the audiovisual approach is found to produce a significant improvement over the unfiltered noisy signal, showing that it is capable of performing some

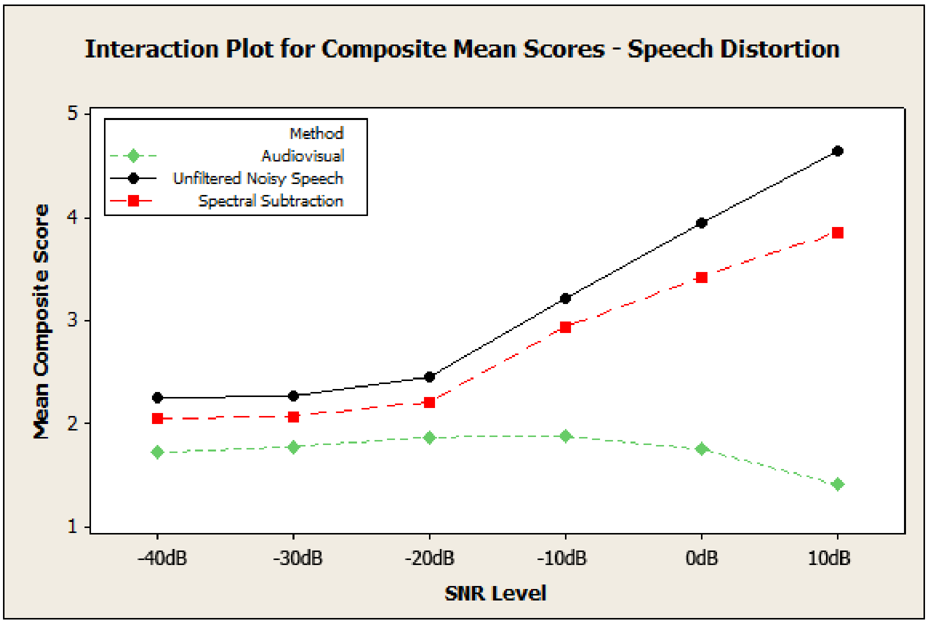


Figure 41: Interaction plot for speech distortion composite objective score of VidTIMIT corpus at varying SNR levels, showing Unfiltered noisy-speech, Spectral Subtraction, and Audiovisual Filtering scores.

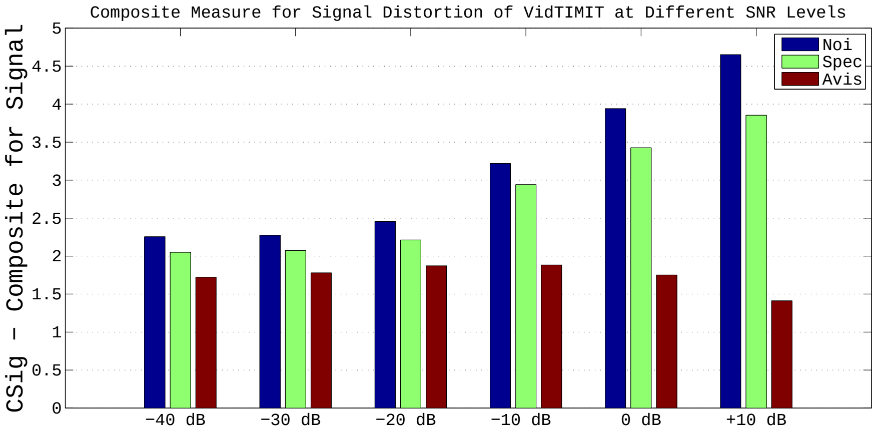


Figure 42: Composite objective mean test scores of VidTIMIT Corpus for noise distortion level for Unprocessed Noisy Signal (Noi), Spectral Subtraction (Spec), AV Enhancement (Avis).

noise removal. However, at higher SNR levels, the difference is not found to be statistically significant at -10dB, and then the audiovisual approach is found to produce significantly worse results at 0dB and +10dB. The findings of the speech and noise composite measures are backed up by the overall score. This is shown in table 14, and the associated interaction plot in figure 45. It can be seen that the noisy unfiltered file produces the best overall results at all SNR levels. Table 14 shows that the difference is not significant at low SNR levels (-40dB, -30dB), but the unfiltered speech score is significantly better than the audiovisual filtering score at higher SNR levels. It can also be seen from table 14 that the unfiltered score also outperforms the spectral subtraction approach.

There are several reasons for these findings. Firstly, the limitations with the audiovisual model, as discussed in section 5.5 are repeated with the results presented in this section. At higher SNR levels, distortion is introduced to the filtering, resulting in poor results at these levels. Secondly, as discussed in the corpus review in chapter 3, the VidTIMIT corpus is recorded in an environment with some background noise present. The objective measures work by performing a comparison of the processed file with the original file. If the original clean speech sentence contains noise, then this will affect the final scores. In this case, noisy sentences are being compared to clean sentences which also contain noise, resulting in higher results for noisy sentences than might be expected. Finally, as stated, these results also confirm the limitations with the visually derived filtering approach as found in section 5.6, and also identify limitations that were not explored in similar work by others (Almajai & Milner [11, 12]).

In addition to the results presented above, a comparison of these results with the findings presented in section 5.5 was also made. Firstly, a comparison of the unfiltered overall composite results for the GRID and VidTIMIT results (presented in section 5.5 and this section respectively) is shown in figure 47.

The overall unfiltered scores in figure 47 are of interest, because it can be seen in this figure, as well as table 15 and the interaction plot in figure 48, that the overall composite score of sentences from the GRID corpus is significantly lower than for sentences from the VidTIMIT corpus. As discussed previously, this is due to the audio quality of the clean sentences in

Table 13: Selected results of Bonferroni multiple comparison of VidTIMIT corpus, showing P-Value results for difference between Unfiltered Speech and Audiovisual Filtering for noise intrusiveness scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -0.532 | 0.035 | -15.14 | 0.000 |
| -30dB | -0.557 | 0.035 | -15.85 | 0.000 |
| -20dB | -0.468 | 0.035 | -13.34 | 0.000 |
| -10dB | 0.094 | 0.035 | 2.682 | 1.000 |
| 0dB | 0.796 | 0.035 | 22.672 | 0.000 |
| +10dB | 1.513 | 0.035 | 43.11 | 0.000 |

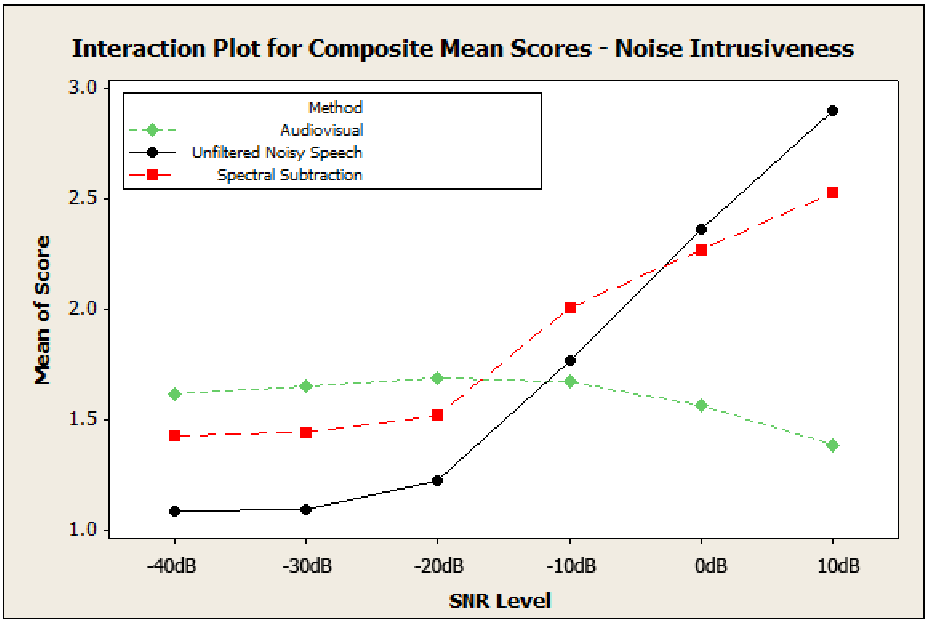


Figure 43: Interaction plot for noise intrusiveness composite objective score of VidTIMIT corpus at varying SNR levels, showing Unfiltered noisy-speech, Spectral Subtraction, and Audiovisual Filtering scores.

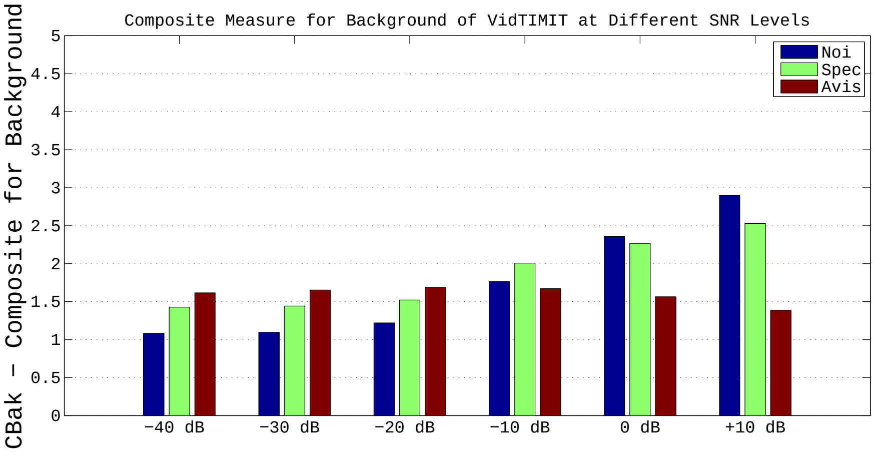


Figure 44: Composite objective mean test scores for speech distortion level of VidTIMIT Corpus for Unprocessed Noisy Signal (Noi), Spectral Subtraction (Spec), AV Enhancement (Avis).

Table 14: Selected results of Bonferroni multiple comparison of VidTIMIT corpus, showing P-Value results for difference between Unfiltered Speech and Audiovisual Filtering for overall composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 0.176 | 0.058 | 3.018 | 0.405 |
| -30dB | -0.011 | 0.058 | -0.190 | 1.000 |
| -20dB | 0.240 | 0.058 | 4.130 | 0.006 |
| -10dB | 1.123 | 0.058 | 19.277 | 0.000 |
| 0dB | 1.958 | 0.058 | 33.619 | 0.000 |
| +10dB | 2.965 | 0.058 | 50.91 | 0.000 |

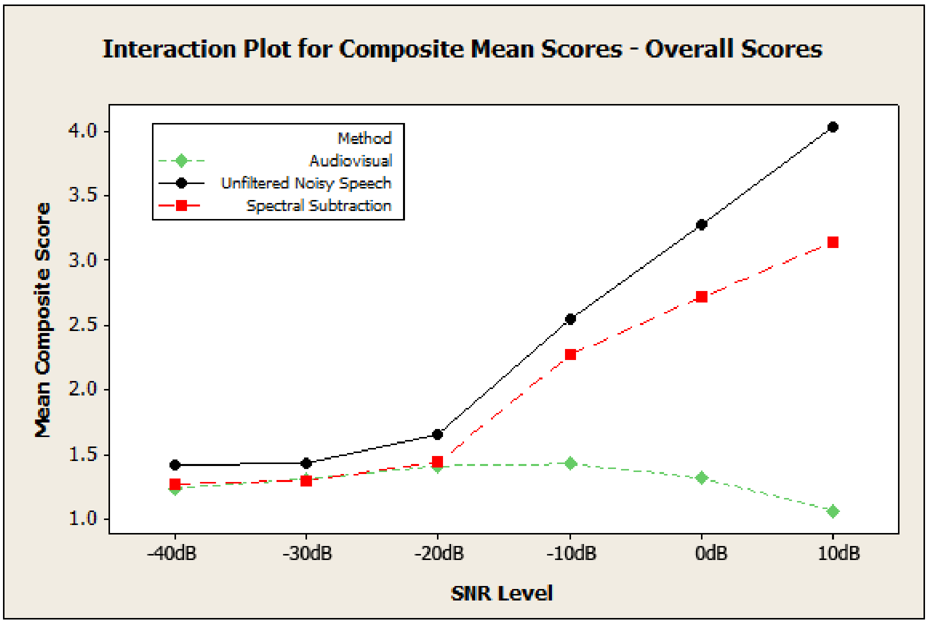


Figure 45: Interaction plot for overall composite objective mean score of the VidTIMIT Corpus at varying SNR levels, showing Unfiltered noisy-speech, Spectral Subtraction, and Audiovisual Filtering scores.

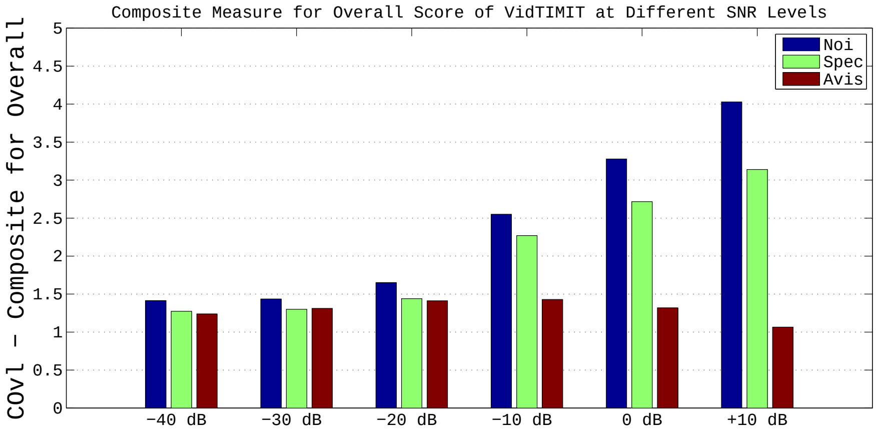


Figure 46: Composite objective mean test scores of VidTIMIT corpus for overall speech quality, for Unprocessed Noisy Signal (Noi), Spectral Subtraction (Spec), AV Enhancement (Avis).

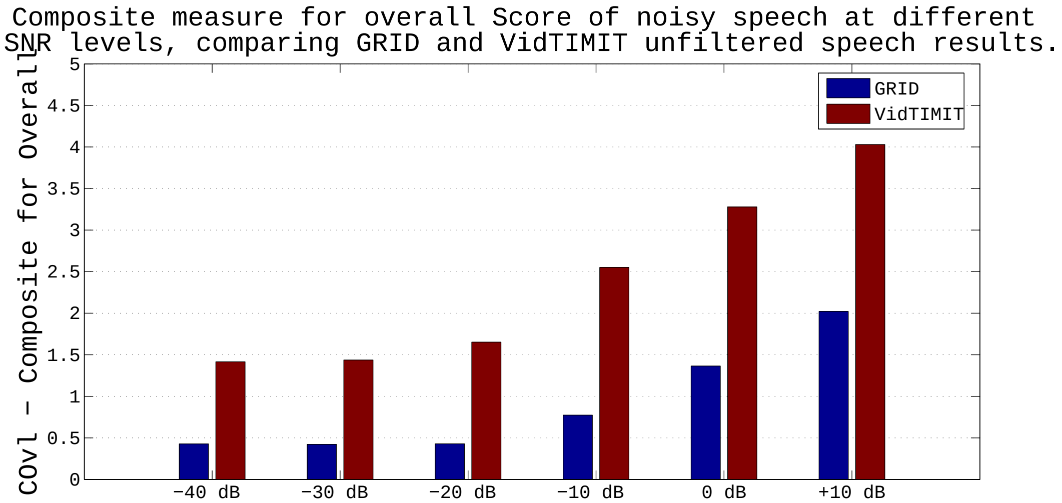


Figure 47: Composite objective mean overall scores, for unfiltered noisy-speech, comparing unfiltered noisy GRID (left) and VidTIMIT (right) mean scores at varying SNR levels.

the VidTIMIT corpus. As there is a degree of noise already present in the clean sentence, then this results in a higher score for noisy sentences than might be expected, and so has an impact on all results. It demonstrates a limitation with using a purely objective result, and is an issue with the corpus and the measurement technique rather than anything to do with any particular filtering approach.

In addition to a comparison of overall composite scores for unfiltered speech, the same comparison was also made of composite overall scores for sentences from the GRID and VidTIMIT corpora filtered using the audiovisual approach presented in this thesis. The results are shown in figure 49.

As was expected, taking into account the limitations of testing the audiovisual speech model on unseen data previously discussed in this section, and also the effect of the noisy corpus on overall scores, sentences from the GRID corpus were found to have a significantly higher mean score than sentences from the VidTIMIT Corpus. This was confirmed by the comparison of means in table 16 and the interaction plot in figure 50. Overall, these results confirm the limitation identified previously, that the system performs poorly when it is tested with data that is unrelated to the chosen training set. This is an expected result, as other similar research that uses a similar visually derived filtering technique such as work by Almajai & Milner [12] makes use of test data that closely matches the training set (i.e. the same single speaker

Table 15: Selected results of Bonferroni Multiple Comparison of unfiltered speech, showing P-Value results for difference between GRID Corpus Unfiltered Speech and VidTIMIT Corpus Unfiltered Speech for overall composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 0.987 | 0.047 | 21.117 | 0.000 |
| -30dB | 1.015 | 0.047 | 21.695 | 0.000 |
| -20dB | 1.224 | 0.047 | 26.175 | 0.000 |
| -10dB | 1.778 | 0.047 | 38.010 | 0.000 |
| 0dB | 1.913 | 0.047 | 40.910 | 0.000 |
| +10dB | 2.008 | 0.047 | 42.940 | 0.000 |

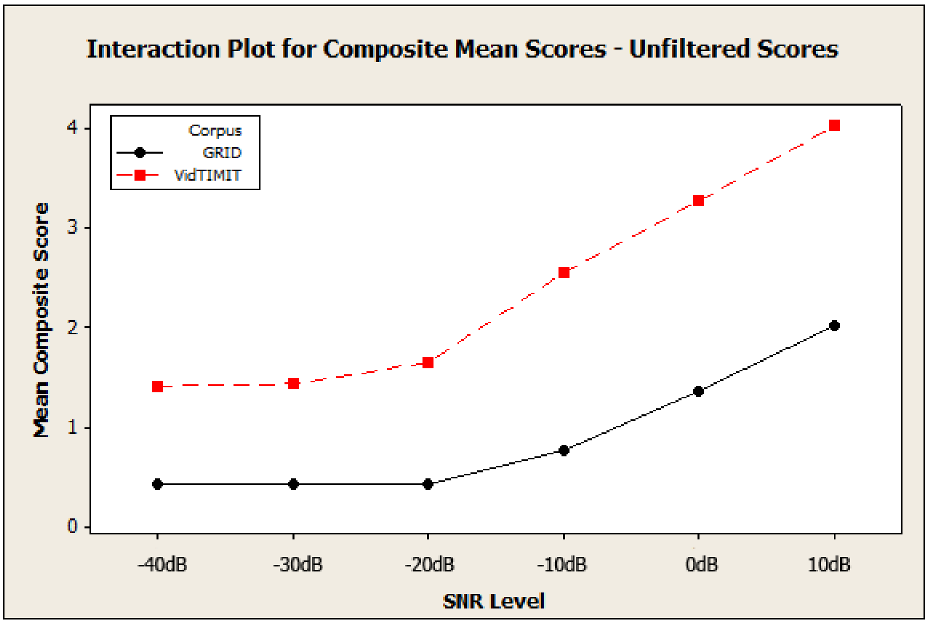


Figure 48: Interaction plot for overall composite objective mean score of GRID and VidTIMIT sentences at varying SNR levels.

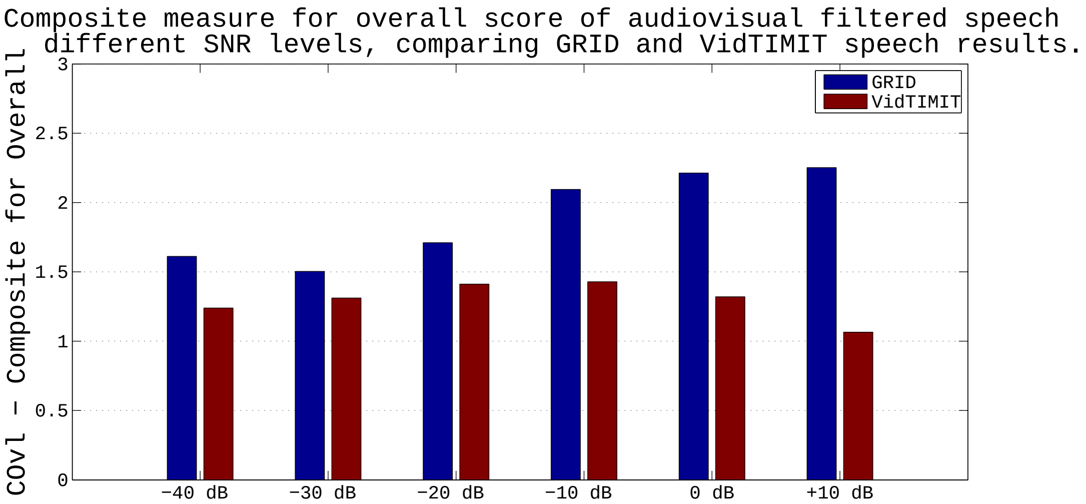


Figure 49: Composite objective mean overall scores, for speech filtered with the audiovisual approach, comparing GRID (left) and VidTIMIT (right) mean enhanced scores at varying SNR levels.

Table 16: Selected results of Bonferroni multiple comparison of audiovisual filtered speech sentences, showing P-Value results for difference between GRID Corpus filtered Speech and VidTIMIT Corpus filtered Speech for overall composite Scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -0.372 | 0.086 | -4.327 | 0.001 |
| -30dB | -0.192 | 0.086 | -2.231 | 1.000 |
| -20dB | -0.299 | 0.086 | -3.478 | 0.035 |
| -10dB | -0.666 | 0.086 | -7.750 | 0.000 |
| 0dB | -0.892 | 0.086 | -10.390 | 0.000 |
| +10dB | -1.186 | 0.086 | -13.81 | 0.000 |

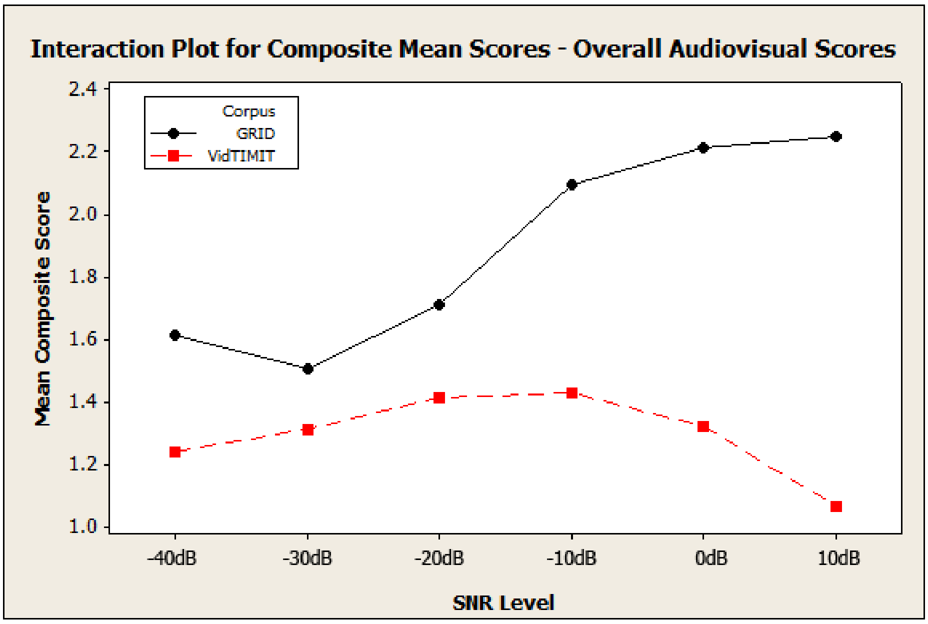


Figure 50: Interaction plot for overall composite objective mean score of the GRID and VidTIMIT corpora at varying SNR levels.

corpus is used for training and testing). In order to improve this aspect of the system, an improved audiovisual model could be considered, making use of more sophisticated filtering and a much more varied training set.

## 5.8 inconsistent audio environment

### 5.8.1 *Problem Description*

Section 5.5 presented results that showed that when the two-stage system developed in this work was presented with a consistent noisy-speech environment, good results were found. This section focuses on a different issue, that of speech filtering in an inconsistent audio environment. Many audio-only source separation algorithms perform well in stable environments, but less strongly when the audiovisual environment is volatile and changeable. A real life example of this is with modern hearing aids that contain sophisticated algorithms to utilise microphones to filter speech, but in changeable environments, output speech quality can suffer.

This section will demonstrate the effects of using a less consistent environment. Rather than a consistent noise as described in section 5.5, a different kind of noise containing silences and loud noises (in the form of inconsistent clapping) is mixed with speech. This simulates an environment where the noise source is inconsistent and changeable. The effects of using the two-stage audiovisual speech enhancement system is compared to the results of using audio only beamforming.

### 5.8.2 *Experiment Setup*

The system was set up in the same manner as described in sections 5.5 and 5.3.

Speech sentences from the GRID Corpus are used again in this experiment. Like in section 5.5, these sentences have not been used for training the audiovisual system directly, but are spoken by the same speakers as used for training. The key difference from other experiments in this chapter is that rather than consistent aircraft noise being used as the noise source, a completely different type of noise is used. Speech sentences from the GRID Corpus are mixed with loud and inconsistent clapping at an SNR of -10dB. The noise is shown in the spectrogram in figure 51 and the waveform in figure 52. It can be seen that the noise contains a period of silence at the start, and then the clapping is irregular and loud. Clean sentences from the corpus, with an example of a spectrogram and waveform shown in figures 53 and 54 respectively, are mixed with this noise to produce the noisy-speech mixture. The resulting combined noisy-speech mixture is shown in figures 55 and 56.

To compare the effects of audio-only and two-stage filtering, the noisy-speech mixture was filtered in two different ways. Firstly, the noisy mixture was processed using the standard audio-only beamformer. Secondly, this was then compared to filtering the same sentence with the two-stage audiovisual approach. In this section, a single example sentence was selected to be representative of all findings, as a very similar outcome was found for all speech sentences.

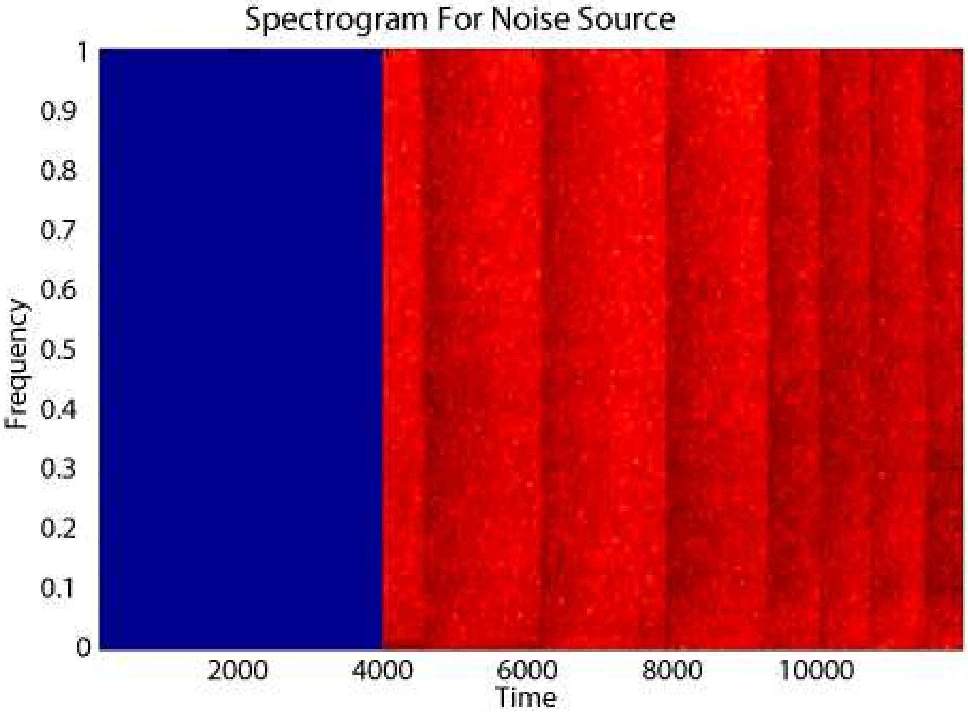


Figure 51: Noise source spectrogram for inconsistent noise environment experiment.

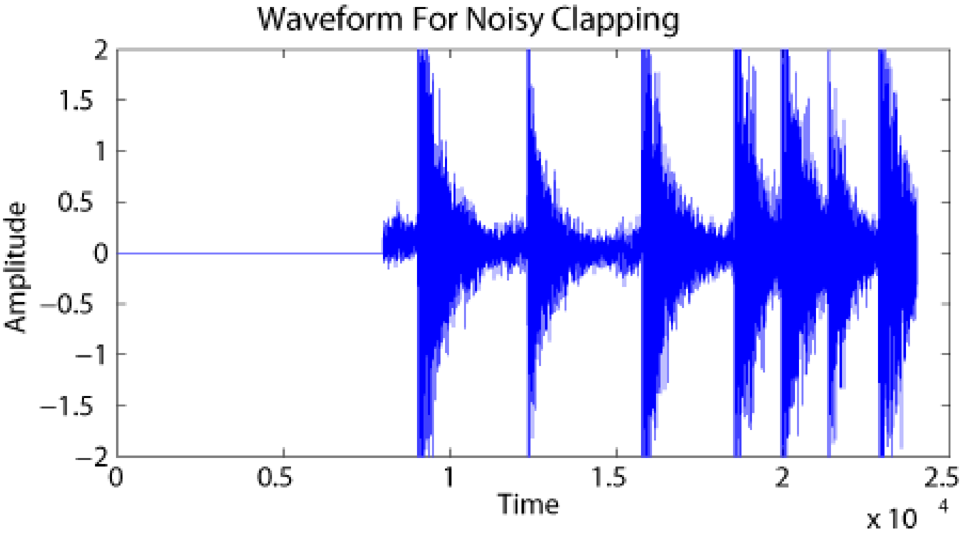


Figure 52: Noise source waveform for inconsistent noise environment experiment.

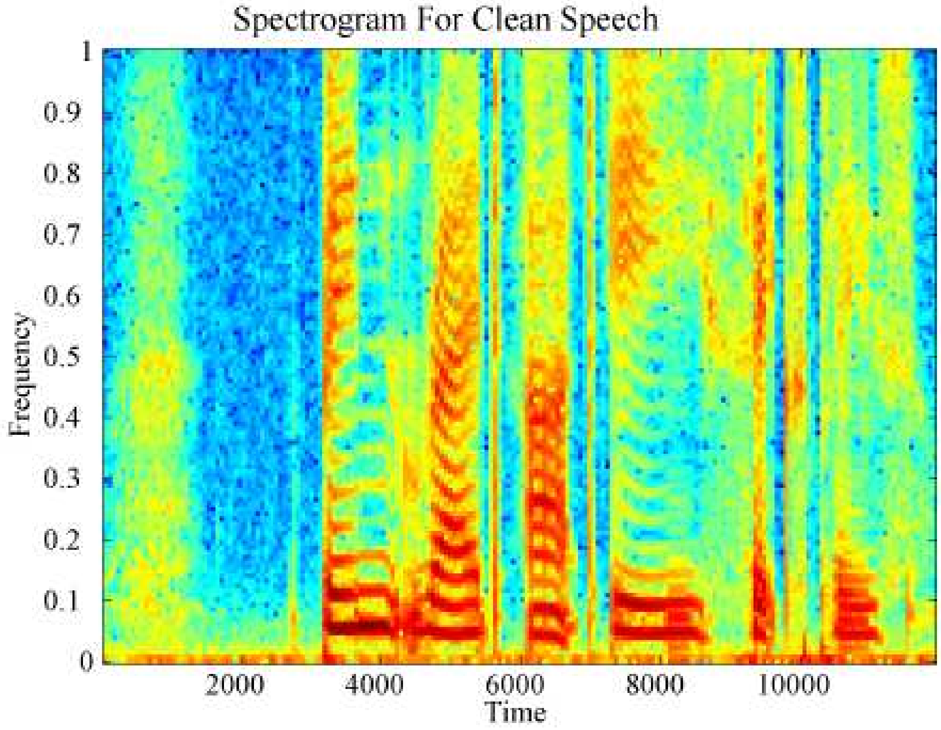


Figure 53: Clean speech spectrogram for inconsistent noise environment experiment.

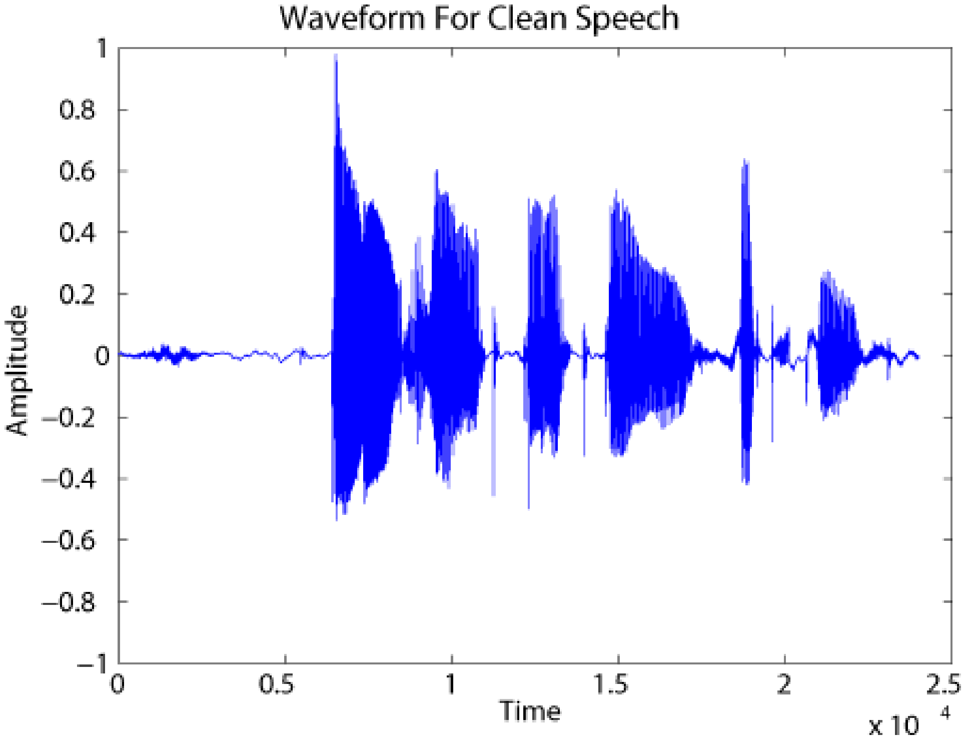


Figure 54: Clean speech waveform for inconsistent noise environment experiment.

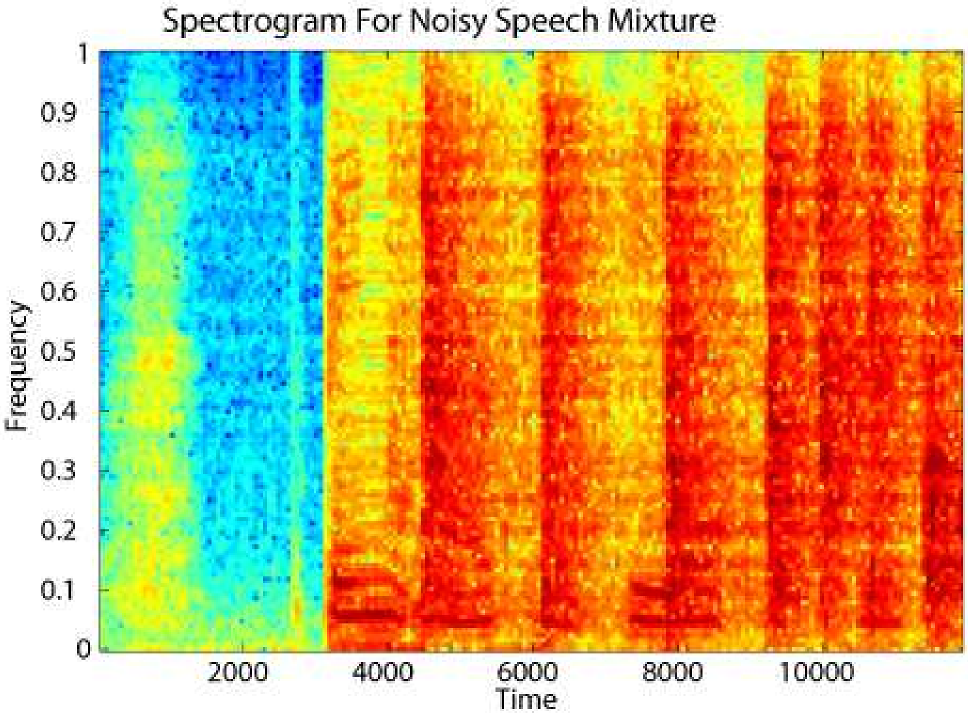


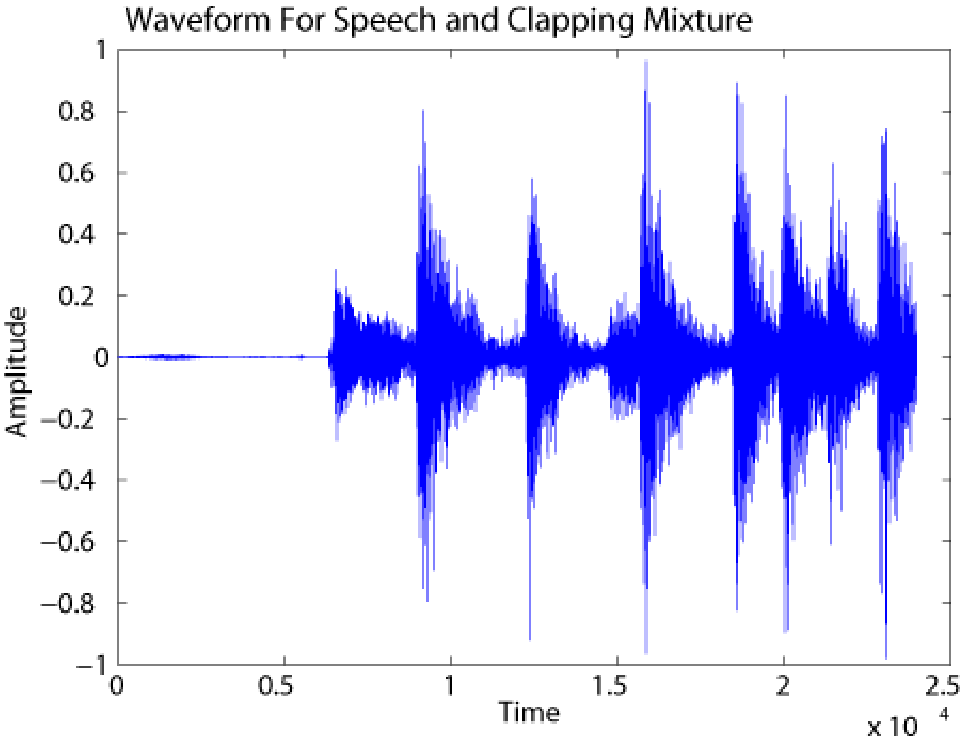
Figure 55: Speech and noise mixture spectrogram for inconsistent noise environment experiment.

### 5.8.3 *Evaluation Approach*

To evaluate this experiment, it was felt that a visual comparison of results would be adequate for representation in this section because the difference between the two filtering approaches was found to be very clear for every test sentence tried, and so it was decided that the most appropriate way to represent this was to show a visual comparison of the outputs in spectrogram and waveform form.

### 5.8.4 *Results and Discussion*

After creating the mixture of speech and noise as described above and shown in figures 55 and 56, this was filtered using two different approaches. Firstly, the noisy-speech mixture was filtered using audio only beamforming. This was then compared to the same sentence, filtered using the two-stage audiovisual approach developed in this work.



cockpit

Figure 56: Speech and noise mixture waveform for inconsistent noise environment experiment.

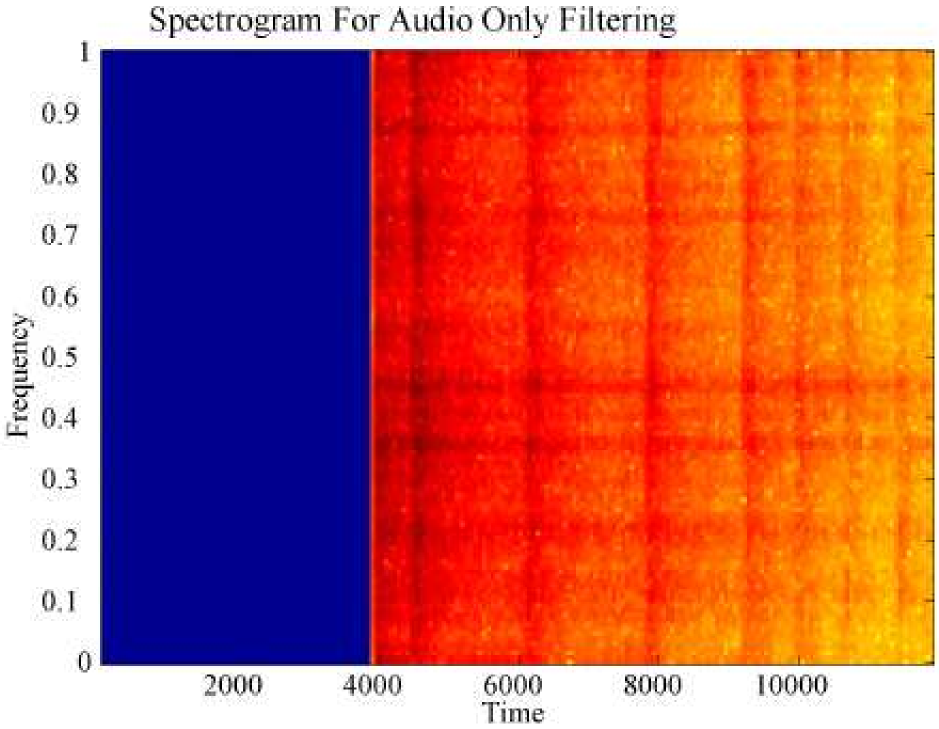


Figure 57: Spectrogram for inconsistent noise environment experiment, result generated by audio only speech filtering.

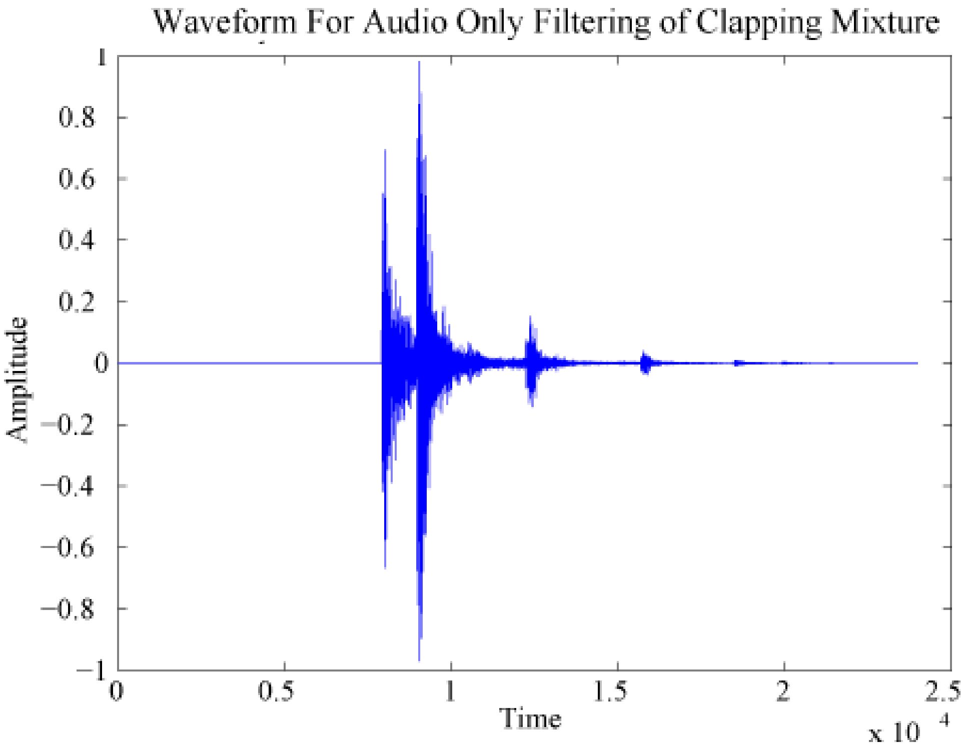


Figure 58: Waveform for inconsistent noise environment experiment, result generated by audio only speech filtering.

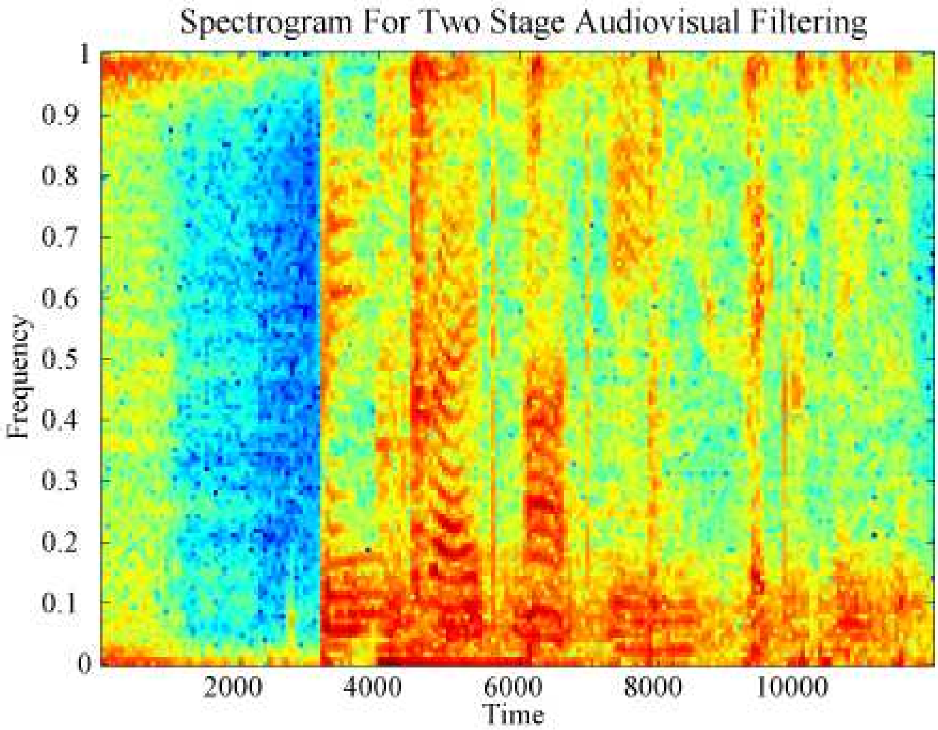


Figure 59: Spectrogram for inconsistent noise environment experiment, result generated by two-stage audiovisual speech filtering.

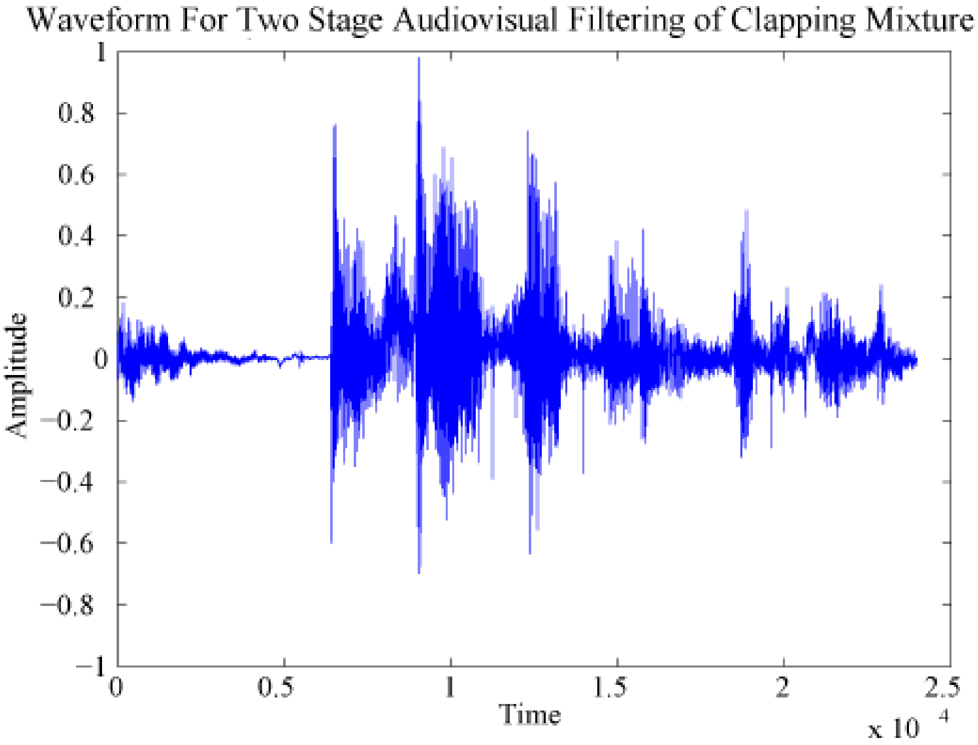


Figure 60: Waveform for inconsistent noise environment experiment, result generated by two-stage audiovisual speech filtering.

An example of the results found with audio-only processing is shown in the spectrogram and waveform in figures 57 and 58. It can be seen from a visual examination of the waveform that the sound completely fails to match the clean speech signal (figure 54). Listening to the file confirmed that the sound produced was not an accurate match for either clean-speech or noise.

In contrast to the simple audio-only processing (figures 57 and 58), when the additional pre-processing is used as part of the audiovisual filtering, then an improved result is produced. This can be seen in the spectrogram in figure 59 and also the waveform in figure 60. The comparison shows that a much-improved sound is produced as a result of filtering, which when compared to the original clean speech source spectrogram in figure 53, and the waveform figure 54, matches much more closely than using simple audio only filtering.

Overall, this section indicates that adding visually derived filtering may increase the flexibility of the system, as shown by the representative example result presented in this section. This shows that the addition of visually derived filtering can overcome limitations with single modality speech filtering, which can make it much more useful in environments where the noise source rapidly changes in SNR, volume, content, or when it can be difficult for audio only filtering to correctly distinguish between speech and noise.

## 5.9 SUMMARY

The established relationship between audio and visual aspects of speech (described in chapter 2) has been exploited by others in recent years to create multimodal speech filtering systems, as summarised in chapter 3. In this thesis, a novel two-stage audiovisual speech enhancement system has been presented, that also exploits the relationship between audio and visual aspects of speech. This new system was described in chapter 4. In this chapter, the results of a thorough evaluation of the performance of this multimodal speech enhancement system were presented. Firstly, the speech evaluation measures, both objective measures and subjective human listening tests, were outlined, and the remainder of the chapter described experiments carried out to thoroughly assess the performance of the multimodal speech enhancement system.

Section 5.3 presented a summary of an investigation into the optimum parameters to use in more formal experiments. This preliminary investigation evaluated the use of different sizes of training sets and the number of GMM components to use to train the audiovisual clean speech estimation model, and concluded that the use of a training set from four different speakers and 12 GMM components delivered the best results. Section 5.5 used objective and subjective testing to present a thorough evaluation of the performance of the two-stage multimodal speech filtering system in very noisy environments, using aircraft noise at a variety of SNR levels, from -40dB to +10dB. These results were compared to the results of using noisy unfiltered speech sentences for evaluation and also audio only spectral subtraction. The results showed that at lower SNR levels, the audiovisual approach was by far the best performing system, and produced promising results. At higher SNR levels however, it was found that the multimodal approach performed poorly.

Three other scenarios were also presented. Firstly, in section 5.6, one limitation of this multimodal system, a failure to function adequately with novel speakers, was identified. This was demonstrated by comparing the results from the main test-set (comprising sentences that the system had not been trained with, but from the same speakers as the training set), and a new test-set, using a completely different speaker from the same corpus that had not previously been encountered during training. A comparison of mean PESQ results confirmed that sentences from an unknown speaker produced worse results than for known speakers, confirming a limitation of using visually derived filtering. In addition, the performance of the system using completely novel data from a different speech corpus was also evaluated in section 5.7, and confirmed the limitation with regard to performance with novel speakers. However, section 5.8 showed one of the strengths of this new system, that of coping with an inconsistent noise environment. An inconsistent clapping noise was used in the experiments described in section 5.8, with periods of silence and varying speeds of clapping to simulate a challenging noisy-speech environment. An audio only beamforming approach was compared to the multimodal system, with a comparison of outputs showing that the multimodal system added a degree of flexibility to speech filtering and produced better results.

Overall, the discussion of results shows that while this new two-stage system has limitations, it produces promising results in challenging environments, and is flexible in comparison to single modality speech enhancement systems. However, there are occasions (such as at high SNR levels) when the additional filtering produces increased distortion, and so there is still scope for refinement. The visually derived Wiener filtering approach could be improved. This work used a relatively simple noise free speech estimation approach, making use of the GMM-GMR approach. This was originally developed for robot arm training by Calinon *et al.* [33] with the emphasis on producing an efficient and smoothed output, and so alternative methods could be considered. Recent work by Almajai & Milner [12] makes use of a more complex noise free speech estimation system, with phoneme specific mixture models used, suggesting that a more sophisticated visual filtering approach can deliver improved results.

The results given in this chapter also suggest that there is not one single speech enhancement approach that is guaranteed to produce optimal results in all circumstances. The results presented in this work show that while this two-stage approach produces improved results in very noisy environments, the use of visual information can sometimes be counterproductive. The potential inconsistent availability of audio and visual information in a more realistic speech environment also has to be considered. This work makes use of a simulated environment, with pre-recorded visual information and static speech and noise sources, but in a more realistic environment which a hearing aid would be expected to function in there may be multiple moving sources or good quality visual information may not always be available. Changes in light, pose, and actions by the speaker such as placing a hand over their mouth may render visual information unusable. Therefore, it is important to consider how a multimodal system can best take advantage of the variation in audio and visual information to deliver the best results on a frame by frame basis.

The extension of this system to become more autonomous, adaptive and context aware is feasible. This would automatically select the most appropriate speech filtering technique on a frame by frame basis, depending on the quality and availability of audio and visual information. For example, in a very noisy environment with good quality visual information readily available, it would be appropriate to use multimodal two-stage audiovisual speech enhancement, whereas if less noise was present, or if visual frames were found to contain inadequate information, it would be better to use audio only speech filtering. Taking inspiration from systems which process speech in different ways depending on the type of noise such as Neurofuzzy methods (Esposito *et al.* [57]), and the decision rules used in commercial hearing aids, the next chapter discusses the extension of the system evaluated in this chapter, to present a novel, proof of concept, multimodal fuzzy logic based speech enhancement framework.

# 6

# TOWARDS FUZZY LOGIC BASED MULTIMODAL SPEECH FILTERING

## 6.1 INTRODUCTION

The main aim of this thesis is to develop a multimodal speech enhancement system. After an investigation of state-of-the-art research in chapter 3, chapter 4 proposed a new two-stage audiovisual speech enhancement system that makes use of both audio and visual information to filter speech. The results of comprehensive testing of this system were presented in chapter 5, and a number of key strengths and weaknesses were identified. It was concluded that although there were strengths with this system, in that good results were found in extremely noisy environments showing that this new system was capable of producing results in challenging environments, and that it demonstrated the feasibility of speech enhancement using visual information, there were also limitations found. For example, the system was found to introduce speech distortion at a high SNR. Chapter 5 concluded by identifying some potential refinements to the system, one of which was to extend the initial system with the use of fuzzy logic to make more cognitively inspired use of audio and visual speech information.

This chapter extends the system evaluated in chapter 5 and presents a multimodal fuzzy logic based speech enhancement framework. Firstly, some limitations with the initial system are discussed, explaining why a single speech enhancement system is not suitable for use in all circumstances. The decision to make use of a fuzzy logic based system is then justified. The requirements of an autonomous, adaptive, and context aware speech enhancement system are discussed, followed by an assessment of why a fuzzy logic based approach is considered to be a suitable method for extending the system initially presented in chapter 4. An introduction to the use of fuzzy logic in the context of this system is also provided. This chapter also provides a brief review of several other potential alternative approaches to fuzzy logic.

The chapter then presents a novel, multimodal, fuzzy logic based speech filtering framework. The utilisation of the audio and visual input data, the inputs to the fuzzy inference system (the detectors), and the resulting fuzzy sets are described. The rules for the fuzzy logic based system, based on these fuzzy sets are then discussed. Finally, the challenges of thoroughly evaluating this initial system are then briefly discussed. While the work presented in this chapter does not represent a final and completed system, it is intended to demonstrate the feasibility of such an approach as an extension of the initial system presented previously in this thesis, showing that making more intelligent use of multimodal information is viable.

The remainder of this chapter is divided as follows. Section 6.2 discusses the limitations of the current system, and the justification for extending the initial system to utilise a fuzzy logic controller is described in section 6.3. This includes the requirements and the suitability of implementing a fuzzy inference system. Some alternative approaches are then briefly discussed in section 6.4. The extended fuzzy logic framework, utilising the feature extraction techniques and speech processing techniques discussed in previous chapters is then described in section 6.5, detailing how fuzzy logic can be integrated into the existing system, and describing the input fuzzy variables and rules in detail, with a discussion of evaluation approaches presented in section 6.6. Finally, section 6.7 sums up the chapter.

## 6.2 limitations of current two-stage system

The two-stage speech enhancement system presented previously in this thesis was shown in chapter 5 to be capable of producing positive results in environments where audio-only speech enhancement techniques were found to perform poorly. However, the results also identified a number of weaknesses with this two-stage approach, most significantly that when the SNR was relatively high, the two-stage approach was outperformed by audio-only approaches due to the distortion introduced in the audiovisual filtering process. There were also limitations that apply more generally to visually aided speech filtering systems. These systems, whether visually derived Wiener filtering approaches (Milner & Almajai [130]) that estimate the noise-free audio signal with visual information, or multimodal beamforming systems (Rivet *et al.* [149]) that use visual information for directional focus, rely on a clean source of visual information. The majority of multimodal speech research in the literature make the assumption that a good quality source of visual information is available at all times. The work presented previously in this thesis makes the same initial assumption. This is acceptable for laboratory simulations, where a pre-recorded speech corpus is often used for research and development. However, in real world environments, such visual information is not always guaranteed to be present.

Visual speech information is particularly vulnerable to corruption. Consider a hypothetical scenario where a person is listening to a speaker. The listener may not always be looking directly at the speaker; their attention may be directed elsewhere at points during the conversation. There may be situations where the speaker turns their head, places a hand over their mouth, or another person walks between the speaker and listener, temporarily blocking the view of the speaker's face. The light level may also change, making it difficult to identify speech. These problems can affect both tracking and filtering, and in a real world scenario, have to be accounted for.

There are also a number of other limitations that are not specific to visual information but apply to the audio domain. These were described in more depth in chapter 3, but essentially there are many examples of certain types of speech filtering being vulnerable to environmental conditions. For example, directional microphones are often only recommended for wind-free environments. If there is wind present, it is recommended that omnidirectional mode is used. Omnidirectional mode (with the directional setting disabled) is also recommended for very quiet environments without background noise present (Chung [44]). There are also scenarios where speech algorithms in hearing aids that reduce gain in frequency channels can perform poorly (for example, in broadband noise, such as in an automobile).

The conclusion that can be drawn from this, and from the results discussed in chapter 5, is that there is no single specific speech processing algorithm that is guaranteed to perform strongly in all scenarios. Different approaches have their own weaknesses, so while visually derived filtering is vulnerable to missing visual data; beamforming is vulnerable to transient noise such as an unexpected loud handclap. Although the system presented in the previous chapters can offset these weaknesses to an extent, it introduces some of its own. To solve problems with real data, as discussed in chapter 3, some commercially available audio-only hearing aids make use of decision rules to determine the extent and type of processing to apply to an input signal, based on various input detectors, and this allows for the adjustment of hearing aid settings to filter the input sound in a suitable manner.

## 6.3 fuzzy logic based model justification

After consideration of the limitations identified with the system presented in the previous chapters and speech enhancement systems generally, it was concluded that no single processing method was considered to be ideal for use in all scenarios, with results showing that audio-only, two-stage, and even no processing were suitable options, depending on the speech environment. It was also concluded that a real world environment is inconsistent and changeable, and a degree of flexibility is desirable with regard to speech processing. In order to take account of this, examples of how real world speech filtering systems deal with this problem were discussed in chapter 3. Chapter 3 described existing commercially available hearing aids, and identified that many state-of-the-art hearing aids are very sophisticated and make use of decision rules to decide on the level of speech filtering to apply. For example, as reported by Tellier *et al.* [171], hearing aids exist that can take account of a number of detectors to analyse the input signal in order to classify the noise. Such an idea can also be seen in neuro-fuzzy systems such as by Esposito *et al.* [57] that again seek to classify noise. Various audio input detectors can be used as an input to a set of decision rules, which then may apply different degrees of filtering, depending on the input. For example, in the case of hearing aids equipped with noise reduction algorithms, varying levels of gain reduction are carried out depending on the input (Chung [44]).

This chapter extends this idea by proposing a switching system that makes use of a set of rules in order to determine the most suitable means of processing individual speech frames when there is a choice between no processing, audio-only beamforming, or two-stage audiovisual speech processing. There are a number of requirements for such a system, and this section discusses these requirements. In addition, it was decided that a suitable choice to implement such a system was fuzzy logic (Zadeh [187]), and so the remainder of this system provides an introduction to fuzzy logic and a justification of the use of this technique. The preliminary proof of concept fuzzy logic based speech enhancement framework presented in this chapter extends the concept of decision rule based hearing aid algorithms to become multimodal, by using a fuzzy logic controller to determine the most suitable processing option to use for a frame of speech. To the best knowledge of the author, this is a novel framework, with no prior examples of multimodal two-stage fuzzy logic based speech enhancement in the literature.

### 6.3.1 *Requirements of Autonomous, Adaptive, and Context Aware Speech Filtering*

As discussed in section 6.2, it can be concluded that the use of one single speech filtering technique is not ideal in all circumstances, with the system presented in chapter 4 performing strongly in some conditions, but not others. Therefore, it is proposed to extend the initial multimodal system to make use of audio and visual information in a more autonomous, adaptive and context aware manner. Any such system has a number of requirements. Firstly, it has to be intelligent and context aware. By this, it is meant that the system takes account of the audio and visual speech environment and varies the processing decision depending on changes in the environment. It also has to be adaptive. Although many speech processing systems are trialled heavily in laboratory environments and clinical trials, as reported in chapter 3, the improvements found in these controlled environments are often not found to match up to real world trials, and so any proposed system has to be able to successfully process unpredictable environments. Any proposed system should also be autonomous. As discussed in the review of directional microphone research in chapter 3, Cord *et al.* [48, 49] reported that if the user was expected to manually determine the most suitable microphone setting (for example, omnidirectional or unidirectional) and switch their hearing aid to it when appropriate, then in the majority of cases, users would simply keep their hearing aids in omnidirectional mode. This was because of the user not being able to perceive any obvious immediate benefits of directional microphones, poor advice during the fitting process, or simply because of the inconvenience. Therefore, any proposed system should be automated, in order to deliver the optimum performance.

It is also important to consider the ease of future development. While this is less important with respect to the early stage system specifically being discussed in this thesis, commercial hearing aids currently make use of a wide range of proprietary detectors in order to determine the most suitable level of gain or directionality to apply at different frequencies (Chung [44]). As discussed in chapter 3, this varies between hearing aids and between manufacturers. Therefore, it should be possible for the system proposed in this thesis to be capable being extended to make use of an increased number of detectors, satisfying a requirement to be scalable. Another requirement of any state-of-the-art speech filtering system is that it should be capable of being tweaked and tuned without great difficulty. Hearing loss can vary widely between individuals, including hearing loss at specific frequencies or frequency ranges, and this wide variety of loss is not always handled well by existing theory behind hearing aids (Allen *et al.* [8]). Accordingly, when a modern programmable hearing aid is provided, patients are expected to undergo fitting sessions, where their hearing aid is programmed to better fit their individual hearing loss and comfort levels. Therefore, any proposed system should contain accessible parameters that can be tweaked and tailored in order to adapt to the hearing ability and preferences of the user.

### 6.3.2 *Fuzzy Logic Based Decision Making*

It was decided that one technique that met these criteria was fuzzy logic. Fuzzy logic was first proposed by Zadeh [187], and allows uncertainty to be represented using the concept that a variable may belong to a set to an extent, but not completely. This moves away from the traditional mathematical concept of data either belonging or not belonging to a set, and uses the concept that a value can partially belong to a set. These sets are then used to create rules based on expert knowledge that can be evaluated to give an output based on uncertain input. Fuzzy inference is used in many applications, such as to the control of radio controlled vehicles (Tanaka *et al.* [170]) and noise cancellation (Chang & Shyu [35]). It is important to clarify the distinction between fuzzy logic and probability. Fuzzy logic is not concerned with mathematically modelling a system (for example, HMMs Ghahramani [68])), and instead uses expert rules. While they both represent uncertainty, the semantics are different. For example, probability is based on the concept that a value has x probability of belonging to a set. This means that it may belong to the set, or it may not. On the other hand, an example of a fuzzy statement would say that a value belongs to a set to x extent. This is an important distinction. This is relevant because this concept can be applied to speech input, for example, audio input conditions can vary, depending on environmental conditions.

As fuzzy logic will be discussed in depth in the next two chapters of this thesis, a number of terms will be formally introduced:

### *Fuzzy Sets*

Classically, values are represented with crisp sets (Hellmann [80]). These are sets of values where the values within it are clear and explicit, meaning that a value can easily be defined as part of a set or not. A crisp set is adequate for many applications, however, in the case of the work discussed in this thesis, it could be argued that there is not a clear-cut off and strict separation between, for example, low noise and high noise environments. Fuzzy logic allows for a relaxation of these crisp sets to create fuzzy sets. A hypothetical example of the difference between fuzzy and crisp sets is shown in the example presented in figure 61. It can be seen in the hypothetical example that the crisp set represents values that either are members of the set or not. However, the fuzzy set shows a different meaning. Values in the example shown range from 1 to 0, just like the crisp set. Values of 1 represent a definite membership of the set, however, values of less than 1 but greater than 0 represent a partial membership of the set.Values of 0 indicate non-membership of that set.

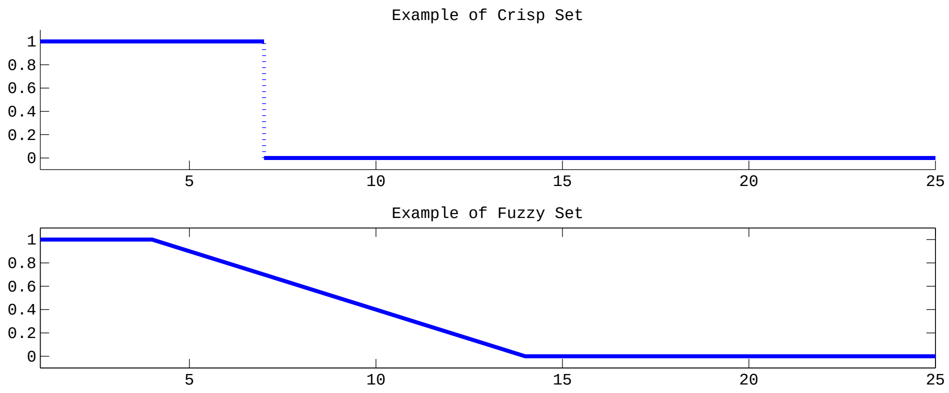


Figure 61: Representation of a traditional crisp set (top), showing values as either belonging to a set, or not belonging, compared to an example of a fuzzy set (bottom), showing that a value can definitely belong to a set (if it has a value of 1), or it may partially belong to that set if it has a value between 0 and 1, or not belong to a set if it has a value of 0.

There can be several membership functions for an input, making it possible for an input to be a partial member of several sets simultaneously. So for example, a hypothetical input audio value can be part of a "low noise" set, part of a "high noise" set, or possibly a partial member of both sets. These membership functions are defined as showing the extent of participation of each input (Hellmann [80]). These membership values are used by the rules to determine outputs.

*Application of Fuzzy Operator*

Fuzzy sets can then be used to perform a number of operations, including intersection (AND), unification (OR), and negation (NOT). This allows for the construction of rules. Rules make use of expert knowledge, and are expressed using variables described by fuzzy sets. Each rule consists of two components, an antecedent and a consequent. The antecedent consists of the part of each rule before the 'Then' (e.g. IF a AND NOT b), and the consequent is the output (e.g. THEN c). These are used to create a series of rules, using each of the fuzzy sets. Each of the rules is evaluated to evaluate the extent to which it belongs to a chosen output. As a fuzzified input may belong to several sets, this means that several rules may be fired.

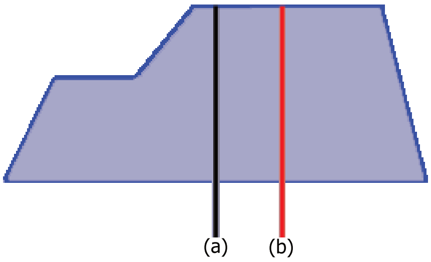


Figure 62: Example of the an aggregated fuzzy output. The output singleton value (a single defuzzified output number) is calculated from this. This can take several forms, including a centroid of the aggregation (a), or the middle of the maximum value (b).

### *Aggregation*

In order to produce an output value, each of the individual rule outputs is aggregated into a single fuzzy set. By this, it is meant that each rule is evaluated, producing the output fuzzy values described above. Each rule is then combined to produce a single fuzzy set that encompasses all the rule outputs that were applicable.

### *Defuzzification*

The final stage of a fuzzy inference system is the determination of a final output. This is known as Defuzzification, and refers to the process of obtaining a single output value from the aggregated fuzzy set. It is generally considered useful to obtain one final decision from the fuzzy-system, and so the aggregated fuzzy set described above must be defuzzified. Figure 62 represents a hypothetical aggregated fuzzy output set. In order to output a single value, one value must be chosen. There are several different methods, including centroid (the centre of the total aggregated area), middle of maximum (the average of the maximum value of the output set), and largest of maximum. An example of a single output value using the centroid method is shown in figure 62 by (a), and an example of middle of maximum is shown by (b).

Since the development of fuzzy inference systems, there have been many areas where fuzzy logic has been applied, such as for the control of vehicles (Tanaka *et al.* [170], Abdullah *et al.* [1]), noise cancellation (Chang & Shyu [35]), speech recognition (Halavati *et al.* [74], Avci & Akpolat [14]), and image enhancement (Choi & Krishnapuram [42]). It is a very versatile approach, combining expert knowledge and uncertainty, and it was determined that this approach was suitable for use as part of the speech filtering system presented in this chapter.

### 6.3.3 *Suitability of a Fuzzy Logic Approach*

Several approaches were considered for use as part of this system, such as making use of Artificial Neural Networks (ANNs), GMMs, HMMs, or a hybrid of these approaches, such as neuro-fuzzy approaches, which use fuzzy inference inputs into a neural network (Esposito *et al.* [57], El-Wakdy *et al.* [55]). As discussed previously, the proposed intelligent system had a range of requirements, and fuzzy logic was considered to be a suitable approach for several reasons. Firstly, a fuzzy based system fulfils the requirement of being context aware, adaptive, and autonomous. Fuzzy logic is an approach that allows for uncertainty to be represented, therefore it is context aware, in that it is capable of responding to different changes in the environment, based on inputs into the system. It is also adaptive, in that it can respond to these inputs, so in the system presented in this chapter, the different inputs (which will be discussed later in this chapter) provide information about the environmental context (such as the level of noise), the fuzzy-system makes a decision regarding the suitable processing choice, depending on this input. Fuzzy logic is also autonomous, and can make decisions without manual input. This can be seen in examples of other fuzzy inference systems, such as for controlling autonomous vehicles (Abdullah *et al.* [1]).

Another reason that fuzzy logic is considered more suitable than other approaches is because of the philosophy behind fuzzy logic, as opposed to probability based approaches. Fuzzy logic is not concerned with mathematically modelling a complete system in the manner of HMMs and GMMs, but instead makes use of expert rules, which is more in line with the approach discussed in this chapter, as it is similar to the rules and processing used in contemporary hearing aids.

One key advantage to using fuzzy logic rather than equivalents such as ANNs and GMMs is that fuzzy logic is based on expert knowledge; this means that there are a number of rules that can be programmed. Rather than requiring complete training, or the use of a mathematical model, there are a number of rules that are defined. This is of relevance, because as discussed above, a future requirement for a practical implementation of this system is that it should be possible to adjust settings and programs to suit individual users. As discussed in chapter 3, current commercial hearing aids offer a range of programs and settings (Chung [44]), and users are expected to attend fitting sessions to customise their hearing aids for their specific hearing loss. The use of a fuzzy based system allows for this customisation to be applied. For example, different users may have a different interpretation of what constitutes a "very noisy environment", and so this can be customised in a fuzzy based system in a more straightforward manner than (for example) training ANNs or GMMs.

The same applies to the scalable potential of using a fuzzy based system. As discussed in chapter 3, hearing aids make use of a number of different detectors for determining gain control or directionality. These can represent inputs into rule based decision systems regarding gain or algorithm adjustment, and the number of detectors and the rules used in specific devices are both proprietary and can vary greatly between manufacturers. The use of fuzzy logic represents a logical extension to this concept, as it represents a system that can easily be extended to take account of additional rules, scenarios, and inputs. The preliminary system represented here makes use of very basic detectors, and could theoretically be represented using a different approach, such as with HMMs. However, it was decided that it was important to demonstrate a fuzzy-system, as it would arguably be more difficult to extend, train and implement a more sophisticated version of a Hidden Markov Model (HMM) based system in future, whereas a fuzzy logic system is easier to refine and extend, due to its use of expert knowledge and the clearly defined rule base. For example, the addition of a wind detector in a future implementation would require the tweaking and addition of rules, rather than complete retraining of the system.

Overall, although there were several feasible alternative approaches, as will be discussed in section 6.4, it was decided that fuzzy logic was a suitable approach for the system proposed in this chapter.

## 6.4 potential alternative approaches

### 6.4.1 *Hidden Markov Models*

HMMs (Ghahramani [68]) are statistical Markov models, and are a type of temporal Bayesian network. While simpler Markov models have directly observable states, in a more complex hidden model, the states are hidden, and only the output values and their probabilities are directly observable. This approach is often used for time series data, where the data does not depend on previous time steps, but is theoretically only limited to that step. As described by Ghahramani [68], HMMs are defined by a set of states, an alphabet of changes between states, a transition probability matrix (giving probabilities of transitions from state to state), and an emission matrix (giving the probabilities of the outputs). There have been many different examples of making use of HMMs in the literature, with application to a range of problems, including speech recognition and enhancement (Bansal *et al.* [17], Hershey & Casey [82]), robot control applications (El-emary *et al.* [54]), multimodal emotion recognition (Zeng *et al.* [190]), and many other tasks. A detailed overview of HMMs is provided by Ghahramani [68].

There are a number of general benefits to making use of HMMs. Firstly, they have a solid statistical grounding, with statisticians able to perform mathematical analysis of the results and manipulate the training process. Also, although hidden, there is also transparency, in that it is possible for the model to be read, so it is not a complete 'black box' solution. Finally, it is possible to incorporate prior knowledge into the model. By this, it is meant that prior knowledge can be used to constrain the training process, and the model can be initialised close to something believed to be the correct final output. This reduces the required training process.

HMMs are considered to be suitable for classification, decoding, and learning problems. In classification problems, given a HMM and an observation sequence, then the probability of the next observation can be calculated. In decoding problems, given the model and sequence, then the most likely sequence of hidden states can then be calculated. Finally, there are learning problems. Given some training sequences and a model, the problem is to determine the parameters most suitable for the model.

The most significant practical limitation is that there is a lack of potential flexibility in a speech filtering system designed using a statistical approach. While an initial system could feasibly be designed and trained, it would be designed for a very specific set of circumstances. As discussed elsewhere in this chapter, hearing aids are customised for individual users in fitting sessions, where individual parameters such as gain control are adjusted to the comfort and hearing loss of that individual. It is hypothesised that with different levels of hearing loss, there are different levels of listener comfort and different processing thresholds that are considered to be suitable, due to very specific differences in how individuals with hearing loss perceive specific sounds, showing that one solution may not be suitable for all users (Allen *et al.* [8]). When hearing aids are fitted, patients are required to undergo at least one fitting session to adjust the hearing aid to their individual hearing loss. A fuzzy logic system makes use of expert knowledge, and it is feasible for individual fuzzy set thresholds to be tweaked in a hypothetical fitting process.

If using a HMM, in order to adjust to the comfort of an individual, a new model would have to be trained to suit, with the training data being labelled in order to match the comfort of an individual listener. This would result in a different model for each listener, and is a far more complex procedure than a domain expert (such as a trained audiologist) being able to tweak thresholds with a rule based model.

There is also an issue with extension of the model. Although the initial system proposed in this chapter is relatively simple, making use of a limited number of detectors, there is scope for future refinement of this system to take account of a different range of detectors, in order to improve performance. For example, processing could vary depending on the type of noise (Esposito *et al.* [57]), rather than just the presence. In such a scenario, the extension of a fuzzy logic system is less complex than extending a statistical approach like using HMMs or GMMs. The extension of a fuzzy logic system would require new rules to be devised and new fuzzy sets to be created, whereas the extension of a statistical model would require a completely new model to be trained with different hidden states, and accordingly, a different transition matrix, which could grow to become unwieldy in a more complex model.

Finally, there is the issue of training a model. Even a relatively simple approach with a limited number of states and outcomes will be expected to take account of a wide range of input data. This necessitates the acquisition of considerable quantities of labelled data, in order for the model to be able to be able to determine the appropriate output. There are also the standard machine learning difficulties of overfitting (see Ghahramani [68]) to consider, as well as the operation of a HMM based approach being potentially slow due to having to evaluate the complete model on a frame-by-frame basis.

Overall, although it is possible for the initial system presented in this chapter to be developed using a statistical approach such as HMMs, and these have been applied to a wide range of domains such as speech recognition Luettin *et al.* [122], Bansal *et al.* [17], emotion recognition (Zeng *et al.* [190]) and speech filtering (Hershey & Casey [82]), it was determined that a system implemented with this approach is less practical when the future proposed usage of this system is taken into account.

### 6.4.2 *Neural Networks*

Artificial Neural Networks (ANNs) (Zurada [193], Zayed *et al.* [188], Haykin [78]) are considered to be a biologically inspired machine learning approach as they are theoretically similar to the structure of the brain (i.e. the biological connections between neurons). Generally, ANNs consist of input and output processing nodes, which are connected to a network using weighted connections. Neurons receive weighted values from incoming nodes, sum the received values, apply an appropriate activation function, and then pass the output to other nodes. The transfer function of an individual neuron refers to the threshold required before the neuron fires an output. There are many different types, including the commonly used logistic or tanh sigmoid function used commonly in Multilayer Perceptron (MLP) networks (Rumelhart *et al.* [156]), which activates when the input exceeds a threshold level, and also Leaky Integrate-and-

Fire (LIF) neurons, first proposed by Adrian [4], which are used in spiking networks (Maass [123]), where the inputs to a neuron are received in the form of input peak signals known as spike trains. If enough cumulative spike inputs are received, then the neuron fires then resets, but if spikes are not received, neuron activity gradually dissipates (hence the term "leaky").

There are many different topologies that are used, meaning that there are a multitude of network designs that have been used in the literature. The most common are what are known as Feedforward Networks (Haykin [78]). These are networks that receive inputs and then pass them through one or more hidden layers, before being output. The values of this network are fully determined by inputs. Training of these networks is carried out with various forms of gradient descent, including error-backpropogation (Rumelhart *et al.* [156]).

Another topology that is used (although less commonly than feedforward networks) is recurrent neural networks. This particular topology is currently the subject of much study, in particular the area of Reservoir Computing (Schrauwen *et al.* [163]). The difference between feedforward network and recurrent networks is that connections exist both backwards and forwards through layers in the networks. There are also delays between connections, and so these networks effectively have a form of memory as information remains circulating inside the networks. This memory makes recurrent networks especially suitable for temporal problems. Recurrent neural networks have been used for a variety of purposes, including instrument classification (Newton & Smith [134]) and robot arm control (Joshi & Maass [97]). A variety of different recurrent network designs have been used, including Echo State Machines (Jaeger [94]), and Liquid State Machines (Maass *et al.* [124]). These basically rely on setting up a neural network with a reservoir of randomly initialised weighted neurons. The reservoir network is then left untrained, with only an output layer trained. An example of this is work by Newton & Smith [134], who have made use of such a network to classify musical instrument sounds based on onset spikes, and multimodal laughter detection by Scherer *et al.* [162]. However, recurrent networks can be difficult to train, due to their nature, and have slow convergence rates (Hammer & Steil [75]).

There have been many different utilisations of neural networks in the literature. A full summary is considered to be outside the scope of this thesis. But neural networks have been used for many different problems in a range of domains such as classification problems (Newton & Smith [134]), decision support (such as for cancer care, which is summarised by Lisboa & Taktak [118]), speech filtering (Hussain & Campbell [89], Esposito *et al.* [57]), and character recognition (Gader *et al.* [66]).

Overall, there are many different topologies of neural networks, each with their own strengths and weaknesses, and with many different applications in a range of fields. Like HMMs, ANNs are good for information processing and tasks such as classification problems, and some topologies (such as echo state networks) can also make use of temporal information. They are capable of solving complex problems, and have been used in the speech processing domain, such as in neuro-fuzzy systems (Esposito *et al.* [57]) and speech enhancement (Hussain & Campbell [89]). It is potentially feasible for a neural network to be developed to make a decision regarding the most suitable processing method when presented with novel information.

However, there are some issues with using this approach. Like the HMM method discussed previously, there is the fundamental issue of customising a neural network based approach for individual listener comfort. While fuzzy logic can be set by adjusting some rules to suit, a neural network based approach may require considerable retraining for each user, using training data that would be difficult to acquire. If the training process is not suitable (and during the fitting of hearing aids, patients often have follow up sessions to tweak their settings), then the network would have to be retrained until suitable. The same issue (lack of training data) applies to many other machine learning approaches such as support vector machines and GMMs. Although they could feasibly solve the particular problem discussed in this chapter, their general purpose utilisation in a future system becomes more problematic. Neural network approaches also have the issue of being effectively a "black box" system. Tweaking and refining is not a simple matter.

In summary, although neural network approaches can theoretically be used to solve the problem outlined in this chapter, it was considered more practical to make use of a fuzzy logic based system.

## 6.5 fuzzy based multimodal speech enhancement framework

In light of the limitations with both the audiovisual system presented earlier in this thesis, and other speech enhancement limitations that have been outlined earlier in this chapter, the initial multimodal two-stage speech enhancement system presented in chapter 4 has been extended to become more sophisticated by developing a fuzzy logic based system. As discussed previously in this section, it was felt that using fuzzy rules represented the most practical solution, and could theoretically be implemented and modified in future designs of hearing aids. To do this, a fuzzy logic controller has been implemented to determine the most suitable method for processing an individual frame of speech. There are several possible processing options, (i) applying no filtering to the speech frame, (ii) audio-only processing with a beamformer, or (iii) the two-stage audiovisual speech filtering approach discussed in chapter 4. To determine the most suitable processing option, a set of rules are used, which receive fuzzy inputs from detectors based on the input data. This initial implementation is evaluated in chapter 7. The remainder of this section discusses the proposed fuzzy logic based framework in detail.

### 6.5.1 *Overall Design Framework of Fuzzy System*

To integrate the fuzzy logic controller into the multimodal two-stage speech enhancement system described in chapter 4, the initial system shown in figure 15 is extended further by the integration of a fuzzy logic controller and the subsequent adjustment of the speech filtering options. The basic components introduced in chapter 4 are unchanged. Visual tracking and feature extraction is handled in the same manner, as is the audio feature extraction process. With regard to speech processing, the two processing options used, visually derived Wiener filtering and audio-only beamforming remain unchanged. However, the difference is that one or both of these stages may be bypassed on a frame-by-frame basis, depending on the inputs received by the fuzzy logic controller. This redesigned framework is shown in figure 63.

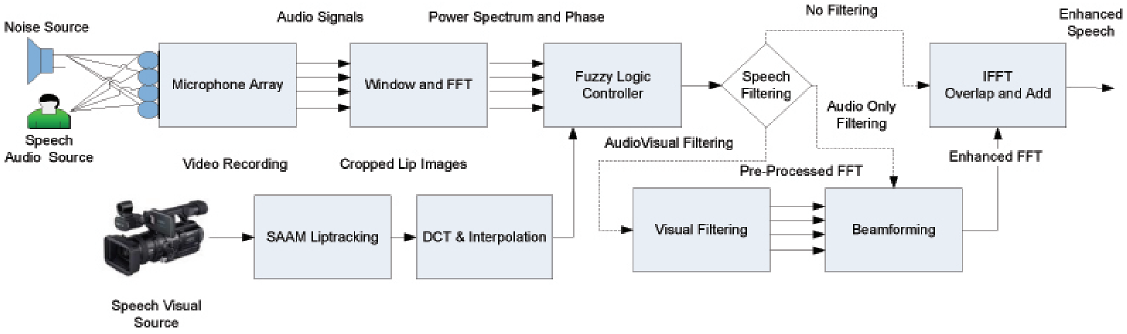


Figure 63: System diagram of proposed fuzzy logic based two-stage multimodal speech enhancement system. This is an extension of figure 15, with the addition of a fuzzy logic controller to receive inputs and decide suitable processing options on a frame-by-frame basis.

The diagram in figure 63 shows the high level extended system diagram with the alternative speech processing options. Depending on the inputs to the fuzzy logic controller, the type of processing performed on the input signal may vary from frame-to-frame. So for example, if it is detected that there is very little audio activity in a particular frame, then it may be decided to leave that frame unfiltered. Alternatively, if a moderate amount of audio energy is detected, then it may be decided that audio-only beamforming is the most appropriate processing method. If however, a lot of audio activity is detected in a particular frame and the visual information is considered to be of good quality, then the full two-stage process as described previously in this thesis may be used.

The decision as to which option is to be used is taken with the aid of a number of detectors applied to the input signal. As previously stated in chapter 3, audio-only hearing aids make use of a wide range of proprietary detectors such as level, wind and modulation detectors. In the initial implementation presented in this chapter, three detectors are used. An audio level detector is utilised. This does not consider speech or noise separately and is not a VAD, but simply considers the level of audio activity in each frame. Similarly, a visual quality detector has been developed to apply to the input DCT vector as an evaluation of the visual signal quality. The final detector used in this system is simply a feedback input of the processing decision made in the previous frame. This aims to minimise the chopping effect if the chosen processing method changes substantially from frame-to-frame. These fuzzy inputs are described in more detail in the next section.

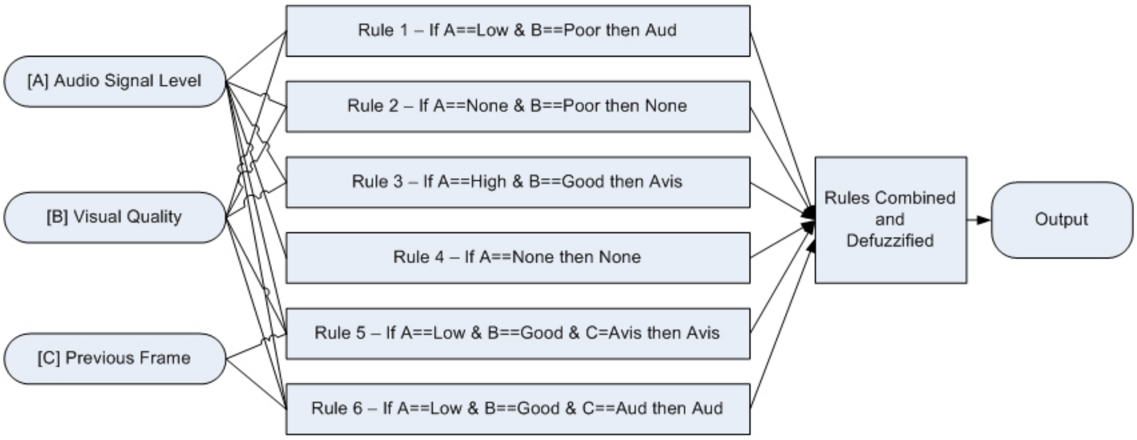

[A] Audio Signal Level
[B] Visual Quality
[C] Previous Frame
Rule 1 – If A==Low & B==Poor then Aud
Rule 2 – If A==None & B==Poor then None
Rule 3 – If A==High & B==Good then Avis
Rule 4 – If A==None then None
Rule 5 – If A==Low & B==Good & C=Avis then Avis
Rule 6 – If A==Low & B==Good & C==Aud then Aud
Rules Combined and Defuzzified
Output

Figure 64: Diagram of fuzzy logic components, showing the three chosen fuzzy inputs and the list of rules to be applied.

Fuzzy logic rules are then used to determine the most suitable processing method, depending on the input. Each individual frame is processed in a manner judged by the fuzzy logic controller to be most suitable. The next sections first discuss the input detectors in detail, and then describe the operation of the fuzzy logic rules used in this novel framework.

### 6.5.2 *Fuzzy Logic Based Framework Inputs*

The fuzzy logic controller builds a relationship between system inputs and the rules used to define the processing selection. In order to accomplish this, it takes a number of input variables and applies these to fuzzy logic rules. As discussed in section 6.3, each input variable must be decomposed into a set of regions (or fuzzy sets), consisting of a number of membership functions. The composition of these membership functions can vary in size and shape, based on the preference of the designer (Bagis [15]), and for the work in this thesis, it was decided to make use of trapezoid membership functions for all inputs in order to ensure consistency. . The fuzzy-system diagram is shown in figure 64, and it can be seen that there are three inputs to consider, audio level, visual quality, and previous frame processing decision.

*Visual Quality Fuzzy Input Variable*

The first input variable is the visual quality. This measures the level of detail found in each cropped ROI. As the system is audiovisual, visual information is a key component of the processing. However, this information can be of varying quality. There are occasions when the entire lip region is visible, but there are also occasions when the lip-tracker returns an incorrect result due to scenarios such as the speaker turning their head. There are also occasions when the lip region may be blurred due to movement, or only a partial ROI is returned. This is not such an issue with regard to the audiovisual speech databases utilised in the previous evaluation (in chapter 5), when a custom corpus is used, there are many more examples of poor visual data to take account of.

As there were many different potential speakers, an approach with as much flexibility as possible was required. One potential approach was to make use of a machine learning technique such as a HMM to create a model to evaluate the ROI and return a score to use as a fuzzy input variable. However, it was felt that this was not required for the initial implementation presented in this thesis. Instead, a simpler approach was devised that made use of the input DCT vector.

In order to determine the most suitable input variable, a custom corpus was recorded using real data from a variety of volunteers. This is discussed in more depth in chapter 7.5, and a number of trial videos were evaluated to calculate the most suitable value, with various variables investigated, such as the DCT input vector, and the tracker parameters of the actual cropped images. An investigation identified that the fourth DCT coefficient was consistently a better representation of the accuracy of the cropped DCT than any other single factor, and so this was used to create a mean value. As the DCT transform represents pixel intensity, it was calculated that while the value of this would vary from image to image, the fourth coefficient value would remain relatively consistent. Therefore, for each frame, the absolute value (converting negative values to positive) of the fourth DCT coefficient was calculated. This was then compared to a moving average of up to the 10 previous frames that were considered

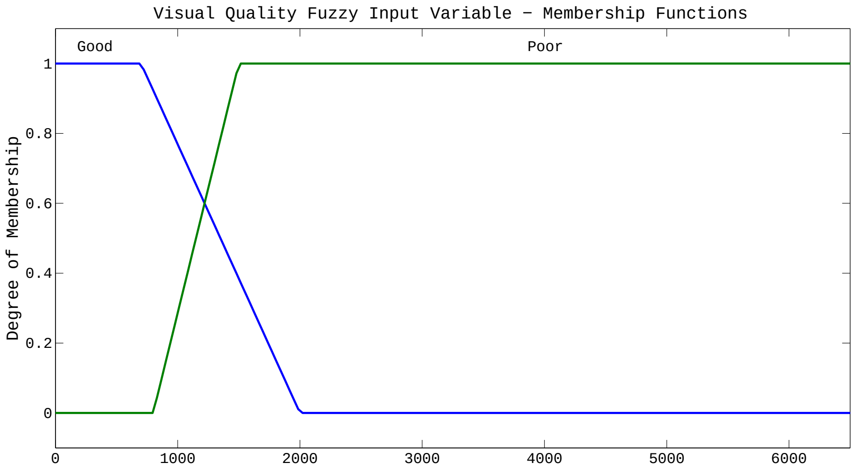


Figure 65: Switching logic input parameter: visual detail level. Depending on the level of visual detail, the estimated parameter can be considered to be 'Good' or 'Poor' to varying extents.

to also be of good quality, and the difference between this moving average and the coefficient represented the visual input variable.

To create this moving average, one assumption was made, that the first value of each sentence was successfully identified with the Viola-Jones detector (Viola & Jones [175]). This first value was used as the initial moving average mean value. For the second frame onwards, the new value was compared to the mean of the moving average. If the new value was considered to be within a threshold (preliminary trials identified an appropriate threshold to be 2000), then this value was considered to be suitable, and so was added to the moving average. To take account of variations in speech from frame-to-frame, only a maximum of the 10 most recent values were considered as part of the moving average. This moving average threshold aims to minimise incidences of incorrect results being added to the moving average.

Preliminary trials found that examples of poor quality visual information tended to result in a greater difference from the mean than good information, and so this approach was found to be suitable. The trapezoidal membership functions are shown in figure 65. Although the choice of membership functions can vary depending on the preference of the designer, it was felt that it would be suitable to use trapezoidal membership functions for all variables in order to ensure consistency.

Figure 65 shows that there are two membership functions, 'Good' and 'Poor'. The lower the input value, the closer to the mean and therefore the better the frame of visual data was considered to be. However, as values for individual speakers could vary depending on factors such as the size of the detected ROI and the degree of emotion in their speech (for example, affecting the size of mouth opening), there was no fixed value that was guaranteed to work for every speaker, and therefore a crisp set was not considered to be suitable. As a fuzzy membership function was used, it was considered that a visual quality value of less than 800 was definitely an example of a good frame of visual information. Between 800 and 2000, then it could be sometimes considered a partial member of the good set in that there was some ambiguity depending on the speaker, and also there were examples of partial frames (where only part of the ROI was accurate). This justified the decision to use fuzzy input variables.

Although alternative techniques could potentially be used, a detailed evaluation of the performance of this input variable (described in chapter 7) demonstrated the suitability of using this input as part of the system.

### *Audio Power Fuzzy Input Variable*

The second input variable to be used is the audio power level. This considers how much acoustic activity there is in an individual frame of speech. The membership functions are shown in figure 66. This variable does not consider the problem of voice activity detection, and so does not attempt to distinguish between speech and noise. One reason for this is that the system is designed to be tested in extremely noisy environments, and audio-only VAD techniques do not always perform well in these environments. As shown in the results in chapter 5, at an extremely low SNR, no speech at all can be identified in noisy input speech mixtures. It is possible to devise an audiovisual VAD (Almajai & Milner [11]), and this could represent future potential development, but it was felt that the most important factor with regard to the proof of concept system presented in this chapter was identifying the level of the audio input as in a real environment the level of noise does not remain consistent, and can change from frame-to-frame. In terms of the various conventional hearing-aid input detectors

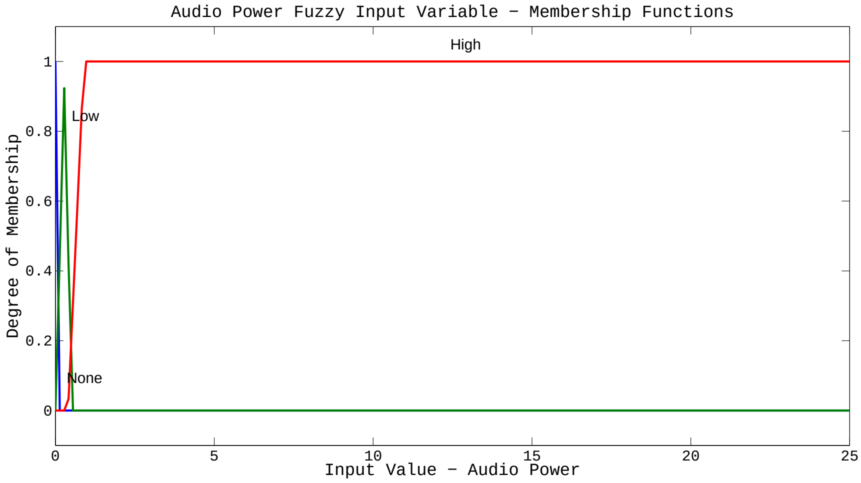


Figure 66: Switching logic input parameter: audio frame power. Depending on the level of audio power, the estimated parameter can be considered to be 'None', 'Low', or 'High'.

mentioned in chapter 3, this input variable functions in a similar manner to a level detector (Chung [44]), and is suitable for calculating the level in very noisy environments.

To calculate the audio power in each input speech frame, the frame is first converted back to the time domain to return the amplitude waveform for that frame of speech. The mean of the absolute values of the frame is then found. This represents the level of the audio power. The fuzzy set that the audio power input variable belongs to is then calculated based on this input, as shown in figure 66.

To take account of extremely noisy input variables, due to the extremely low SNR that the system is tested with, it can be seen from figure 66 that the largest trapezoidal membership function is the 'High' value, which has a maximum value of 25. Figure 67 shows the same membership functions, but shows only the fuzzy membership functions for values less than 1.5.

Figure 67 shows that if the level is recorded as being very low (less than 0.015), the input level is considered to belong to the 'None' membership function. However, as the level detector is very sensitive, it can be seen that any positive level (ranging from 0.009 to 0.5) is also part of the 'Low' fuzzy-set to an extent. Finally, any values greater than 0.4 were considered to be a member of the 'High' set to an extent, and values greater than 0.9 were considered to fully belong to the 'High' set. These values were set by using trial data, and in the evaluations

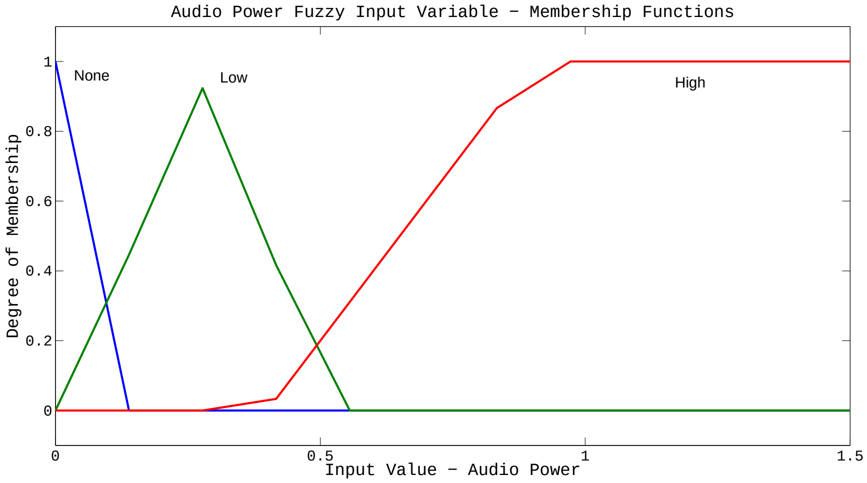


Figure 67: Switching logic input parameter: audio frame power, showing only membership functions for values ranging from 0 to 1.5. Depending on the level of audio power, the estimated parameter can be considered to be 'None', 'Low', or 'High'.

presented in chapter 7, the effect of different levels of audio power (represented by varying the type of noise and the SNR on the fuzzy output is discussed in section 7.7.7 of chapter 7.

*Previous Frame Fuzzy Input Variable*

The third input variable is the previous frame fuzzy logic output. This is simply a feedback variable that passes in the fuzzy logic controller output from the previous frame. As stated, there are three different processing options, and this can be seen in figure 68, which is valid for the representation of both the controller output and the third input. The reason for this third input is to act as a smoothing function in marginal cases. For example, the audio and visual inputs may produce input variables that lie near the thresholds between two possible processing options. Small changes in subsequent frames may produce a radically different processing decision from frame-to-frame. As a consequence, the output sound quality may be of poor listening comfort (as is sometimes found in conventional hearing aids when the engaging/adaption/attack configuration is set poorly, resulting in a 'choppy' sound, as discussed by Chung [44]). The use of the previous frame in marginal scenarios is designed to limit this. This input performs the role of engaging/adaption/attack configuration in this preliminary system, as it introduces what is effectively a small delay into processing changes. An evaluation of the role of this input variable on the output decision is discussed in chapter 7,

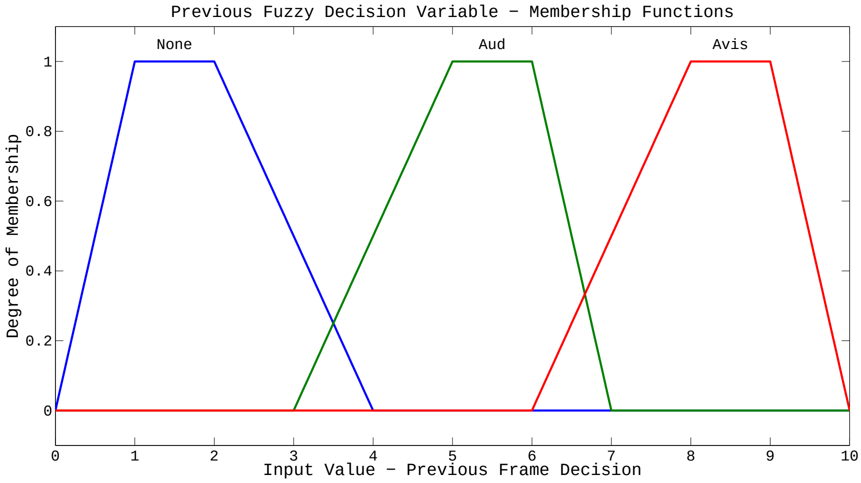


Figure 68: Switching logic input parameter: previous frame output. This input variable considers the processing method chosen in the previous frame. Therefore, this input fuzzy set diagram matches the output choice. This input is useful in marginal examples, as will be discussed later in this chapter.

demonstrating that this input variable plays an important role in limiting changes in the fuzzy output. In addition to this evaluation of using the previous frame, the use of a mean of several frames as part of an input variable was also considered, and the results of an evaluation of using a different number of frames as an input variable is presented in chapter 7. It was concluded from this evaluation that there was no noticeable improvement when using a mean of 3, 5, or 10 previous frames. Therefore, it was considered suitable to use the single previous output value as an input variable.

Figure 68 shows the three trapezoidal membership functions (again using this shape for reasons of consistency) representing the possible processing choices.

There are three membership functions, with each one corresponding to a processing decision 'None' (meaning to leave the frame unprocessed), 'Aud' (meaning to use audio-only beamforming), and 'Avis', meaning to use the audiovisual approach. These match the output decision fuzzy sets. For a sentence of audiovisual data, each frame is processed to extract the three input variables, these inputs are passed into the fuzzy logic controller, and fuzzy rules are then applied to produce the processing decision.

### 6.5.3 *Fuzzy Logic Based Switching Supervisor*

#### *Fuzzy Rules for the Switching Decision*

Fuzzy logic control can be used for intelligent switching as in, for example, the work of Tanaka *et al.* [170], who used fuzzy logic for stable control of a radio controlled hovercraft. In the preliminary framework presented in this chapter, the fuzzy logic controller is used to determine the most suitable speech processing method to apply to an individual frame of speech, based on the fuzzy input variables defined in the previous section. The rules were decided after preliminary experimentation and use the three input variables defined in the previous section, with the fuzzy sets as shown in figure 65, figure 66, and figure 68. As discussed in section 6.3, one difference between simple rules with crisp sets and fuzzy-based rules is that the rules are fired to varying degrees, depending on the extent to which the input variables are part of potentially overlapping membership functions. If an input variable is part of more than one membership function (for example, the audio level may be considered to be part of both the 'None' and 'Low' sets, then contrasting rules may be fired, with the strength of each rule depending on the extent to which the input variable is part of the relevant fuzzy set. As discussed in section 6.3, these rule outputs are aggregated to produce one fuzzy output set encompassing all of the rules that were fired. Finally, this is defuzzified (again, as described in section 6.3) to produce one single fuzzy output decision value. In this work, the centroid value was used (see figure 62 for an example of this).

The input variables to the fuzzy-system are described above and are the audio level (audSigPow), visual quality (visQuality), and the previous frame controller output (prevFrame). The input variables and the possible membership functions that the values may belong to are shown in table 17. An input variable may simultaneously belong to more than one fuzzy set to varying extents.

The processing output options are no processing (a), audio-only processing (b), or two-stage audiovisual processing (c). The complete set of rules used in this system is listed as follows:

- Rule 1: IF audioSigPow IS low AND visQuality IS poor THEN process is b

Table 17: Fuzzy input variables and possible membership functions.

| Input Variable | Potential Membership Functions | | |
|---|---|---|---|
| visQuality | Good | Poor | |
| audSigPow | None | Low | High |
| prevFrame | None | Aud | Avis |

- Rule 2: IF audioSigPow IS none AND visQuality IS poor THEN process is a
- Rule 3: IF audioSigPow IS high AND visQuality IS good THEN process is c
- Rule 4: IF audioSigPow IS none THEN process is a
- Rule 5: IF audioSigPow IS low AND visQuality IS Good AND prevFrame IS avis THEN process is c
- Rule 6: IF audioSigPow IS low AND visQuality IS Good AND prevFrame IS aud THEN process is b

Chapter 7 presents an evaluation of the system, and in particular, section 7.7.7 focuses on an investigation of the effect of input variables on the fuzzy output decision. This evaluation found that the system functioned as expected. Rule 1 activates audio-only processing if the audio input variable belongs to the 'Low' fuzzy set and the visual quality is defined as being 'Poor'. Rules 2 and 4 ensure that the frame is left unfiltered if the audio level is found to be so low that the audio level is defined as being 'None'. Rule 3 activates audiovisual processing if there is a sufficient level of noise, and if visual information of an adequate quality is available. Rules 5 and 6 are designed to take effect in scenarios where the potential choice of processing algorithm is ambiguous. If the audio level is defined as 'Low', but 'Good' quality visual information is available, then the previous frame input is also considered. Rule 5 activates audiovisual processing if the previous frame output was also audiovisual, and rule 6 activates audio-only processing if the previous frame decision was audio-only. This is intended to ensure continuity between frames and prevent rapid frame-by-frame changes that act as an irritant to listeners.

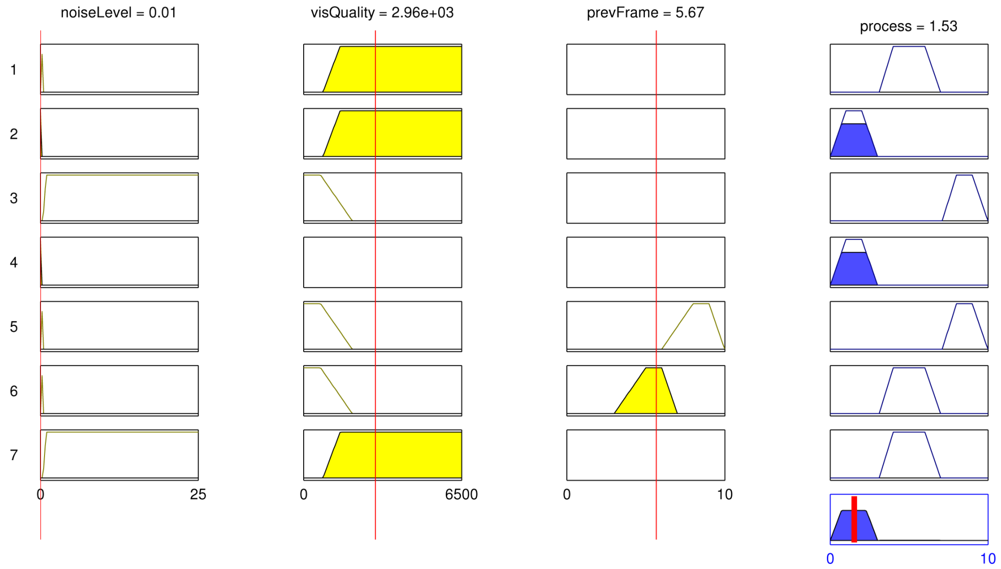


Figure 69: Demonstration of rule selection for an example where the audio level is very low, resulting in a decision to use no speech filtering.

*Fuzzy Inference Procedures for Switching Logic*

This section provides some examples of the theoretical operation of the rules defined above. At each frame of speech, the three input variables are calculated and input to the fuzzy rules. The first example is shown in figure 69. This is a scenario where the audio frame power is recorded as being very low (0.01), the visual quality is defined as being 'Poor' (2960), and the previous frame value is set to 5.67. In this case, rules 2 and 4 are fired, indicated that no processing is to be used. The second example presented has an audio power of 0.45, a visual quality value of 951, and a previous frame value of 1.68. These parameters are a good example of fuzzy logic in action because both rule 1 and 3 are fired. As can be seen in figure 70, the audio power falls within the overlap area for both high and low fuzzy sets, and the visual information quality is also applicable to both 'Good' and 'Poor' membership functions. Despite both rules firing, figure 70 shows that rule 1 is dominant, and so the defuzzified output indicates that audio-only processing should be used. The third example, shown in figure 71, is much more clear-cut. With an audio level of 4.06 and a visual quality value of 156, only rule 3 is fired, indicating that audiovisual processing should be used with this hypothetical speech frame.

The final two examples demonstrate the purpose of the previous frame input variable. Both examples, shown in figure 72 and figure 73 have an audio input variable which is defined as

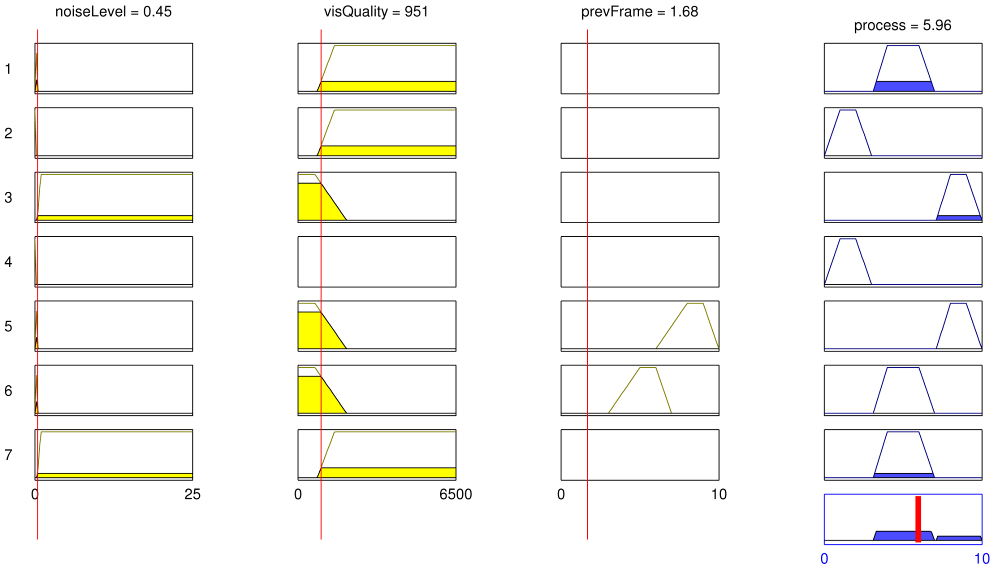


Figure 70: Demonstration of rule selection for an example where the audio level is estimated to be at a moderate level with a low quality of visual information, with the outcome that the system selects audio-only speech processing.

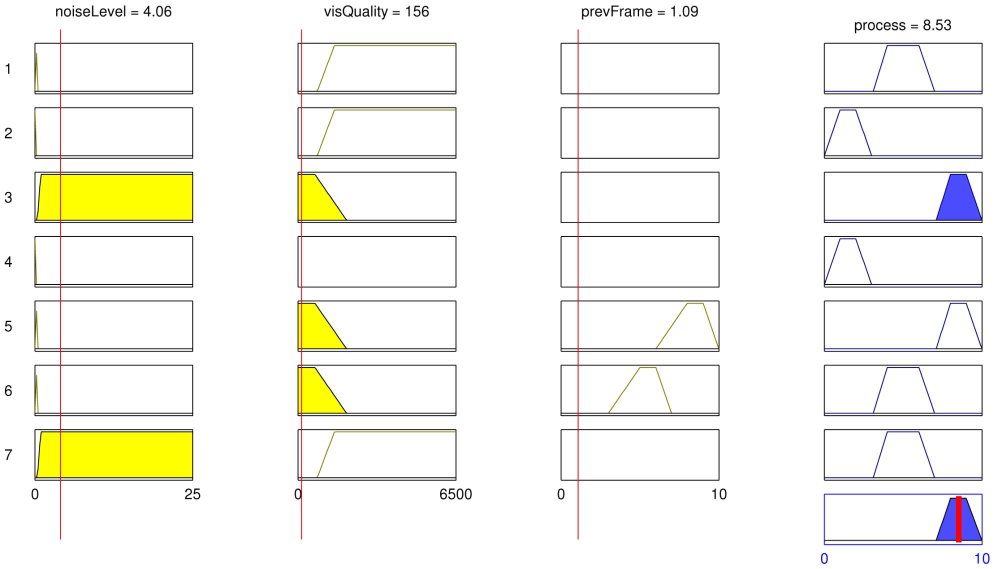


Figure 71: Demonstration of rule selection for an example where the audio level is very high, and good quality visual information is available, resulting in the use of audiovisual processing.

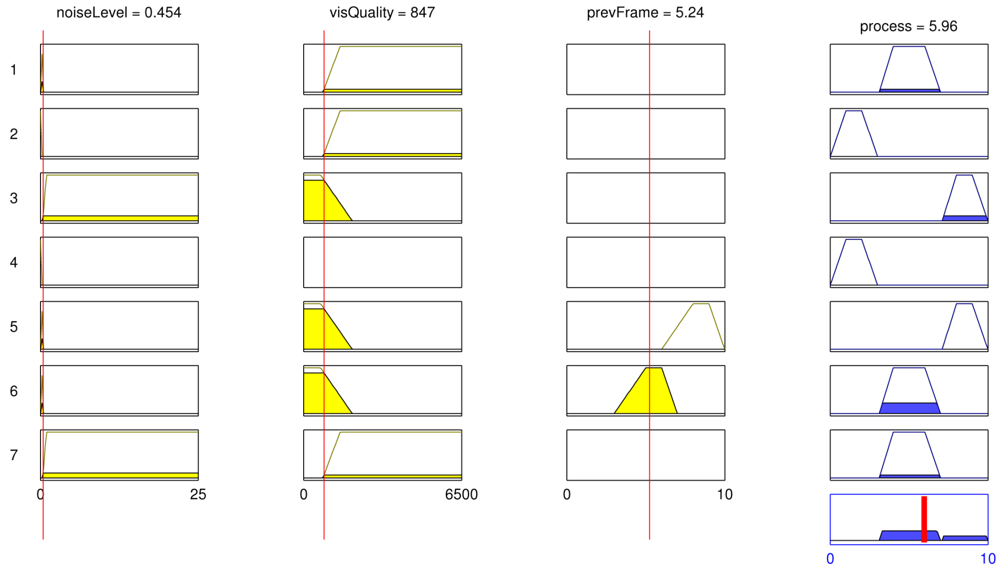


Figure 72: Demonstration of rule selection for a marginal example. Good quality visual information is available, and the audio level is measured as being applicable to both moderate and high levels. In this case, the previous frame output information is used, which in this example was audio-only processing, and so the audio-only option is chosen again.

being 0.454. This is a value which belongs to two potential fuzzy sets. The visual quality level is set to 847 in both examples, meaning that this visual information is considered to be a member of both the 'Good' and 'Poor' membership functions. As the examples show, both rules 1 and 3 are fired, and both have a similar level of dominance when it comes to establishing the output. These values mean that small changes in the audio variable in successive frames may result in an entirely different type of processing being used from frame-to-frame, despite there being potentially only a very small change in environmental conditions. This is undesirable because to the listener, rapid and unnecessary switching between processing methods can often result in an unpleasant listening experience. As can be seen in figure 72, although rules 1 and 3 are fired, rule 6 is also fired. Rule 6 considers the previous frame, and as this is defined in this example as using audio-only processing, this becomes dominant, and so when defuzzification takes place, audio-only processing is chosen. Figure 73 uses the same audio and visual variables, but with the previous frame input variable being defined as audiovisual. In this case, the audiovisual technique is chosen by the fuzzy logic controller.

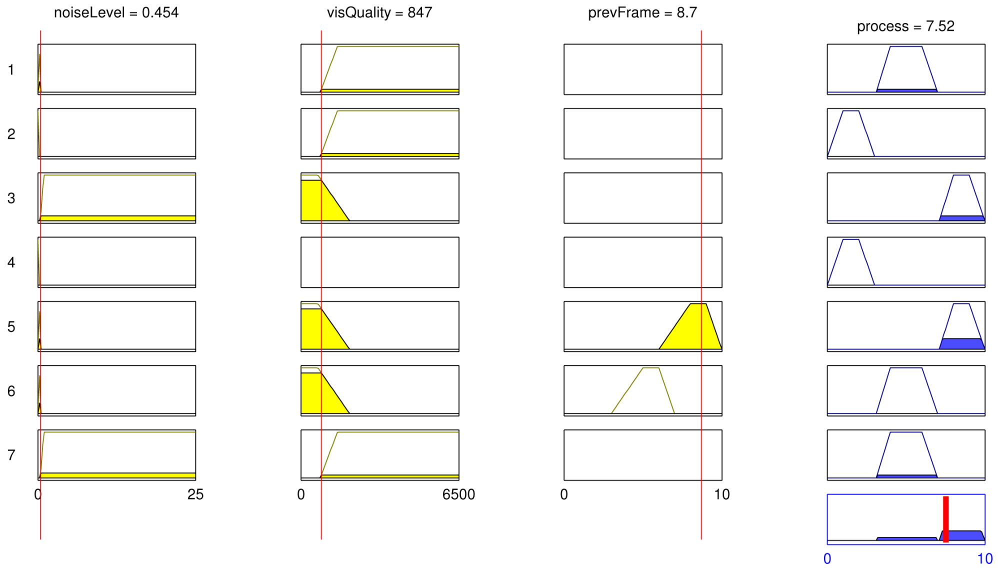


Figure 73: Demonstration of rule selection for a marginal example. Good quality visual information is available, and the audio level is measured as being applicable to both moderate and high levels. In this case, the previous frame output information is used, which in this example was audiovisual processing, and so the audiovisual option is chosen again in this frame.

## 6.6 EVALUATION APPROACH

In order to evaluate the capability and potential of the system presented in this chapter, an evaluation needs to be carried out with challenging real world data. Any evaluation should use scenarios that are as natural as possible. By this, it is meant that the test data should use data with varying noise levels, with speakers who are moving, and with data not trained previously with the system. So for example, the quality of visual data must vary, representing the speaker turning their head, moving quickly, or changes in light or visibility. With regard to audio information, in a real world scenario, the noise level is not always consistent and so the use of the consistent aircraft cockpit noise, as used in chapter 5, is not sufficient. There is also the Lombard Effect to consider (Lombard [120]), when the vocalisation of a speaker changes in response to environmental conditions.

Because of these issues, it could be argued that the corpora used in this thesis, GRID (Cooke *et al.* [46]) and VidTimit (Sanderson [157]) are not suitable for further testing. Both are recorded in a clean visual environment without distractions. Speakers from both corpora speak only single sentences, without moving their head to any great degree, and there are no frames where the visual information is of poor quality or unavailable. Therefore, the corpora would need to be artificially edited in order to be useful for testing the fuzzy logic system presented in this chapter.

Another issue with the corpora is with the audio data. The GRID corpus has good quality audio recording available for each sentence, with no background noise. While there is a small level of noise present in the background of the VidTimit recordings, this is at a consistent low level. While noise has been added during previous testing, as shown in the simulated room environment in chapter 5, this is artificial, and means that the speaker has not adjusted their speech and mannerisms to take account of noise in accordance with the Lombard Effect. The sentences are also very short, and do not take account of scenarios such as longer discussions and turn taking. To successfully evaluate the system, speech recorded in more natural conditions is required. This will be discussed in more depth in the next chapter.

## 6.7 SUMMARY

The goal of this thesis is to present a multimodal two-stage speech enhancement framework that works towards being autonomous, adaptive, and context aware. After an investigation of state-of-the-art research in chapter 3, an initial audiovisual speech enhancement system was presented in chapter 4 that filtered noisy speech in very difficult environments. Although results were positive, a thorough evaluation of the system identified a number of limitations with the system, such as distortion introduced at a high SNR. This, along with more general weaknesses with speech enhancement systems, suggested that there was still scope for improvement.

This chapter extended the system first presented in chapter 4 by integrating a fuzzy logic controller into the system and providing alternative processing options for each frame of speech, meaning that depending on the audio and visual input data, the fuzzy logic controller will determine whether to process each individual speech frame using the full audiovisual two-stage approach presented previously, audio-only beamforming, or to leave the frame entirely unfiltered. This chapter first discussed the limitations of the two-stage system current implemented in section 6.2, as well as more general limitations with speech enhancement systems, and justified the use of a fuzzy logic controller to overcome these limitations in section 6.3. Some alternative approaches were discussed in section 6.4, concluding that although there is potential for making use of alternative approaches such as neural networks or HMMs, it was considered that making use of a fuzzy logic system represents a solution that is potentially easier to implement and directly use in future hearing aid technology and so was chosen for use in the framework presented in this chapter. The preliminary framework presented in this chapter was then described, specifying the inputs used, the associated fuzzy membership functions, and the rules for deciding the appropriate filtering option to use. It was emphasised that this system is a proof of concept, with much scope for further development, and so does not represent a finalised system.

To evaluate the system, while limited experimental scenarios with artificial adjustments made to pre-recorded corpora are of interest, this does not adequately simulate the conditions that a real world fuzzy logic based system would be expected to process, and where noise has an effect on both audio and visual modalities. The next chapter describes an evaluation of the system, using more natural recorded data.

# 7

# EVALUATION OF FUZZY LOGIC PROOF OF CONCEPT

## 7.1 INTRODUCTION

This thesis presented a novel two-stage speech enhancement system in chapter 4, which was then thoroughly tested in chapter 5. As a result of these tests, although promising results were found, some limitations with both this system and speech enhancement systems generally were identified, including limitations with the visually derived filtering utilised in this thesis. This resulted in the development of a proposed fuzzy logic based two-stage speech processing system that uses fuzzy input variables to determine the most appropriate method of processing an individual frame of speech. This preliminary system was described in depth in chapter 6, along with a review of other alternatives to using fuzzy logic, including machine learning techniques such as ANNs and HMMs.

This chapter presents an evaluation of this preliminary system, firstly discussing the need for testing, and the requirements of for testing a system that is designed to deal with more realistic speech processing scenarios. However, there are a number of limitations with the system in its current implementation that prevent a full testing programme from being successfully concluded, including limitations with the system in its current state, issues with the individual speech processing techniques, as discussed in chapter 5, and also hardware limitations. To present a full description of the limitations with the system in its current preliminary state, and to better meet the requirements of more realistic test scenarios, a new corpus is used, recorded specifically to provide challenging audiovisual data. This corpus is then used as part of a series of challenging experiments in order to evaluate the performance of this system.

The remainder of this chapter is divided as follows. Section 7.2 provides a brief reminder of the fuzzy logic system. This is then followed by a description of the requirements for testing this system in a more real world environment in section 7.3, and in section 7.4, the difficulties with applying these tests to the proposed system are outlined. The next section describes the recording of a novel corpus with more challenging audiovisual speech data. This corpus is then used to test the fuzzy input variables used in the system in section 7.6, followed by noisy speech processing tests in section 7.7. The results of the testing process are discussed in section 7.8, which is followed by the summary of the chapter in section 7.9.

## 7.2 system overview

The fuzzy logic system presented in chapter 6 is evaluated in this chapter. This system is similar to that presented in chapter 4, with automatic lip-tracking, audio-only beamforming and two-stage audiovisual speech processing. However, as discussed in chapter 6, the key difference is that the initial system was extended to incorporate a fuzzy logic controller. This controller was intended to add flexibility to the initial system by evaluating and adjusting the chosen speech filtering decision on a frame-by-frame basis, depending on the values of the fuzzy input variables.

The quality of each frame of input data was determined by making use of three fuzzy input variables. There were an audio level detector, a visual quality indicator, and the previous fuzzy output decision. These were used to activate a number of rules to different extents, depending on the fuzzy input values, which were then aggregated and defuzzified, as described in chapter 6. This produced a single output value between 0 and 10. A value of less than 4 resulted in the frame being left unprocessed. If the value was greater than 7, the frame was processed using the two-stage audiovisual approach, and a value between 4 and 7 meant the audio-only approach was used without visual information being utilised. A full description of this system is given in chapter 6.

## 7.3 testing requirements

To test the system presented in chapter 6, it is not possible to simply use sentence from the GRID corpus as utilised in chapter 5. This corpus does not contain the suitable variation in audio and visual data quality that is required to fully test a fuzzy logic based system, and this limitation is shared by many other publicly available corpora (such as VidTIMIT). Therefore, there are two possible approaches to test this initial system, limited experimental scenarios involving artificial modification of sentences from existing corpora, or recording new audiovisual speech data that represents more realistic scenarios.

### 7.3.1 *Limited Experimental Scenarios*

One option for testing the fuzzy logic based system is to use limited experimental scenarios. These scenarios could utilise the existing corpora used previously in this thesis (such as the VidTIMIT (Sanderson [157]) and GRID (Cooke *et al.* [46]) audiovisual speech databases), but would require artificial modification in order to make them more suitable for testing the fuzzy logic based system. So for example, the audio modality may require artificial noise to be added, and the visual data may need individual cropped frames altered in order to simulate the effects of data variation such as the user turning their head, or changes in light conditions.

Using such an approach has the advantage that it uses existing corpora, so can be compared to other results. It also allows more precise control over environmental conditions. Also, if the corpus has been used previously to train the system, then as shown in chapter 5, the audiovisual filtering performs better with new sentences from the same corpus, and so it allows a focus more on the testing of the fuzzy-based selection, without the issue of poor results due to poor audiovisual filtering performance.

However, using limited artificial scenarios severely limits the extent of testing. Simply editing frames is not fully representative of the conditions a tracker will be expected to process, and short artificial sentences without interruption are not representative of data that a system is likely to be expected to process successfully, where there may be examples of turn-taking, long monologues, interruptions, emotional speech, and the Lombard Effect. Therefore, more realistic speech data is required.

### 7.3.2 *Realistic Speech Data*

By realistic speech data, it is meant that the audiovisual data should be a closer match to scenarios that a finalised hardware implementation of this system would be expected to successfully process. This means that the audio data should ideally contain noise of different levels, with the speaker then adjusting their speech correspondingly as they would in a real situation (as explained by the Lombard Effect). Conversational scenarios should also be of longer length than simple two or three second sentences, with possible examples of overlapping speakers, silences, turn-taking, and emotional speech. The visual data should also be of variable quality, with examples such as speakers who have moved their head away from the camera, put their hands over their mouth, or are speaking in a more natural manner.

The key benefit of using more realistic data is that it can be used as part of an extremely rigorous analysis of a speech processing system. It represents data which a hypothetical fuzzy logic based system would be expected to process, and this makes it easier to identify limitations with the system.

However, given the current preliminary implementation of the system (as will be discussed in more depth in section 7.4), it is extremely challenging to fully evaluate the system with more realistic speech data. With the fuzzy logic based system presented in this chapter still in a preliminary stage of development, the output filtered speech results using real data should be interpreted with a degree of caution. Some prior examples of audiovisual data tests in the literature include work by Almajai & Milner [11], who train and test with the same single speaker corpus using consistent broadband noise at a relatively high SNR, and audiovisual source separation by Naqvi *et al.* [133]. Although Naqvi *et al.* [133] make use of more realistic data, they make a number of assumptions regarding room size and availability of visual information (i.e. good quality visual information is always available). The prior work in the literature demonstrates that audiovisual speech filtering systems at an early stage of development are often tested with fairly limited parameters. Using more realistic speech data also makes it more challenging to present a precise hypothesis of the exact processing option the system is expected to choose from frame-to-frame, which can be useful for observing fuzzy process switching performance. Overall, the use of realistic speech data is recommended when possible, but it is more suited to a system in a more advanced state of development than the preliminary system presented in this chapter.

## 7.4 experimentation limitations

In order to test the system with more realistic speech data, the recording of a suitable novel corpus was required. This was anticipated to be an extremely challenging process, and so the potential issues with this were investigated. An initial test scenario was attempted to record data in a fully real environment. To achieve this, a simple scenario was devised. This involved a volunteer speaking a number of sentences in a quiet room. Two microphones placed in slightly different locations were used to record audio data, and a webcam connected to a pc via usb was used for the equivalent video data. The intention of this test was to establish the level of difficulty involved with recording real data. In this early trial, additional noise sources were not considered, which would be required to fully simulate real noise from different locations. Hardware limitations (in this case, the quality of audio equipment available for use) prevented this from being fully tested. The recorded audio and visual data was then processed. A number of hardware and software issues were identified with this process.

Firstly, as mentioned above, there were some issues with the availability of equipment available for use during this research project. In order to simulate the discrete noise sources with real data required for beamforming to be successfully carried out, a number of output devices capable of producing extremely loud noises were required, but were unavailable. This meant that any noise that the beamformer could potentially remove had to be added in the simulated convolved room scenario described in chapter 4. Using the simulated room software to add noise meant that the impulse responses could easily and accurately be calculated, which, as described in chapter 4, is a requirement for beamforming to be successfully performed. Recording of noisy data by adding a noise source was attempted, but as expected, the system was not able to successfully and reliably process this data, and so this data was not used in the final implementation of the system. Another related issue concerned data synchronisation. Because multiple microphones were required, the data was required to be synchronised for accuracy. As the microphones and camera were recorded individually, the separate audio and visual streams were unsynchronised, and although the tracks could be synchronised by hand (matching the audio and visual data by inspection and adjustment), it proved to be an extremely time consuming task. Some other potential issues with the system in its current early stage of implementation were also identified. Firstly, as the system is currently implemented in MATLAB and is not capable of functioning in real time, data had to be loaded into memory to be processed, which was particularly resource intensive when processing large video files. Therefore, smaller individual speech clips were used for the final speech dataset (which will be described in section 7.5).

There were also some limitations identified with the initial system presented in chapter 4 and tested in chapter 5. Firstly, as discussed previously, there are limitations with the visually derived filtering. Although positive results were found, it was also confirmed that when the system was tested with novel speakers from different audiovisual corpora, it did not generalise well, and so when tested with real data, it is expected to perform poorly, based on the results discussed in chapter 5. Another issue is that the simulated room environment (with mixtures of speech and noise sources) was designed explicitly to demonstrate the effectiveness of beamforming, and so tests using this scenario provide an artificial benefit to using beamforming that would not be expected in a real environment (a similar issue to the improvement in results with using directional microphones in laboratory environments rather than in real environments, as discussed in chapter 3). Overall, it is understandable why other similar speech enhancement work such as by Almajai & Milner [10] and Rivet *et al.* [148] makes use of very limited experimental scenarios.

The system in its current implementation is an extension of the thoroughly tested two-stage filtering approach discussed in chapter 5. Like other early-stage audiovisual speech enhancement work, such as by Almajai & Milner [12], it is currently in a very preliminary stage of implementation. Therefore, any evaluation with real data will provide limited results. To demonstrate this, section 7.7 presents a series of evaluations of this system with newly recorded challenging audiovisual speech data.

## 7.5 recording of challenging audiovisual speech corpus

In order to demonstrate the performance of the fuzzy logic based system presented in chapter 5, it was considered a requirement to demonstrate the effect of making use of more challenging real world speech data. For this, it was considered necessary to record novel data, as none of the corpora reviewed in chapter 3 were considered to be suitable for this purpose.

### 7.5.1 *Requirements of Corpus*

As discussed in section 7.3, the corpus used previously in this thesis for evaluation were not considered to be entirely suitable, due to limitations in content and variation of quality. Therefore, it was felt that it was appropriate to record some real data in order to evaluate the preliminary system presented in chapter 6. The recording of new audiovisual speech data means that several of the limitations present in the existing corpora could be avoided. Firstly, the requirement of the corpus to have variable, yet natural visual data could be met. Rather than artificially replacing single frames, the speaker can be allowed to move in a more natural manner, performing actions such as turning their head and placing their hand over their mouth.

This requirement for natural data can also be extended to the audio aspect. As discussed in chapter 2, the Lombard Effect means that people change their style of speech in the presence of noise, which is not reflected when noise is artificially added to the corpus afterwards.

Recording of a natural corpus potentially allows for this effect to be taken into consideration. There are also real world events such as pauses for turn-taking, multiple speakers, and emotional speech that can be taken account of.

### 7.5.2 *Corpus Configuration*

*Scenarios*

In order to provide a diverse range of audiovisual speech data, and to provide challenging data that the pre-existing corpora used previously in this thesis (GRID and VidTIMIT) fail to supply, volunteers were asked to perform two tasks. Firstly, a reading task, where they read either a short story or a news article. For this task, they were recorded reading for a minute in a quiet environment, and then a minute in an environment with a variable level of noise (a mix of music tracks, with the volume varied randomly). This allowed for both good quality audio data, and also poorer quality raw data to be collected (with the Lombard Effect having an impact on resulting visual data). As this was a continuous reading task, the speech data gathered from each speaker was longer than the approximately three second clips provided in the GRID and VidTIMIT corpora.

However, it was found that using the noisy recorded speech data presented problems, in that the results were found to vary wildly, and the beamformer did not remove noise when (as expected), therefore these recordings were not used. The second scenario was a conversational task, where volunteers were encouraged to speak in a more natural manner. Volunteers were recorded in pairs at a table facing each other, with one speaker recorded at a time. By this it is meant that while the speakers were facing each other and making conversation, the camera was only pointed at one speaker. This allowed more natural and relaxed speech, and the volunteers were also told that they were allowed to move freely and did not have to look directly into the camera at all times. This allowed for more challenging noisy data such as head turning, speakers placing their hands over their mouths, and blurring in individual frames due to motion. Volunteers were given a choice of topics to choose from, such as a TV programme, something that interested them in the media, or could choose their own conversation topic. This resulted in a wide range of conversations, from gossip about friends, to recent events in the media. Due to this being a conversation rather than continuous speech from a single recorded speaker, there were occasional silences, or speech from the other participant in the conversation. This provided challenging data which the system has not been trained with. Again, similarly to the last scenario, each speaker was asked to speak for one minute in a quiet environment, and one minute in a noisy environment, although the noisy data was subsequently not used.

Both scenarios provided extremely challenging data. They used novel speakers, and contained considerably different sentences from those that the system had been trained with, both in length and in content. The resulting visual data was of variable quality, with examples of turning and movement, as well as varying audio quality due to noise. Overall, it was considered that this new audiovisual speech dataset represented extremely challenging data for a speech filtering system to process successfully.

### *Equipment and Recording*

To record volunteers carrying out the tasks described above, a single camera was used with an integrated microphone. A Microsoft VX2000 Lifecam was used to record speech in a quiet room, without any noticeable background noise. Both audio and visual streams were recorded with this single camera as this ensured that there were no issues with synchronisation of data. The visual data was recorded at a resolution of 640 x 480. As discussed in the previous section, there were two scenarios that volunteers were asked to record, and the recording took place in pairs to encourage more natural and relaxed speech. Each speaker was asked to read for one minute in a quiet environment, and then one minute in an environment with variable background music. The second scenario was a conversation task where speakers were explicitly told that they did not have to remain still or look directly at the camera. Again, they were recorded for one minute without noise, and one minute with noise. This meant that for each speaker, there were four minutes of initial raw data theoretically available, although as discussed, subsequent trials identified that the noisy data was unsuitable for use, so in reality, only two minutes were available.

However, there were some issues with the recording process. Firstly, the video camera had automatic brightness adjustment enabled, and so a small number of frames were considerably darker due to occasional automatic readjustment. An example of this can be seen in the lower image in figure 74. There were also a number of glitches in the recording that were discovered afterwards during the review of the data. An example of this can be seen in the top image in figure 74. In this image, the camera has not recorded the head of the speaker in a single frame, although in subsequent and preceding frames, the head is not missing. One other issue was that the recording did not function correctly for one speaker, with some synchronisation issues between audio and visual data. For this reason, there is limited data available from one pair of volunteers.

### *Finalised Corpus Description*

The final corpus contained data from eight speakers, four male, four female. Six of the eight speakers spoke English (five with a Scottish accent and one English), and two were recorded speaking Bulgarian. For each speaker, four minutes of raw data were theoretically available, one minute of quiet conversation, one minute of variable noisy conversation, and then one minute each of noisy and quiet reading. Some example frames of the recorded volunteers are shown in figure 75.

As discussed previously, there were some issues with the recording in the form of glitches and light adjustment, as shown in figure 74. Also, as part of the requirement for the visual data to be challenging and of variable quality, speakers were expected to move naturally. This led to variable quality visual data, with some examples shown in figure 76. The top image shows an example of the speaker in the process of moving their hand in front of their mouth, meaning that lip information is not available, and the ROI therefore cannot be correctly identified. The lower image shows an example of the speaker turning their head to one side. This is a challenge to the tracker, and is also an example of data that was not used in the

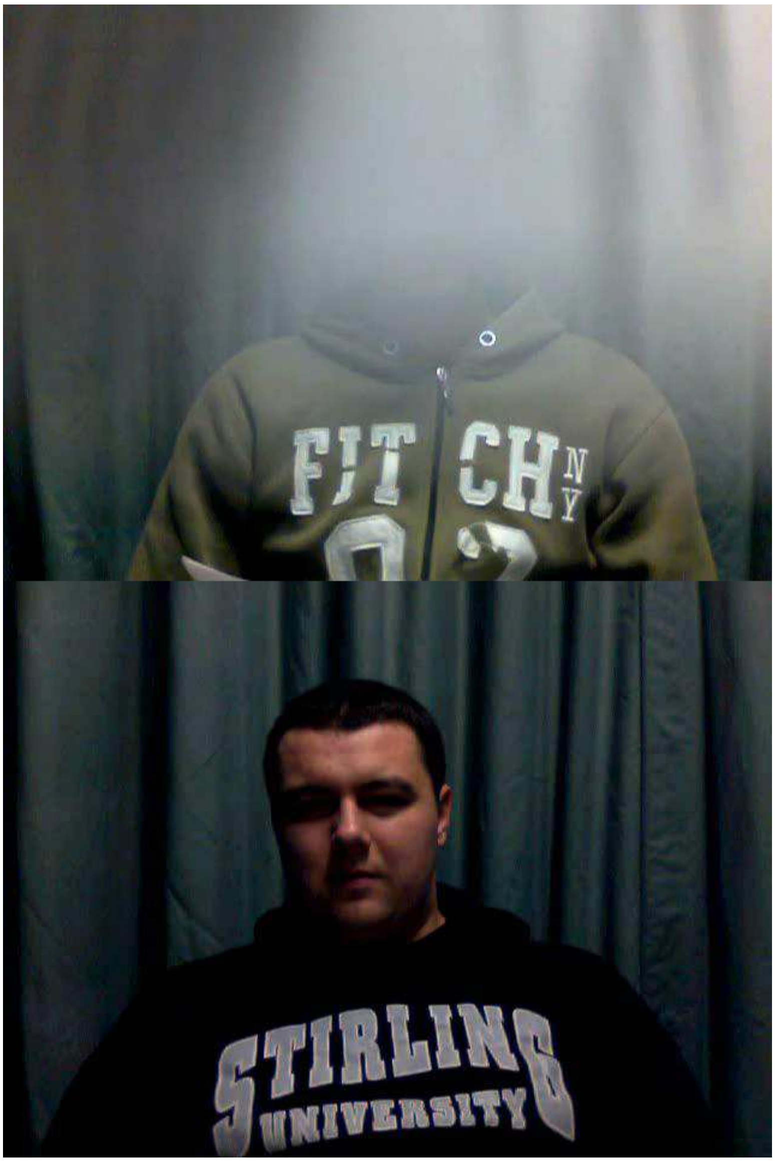


Figure 74: Examples of poor quality visual data due to issues with recording. The top image shows an example of a glitch during recording, resulting in the face region being removed. The bottom image shows a situation where light conditions have changed, resulting in a temporarily darker image.

Figure 75: Speakers from recorded corpus, using sample frames taken from videos.

initial training of the visually derived filtering process. There is also blurring present in this image due to movement.

Of the initial 32 minutes of raw data, approximately 6 minutes was unavailable due to the recording problems described earlier. The data was divided into 20 second clips because of processing and testing requirements. This sentence length was significantly longer than available in the pre-recorded corpora, and was felt to be long enough to test the operation of the fuzzy-system, while still being short enough to process relatively efficiently. A number of these 20 second clips were then chosen for use as part of the testing process. These were chosen to represent a mixture of different conditions and data quality. As mentioned previously, only the sentences without noise were used.

## 7.6 fuzzy input variable evaluation

The previous chapter presented a description of a fuzzy logic based speech processing system, which made use of three fuzzy input variables, the audio power within a frame, visual data quality, and also the output decision of the previous frame, which is fed back in as an input. While the audio input variable is very closely related to the audio signal, the effectiveness of the other two input variables is of great interest. This section presents an evaluation of the suitability and performance of these variables, and justifies the decision to use them as part of the fuzzy-based system.

### 7.6.1 *Visual Quality Fuzzy Indicator*

*Problem Description*

As described in section 6.5, one input variable used in the system was the visual quality variable. There was an assumption made that the initial frame was accurately detected, and subsequent frames were calculated in terms of the difference from the mean of the absolute value of the fourth DCT coefficient. To take account of natural movement over time, a moving

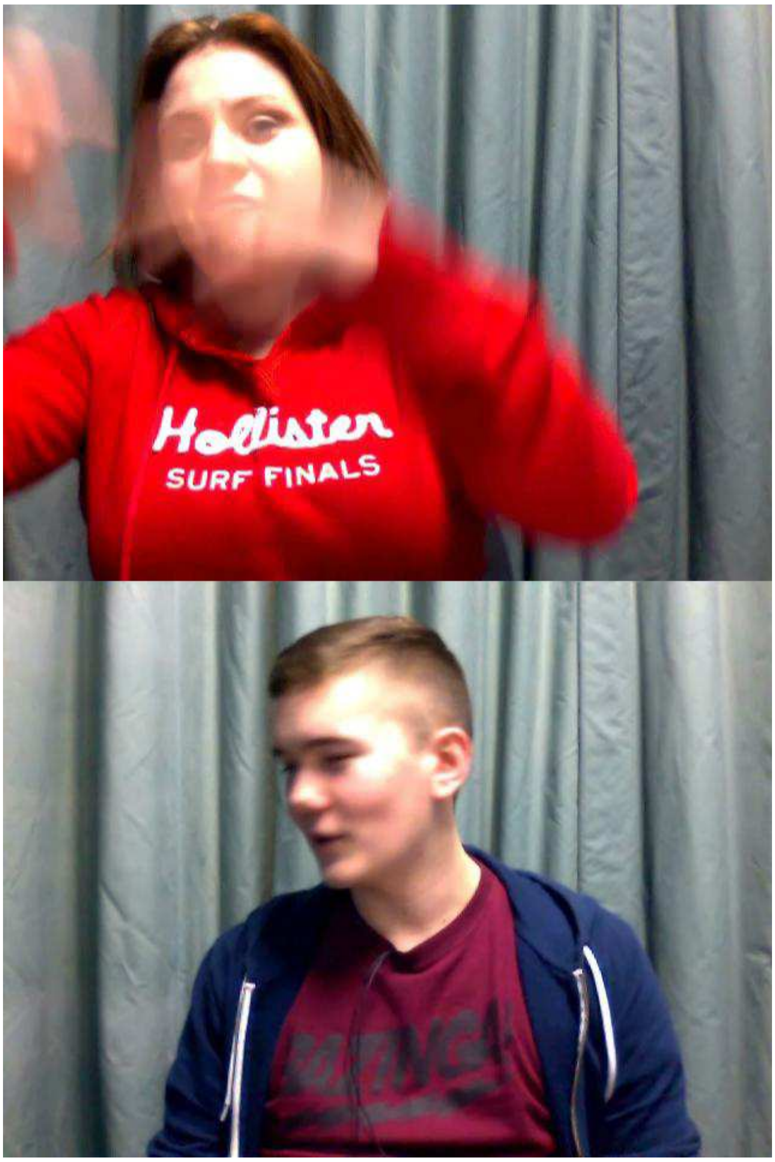


Figure 76: Examples of poor quality visual data due to speaker actions. The top image shows a frame where the speaker has their hand over their mouth due to gesturing during emotional speech. The bottom speaker is in the process of quickly turning their head, and as a result the mouth region is partially obscured, and the face is blurred.

average of the previous 10 frames was used, with only frames that were considered to be within a threshold added to the moving average. This value was then used as the visual fuzzy input value.

The assumption made was that if the absolute value of the fourth coefficient was similar to the mean, then the lip image was likely to be very similar, and therefore a good quality image. It was then calculated that the higher the difference from the mean, the less likely the image was to be a good quality lip image. This section will evaluate the performance of this approach by comparing the generated fuzzy input values to a manual estimation of the lip image quality, and evaluating the accuracy of this approach with a variety of different input values.

### *Experiment Setup*

20 sentences from the corpus described in section 7.5 were used for evaluation. This included 10 sentences recorded in a quiet environment, and 10 recorded in an environment with some noise present. In addition, to ensure that a range of different visual challenges was represented, 10 reading examples, and 10 conversation examples were used, from a number of different speakers. This ensured that challenging data was used and provided a rigorous test of this fuzzy input variable.

For each sentence, a manual review of each cropped lip image was performed. This involved inspecting each frame and assigning it a value. A frame that was considered to be of good quality (in that it showed the whole lip-region) was given a score of 1. An image that was considered to be of lower quality (either showing only part of the lip-region or the wrong region) was given a score of 2. Finally, an extremely poor result (one where no ROI at all was identified) was given a score of 3. This was then compared to the fuzzy input variable.

### *Evaluation Approach*

The manual input estimation of every frame of each sentence was compared to the equivalent fuzzy input variable. As the variable can vary in value between 0 and 6000+, with a lower value indicating less difference from the mean, then based on preliminary trials, a value of less

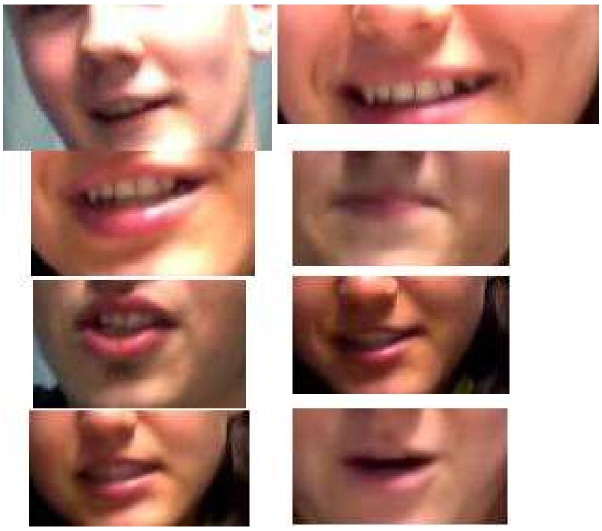


Figure 77: Examples of lip images regarded to be successfully detected. It can be seen that the images are of varying dimensionality, and also include varying levels of additional facial detail depending on the results of the Viola-Jones lip detecter.

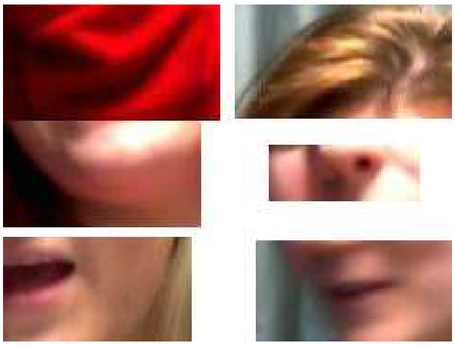


Figure 78: Examples of lip images regarded to be unsuccessfully detected. It can be seen that the images are of varying dimensionality, with issues such as identifying the wrong area of an image as the ROI, tracking only part of the lip-region, or poor quality information due to blurring and head motion.

than 1000 was given a score of 1 (some examples of this are shown in figure 77), a value of less than 4500 but greater than 1000 was given a score of 2 (as shown by the examples in figure 78), and anything greater than 4500 was given a score of 3, representing examples where no ROI was identified, as shown in figure 79 . These values represented similar boundaries to those used in the relevant fuzzy set. This allowed the visual input variable output to be mapped to the manual estimation.

For each sentence, to ensure consistency, the interpolated number of frames was used for comparison, and the fuzzy score was compared to the manually estimated value. The difference between the estimation score and the actual score was then calculated.

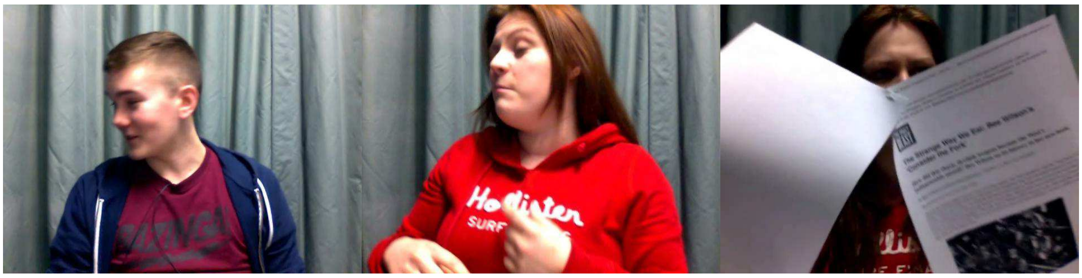


Figure 79: Examples of lip images where noROI was identified and cropping was not successful. It can be seen that this is due to the speaker turning their head or obscuring their face.

Table 18: Overall perfomance of visual quality fuzzy input variable compared to manual scoring, considering each frame of all 20 speech sentences.

| | Number of Frames | Percentage |
|---|---|---|
| Correct | 36836 | 92.15% |
| Incorrect | 3139 | 07.85% |
| Total | 39975 | 100% |

*Summary of Results*

Firstly, when taking 20 all sentences into account (whether recorded in a quiet or noisy environment, or as part of a reading or conversation task), after interpolation there were a total of 39975 frames of data. Of these, 92.15% produced a correct result (one where the fuzzy manual score matched), and 7.85% produced what was considered to be an incorrect result, as shown in table 18. Taking into account that 10 of the 20 sentences consisted of active conversation, this was a considered to be a good overall result.

To analyse the results in more detail, a comparison of the number of frames assigned each score is shown in table 19. This table shows the number and percentage of frames assigned each score both manually and using the fuzzy input variable. When observing the manually categorised frame scores, 90.89% were considered to be good frames, 7.93% were considered to be incorrectly assigned frames, and 1.18% of frames were considered to have identified no correct ROI. In comparison, the estimated fuzzy scores were calculated slightly differently. 94.51% of frames were considered to be good frames, 4.38% were estimated to be incorrect, and 1.12% were considered to have identified no correct ROI.

Table 19: Comparison of assigned values for overall 20 sentence dataset, showing difference in estimated value for manual inspection and fuzzy logic variable.

| Method | Assigned Value | Number of frames | Percentage of total |
|---|---|---|---|
| Manual | 1 | 36334 | 90.89% |
| Manual | 2 | 3168 | 7.93% |
| Manual | 3 | 473 | 1.18% |
| Fuzzy | 1 | 37779 | 94.51% |
| Fuzzy | 2 | 1749 | 4.38% |
| Fuzzy | 3 | 447 | 1.12% |

Table 20: Error between estimated visual fuzzy input and manual value for each frame of all 20 speech sentences.

| Estimated Value | Manual Est. | Fuzzy Est. | Difference | Difference Percentage |
|---|---|---|---|---|
| 1 | 36334 | 37779 | 1445 | 3.977% |
| 2 | 3168 | 1749 | 1419 | 44.79% |
| 3 | 473 | 447 | 26 | 5.497% |

Table 19 shows that the number of frames considered to have no ROI were very similar, with the greatest difference being that a higher number of fuzzy scores were estimated to be suitable than for a manual inspection. This is unsurprising due to the variation between speakers, sentences, cropped ROI dimensionality, and represents a justification for the use of a fuzzy logic variable. The difference in estimated values between the manual and the fuzzy approach is shown in table 20.This table shows that 3.98% of frames were incorrectly categorised as being good values (i.e. the difference between the ground truth and automatic values), 5.5% were incorrectly estimated to identify no ROI, and 44.8% were estimated to incorrectly be estimated as having a value of 2 (i.e. an incorrect/blurry/partial region). This was unsurprising as the difference between good and poor values could sometimes be very small, and indicates that the detector may have limitations with regard to precise identification of incorrect but partial regions.

Table 21: Overall perfomance of visual quality fuzzy input variable compared to manual scoring, considering each subset of 10 sentences, for reading task .

| Reading Task | Number of Frames | Percentage |
|---|---|---|
| Correct | 19513 | 97.64% |
| Incorrect | 472 | 2.36% |
| Total | 19985 | 100% |

Table 22: Overall perfomance of visual quality fuzzy input variable compared to manual scoring, considering each subset of 10 sentences, for conversation task.

| Conversation Task | Number of Frames | Percentage |
|---|---|---|
| Correct | 17323 | 86.66% |
| Incorrect | 2667 | 13.34% |
| Total | 19990 | 100% |

To analyse the incorrect classification results shown in table 20, the complete dataset of 20 sentence was divided into subsets. Firstly, the sentence was divided into two subsets, one for the conversation task, and the other for the reading task. This was expected to make a difference to the results, as it was expected that there would generally be less challenging visual data for the reading task. The overall results for the 10 reading task sentences are shown in table 21.

Table 21 shows that 97.64% of fuzzy frames were estimated to correctly classified to match the manual score, with only 472 out of 19985 frames not considered to be correct. This is a very low error. In comparison, the overall score for the conversation subset is shown in table 22.

It is very clear from the values in table 22 that the conversation task had a much lower correct score, suggesting that there are more misclassified frames for active conversation, as expected. To analyse this in more detail, tables 23 and 24 show the classification difference between fuzzy estimation and manual for the conversation and reading tasks respectively.

Table 23: Comparison of assigned values for 10 sentence reading dataset, showing difference in estimated value for manual inspection and fuzzy logic variable.

| Method | Assigned Value | Number of frames | Percentage of total |
|---|---|---|---|
| Manual | 1 | 19607 | 98.11% |
| Manual | 2 | 189 | 0.95% |
| Manual | 3 | 189 | 0.95% |
| Fuzzy | 1 | 19494 | 97.54% |
| Fuzzy | 2 | 337 | 1.6% |
| Fuzzy | 3 | 154 | 0.77% |

Table 24: Comparison of assigned values for 10 sentence conversation dataset, showing difference in estimated value for manual inspection and fuzzy logic variable.

| Method | Assigned Value | Number of frames | Percentage of total |
|---|---|---|---|
| Manual | 1 | 16727 | 83.68% |
| Manual | 2 | 2979 | 14.90% |
| Manual | 3 | 284 | 1.42% |
| Fuzzy | 1 | 18285 | 91.47% |
| Fuzzy | 2 | 1412 | 7.06% |
| Fuzzy | 3 | 293 | 1.47% |

Table 25: Error between estimated visual fuzzy input and manual estimated value for each frame of 10 speech sentence conversation task subset.

| Estimated Value | Manual Est. | Fuzzy Est. | Difference | Difference Percentage |
|---|---|---|---|---|
| 1 | 16727 | 18285 | 1558 | 9.31% |
| 2 | 2979 | 1412 | 1567 | 52.60% |
| 3 | 284 | 293 | 9 | 3.17% |

Table 26: Error between estimated visual fuzzy input and manual value for each frame of 10 speech sentence reading task subset.

| Estimated Value | Manual Estimation | Fuzzy Estimation | Difference | Difference Percentage |
|---|---|---|---|---|
| 1 | 19607 | 19494 | 113 | 0.58% |
| 2 | 189 | 337 | 148 | 78.31% |
| 3 | 189 | 154 | 35 | 18.52% |

Tables 23 and 24 show a very clear division between the data subsets. The conversation task has 86.68% of frames manually estimated to be of good quality, compared to 98.11% for the reading task. This can be partly explained by the additional movement and emotional speech in the conversational task resulting in more tracking errors. However a score of 86.66% accuracy in even the most challenging of scenarios suggests that the visual tracker is performing as well as could be expected. In the reading task, a very small number of frames when inspected did not produce a satisfactory result, and in the majority of cases, this can be explained by examples such as the mouth of the speaker being briefly covered.

The difference between the estimated classifications is shown in table 25 and 26 for conversation and reading tasks respectively.

Tables 25 and 26 show that there are some differences in classification errors. A comparison of these percentage errors for each task with the overall values given in table 20 is shown in table 27.

Firstly, it can be seen in table 27 that when it comes for incorrect estimation of a good frame, the conversation task has a much higher percentage error of 9.31%, compared to the

Table 27: Percentage error difference comparison between visual fuzzy input variable and manual estimated value for overall dataset, and reading and conversation subsets.

| Est. Value | Overall Percent Error | Reading Percent Error | Conversation Percent Error |
|---|---|---|---|
| 1 | 3.98% | 0.58% | 9.31% |
| 2 | 44.79% | 78.31% | 52.60% |
| 3 | 5.50% | 18.52% | 3.17% |

reading task error of 0.58%. This suggests that in conversation tasks, the fuzzy input variable is incorrectly categorising a number of frames as being of good quality rather than poor quality. This can be partly explained by the relative simplicity of the fuzzy input variable, and also that this input variable could possibly be further improved by a more sophisticated technique possibly using some form of machine learning, such as a HMM. It is important to consider that the input variable is a generalised value and due to the differences between speakers that there will be errors in classification. With regard to the error percentages in classifying a value as 3, or 2, the numbers are very small as shown in tables 25 and 26. Overall, it shows that when it comes to correctly classifying a frame, the system performs well, although there are some issues with identifying a frame as being of poor quality (wrong area/partial).

Table 25 shows that 2979 conversation frames were manually identified as being partial or incorrect, whereas the fuzzy-system identified only 1412 frames. The equivalent values for the reading task in table 26 show that 189 frames were manually identified as being partial or incorrect, whereas the fuzzy-system identified 337 frames. One hypothesis for this is that the larger number of incorrect frames in the conversational task comes from the tracker selecting an incorrect area and taking a number of frames to return, whereas this was not such an issue for the reading task due to the speaker moving less. Again, this is challenging to fully take account of due to differences between speakers and sentences, without manual inspection, and so the fuzzy variable is designed to take account of this uncertainty. Overall, the error is extremely small for the reading task as might be expected, and while larger for the conversation task, is still considered to be acceptable, with scope for improvement.

Table 28: Comparison of assigned values for 10 sentence reading dataset, showing difference in estimated value for manual inspection and fuzzy logic variable.

| Sentence | No. Correct | Perc. Correct | No. Incorrect | Perc. Incorrect | Total Frames |
|---|---|---|---|---|---|
| 1 | 1933 | 96.70% | 66 | 3.30% | 1999 |
| 2 | 1992 | 99.65% | 7 | 0.35% | 1999 |
| 3 | 1974 | 98.75% | 25 | 1.25% | 1999 |
| 4 | 1985 | 99.30% | 14 | 0.70% | 1999 |
| 5 | 1880 | 94.05% | 119 | 5.95% | 1999 |
| 6 | 1926 | 96.59% | 68 | 3.41% | 1994 |
| 7 | 1952 | 97.65% | 47 | 2.35% | 1999 |
| 8 | 1915 | 95.80% | 84 | 4.20% | 1999 |
| 9 | 1957 | 97.90% | 42 | 2.10% | 1999 |
| 10 | 1999 | 100% | 0 | 0% | 1999 |

In addition to a comparison between different tasks, it was also considered to be of interest to compare the error in individual sentences in order to identify if differences between the fuzzy estimation and the manual evaluation were evenly split, or were concentrated in specific sentences. Each of the 10 sentences in each conversation subset was evaluated to compare the difference in results. Considering the reading task first, the results are shown in table 28.

Table 28 shows that as expected, the percentage of matching fuzzy and ground truth values predicted is above 94% in all cases, with only a very small number of results where the fuzzy estimation does not match the manual evaluation. In comparison, table 29 shows the match between the fuzzy estimation and the manual evaluation for the 10 sentences chosen for the conversation task.

Table 29 shows that the variation between individual sentences is much higher, which is to be expected considering the issues the tracker faces with conversational speech. Although 6 of the 10 conversational sentences have a higher correct percentage than 90%, there is particular error concentrated in one sentence, with 66.18% of frames showing a difference between the

Table 29: Comparison of assigned values for 10 sentence conversation dataset, showing difference in estimated value for manual inspection and fuzzy logic variable.

| Sentence | No. Correct | Perc. Correct | No. Incorrect | Perc. Incorrect | Total Frames |
|---|---|---|---|---|---|
| 1 | 1836 | 91.85% | 163 | 8.15% | 1999 |
| 2 | 1432 | 71.64% | 567 | 28.36% | 1999 |
| 3 | 1999 | 100% | 0 | 0% | 1999 |
| 4 | 1947 | 97.40% | 52 | 2.60% | 1999 |
| 5 | 1840 | 92.05% | 159 | 7.95% | 1999 |
| 6 | 1930 | 96.55% | 69 | 3.45% | 1999 |
| 7 | 676 | 33.82% | 1323 | 66.18% | 1999 |
| 8 | 1689 | 84.49% | 310 | 15.51% | 1999 |
| 9 | 1978 | 98.95% | 21 | 1.05% | 1999 |
| 10 | 1996 | 99.85% | 3 | 0.15% | 1999 |

manual and fuzzy estimation. An inspection of this specific cropped image sequence identified that the reason for this was the performance of the tracker. While the tracker initially identifies a correct ROI, there is an issue in that due to the specific features of this face, a large number of frames are considered to be partial and only show a percentage of the mouth. While a manual inspection resulted in these being classified as partial results, the majority of the mouth was shown in these frames, as shown in figure 80, and so the difference was relatively small, resulting in the fuzzy value assigning these a score that was within the range of being considered good quality data. This indicates the difficulties with giving a precise score of 1, 2, or 3. Again, this is a justification for using a fuzzy variable rather than a crisp set. Overall, there is a larger error for conversation tasks as expected, with a greater variation between sentences.

In summary, the visual input fuzzy variable was considered to be very accurate, with the majority of frames being correctly classified. It can be seen that the majority of errors were found when conversation data was used, where there was expected to be a greater variation in

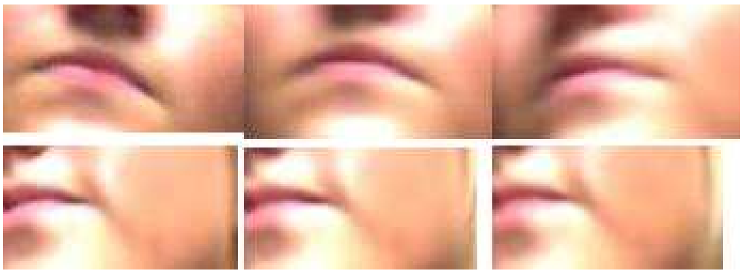


Figure 80: Examples of lip tracker extracting an incorrect image for a sequence of frames. These frames were consecutive frames from a single sentence and show that while a manual investigation may identify this as a partial result, the fuzzy input may be more nuanced, due to most of the mouth being present.

data due to factors such as changes in mouth shape, and more emotional speech. In particular, one specific sentence in the test-set was shown to have a greater error than any other sentence, and an inspection of the data demonstrated that this could be identified as due to potential ambiguity over the quality of the visual data, thus justifying the use of fuzzy logic rather than crisp sets, and demonstrating that the chosen thresholds are reasonably accurate and lead to correct classification in the majority of cases. There is scope for improvement using a form of machine learning such as a HMM to build a classification model, but it was felt that the technique used to calculate the input variable was suitable for this project, as shown by the results presented in this section.

### 7.6.2 *Previous Frame Fuzzy Input Variable*

*Problem Description*

As described in section 6.5, one input variable used in the system was the previous frame fuzzy output decision. As discussed in chapter 6, the aim of this variable was to prevent rapid switching from frame-to-frame when the input data could theoretically be processed by more than one processing option and there were very small differences from frame-to-frame, meaning that a small change in environmental conditions may result in rapid changes in processing decision from frame-to-frame. Rapid oscillation between processing options can reduce listener comfort, and should be minimised. Although it was decided to make use of the previous single output decision, it was also possible to make use of a mean of several previous decisions. It was possible that using a moving average of the previous outputs could be more effective in reducing switching than using a single value, and so this section investigates the effect of making use of the single previous output and compares this to using a mean of the previous 3, 5, and 10 previous output decisions.

In addition, the aim of using this input variable was to reduce oscillation, and so the effectiveness of using this variable is evaluated by comparing the output from the system when the rules relating to this input variable were enabled to the equivalent output when the rules were disabled (and so ignoring the previous frame input variable entirely). It is important to note that this section is not focused on analysing the output signal quality, but on evaluating the difference in processing decision from frame-to-frame.

*Experiment Setup*

To evaluate the effect of varying the previous frame input variable, a small dataset of 3 sentences from the corpus described in section 7.5 was used for evaluation. Broadband machine noise was added to these sentences using the simulated room environment at varying SNR levels to produce 18 noisy speech sentences with a range of audio and visual fuzzy input variables. In addition to this 3 sentences that did not have noise added to them, but were recorded in a noisy environment were also used, producing a total of 21 sentences. This input was then evaluated using the fuzzy logic system, and the output decision for each frame was recorded. In addition, the fuzzy rules pertaining to the previous frame were disabled, and the 21 sentences were evaluated again, and the decision (this time effectively only using two input variables) for each frame was also recorded.

*Evaluation Approach*

The 21 sentences were evaluated four times using the fuzzy logic system, using the single previous output decision, the mean of the value for the previous 3 outputs, the mean of the previous 5 outputs, and the mean of the previous 10 outputs as the input variable. The resulting output processing decision from the fuzzy logic system was then compared to the decision from the previous frame to calculate the difference between frames. As the system is fuzzy, it is possible for the output decision to vary very slightly from frame-to-frame, without the difference being large enough to affect the processing decision (i.e. no processing, audio-only, or audiovisual), and so it was felt of more relevance to focus on frames where there was a difference in output decision from the previous frame greater than ± 1.

*Summary of Results*

Initially, an investigation is carried out into whether using the single previous output decision, or whether a moving average of the previous 3, 5, or 10 output decision values will result in the greatest reduction in difference between frames. The 21 sentences described previously were used to evaluate the system using a different previous frame input variable. The fuzzy output decision for the same sentence can vary depending on the SNR that the speech and noise are mixed at. This means that the output decision of the fuzzy-system can be different when the SNR is different, even if the type of noise and the visual input variable is the same. This will be investigated in more depth in section 7.7.7. Again, the aim of this section is to investigate the output decision of the fuzzy logic system, rather than the output audio value.

To demonstrate the difference in fuzzy output decision depending on the difference in SNR, figure 81 shows an example of the same sentence mixed with noise at four different SNR levels. The output decision (which can range from 0 to 10, as discussed previously) of each frame is shown. (a) Shows the sentence with a moderate amount of noise. Although there are small changes from frame-to-frame, as shown by the axis, there is a very small difference, with the output decision varying between 5.03 and 5.05, which means audio only processing is chosen for all frames. (b) Shows a sentence with a high SNR. Audio only processing is chosen for the majority of frames, but it can also be seen that the fuzzy output decision also drops to a much lower value for some frames, meaning that the system is changing between audio-only processing and no filtering. (c) Shows a sentence with a lower SNR. It can be seen that the output decision varies between audiovisual processing and audio-only processing. Finally, (d) shows the same sentence with the lowest SNR, and it can be seen that the output decision is predominantly high, which means audiovisual processing is used. This demonstrates that a change in SNR can have a big difference on the processing decision, and this will be examined in more depth in section 7.7.7. In this section, the difference in output value from frame-to-frame is of the most interest.

Figure 81 shows that there are occasions when a change in fuzzy output decision does not result in a change in processing option, due to the change being very small, but it can also be seen that there are occasions when the processing decision will change rapidly between frames. The use of the previous fuzzy output decision as an input into the subsequent frame is intended to reduce this effect. This section evaluates whether the difference in output decision between frames is affected by using the previous value alone, or a mean of the previous 3 outputs, the previous 5 outputs, or the previous 10 outputs. Firstly, table 30 shows the number of frames where a difference of any value is found from the previous frame, showing the total number of frames with a difference and the percentage of the total frames, for the four different previous input variables.

Table 30 shows that the number of frames with a change in decision varies widely , with sentence 16 demonstrating a difference in almost every frame, whereas sentence 7 only had a difference in 12% of frames. This was to be expected, and it demonstrates that the fuzzy inference system is capable of reacting to changes in input data using the fuzzy input variables. However, as discussed previously, the magnitude of the difference is also relevant. As the processing choice is determined by the output value, it can be argued that a small change in the output decision (for example, from 5.03 to 5.05) is not generally expected to make a difference to the processing decision. Therefore, it was decided to filter the data by only considering values where from the previous frame is greater to or equal plus or minus 1. Table 31 shows the same sentences as processed in table 30, but only considering frames with a difference greater than or equal to ± 1.

Table 31 shows that compared to table 30, the number of frames where a difference is recorded is reduced considerable, suggesting that although the fuzzy output decision frequently changes, the majority of changes are not expected to result in a change in processing decision from the previous frame. Table 31 indicates that there was a difference between frames that may result in a change in processing method from the previous frame on a relatively low number of occasions, as low as 0%, and as high as 8.7%. Again, the difference

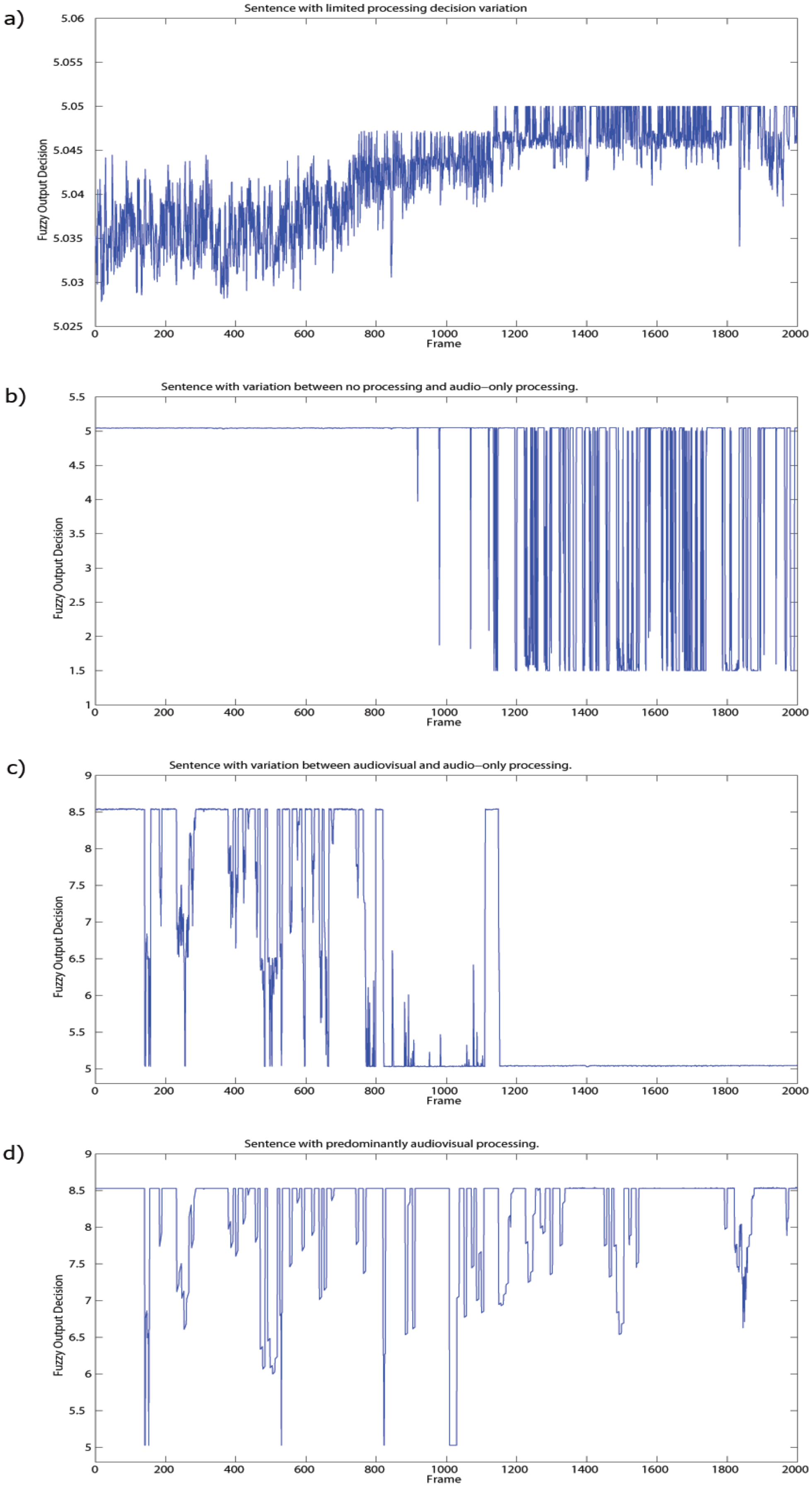


Figure 81: Difference in fuzzy output system for same sentence, with different levels of noise added. (a) represents a system with a moderate level of noise, (b) shows a sentence with a low level noise, and (c) and (d) show the fuzzy decision with lower SNR levels.

Table 30: Number and percentage of frames with any difference in fuzzy output decision compared to previous frame.

| | Prev. Frame | | Mean of 3 Frames | | Mean of 5 Frames | | Mean of 10 Frames | |
|---|---|---|---|---|---|---|---|---|
| Sent. | No. Diff | % Diff | No. Diff | % Diff | No. Diff | % Diff | No. Diff | % Diff |
| 1 | 1908 | 94.45% | 1894 | 94.75% | 1889 | 94.50% | 1888 | 94.45% |
| 2 | 1896 | 94.85% | 1859 | 93.00% | 1847 | 92.40% | 1852 | 92.65% |
| 3 | 1889 | 94.50% | 1880 | 94.05% | 1875 | 93.80% | 1868 | 93.45% |
| 4 | 1996 | 99.85% | 1996 | 99.85% | 1996 | 99.85% | 1995 | 99.80% |
| 5 | 1834 | 91.75% | 1834 | 91.75% | 1834 | 91.75% | 1834 | 91.75% |
| 6 | 1035 | 51.78% | 1035 | 51.78% | 1035 | 51.78% | 1035 | 51.78% |
| 7 | 252 | 12.61% | 252 | 12.61% | 252 | 12.61% | 252 | 12.61% |
| 8 | 1962 | 98.15% | 1957 | 97.90% | 1955 | 97.80% | 1952 | 97.65% |
| 9 | 1724 | 86.24% | 1690 | 84.54% | 1679 | 83.99% | 1671 | 83.59% |
| 10 | 1978 | 98.95% | 1978 | 98.95% | 1978 | 98.95% | 1978 | 98.95% |
| 11 | 1998 | 99.95% | 1998 | 99.95% | 1998 | 99.95% | 1998 | 99.95% |
| 12 | 1561 | 78.09% | 1561 | 78.09% | 1561 | 78.09% | 1561 | 78.09% |
| 13 | 1048 | 52.43% | 1048 | 52.43% | 1048 | 52.43% | 1048 | 52.43% |
| 14 | 1937 | 96.90% | 1902 | 95.15% | 1901 | 95.10% | 1904 | 95.248% |
| 15 | 1586 | 79.34% | 1546 | 77.34% | 1535 | 76.79% | 1520 | 76.038% |
| 16 | 1998 | 99.95% | 1998 | 99.95% | 1998 | 99.95% | 1998 | 99.950% |
| 17 | 1574 | 78.74% | 1574 | 78.74% | 1574 | 78.74% | 1574 | 78.739% |
| 18 | 886 | 44.32% | 886 | 44.32% | 886 | 44.32% | 886 | 44.322% |
| 19 | 225 | 11.26% | 225 | 11.26% | 225 | 11.26% | 225 | 11.256% |
| 20 | 1978 | 98.95% | 1977 | 98.90% | 1977 | 98.90% | 1977 | 98.899% |
| 21 | 1959 | 98.00% | 1922 | 96.15% | 1913 | 95.70% | 1908 | 95.448% |

Table 31: Number and percentage of frames with a difference in fuzzy output decision greater than or equal to ± 1, compared to previous frame.

| | Prev. Frame | | Mean of 3 Frames | | Mean of 5 Frames | | Mean of 10 Frames | |
|---|---|---|---|---|---|---|---|---|
| Sent. | No. Diff | % Diff | No. Diff | % Diff | No. Diff | % Diff | No. Diff | % Diff |
| 1 | 40 | 2.00% | 39 | 1.95% | 40 | 2.00% | 40 | 2.00% |
| 2 | 119 | 5.95% | 120 | 6.00% | 120 | 6.00% | 122 | 6.10% |
| 3 | 34 | 1.70% | 34 | 1.70% | 34 | 1.70% | 34 | 1.70% |
| 4 | 9 | 0.45% | 16 | 0.80% | 17 | 0.85% | 18 | 0.90% |
| 5 | 10 | 0.50% | 13 | 0.65% | 14 | 0.70% | 18 | 0.90% |
| 6 | 24 | 1.20% | 30 | 1.50% | 30 | 1.50% | 29 | 1.45% |
| 7 | 22 | 1.10% | 22 | 1.10% | 22 | 1.10% | 22 | 1.10% |
| 8 | 109 | 5.45% | 112 | 5.60% | 110 | 5.50% | 110 | 5.50% |
| 9 | 120 | 6.00% | 118 | 5.90% | 118 | 5.90% | 121 | 6.05% |
| 10 | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| 11 | 43 | 2.15% | 68 | 3.40% | 74 | 3.70% | 64 | 3.20% |
| 12 | 64 | 3.20% | 77 | 3.85% | 84 | 4.20% | 87 | 4.35% |
| 13 | 48 | 2.40% | 48 | 2.40% | 48 | 2.40% | 48 | 2.40% |
| 14 | 167 | 8.35% | 169 | 8.45% | 171 | 8.55% | 172 | 8.60% |
| 15 | 174 | 8.70% | 174 | 8.70% | 174 | 8.70% | 174 | 8.70% |
| 16 | 4 | 0.20% | 4 | 0.20% | 4 | 0.20% | 4 | 0.20% |
| 17 | 12 | 0.60% | 16 | 0.800% | 17 | 0.85% | 15 | 0.75% |
| 18 | 11 | 0.55% | 11 | 0.550% | 11 | 0.55% | 11 | 0.55% |
| 19 | 8 | 0.40% | 8 | 0.400% | 8 | 0.40% | 8 | 0.40% |
| 20 | 4 | 0.20% | 4 | 0.200% | 4 | 0.20% | 4 | 0.20% |
| 21 | 110 | 5.50% | 108 | 5.403% | 108 | 5.40% | 109 | 5.45% |

Table 32: Number and percentage of frames with any difference in fuzzy output decision compared to previous frame, showing mean difference of all sentences (41980 frames).

| | Prev. Frame | | Mean of 3 Frames | | Mean of 5 Frames | | Mean of 10 Frames | |
|---|---|---|---|---|---|---|---|---|
| Diff. | No. Diff | Perc Diff. | No. Diff | Perc Diff. | No. Diff | Perc Diff. | No. Diff | Perc Diff. |
| $> \pm 0$ | 33224 | 79.142% | 33012 | 78.637% | 32956 | 78.584% | 32924 | 78.428% |
| $\geqslant \pm 1$ | 1132 | 2.697% | 1191 | 2.837% | 1208 | 2.878% | 1210 | 2.882% |

between individual sentences is to be expected considering the different noise conditions. Subsequent tables will consider mean values of all sentence frame differences. The means of the sentences used in tables 31 and 30 are compared in table 32 for different sizes of input variable

It can be seen in table 32 that considering any change in decision, there is a slight drop from 79.14% when using a single value, to 78.43% when a mean of 10 frames is used. However, when only considering larger changes, table 32 shows that increasing the number of previous decisions used as part of the mean input variable actually results in a very small increase in difference. When only the single previous output decision is used as the input variable, 1132, or 2.7% of the total 41980 frames show a change in decision. Using a mean of the 3 previous decisions results in a change of 2.8%, increasing to 2.9% when a mean of 5 previous decisions, and then finally 2.9% when a mean of the 10 previous decisions is used. Overall, the difference between frames when using an increased number of previous decisions as part of the input mean variable was considered to be so small that it had no particularly noticeable difference. Therefore, it was felt that it was suitable to use only the previous decision as an input variable into the fuzzy logic system.

The second aspect of this evaluation concerned the impact that this fuzzy input variable had on reducing the oscillation from frame-to-frame. To investigate this, the test-set described above was evaluated with the system. The fuzzy logic system was adjusted to disable the rules concerning the previous input variable, in effect meaning that the system made use of only the audio and visual input variables at all times. The mean results of this evaluation are shown in table 33.

Table 33: Number and percentage of frames with any difference in fuzzy output decision compared to previous frame, showing mean difference of all sentences (41980 frames), evaluated when previous frame rule is disabled in fuzzy-system.

| | Prev. Frame | | Mean of 3 Frames | | Mean of 5 Frames | | Mean of 10 Frames | |
|---|---|---|---|---|---|---|---|---|
| Diff. | No. Diff | Perc Diff. | No. Diff | Perc Diff. | No. Diff | Perc Diff. | No. Diff | Perc Diff. |
| > ± 0 | 13239 | 31.536% | 13239 | 31.536% | 13239 | 31.536% | 13239 | 31.536% |
| ⩾ ± 1 | 2460 | 5.856% | 2460 | 5.856% | 2460 | 5.856% | 2460 | 5.856% |

The results presented in table 33 are of interest for several different reasons. Firstly, because no previous frame rules are enabled, there is no change at all when a different number of previous decisions are part of the mean input variable. This confirms that this input variable has a role in affecting the output decision. Another difference of interest is that removing the rules pertaining to this input variable results in only 31.54% of frames showing a difference from the previous frame. This is lower than the 79% shown when using various versions of the previous frame input variable with the rules enabled, again, showing that this input variable is having an effect on the output decision.

However, this figure is for all differences, including those that are very small. When only a difference equal to or greater plus or minus 1 is considered, it can be seen that whereas with the rules enabled, as shown in table 32, the frames with a recorded difference varies from 2.7% to 2.88%. With this input variable not used, 5.86% of frames record a difference in output decision from the previous frame. Therefore, it can be concluded that the use of this input variable successfully limits processing decision variation from frame-to-frame, justifying the decision to make use of it as an input into the fuzzy logic speech filtering system presented in this thesis.

## 7.7 DETAILED SYSTEM EVALUATION

### 7.7.1 *Problem Description*

The initial two-stage system presented in the previous chapters of this thesis was found to be effective at performing speech enhancement in noisy environments with a very low SNR. However, examples were also highlighted of situations where both conventional single modality speech processing and the two-stage audiovisual speech enhancement system were found to be inadequate. The previous chapter proposed an extension of the initial two-stage system to utilise a fuzzy logic based system that aimed to overcome some of these limitations. The proposed novel framework is intended to add increased flexibility and versatility to the initial audiovisual speech enhancement approach that was presented in chapter 4. The previous section justified the use of the fuzzy input variables, demonstrating through detailed evaluation that it was possible for the visual input to be accurately fuzzified for a range of data. It was also demonstrated that using the previous fuzzy decision output as an input variable for the subsequent frame could also reduce the degree of oscillation between different processing options.

This section focuses on the evaluation of this multimodal fuzzy logic based speech enhancement framework. Due to the limitations discussed in section 7.4, there were some conditions. Firstly, in order to adequately represent conditions that a fuzzy-based system would be expected to successfully process, real audiovisual speech data was required, with examples of variation in audio and visual conditions. This was achieved by making use of the newly recorded corpus discussed in section 7.5. Again, due to the limitations discussed previously, it was not possible to record a truly noisy corpus that was compatible with this preliminary system, and so noise was added to the system using the simulated room environment discussed in chapter 4. This has the advantage that it is compatible with the system, and so results can be achieved. It also allows for the fuzzy switching system to be demonstrated. However, the limitation with this approach is that the simulated room with clearly defined noise and speech sources represents an ideal scenario that an audio-only beamformer would be expected to process without difficulty. Because of this, and also the limitations with the audiovisual approach when presented with novel data (as discussed in chapter 5), the results presented in this section should be interpreted with a degree of caution. Noises were added to speech sentences, with the SNR varying between -40dB and +10dB, and the visual data was processed using the automatic lip-tracking approach. There are some limited examples of the Lombard Effect, in that the speaker is taking part in an animated conversation and so adjusts their voice to take account of both parties talking, but this is not considered to be a major part of the chosen test-set.

This noisy speech mixture was then processed by the fuzzy logic based framework and evaluated. The fuzzy logic based output processed by this proof of concept framework is compared to the equivalent output using other techniques, including the two-stage system demonstrated previously. The performance of the fuzzy logic controller is also evaluated by investigating the fuzzy switching process.

### 7.7.2 *Experiment Setup*

The system described in chapter 6 was evaluated by making use of 10 sentences selected from the newly recorded audiovisual corpus described in section 7.5. To provide a range of data, 5 of the 10 sentences were taken from the reading task, and 5 from the conversation task. Each sentence represented a 20 second snippet of conversation or reading, with different content and visual information depending on the specific sentence. The sentences were chosen on the basis that they provided a wide range of content, and the only assumption made was that the lip-region could be successfully identified in the initial frame. As discussed in section 7.6, there are tracking errors, as expected. Each sentence was split into frames, producing approximately 1999 frames per sentence (the precise number of frames varied slightly depending on the sentence).

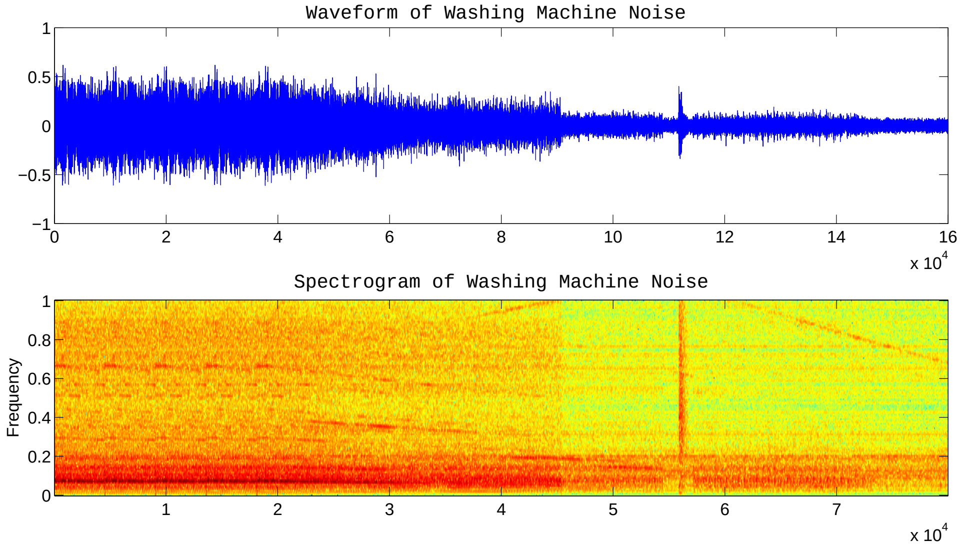


Figure 82: Waveform (top), and spectrogram (bottom) of broadband washing machine noise used for objective and subjective tests.

In a similar manner to chapter 5, noise is to speech with the simulated room environment at a range of SNR levels, from -40dB to +10dB. Two different noises are used. Firstly, broadband noise is added to the speech. This consisted of a recording of a washing machine, as shown in figure 82. It can be seen in figure 82 that the amplitude of the signal varies over time, with a gradual decrease in amplitude throughout the recording segment. This noise was used for the objective tests described in section 7.7.4, and also for the subjective listening tests in section 7.7.5.

To compare the results of using broadband noise with using a different noise source, an inconsistent clapping noise is also used. As discussed in chapter 5 previously, this inconsistent clapping noise is designed to be difficult for the beamformer to process due to the use of silences and transient sounds. This is used for the objective tests discussed in section 7.7.6, and is shown in figure 83.

A number of different versions were compared of each speech sentence. For the objective tests, the audiovisual approach presented in chapter 4 was used, along with an audio-only beamforming approach to serve as a comparison. In addition to this, the spectral subtraction approach used in chapter 5 and the unfiltered noisy signal were also used. These results were compared with results from the fuzzy logic based system. For the subjective listening tests,

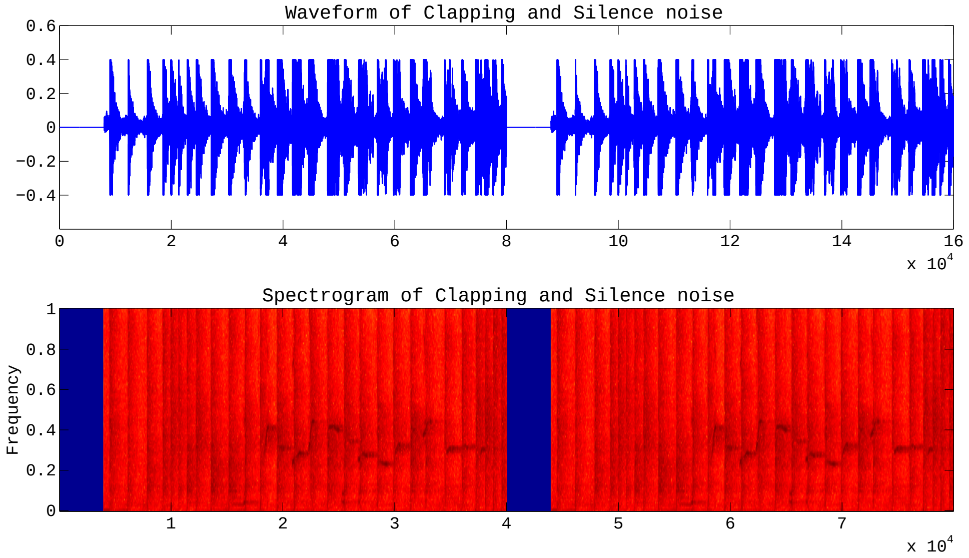


Figure 83: Waveform (top), and spectrogram (bottom) of clapping noise used for objective tests.

the three processing methods used for comparison were the two-stage audiovisual approach, the beamforming approach, and the fuzzy logic approach (only three methods were used in order to prevent listener fatigue).

### 7.7.3 *Evaluation Approach*

As the input variables were evaluated individually in the previous section, this section focused on the audio performance of the fuzzy switching system. To do this, the composite measures (Hu & Loizou [87]) used in chapter 5 are used to perform a detailed evaluation. This provides composite mean values that can be analysed. The output of using the fuzzy logic processing system was compared to mean values calculated by using a number of other techniques, including spectral subtraction, the two-stage audiovisual approach, audio-only beamforming, and the unfiltered noisy speech.

In addition to this, it was felt that it would be suitable to run a number of listening tests to evaluate the subjective quality of the speech. As the speech sentences were considerably longer than the original speech sentences used for evaluation in previous chapters (20 seconds rather than 3 seconds), there were concerns that testing the entire dataset would be challenging due to listener fatigue (10 sentences of 20 seconds length, with each sentence being processed in 5 different ways at 6 different SNR levels). Therefore, to reduce listener fatigue, only three versions of each conversation snippet were evaluated, audiovisual, audio-only, and the fuzzy logic approach, and the number of 20 second conversation snippets was reduced from 10 to 5. These are then tested with volunteers to produce suitable MOS. The results of these listening tests are discussed in section 7.7.5.

In addition to the audio output, the fuzzy switching is also evaluated. To do this, individual sentences are inspected to assess the effect of various factors, such as adjusting the SNR, the effect of poor visual information, and the difference between individual sentences. This is evaluated by a visual inspection and comparison. The fuzzy output decision on a frame-by-frame basis is then inspected. This is discussed in section 7.7.7.

### 7.7.4 *Objective Testing With Broadband Noise*

As discussed above, each 20 second snippet of either conversation or reading had broadband machine noise (as shown in figure 82) added to it at different SNR levels, ranging from -40dB to +10dB. Each mixture of speech and noise was then evaluated with the composite objective measures developed by Hu & Loizou [87]. Five versions of each sentence were compared, firstly, the audiovisual two-stage system presented in chapter 4. As this approach was shown in chapter 5 to perform poorly with completely novel speakers, then it was expected that this approach would perform poorly when tested with the newly recorded corpus. In addition to this, the results of performing audio-only beamforming are also presented. As the simulated room is designed to demonstrate the performance of the beamformer, it is expected that the results of using this technique will be extremely good. The noisy unfiltered sentence is also used, along with the spectral subtraction approach used for evaluation in chapter 5. These are compared to the results of using the fuzzy-based system presented in the previous chapter. The means of the composite overall, speech distortion, and background distortion at different SNR levels are provided in tables 34, 35, 36, and figures 84, 44, 45 respectively.

Table 34: Composite objective mean test score table for overall speech quality for speech with washing machine noise added, for audiovisual speech, audio-only beamforming, fuzzy-based processing, audio-only spectral subtraction, and unprocessed speech.

| Level | Avis | Beamforming | Fuzzy | Spectral | Noisy |
|---|---|---|---|---|---|
| -40dB | 1.482 | 3.5078 | 1.110 | 2.136 | 2.557 |
| -30dB | 1.672 | 3.802 | 1.108 | 2.341 | 2.445 |
| -20dB | 1.798 | 3.994 | 2.054 | 1.904 | 2.233 |
| -10dB | 1.720 | 4.063 | 3.534 | 1.806 | 1.818 |
| 0dB | 1.315 | 4.089 | 3.903 | 2.573 | 2.485 |
| +10dB | 0.665 | 4.102 | 3.800 | 3.117 | 3.083 |

Table 35: Composite objective mean test score table for speech score speech quality for speech with washing machine noise added, for audiovisual speech, audio-only beamforming, fuzzy-based processing, audio-only spectral subtraction, and unprocessed speech.

| Level | Avis | Beamforming | Fuzzy | Spectral | Noisy |
|---|---|---|---|---|---|
| -40dB | 1.649 | 4.415 | 1.373 | 2.121 | 2.561 |
| -30dB | 1.786 | 4.642 | 1.401 | 2.292 | 2.490 |
| -20dB | 1.874 | 4.790 | 2.391 | 2.059 | 2.394 |
| -10dB | 1.729 | 4.846 | 4.226 | 2.199 | 2.253 |
| 0dB | 1.128 | 4.870 | 4.676 | 3.021 | 3.000 |
| +10dB | 0.115 | 4.882 | 4.536 | 3.625 | 3.682 |

Table 36: Composite objective mean test score table for noisy speech quality for speech with washing machine noise added, for audiovisual speech, audio-only beamforming, fuzzy-based processing, audio-only spectral subtraction, and unprocessed speech.

| Level | Avis | Beamforming | Fuzzy | Spectral | Noisy |
|---|---|---|---|---|---|
| -40dB | 1.842 | 2.770 | 1.630 | 1.995 | 1.957 |
| -30dB | 1.910 | 3.001 | 1.591 | 2.116 | 1.889 |
| -20dB | 1.917 | 3.331 | 2.101 | 1.847 | 1.753 |
| -10dB | 1.835 | 3.750 | 3.285 | 1.774 | 1.476 |
| 0dB | 1.620 | 3.799 | 3.592 | 2.224 | 1.898 |
| +10dB | 1.359 | 3.816 | 3.429 | 2.519 | 2.341 |

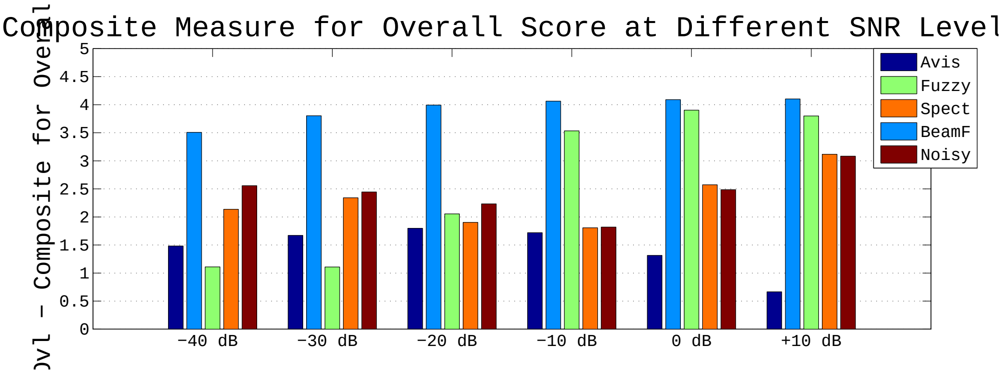


Figure 84: Composite objective mean test scores for overall speech quality for speech with washing machine noise added, for audiovisual speech, audio-only beamforming, fuzzy-based processing, audio-only spectral subtraction, and unprocessed speech.

Considering the overall score first, the audio-only beamformer produced the best overall score, which was expected. The unfiltered and spectral subtraction scores are very similar, which matches expectations based on the results in chapter 5. It can also be seen that the audiovisual approach is the worst performing method, which again matches expectations. As seen in chapter 5, as the SNR increases, the audiovisual score decreases, which matches the results when tested on the VidTIMIT corpus in chapter 5.

The performance of the fuzzy-based system is of interest. The interaction plot for the overall score is shown in figure 87 and the results of Bonferroni multiple comparison for the difference between the audiovisual and fuzzy logic approach, and the audio-only and fuzzy approach are given in tables 37 and 38. The difference of means in table 37 shows that at a very low

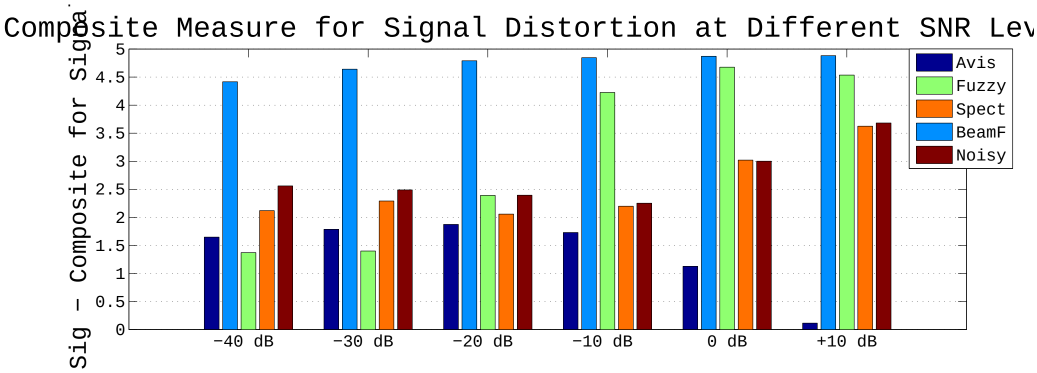


Figure 85: Composite objective mean test scores for speech distortion level for speech with washing machine noise added, for audiovisual speech, audio-only beamforming, fuzzy-based processing, audio-only spectral subtraction, and unprocessed speech.

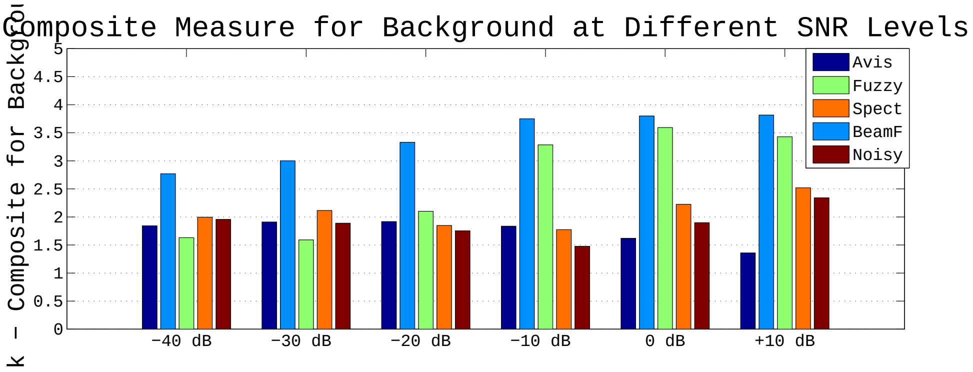


Figure 86: Composite objective mean test scores for noise distortion level for speech with washing machine noise added, for audiovisual speech, audio-only beamforming, fuzzy-based processing, audio-only spectral subtraction, and unprocessed speech.

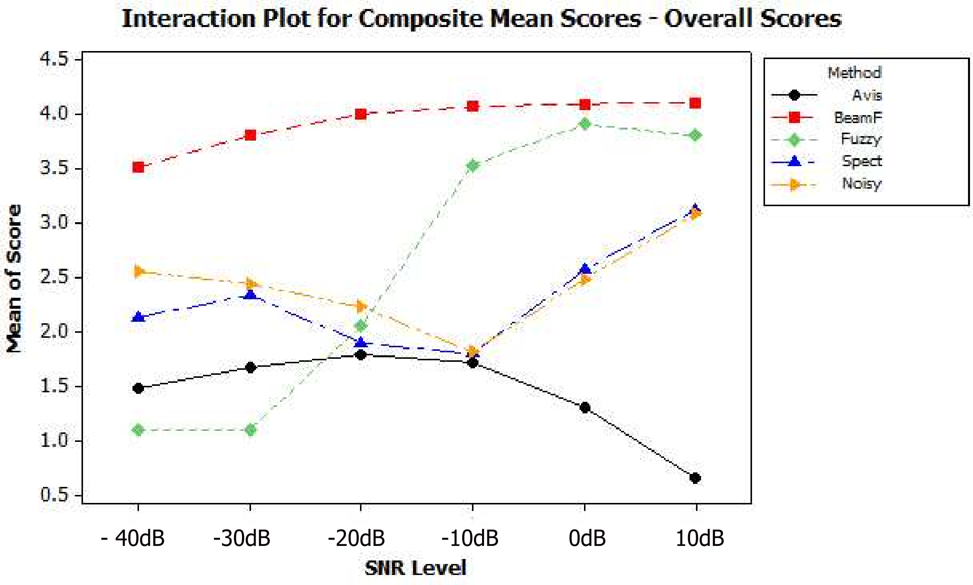


Figure 87: Interaction plot for overall composite objective mean score of speech with washing machine noise added at varying SNR levels, showing audiovisual speech (black and circle markers), audio-only beamforming (red with square markers), fuzzy-based system (green with diamond markers), spectral subtraction (blue with triangles), and unfiltered noisy speech (orange with triangles).

SNR (at SNR levels of -40dB, -30dB, -20dB), the fuzzy logic approach is the worst performing approach. However, although it is the worst performing approach the difference between the audiovisual and fuzzy approaches was not statistically significant (p>0.05). This suggests that as the noise level is extremely high, the fuzzy logic system makes use of the audiovisual method, which explains the lack of difference.

At higher SNR levels, when there is less noise, the fuzzy-system makes more use of the audio-only approach, and so as shown by the comparison of means in table 38, the difference between the fuzzy-system at these higher SNR levels is not statistically significant (p>0.05). However, the scores do not match exactly. This is because, as will be discussed in section 7.7.7, the fuzzy-system does not make use of the same approach in all frames, as it switches in response to precise changes in input variables. Similar results can be seen for the composite speech distortion score, with the interaction plot shown in figure 88 and the comparison of means in tables 39 and 40. The speech distortion introduced by the audiovisual approach is reflected in the difference in the audio-only score and the fuzzy logic approach at -10dB being significantly different (p<0.05), which is not the case for the overall composite score. Likewise,

Table 37: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audiovisual Filtering and Fuzzy Processed Speech with washing machine noise added for overall composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -0.372 | 0.144 | -2.589 | 1.000 |
| -30dB | -0.564 | 0.144 | -3.926 | 0.053 |
| -20dB | 0.256 | 0.144 | 1.782 | 1.000 |
| -10dB | 1.814 | 0.144 | 12.639 | 0.000 |
| 0dB | 2.588 | 0.144 | 18.030 | 0.000 |
| +10dB | 3.135 | 0.144 | 21.84 | 0.000 |

at -20dB, the difference between the means of the audiovisual and fuzzy-based approach is statistically significant ($p<0.05$), which reflects the improvement seen when the fuzzy-based approach uses a greater number of audio-only frames.

The composite noise distortion score is very similar, as shown in the interaction plot in figure 89 and the comparison of means in tables 41 and 42. The difference between the means of the audio-only and fuzzy-based approach is significant ($p<0.05$) at SNR levels of -40dB to -10dB and also at +10dB, suggesting that at +10dB, the system is using the unfiltered approach in a number of frames, so some background noise is included in the output.

To verify these results, it was felt suitable to conduct listening tests of this data, and this will be discussed in section 7.7.5.

### 7.7.5 *Subjective Testing with Broadband Noise*

The previous section discussed the results of objective tests of speech mixed with broadband machine noise. This section reports the results of listening tests performed on this dataset (the same sentences with the same noise added as shown in figure 82). The listening tests were conducted in a manner similar to those discussed in chapter 5. 10 volunteers took part in listening tests in a quiet room, using noise cancelling headphones. All of the volunteers spoke

Table 38: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audio-only beamforming and Fuzzy Processed Speech with washing machine noise added for overall composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -2.398 | 0.1435 | -16.70 | 0.0000 |
| -30dB | -2.694 | 0.1435 | -10.18 | 0.0000 |
| -20dB | -1.940 | 0.1435 | -13.52 | 0.0000 |
| -10dB | -0.529 | 0.1435 | -3.68 | 0.1317 |
| 0dB | -0.187 | 0.1435 | -1.31 | 1.0000 |
| +10dB | -0.302 | 0.1435 | -2.104 | 1.0000 |

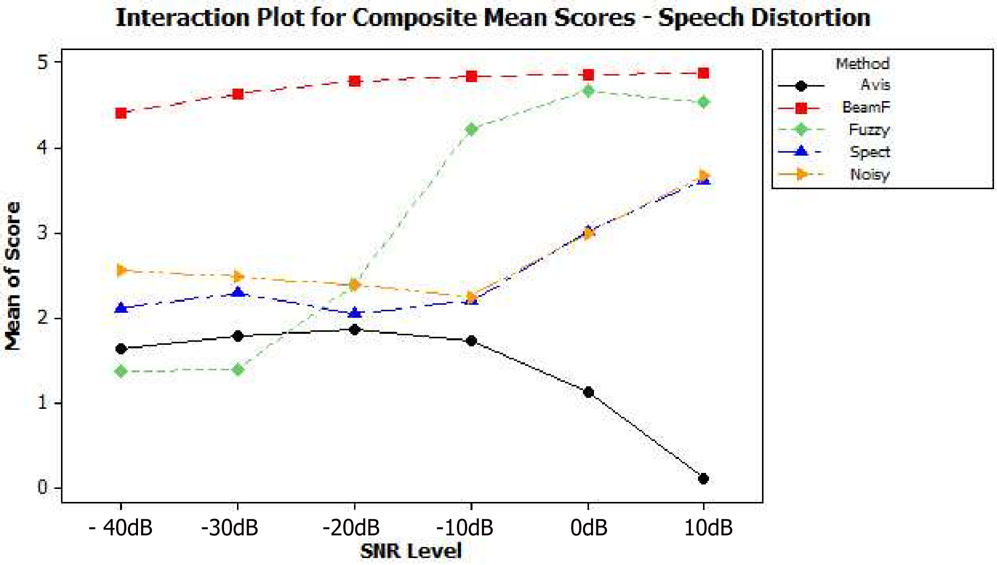


Figure 88: Interaction plot for speech distortion composite objective mean score of speech with washing machine noise added at varying SNR levels, showing audiovisual speech (black and circle markers), audio-only beamforming (red with square markers), fuzzy-based system (green with diamond markers), spectral subtraction (blue with triangles), and unfiltered noisy speech (orange with triangles).

Table 39: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audio-only beamforming and Fuzzy Processed Speech with washing machine noise added for speech distortion composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -3.043 | 0.129 | -23.59 | 0.000 |
| -30dB | -3.241 | 0.129 | -25.12 | 0.000 |
| -20dB | -2.399 | 0.129 | -18.60 | 0.000 |
| -10dB | -0.620 | 0.129 | -4.81 | 0.001 |
| 0dB | -0.195 | 0.129 | -1.51 | 1.000 |
| +10dB | -0.346 | 0.129 | -2.680 | 1.000 |

Table 40: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audiovisual Filtering and Fuzzy Processed Speech with washing machine noise added for speech distortion composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -0.276 | 0.1290 | -2.14 | 1.0000 |
| -30dB | -0.386 | 0.1290 | -2.99 | 1.0000 |
| -20dB | 0.516 | 0.1290 | 4.00 | 0.0398 |
| -10dB | 2.497 | 0.1290 | 19.36 | 0.0000 |
| 0dB | 3.548 | 0.1290 | 27.501 | 0.0000 |
| +10dB | 4.421 | 0.1290 | 34.27 | 0.0000 |

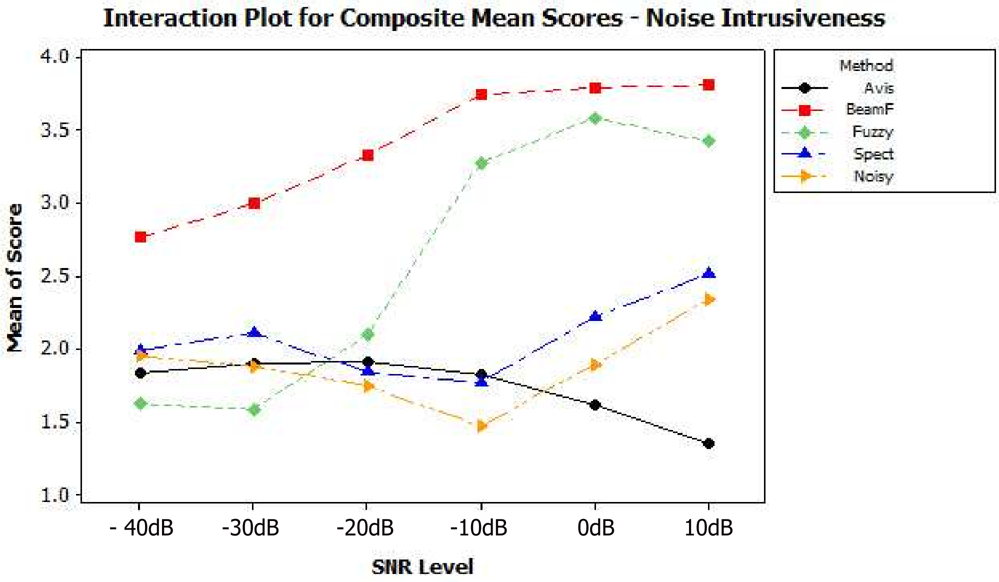


Figure 89: Interaction plot for noise intrusiveness composite objective mean score of speech with washing machine noise added at varying SNR levels, showing audiovisual speech (black and circle markers), audio-only beamforming (red with square markers), fuzzy-based system (green with diamond markers), spectral subtraction (blue with triangles), and unfiltered noisy speech (orange with triangles).

Table 41: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audio-only beamforming and Fuzzy Processed Speech with washing machine noise added for noise intrusiveness composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -1.140 | 0.098 | -11.62 | 0.000 |
| -30dB | -1.410 | 0.098 | -14.38 | 0.000 |
| -20dB | -1.231 | 0.098 | -12.55 | 0.000 |
| -10dB | -0.466 | 0.098 | -4.75 | 0.002 |
| 0dB | -0.207 | 0.098 | -2.11 | 1.000 |
| +10dB | -0.387 | 0.098 | -3.95 | 0.049 |

Table 42: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audiovisual Filtering and Fuzzy Processed Speech with washing machine noise added for noise intrusiveness composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -0.212 | 0.098 | -2.164 | 1.000 |
| -30dB | -0.318 | 0.098 | -3.247 | 0.605 |
| -20dB | 0.184 | 0.098 | 1.873 | 1.000 |
| -10dB | 1.450 | 0.098 | 14.789 | 0.000 |
| 0dB | 1.972 | 0.098 | 20.115 | 0.000 |
| +10dB | 2.070 | 0.098 | 21.11 | 0.000 |

English as a first language, and none reported any abnormalities with their hearing. There were 6 male subjects and 4 female subjects, with an age range between 21 and 37. Listeners were played sentences randomly from the test-set, and were asked to score each between 0 and 5 based on the same criteria as in 5. They were asked to score speech signal distortion, noise intrusiveness level, and overall quality.

As there were concerns over listener fatigue due to the potential duration of listening tests using the entire dataset tested in section 7.7.4, a smaller subset of the test-set was used. 5 sentences were selected (again, a mix of reading and conversation tasks), from different speakers, and broadband noise was added at 6 different SNR levels. 3 different processing methods were used, the audiovisual approach, the audio-only approach, and the fuzzy-based system. The overall, speech distortion, and noise intrusiveness MOS results are shown in tables 43, 44, and 45, with the equivalent data shown in figures 90, 91, and 92 respectively.

An inspection of figures 90, 91, 92 shows that the scores for subjective listening tests look very similar to the results presented in section 7.7.4. Just as in reported in results in the previous section, the audiovisual approach is consistently identified to have the worst output scores, and the audio-only technique returns the best results. The fuzzy-based approach

Table 43: MOS table for overall quality for speech with washing machine noise added, for audiovisual speech, audio-only beamforming, and fuzzy-based processing.

| Level | Avis | Beamforming | Fuzzy |
|---|---|---|---|
| -40dB | 1.392 | 3.400 | 1.196 |
| -30dB | 1.650 | 4.130 | 1.604 |
| -20dB | 1.774 | 4.374 | 1.988 |
| -10dB | 1.950 | 4.364 | 3.650 |
| 0dB | 1.436 | 4.370 | 3.986 |
| +10dB | 0.872 | 4.414 | 3.787 |

Table 44: MOS table for speech distortion for speech with washing machine noise added, for audiovisual speech, audio-only beamforming, and fuzzy-based processing.

| Level | Avis | Beamforming | Fuzzy |
|---|---|---|---|
| -40dB | 0.940 | 2.840 | 0.752 |
| -30dB | 1.470 | 3.800 | 1.274 |
| -20dB | 1.730 | 4.370 | 2.020 |
| -10dB | 2.050 | 4.330 | 3.662 |
| 0dB | 1.740 | 4.440 | 4.460 |
| +10dB | 1.130 | 4.650 | 4.370 |

Table 45: Composite MOS table for noise intrusiveness for speech with washing machine noise added, for audiovisual speech, audio-only beamforming, and fuzzy-based processing.

| Level | Avis | Beamforming | Fuzzy |
|---|---|---|---|
| -40dB | 2.730 | 4.620 | 2.740 |
| -30dB | 2.350 | 4.700 | 2.336 |
| -20dB | 2.210 | 4.380 | 2.190 |
| -10dB | 2.005 | 4.190 | 3.630 |
| 0dB | 1.260 | 4.110 | 3.660 |
| +10dB | 1.000 | 4.390 | 3.450 |

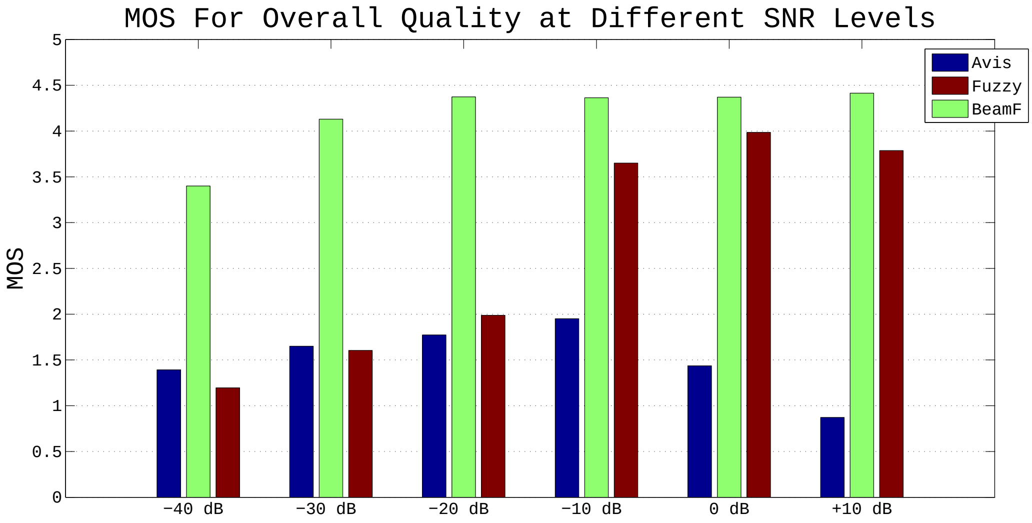


Figure 90: Mean Opinion Score for overall speech quality for speech with washing machine noise added, for audiovisual speech, audio-only beamforming, and fuzzy-based processing.

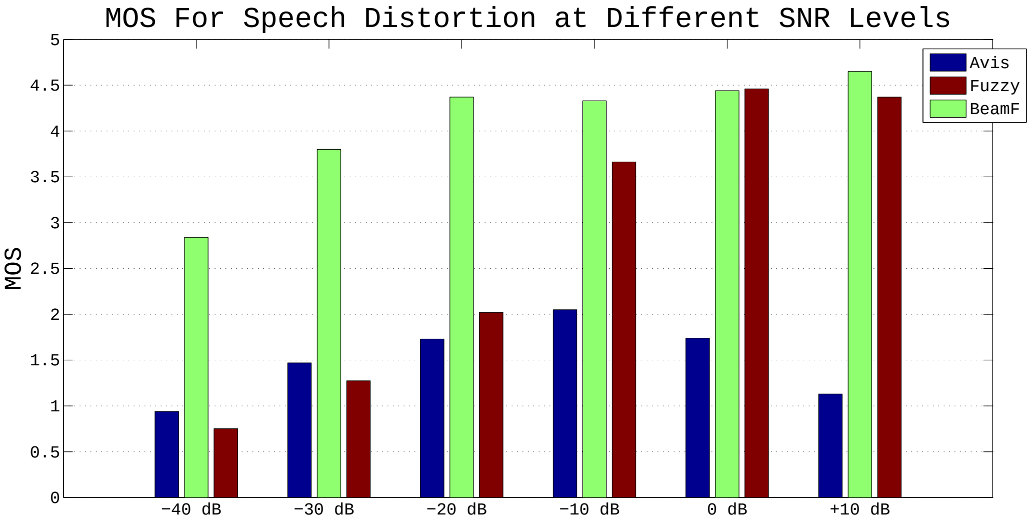


Figure 91: Mean Opinion Score for speech distortion level for speech with washing machine noise added, for audiovisual speech, audio-only beamforming, and fuzzy-based processing.

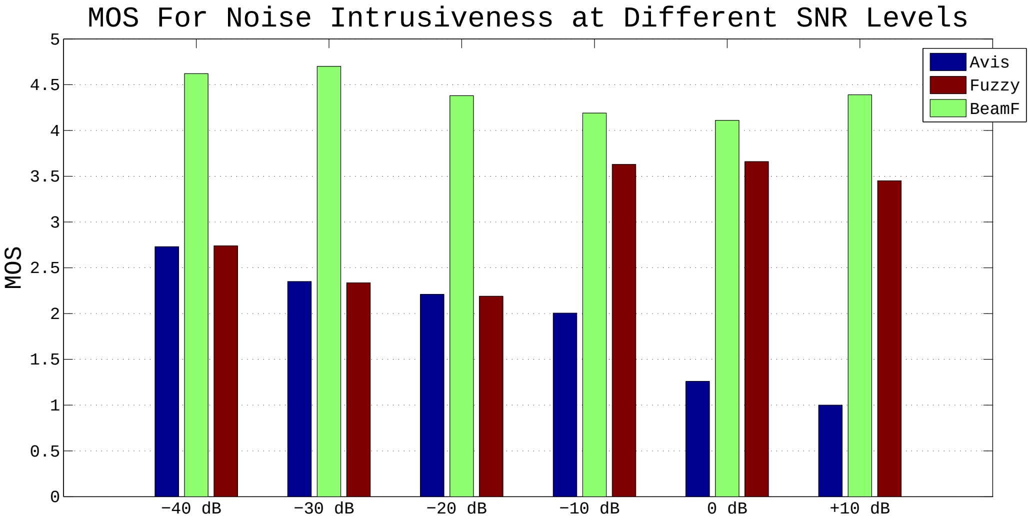


Figure 92: Mean Objective Scores for noise intrusiveness level for speech with washing machine noise added, for audiovisual speech, audio-only beamforming, and fuzzy-based processing.

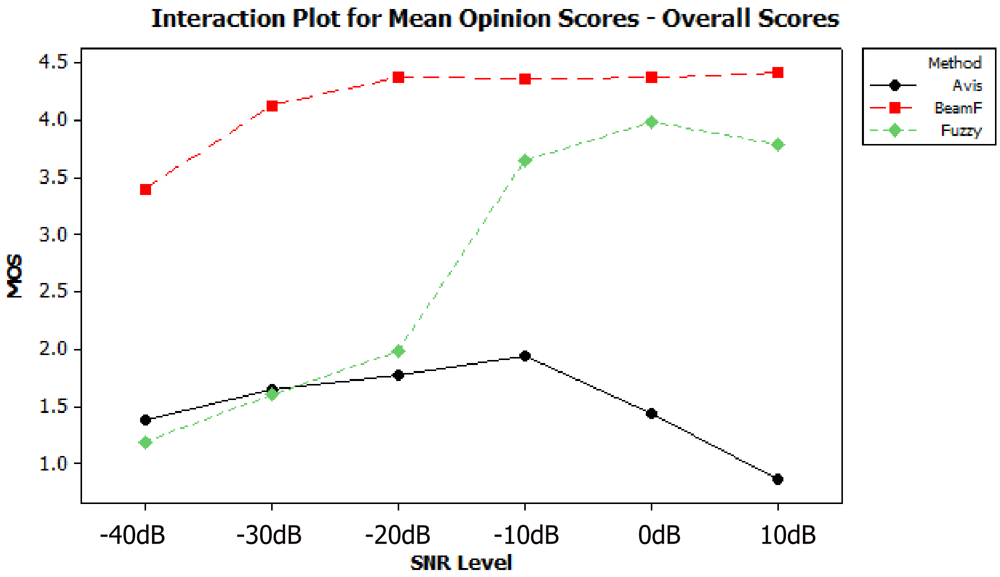


Figure 93: Interaction plot for overall MOS at varying SNR levels, showing audiovisual speech (black and circle markers), audio-only beamforming (red with square markers), and fuzzy-based system (green with diamond markers).

performs poorly at a very low SNR, but has an improved output at a higher SNR. A more detailed analysis is conducted on the results, using Bonferroni multiple comparison. Figures 93 and 94 show the interaction plots for overall and speech distortion MOS, and the difference of means is given in tables 46, 47, 48, and 49.

It can be seen that the trend of results is very similar to the objective scores discussed in section 7.7.4. At a lower SNR, the audiovisual and fuzzy-based scores are very similar, with no significant difference. This signifies that there was a far greater preference by listeners for the sentences processed with audio-only beamforming. When the SNR is increased, the fuzzy-based approach produced an improved score, with a similar output to the audio-only approach, with the results of Bonferroni multiple comparison showing that at SNR levels of -10dB, 0dB, and +10dB, the overall and speech distortion scores were not significantly different (p>0.05). This indicates that listeners found these sentences to be very similar in terms of overall results, and also when considering the speech in isolation.

The MOS for noise intrusiveness are also very similar to the equivalent objective results presented in the previous section. The interaction plot is shown in figure 95, and the results of

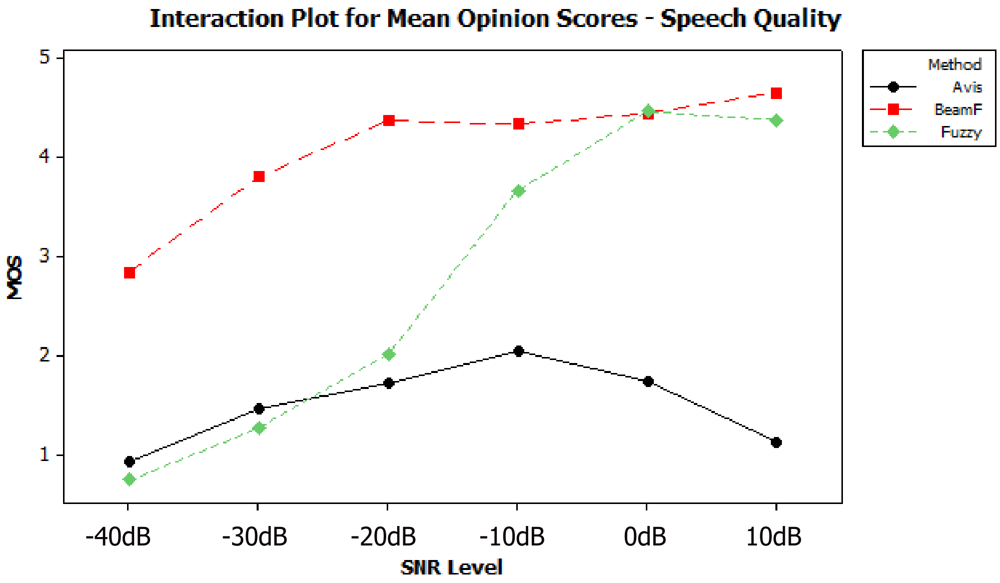


Figure 94: Interaction plot for speech quality MOS at varying SNR levels, showing audiovisual speech (black and circle markers), audio-only beamforming (red with square markers), and fuzzy-based system (green with diamond markers).

Table 46: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audio-only beamforming and Fuzzy Processed Speech for overall subjective scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -2.204 | 0.180 | -12.23 | 0.000 |
| -30dB | -2.526 | 0.180 | -14.02 | 0.000 |
| -20dB | -2.386 | 0.180 | -13.24 | 0.000 |
| -10dB | -0.714 | 0.180 | -3.96 | 0.012 |
| 0dB | -0.384 | 0.180 | -2.13 | 1.000 |
| +10dB | -0.627 | 0.180 | -3.479 | 0.081 |

Table 47: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audiovisual Filtering and Fuzzy Processed Speech for overall subjective scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -0.196 | 0.180 | -1.088 | 1.000 |
| -30dB | -0.046 | 0.180 | -0.255 | 1.000 |
| -20dB | 0.214 | 0.180 | 1.188 | 1.000 |
| -10dB | 1.700 | 0.180 | 9.434 | 0.000 |
| 0dB | 2.550 | 0.180 | 14.151 | 0.000 |
| +10dB | 2.915 | 0.180 | 16.18 | 0.000 |

Table 48: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audio-only beamforming and Fuzzy Processed Speech for speech quality subjective scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -2.088 | 0.169 | -12.34 | 0.000 |
| -30dB | -2.526 | 0.169 | -14.93 | 0.000 |
| -20dB | -2.350 | 0.169 | -13.89 | 0.000 |
| -10dB | -0.668 | 0.169 | -3.95 | 0.013 |
| 0dB | 0.020 | 0.169 | 0.12 | 1.000 |
| +10dB | -0.280 | 0.169 | -1.655 | 1.000 |

Table 49: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audiovisual Filtering and Fuzzy Processed Speech for speech quality subjective scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -0.188 | 0.169 | -1.111 | 1.000 |
| -30dB | -0.196 | 0.169 | -1.158 | 1.000 |
| -20dB | 0.290 | 0.169 | 1.714 | 1.000 |
| -10dB | 1.612 | 0.169 | 9.527 | 0.000 |
| 0dB | 2.720 | 0.169 | 16.076 | 0.000 |
| +10dB | 3.240 | 0.169 | 19.15 | 0.000 |

a comparison of means, comparing the audio-only and fuzzy-based MOS is shown in table 50, with the fuzzy and audiovisual comparison in table 51.

There is one particular feature of interest that is also reflected in the objective tests. Regarding the noise intrusiveness score, table 50 shows that the noise intrusiveness MOS for the fuzzy logic based system at a SNR of +10dB is significantly different (p<0.05), which reflects the increased use of the unfiltered processing option. Although this lower MOS is reflected in the overall score, it is not reflected in the speech distortion comparison of means, suggesting that the addition of noise at this SNR does not have an impact on speech intelligibility.

Overall, these results confirm the validity of the objective test results. The overall, speech distortion, and noise intrusiveness scores are very similar to the objective scores, and the interaction plots and comparisons of means confirm that the audio-only approach significantly outperforms the audiovisual approach. They also confirm that the fuzzy-based system performs as expected. At lower SNR levels -40dB to -20dB), the MOS is very similar to the audiovisual MOS, with small but not significant differences, as shown by the results of a comparison of means. At SNR levels of -10dB and 0dB, the audio-only and fuzzy-based results are very similar, suggesting that audio-only processing is used more often. At a high SNR, the significantly different noise intrusiveness score between audio-only and fuzzy-based scores shows that the system is making use of some unfiltered data. However, the results also show

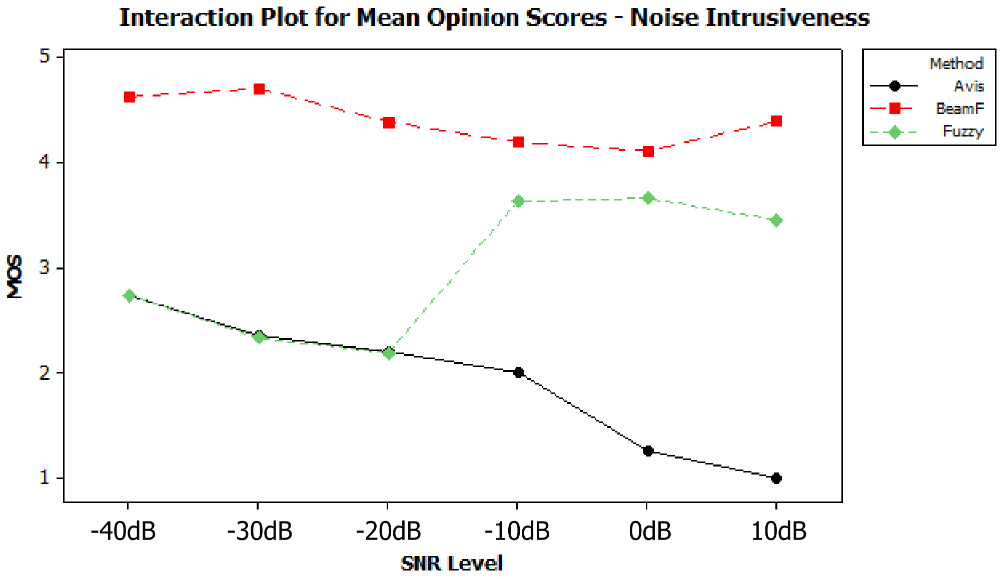


Figure 95: Interaction plot for noise intrusiveness MOS at varying SNR levels, showing audiovisual speech (black and circle markers), audio-only beamforming (red with square markers), and fuzzy-based system (green with diamond markers).

Table 50: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audio-only beamforming and Fuzzy Processed Speech for noise intrusiveness subjective scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -1.880 | 0.213 | -8.82 | 0.000 |
| -30dB | -2.364 | 0.213 | -11.09 | 0.000 |
| -20dB | -2.190 | 0.213 | -10.28 | 0.000 |
| -10dB | -0.560 | 0.213 | -2.63 | 1.000 |
| 0dB | -0.450 | 0.213 | -2.11 | 1.000 |
| +10dB | -0.940 | 0.213 | -4.410 | 0.002 |

Table 51: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audiovisual Filtering and Fuzzy Processed Speech for noise intrusiveness subjective scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 0.010 | 0.213 | 0.047 | 1.000 |
| -30dB | -0.014 | 0.213 | -0.066 | 1.000 |
| -20dB | -0.020 | 0.213 | -0.094 | 1.000 |
| -10dB | 1.625 | 0.213 | 7.624 | 0.000 |
| 0dB | 2.400 | 0.213 | 11.261 | 0.000 |
| +10dB | 2.450 | 0.213 | 11.50 | 0.000 |

that similarly to the objective results in the previous section, the audiovisual MOS is the worst performing technique, and the audio-only approach far outperforms this method. However, these results should be interpreted with a degree of caution. As discussed in this chapter, the audio-only beamforming was expected to perform well, as the simulated room environment is designed specifically to demonstrate the performance of this technique. Section 7.7.6 presents the results of objective tests of speech mixed with a different noise, one designed to challenge the audio-only approach.

### 7.7.6 *Objective Testing with Inconsistent Transient Noise*

The objective and subjective testing in the previous sections (sections 7.7.4, and 7.7.5) identified that the audio-only beamforming approach produced the strongest results. As expected, the audiovisual approach performed poorly when tested with novel data that it had not been trained with, and the fuzzy logic approach produced output that resulted in a poorer score than the audio-only approach due to the fuzzy switching system. However, as mentioned previously, there should be a degree of caution in interpreting these results. Firstly, although the output audio quality for the fuzzy logic processing approach produces lower objective and subjective scores, this is due to limitations with the audiovisual processing approach,

Table 52: Composite objective mean test score table for overall speech quality for speech with transient clapping noise added, for audiovisual speech, audio-only beamforming, fuzzy-based processing, audio-only spectral subtraction, and unprocessed speech.

| Level | Avis | Beamforming | Fuzzy | Spectral | Noisy |
|---|---|---|---|---|---|
| -40dB | 1.016 | 0.194 | 1.018 | -0.208 | 0.367 |
| -30dB | 1.072 | 0.194 | 0.609 | -0.067 | 0.443 |
| -20dB | 0.937 | 0.194 | 0.557 | 0.026 | 0.632 |
| -10dB | 0.833 | 0.194 | 0.500 | 0.631 | 1.295 |
| 0dB | 0.939 | 0.194 | 1.159 | 1.421 | 2.003 |
| +10dB | 0.710 | 0.194 | 0.655 | 2.302 | 2.695 |

rather than with the fuzzy switching. Secondly, although the audio-only results have been identified as producing the strongest results, this is in a scenario with broadband noise from a fixed source, where a beamformer would be expected to perform well.

In this section, a different noise is used, one with silence and clapping, that represents a greater challenge. A mixture of clapping and silence is used as the noise source (as shown in figure 83), and the 10 speech sentences described above are mixed with the noise source at a range of SNR levels, from -40dB to +10dB. These noisy sentences are then processed using the techniques also used in section 7.7.4. The objective composite measures used in section 7.7.4 are also used to evaluate the filtered speech sentences.

The means of the composite overall, speech distortion, and background distortion at different SNR levels are provided in tables 52, 53, 54, and figures 96, 97, 98 respectively.

Considering the overall scores first, it can be seen that the audio-only beamformer returns the same score at all SNR levels. This is shown in the interaction plot in figure 99. Listening to the audio output confirmed that the reason this score was so low and so consistent was because no audio signal was returned. The audiovisual score was also poor, but listening to the output confirmed that an audio signal could be heard, hence the higher score. However, as can be seen in figure 99, the overall score is still very low, as expected. As discussed in the previous sections, the audiovisual approach performs poorly with novel data. The results

Table 53: Composite objective mean test score table for speech score speech quality for speech with transient clapping noise added, for audiovisual speech, audio-only beamforming, fuzzy-based processing, audio-only spectral subtraction, and unprocessed speech.

| Level | Avis | Beamforming | Fuzzy | Spectral | Noisy |
|---|---|---|---|---|---|
| -40dB | 1.039 | -0.446 | 0.292 | -0.749 | 0.049 |
| -30dB | 1.018 | -0.446 | -0.146 | -0.598 | 0.153 |
| -20dB | 0.763 | -0.446 | -0.233 | -0.349 | 0.440 |
| -10dB | 0.446 | -0.446 | -0.279 | 0.486 | 1.308 |
| 0dB | 0.439 | -0.446 | 0.365 | 1.548 | 2.280 |
| +10dB | -0.515 | -0.446 | 0.105 | 2.640 | 3.200 |

Table 54: Composite objective mean test score table for noisy speech quality for speech with transient clapping noise added, for audiovisual speech, audio-only beamforming, fuzzy-based processing, audio-only spectral subtraction, and unprocessed speech.

| Level | Avis | Beamforming | Fuzzy | Spectral | Noisy |
|---|---|---|---|---|---|
| -40dB | 1.617 | 1.700 | 2.017 | 0.984 | 1.203 |
| -30dB | 1.667 | 1.700 | 1.830 | 1.062 | 1.239 |
| -20dB | 1.593 | 1.700 | 1.802 | 1.080 | 1.315 |
| -10dB | 1.593 | 1.700 | 1.830 | 1.319 | 1.575 |
| 0dB | 1.606 | 1.700 | 2.199 | 1.649 | 1.859 |
| +10dB | 1.730 | 1.700 | 1.814 | 2.100 | 2.223 |

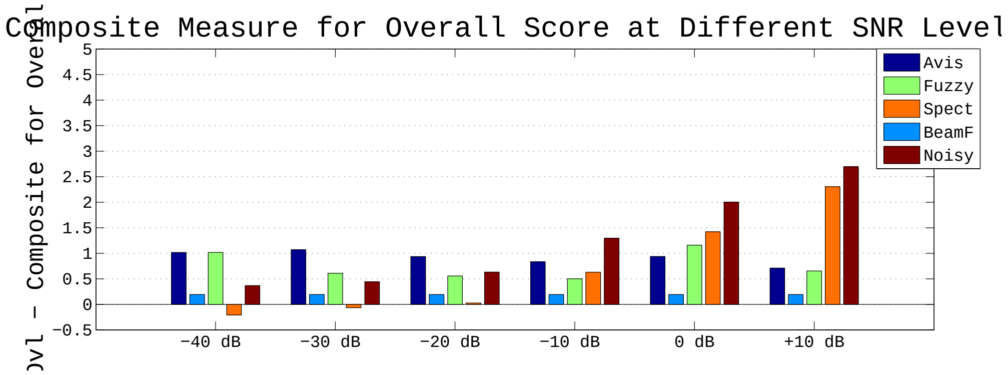


Figure 96: Composite objective mean test scores for overall speech quality for speech with transient clapping noise added, for audiovisual speech, audio-only beamforming, fuzzy-based processing, audio-only spectral subtraction, and unprocessed speech.

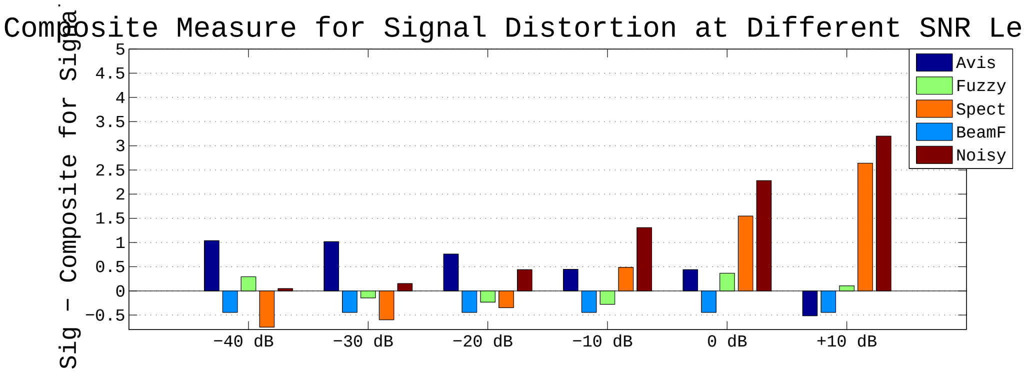


Figure 97: Composite objective mean test scores for speech distortion level for speech with transient clapping noise added, for audiovisual speech, audio-only beamforming, fuzzy-based processing, audio-only spectral subtraction, and unprocessed speech.

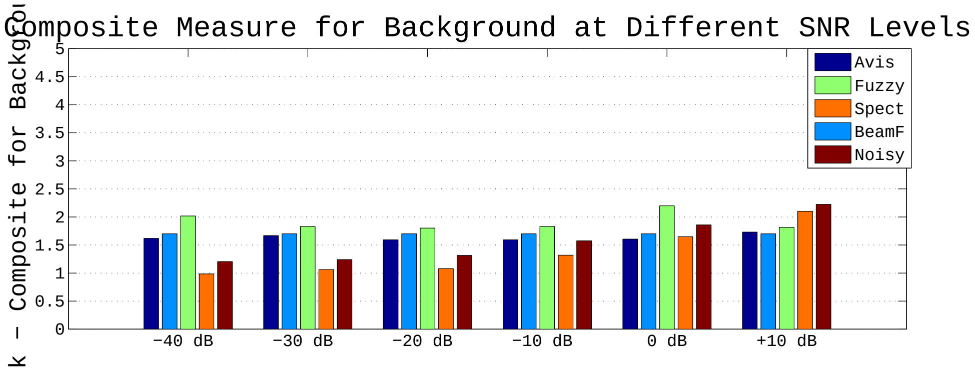


Figure 98: Composite objective mean test scores for noise distortion level for speech with transient clapping noise added, for audiovisual speech, audio-only beamforming, fuzzy-based processing, audio-only spectral subtraction, and unprocessed speech.

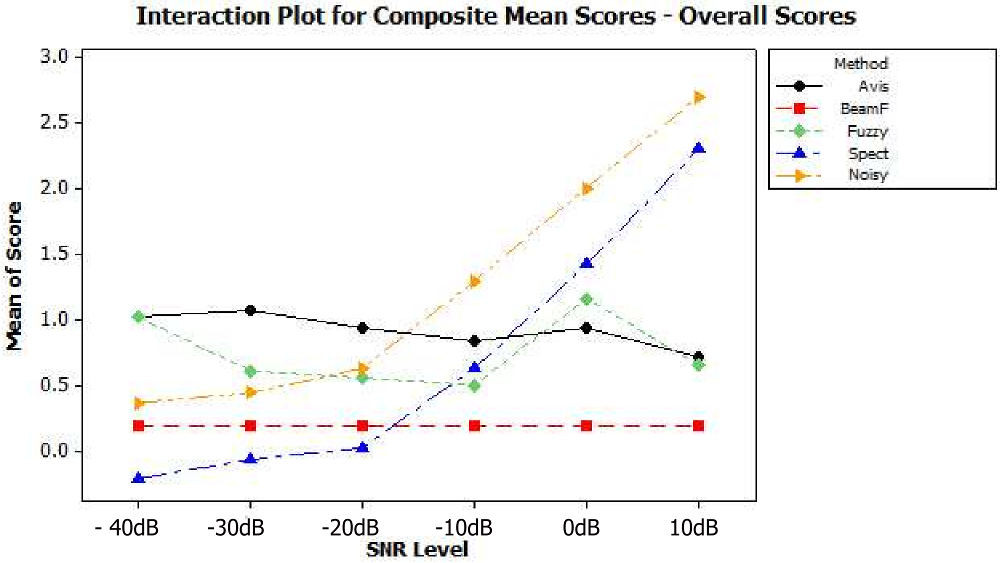


Figure 99: Interaction plot for overall composite objective mean score of speech with transient clapping noise added at varying SNR levels, showing audiovisual speech (black and circle markers), audio-only beamforming (red with square markers), fuzzy-based system (green with diamond markers), spectral subtraction (blue with triangles), and unfiltered noisy speech (orange with triangles).

of Bonferroni multiple comparison, as shown in tables 55 and 56 show that despite the lack of output signal, the difference between the fuzzy output and the audio-only output is only significant at a SNR of -40dB, and 0dB (where $p<0.05$). The difference between the audiovisual and fuzzy output scores was not significant at any SNR level.

This low score is clearly reflected in the speech distortion composite scores, as shown in figure 97 and also in the interaction plot in figure 100. It can be seen that the scores are extremely low, confirming that there was consistently no speech identified in the filtered output value. Despite the high noise level of the input signal (when audiovisual processing would be expected to be used), the results of Bonferroni multiple comparison in table 55 showed that the difference was not significant ($p>0.05$) at most SNR levels.

Again, similar results were found for the noise composite scores, as shown in the interaction plot in figure 101 and the results of Bonferroni multiple comparison in tables 59 and 60.

Overall, the results demonstrated that the audio-only beamforming results presented in the previous sections should be interpreted with a degree of caution. The fuzzy logic based results presented in this section are very dependent on the techniques used for processing speech.

Table 55: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audio-only beamforming and Fuzzy Processed Speech with transient clapping noise added for overall composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 0.824 | 0.148 | 5.585 | 0.000 |
| -30dB | 0.416 | 0.148 | 2.816 | 1.000 |
| -20dB | 0.363 | 0.148 | 2.462 | 1.000 |
| -10dB | 0.307 | 0.148 | 2.079 | 1.000 |
| 0dB | 0.965 | 0.148 | 6.543 | 0.000 |
| +10dB | 0.461 | 0.148 | 3.124 | 0.905 |

Table 56: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audiovisual Filtering and Fuzzy Processed Speech with transient clapping noise added for overall composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 0.002 | 0.148 | 0.011 | 1.000 |
| -30dB | -0.463 | 0.148 | -3.138 | 0.864 |
| -20dB | -0.380 | 0.148 | -2.577 | 1.000 |
| -10dB | -0.333 | 0.148 | -2.254 | 1.000 |
| 0dB | 0.220 | 0.148 | 1.490 | 1.000 |
| +10dB | -0.056 | 0.148 | -0.379 | 1.000 |

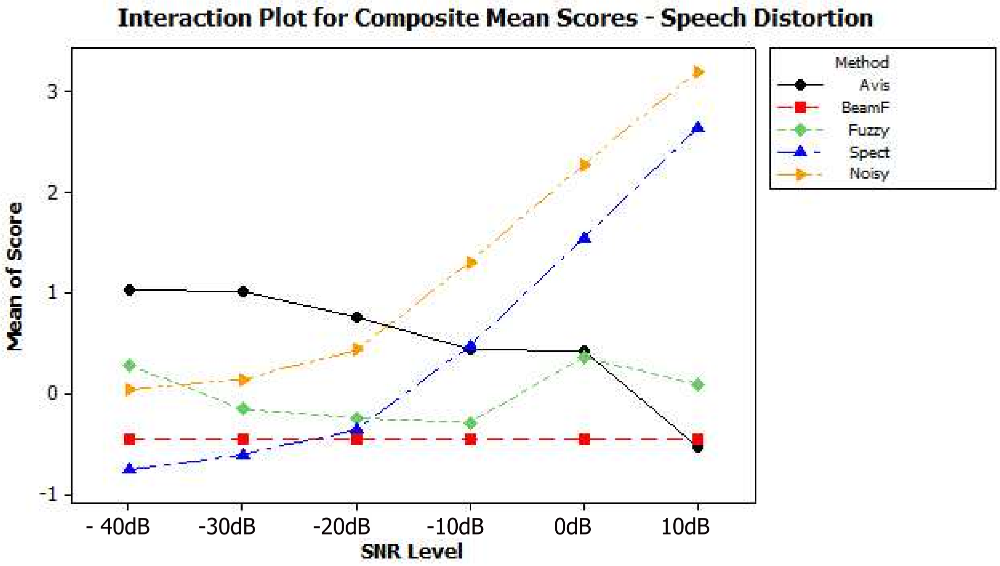


Figure 100: Interaction plot for speech distortion composite objective mean score of speech with transient clapping noise added at varying SNR levels, showing audiovisual speech (black and circle markers), audio-only beamforming (red with square markers), fuzzy-based system (green with diamond markers), spectral subtraction (blue with triangles), and unfiltered noisy speech (orange with triangles).

Table 57: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audio-only beamforming and Fuzzy Processed Speech with transient clapping noise added for speech distortion composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 0.738 | 0.152 | 4.856 | 0.001 |
| -30dB | 0.300 | 0.152 | 1.973 | 1.000 |
| -20dB | 0.213 | 0.152 | 1.399 | 1.000 |
| -10dB | 0.167 | 0.152 | 1.099 | 1.000 |
| 0dB | 0.811 | 0.152 | 5.336 | 0.000 |
| +10dB | 0.550 | 0.152 | 3.621 | 0.166 |

Table 58: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audiovisual Filtering and Fuzzy Processed Speech with transient clapping noise added for speech distortion composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | -0.746 | 0.152 | -4.91 | 0.001 |
| -30dB | -1.164 | 0.152 | -7.66 | 0.000 |
| -20dB | -0.996 | 0.152 | -6.553 | 0.000 |
| -10dB | -0.725 | 0.152 | -4.767 | 0.002 |
| 0dB | -0.074 | 0.152 | -0.485 | 1.0000 |
| +10dB | 0.620 | 0.152 | 4.078 | 0.030 |

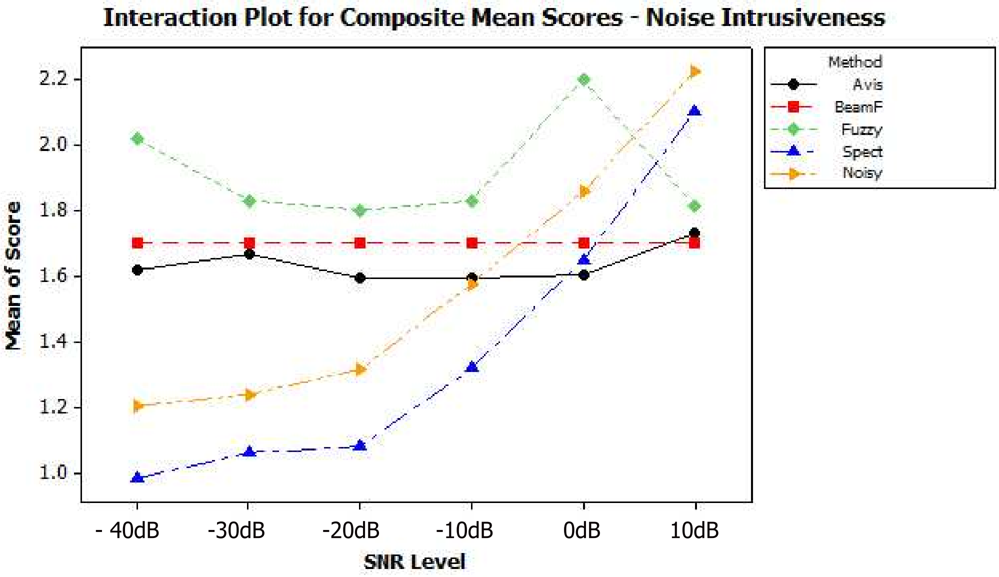


Figure 101: Interaction plot for noise intrusiveness composite objective mean score of speech with transient clapping noise added at varying SNR levels, showing audiovisual speech (black and circle markers), audio-only beamforming (red with square markers), fuzzy-based system (green with diamond markers), spectral subtraction (blue with triangles), and unfiltered noisy speech (orange with triangles).

Table 59: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audio-only beamforming and Fuzzy Processed Speech with transient clapping noise added for noise intrusiveness composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 0.316 | 0.093 | 3.404 | 0.356 |
| -30dB | 0.130 | 0.093 | 1.401 | 1.000 |
| -20dB | 0.101 | 0.093 | 1.089 | 1.000 |
| -10dB | 0.130 | 0.093 | 1.398 | 1.000 |
| 0dB | 0.499 | 0.093 | 5.368 | 0.000 |
| +10dB | 0.114 | 0.093 | 1.227 | 1.000 |

Table 60: Selected results of Bonferroni Multiple Comparison, showing P-Value results for difference between Audiovisual Filtering and Fuzzy Processed Speech with transient clapping noise added for noise intrusiveness composite scores.

| Level | Difference of Means | SE of Difference | T-Value | Adjusted P-Value |
|---|---|---|---|---|
| -40dB | 0.400 | 0.093 | 4.300 | 0.012 |
| -30dB | 0.163 | 0.093 | 1.758 | 1.000 |
| -20dB | 0.209 | 0.093 | 2.245 | 1.000 |
| -10dB | 0.237 | 0.093 | 2.549 | 1.000 |
| 0dB | 0.594 | 0.093 | 6.386 | 0.000 |
| +10dB | 0.085 | 0.093 | 0.909 | 1.000 |

Although previous sections reported that the audio-only approach produced clearly better results, this was when the noise was one which the beamformer was capable of processing. Likewise, the audiovisual results were shown to be limited due to the system not being trained with data similar to that used for testing. Therefore, although the fuzzy logic system is functioning as expected and is switching between techniques, the results are limited by limitations in the specific speech processing techniques. In addition to an evaluation of the output audio signal, the next section presents an investigation of the specific fuzzy switching performance.

### 7.7.7 *Detailed Fuzzy Switching Performance*

In addition to an analysis of the audio output, it is of interest to assess the performance of the fuzzy switching system. As discussed previously in this chapter, it can be seen that the fuzzy logic output varies depending on factors such as the SNR level and the previous output decision value, and the results of subjective and objective tests show that the output mean scores are often similar, but not identical to either the audio-only output scores or the audiovisual scores. However, as a range of sentences (with different associated visual quality fuzzy values), noises, and SNR levels were tested, it was felt suitable to examine the performance of the fuzzy switching approach in detail. This section reports the results of a detailed inspection of the fuzzy logic switching system.

#### *Fuzzy Switching with Varying Noise Type*

Firstly, the difference between sentences mixed with the two different noises used in this chapter is examined. To do this, two sentences are compared, with different noise added. The fuzzy output decision from frame-to-frame of a sentence with transient noise is compared to the frame-by-frame output decision of the same sentence, except with the machine noise added at the same SNR. Firstly, noise was added at a SNR of 0dB to the sentence, and the output is shown in figure 102. In order to ensure that good quality visual information was available at all times, an example of a sentence from the reading task was chosen.

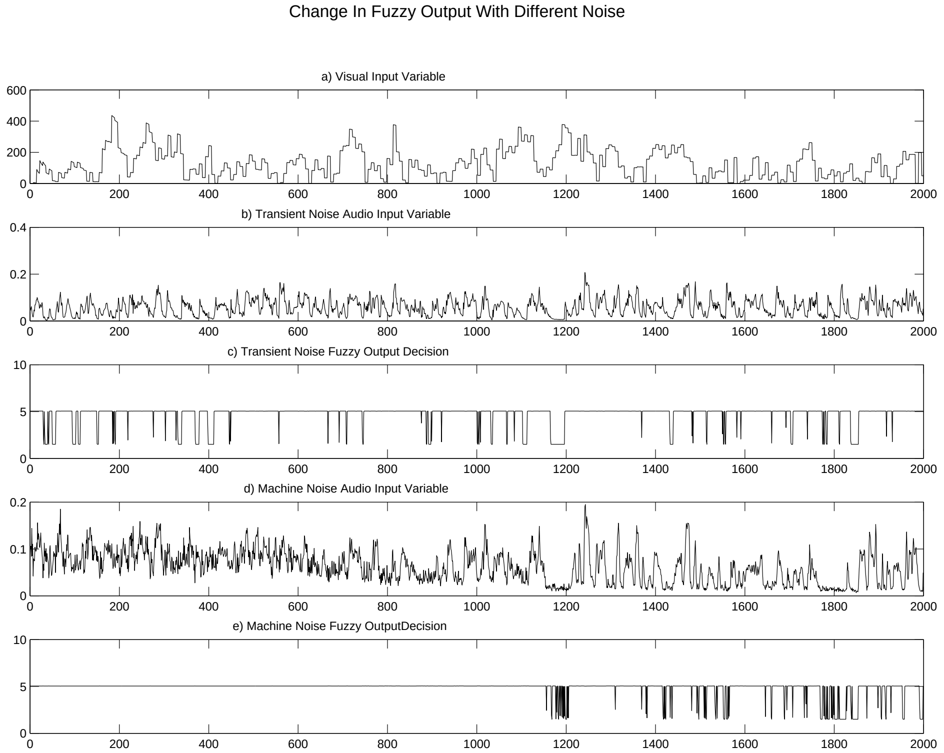


Figure 102: Comparison of fuzzy logic output decision depending on noise type at SNR of odB. (a) shows the input visual information. It can be seen that all values are below 600, therefore every frame is considered to be good quality. As the visual information is unchanged, then this is the same for both transient and machine noise speech mixtures. (b) shows the transient mixture fuzzy input variable. (c) shows the associated transient noise mixture output processing decision. (d) shows the machine noise mixture fuzzy input variable. (e) shows the machine noise mixture output processing decision.

Figure 102 shows the difference in the fuzzy output decision, depending on the input noise variable. As the visual information, SNR, and sentence content was the same for both values, the only difference was the noise type. In figure 102, (c) Shows the fuzzy output decision, based on the visual input variable in (a), and the audio input variable in (b). It can be seen that the noise is of a relatively low level, and so the system alternates between making use of the audio-only and the unprocessed speech options, which is to be expected when it is considered that this noise consists of handclaps and silences. (e) Shows the fuzzy output decision, based on the visual information in variable (a), and the audio input in variable (d). It can be seen that the fuzzy decision is different, as the noise input variable is different. The machine noise is a broadband noise, and so there is more noise present. As can be seen in figure 82, the noise amplitude gradually decreases over time, and this is reflected in the fuzzy output, which uses the audio-only output decision consistently, and as the noise level decreases, the unfiltered output is chosen on some occasions. This is in line with expectations and shows that the system is performing as expected with different noise types. To confirm this, the same sentence and noises are compared again in figure 103, except with the speech and noise mixed at a SNR of -20dB.

Again, the key information is shown in the fuzzy output decisions in (c) and (e) of figure 103. With the transient noise, it can be seen in (c) that there are two large quiet periods, which are also shown in figure 83. In these periods, either the unfiltered or audio-only options are chosen, otherwise, the audiovisual output is chosen as expected. In (e), although the noise is gradually decreasing as shown in (d), as the SNR is very low the audiovisual output is chosen in all frames.

In summary, it can be seen that the fuzzy output decision varies based purely on the noise type. Figures 102 and 103 show that when the same speech sentence with the same quality of visual information is mixed with noise at the same SNR, with the only difference being the type of noise, the frame-by-frame fuzzy output decision is different. This demonstrates that the fuzzy-based system is capable of adapting to different noise types.

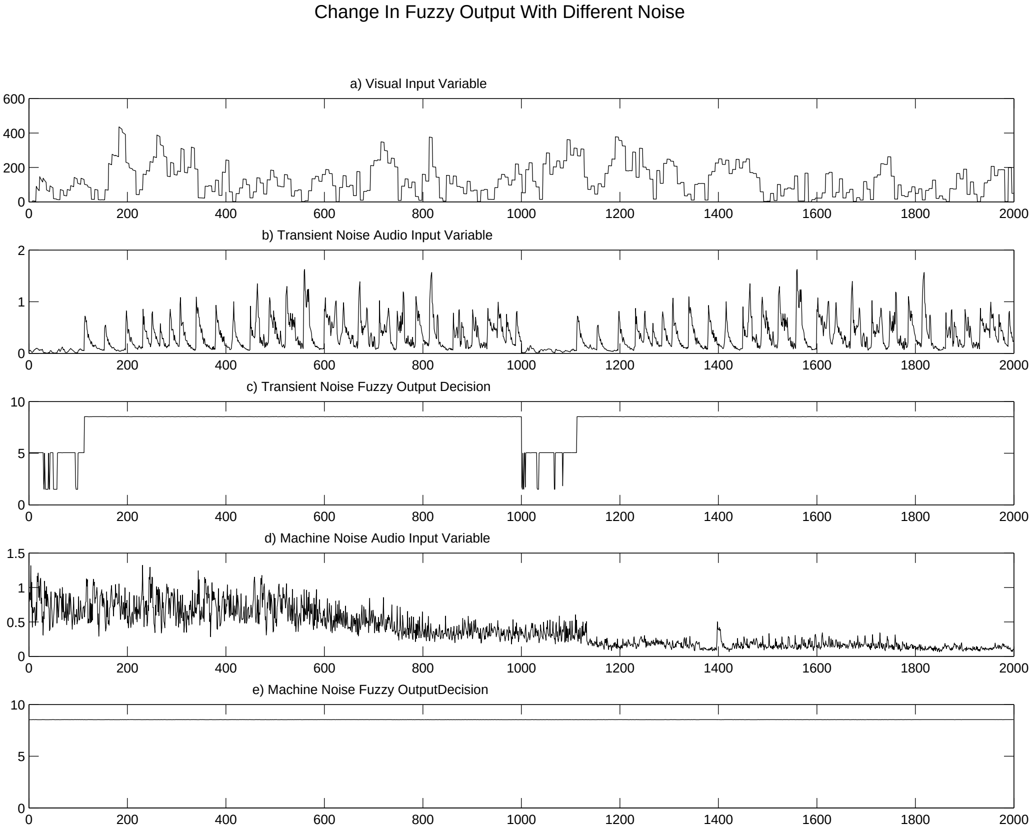


Figure 103: Comparison of fuzzy logic output decision depending on noise type at SNR of -20dB. (a) shows the input visual information. It can be seen that all values are below 600, therefore every frame is considered to be good quality. As the visual information is unchanged, then this is the same for both transient and machine noise speech mixtures. (b) shows the transient mixture fuzzy input variable. (c) shows the associated transient noise mixture output processing decision. (d) shows the machine noise mixture fuzzy input variable. (e) shows the machine noise mixture output processing decision.

### *Fuzzy Switching with Varying Visual Information*

The previous examples considered a sentence with good quality visual information available at all times, but it was also considered to be of interest to observe the effect that varying the quality of visual information had on the fuzzy decision. As discussed in chapter 6, if the audio input level was considered to be high, then the fuzzy logic system would use audiovisual processing, but only if the visual information was considered to be of good quality (i.e. the visual input fuzzy variable was low). To test this, a number of different sentences are compared, and the fuzzy outputs compared. These are shown in figures 104, 105, 106, and 107. In all sentences, machine noise is mixed with the speech signal at a SNR of -30dB to ensure consistency, and provide a noise where audiovisual processing would be expected to be chosen for all frames if good quality visual information is available. Figure 104 shows an example of a sentence with good quality visual information available at all frames.

In figure 104 , (a) represents the visual input variable, (b) represents the audio input variable and (c) shows the fuzzy output decision. As can be seen, the visual information quality is considered to be good for all frames, and so audiovisual processing is chosen at all frames. However, figures 105, 106, and 107 show different sentences with all other conditions kept the same.

Again, in 105, 106, and 107, (a) represents the visual input variable, (b) represents the audio input variable, and (c) shows the fuzzy output decision. It can be seen that despite the noise type and SNR being the same in each figure, the visual input variable (which was shown to be accurate in section 7.6) varies, and so the system only uses audiovisual processing when it is considered to be suitable. This demonstrates that the system only uses audiovisual information when it is considered to be appropriate, and adapts to different sentences.

### *Fuzzy Switching with Varying SNR Level*

In addition to considering the effect of noise type and visual information, the effect of mixing the speech and noise sources at varying SNR levels is of interest. For this example, one sentence was chosen, with a small number of frames with poor quality visual information, and the

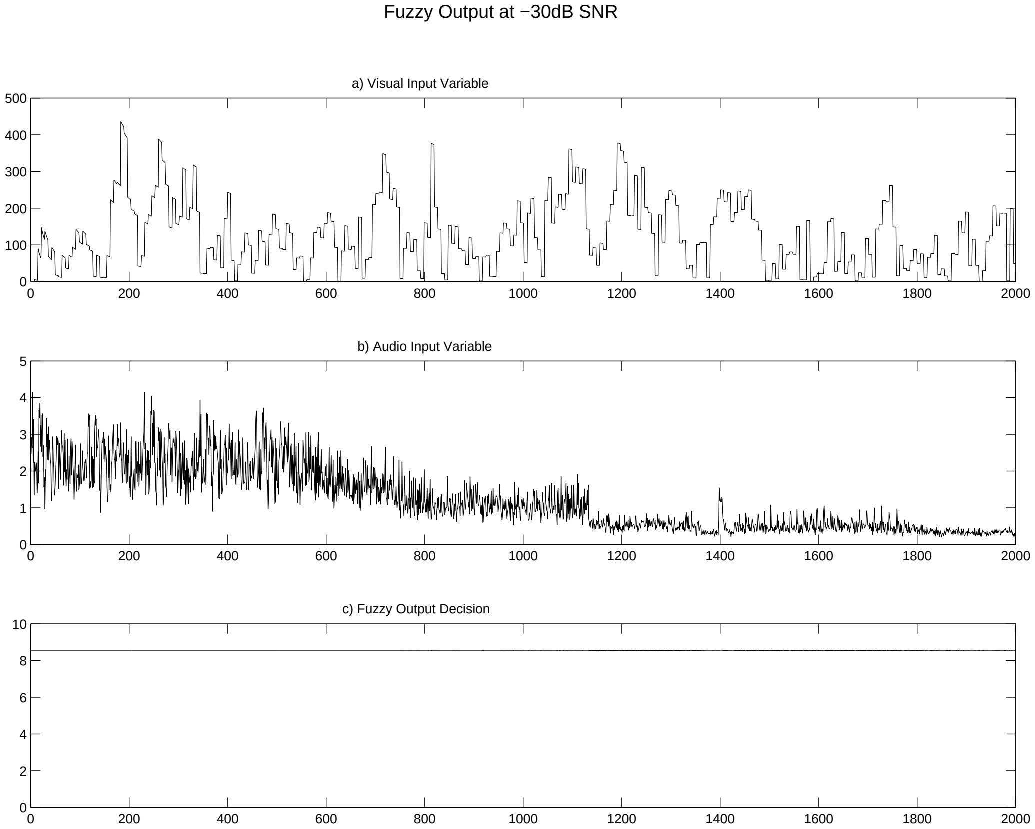


Figure 104: fuzzy logic output decision depending on quality of visual information, for sentence with no frames considered to be of poor quality. (a) shows the input visual variable. It can be seen that all values are below 600, therefore every frame is considered to be good quality. (b) shows the audio input variable, with machine noise added to speech at an SNR of -30dB. (c) shows the fuzzy output processing decision.

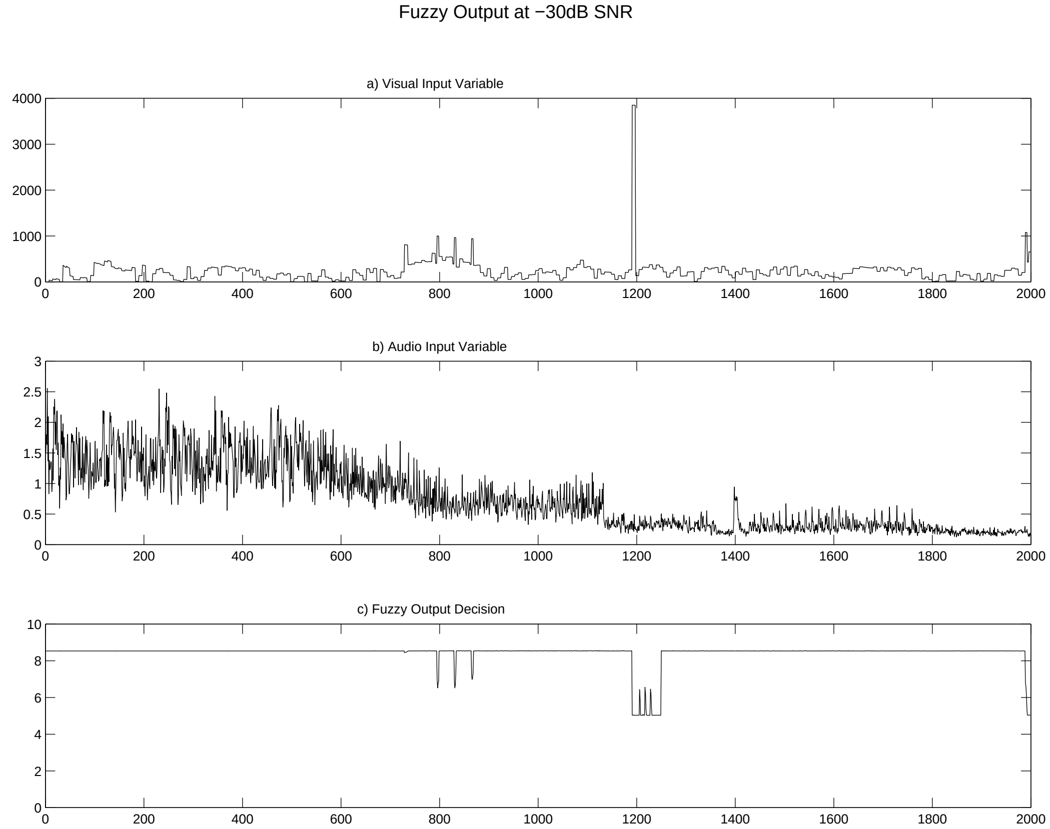


Figure 105: fuzzy logic output decision depending on quality of visual information, for sentence with several frames considered to be of poor quality. (a) shows the input visual variable. It can be seen that there are a small number of frames where there is considered to be poor visual input. (b) shows the audio input variable, with machine noise added to speech at an SNR of -30dB. (c) shows the fuzzy output processing decision.

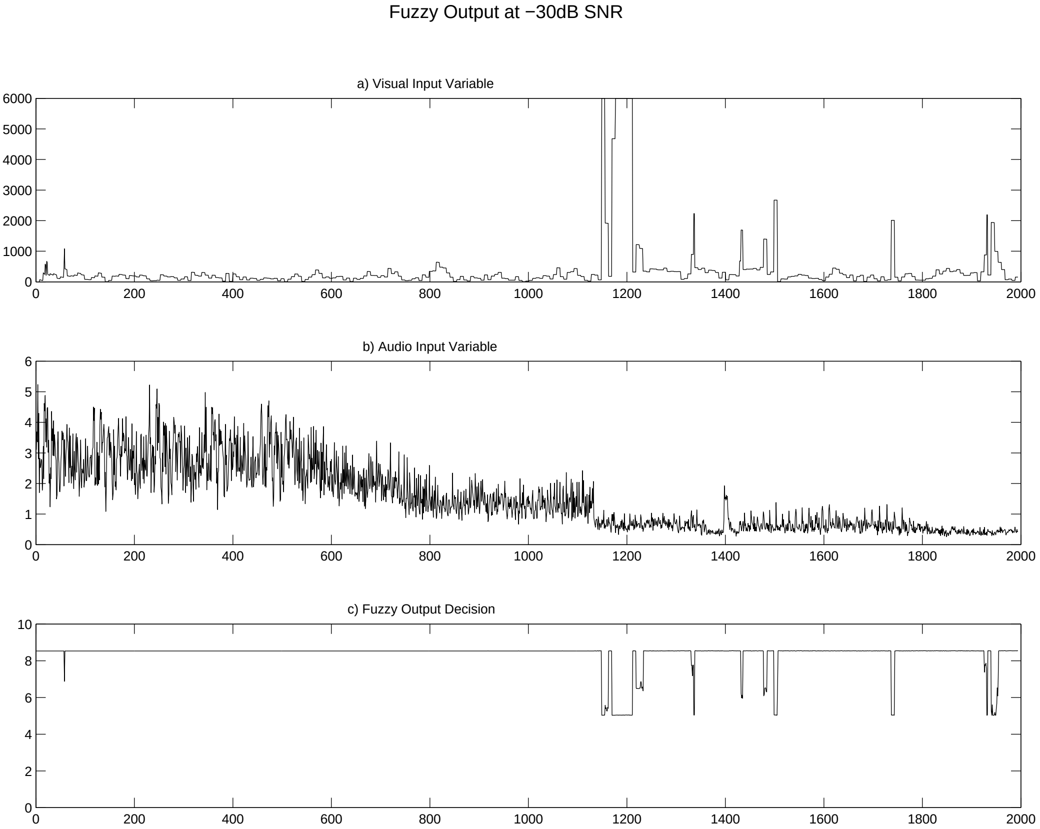


Figure 106: fuzzy logic output decision depending on quality of visual information, for sentence with several frames considered to be of poor quality. (a) shows the input visual variable. It can be seen that there are a number of frames where there is considered to be poor visual input. (b) shows the audio input variable, with machine noise added to speech at an SNR of -30dB. (c) shows the fuzzy output processing decision.

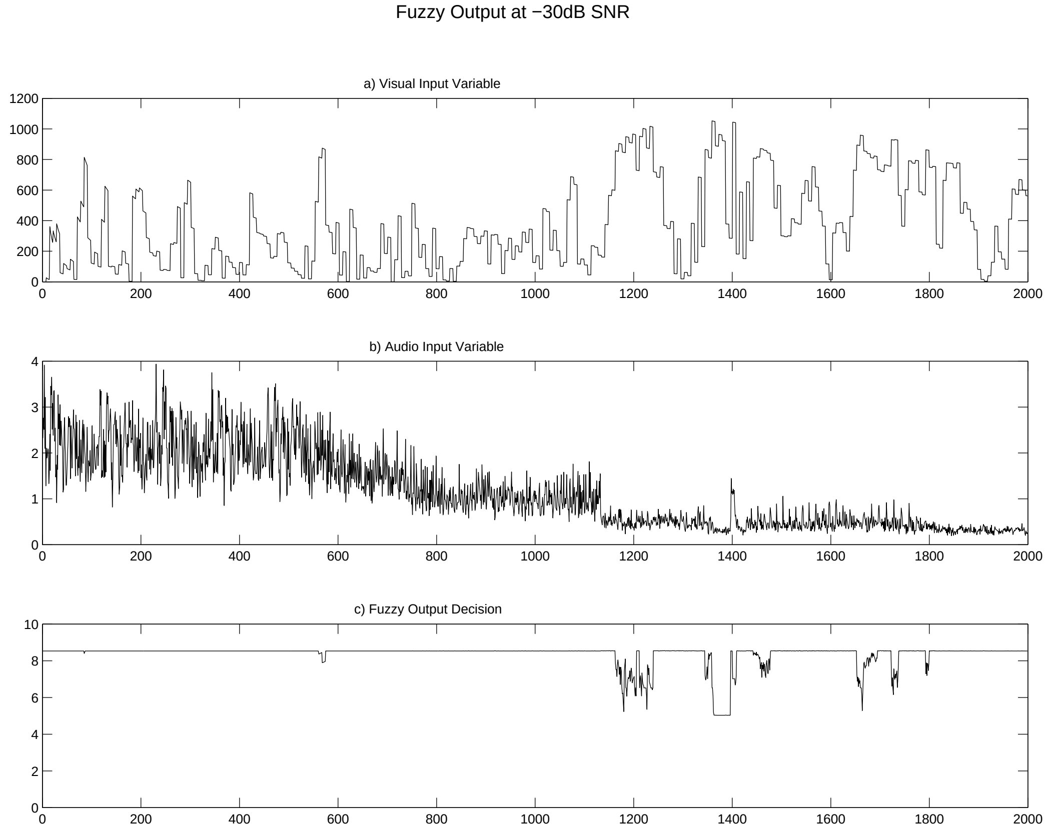


Figure 107: fuzzy logic output decision depending on quality of visual information, for sentence with several frames considered to be of poor quality. (a) shows the input visual variable. It can be seen that there are a small number of frames where there is considered to be poor visual input. (b) shows the audio input variable, with machine noise added to speech at an SNR of -30dB. (c) shows the fuzzy output processing decision.

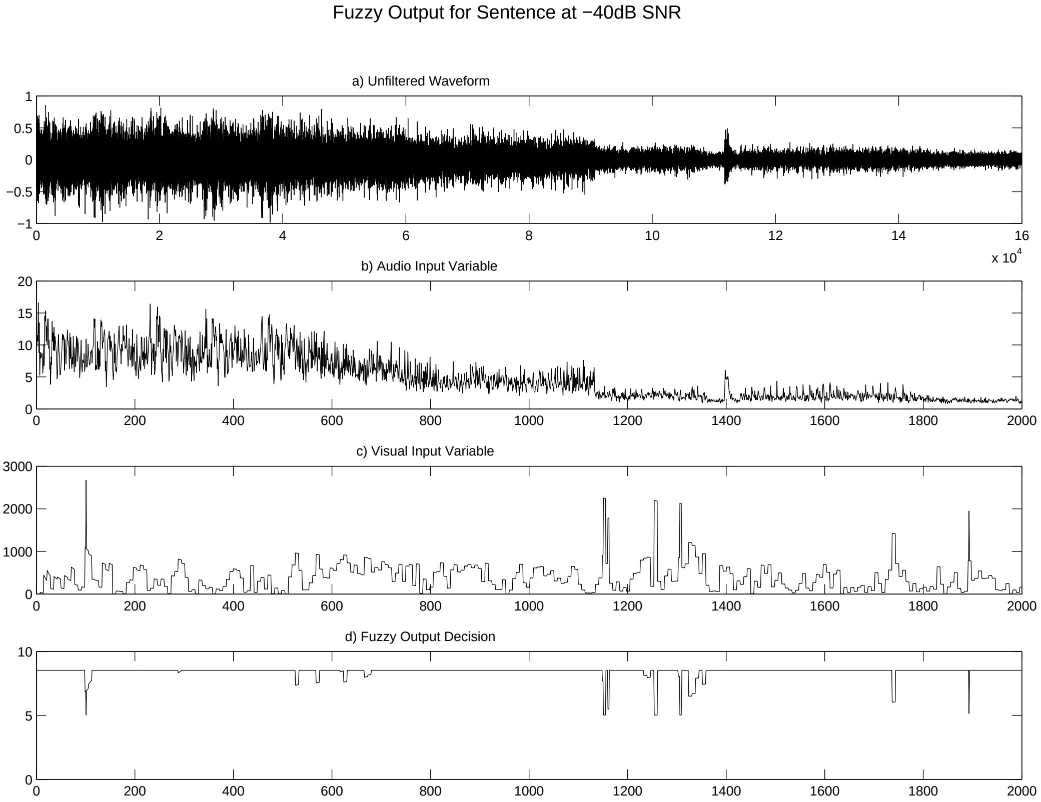


Figure 108: fuzzy logic output decision depending on SNR level. (a) shows the input audio waveform, with speech and noise mixed at a SNR of -40dB. (b) shows the audio input variable. (c) shows the visual input variable, with a small number of frames considered to be of low quality. Finally, (d) shows the fuzzy output decision.

noise source was the broadband machine noise. The sentences were then mixed at different SNR levels, varying from -40dB to +10dB. Figure 108 shows the effect of mixing the sources at an SNR of -40dB.

In figure 108, (a) represents the mixed audio waveform, and (b) the associated fuzzy input variable. (c) Shows the visual input variable and (d) shows the fuzzy processing decision output. It can be seen that as the noise is considered to be consistently high, audiovisual information is used whenever good quality visual information is available. This is very similar to the output for the same sentence, but with a SNR of -30dB. This is shown in figure 109.

Figure 109 shows that the output decision (d), is very similar to figure 108. This is to be expected. If the two figures are compared closely, it can be seen that there is a slight difference in that slightly more frames result in an audio-only processing decision, but this

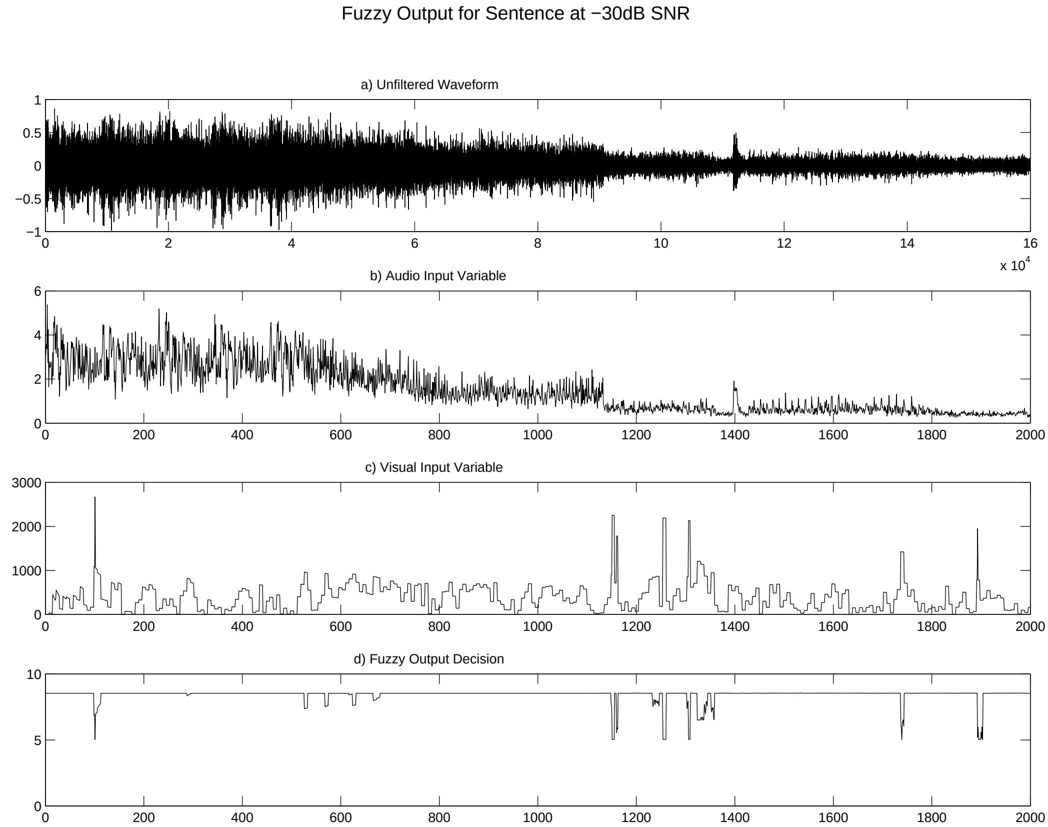


Figure 109: fuzzy logic output decision depending on SNR level. (a) shows the input audio waveform, with speech and noise mixed at a SNR of -30dB. (b) shows the audio input variable. (c) shows the visual input variable, with a small number of frames considered to be of low quality. Finally, (d) shows the fuzzy output decision.

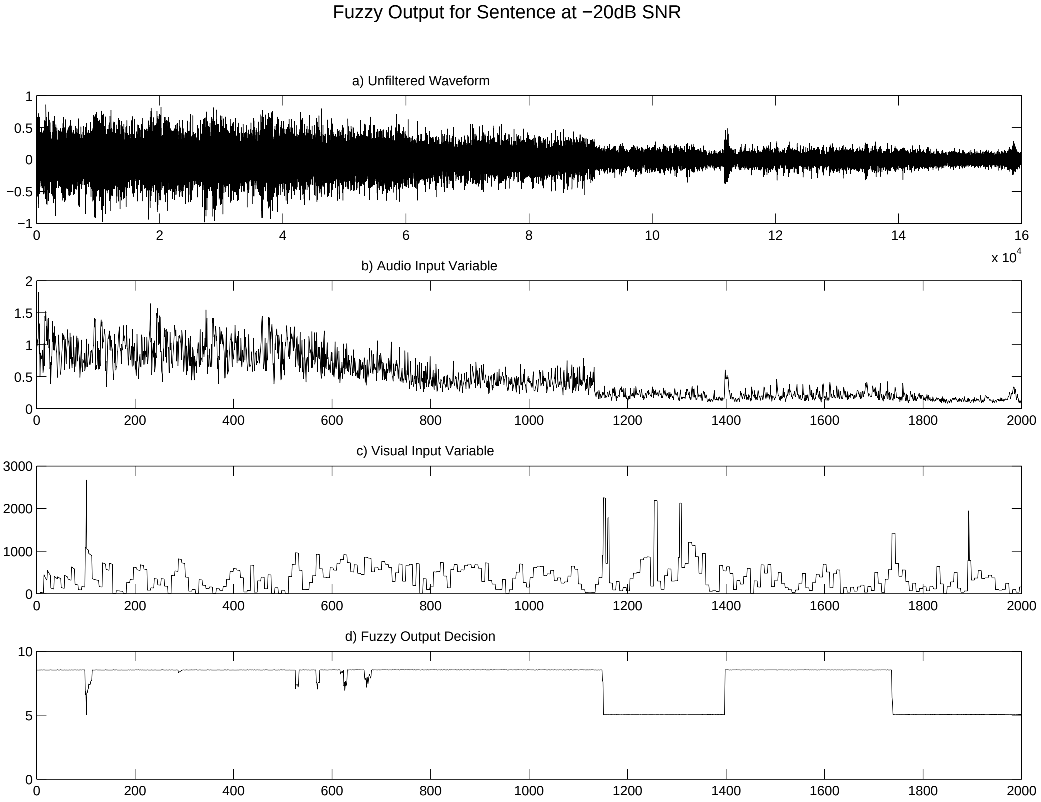


Figure 110: fuzzy logic output decision depending on SNR level. (a) shows the input audio waveform, with speech and noise mixed at a SNR of -20dB. (b) shows the audio input variable. (c) shows the visual input variable, with a small number of frames considered to be of low quality. Finally, (d) shows the fuzzy output decision.

is an extremely small difference. In figure 109, which shows the same sentence and noise mixture, but at a SNR of -20dB, there is a much more noticeable difference.

Again, in figure 109, (a) represents the mixed audio waveform, and (b) the associated fuzzy input variable. (c) Shows the visual input variable and (d) shows the fuzzy processing decision output. It can be seen that initially, the audiovisual processing option is chosen where appropriate. Later in this sentence though, when there is considered to be lower quality visual information available, the system chooses audio-only processing. Unlike figure 109, the decision does not quickly change back to audiovisual processing, but continues to choose audio-only processing for a much greater number of frames. This is because of the increased SNR, demonstrating that the fuzzy logic system adapts to different noise inputs. This adaptability is also shown in figure 111.

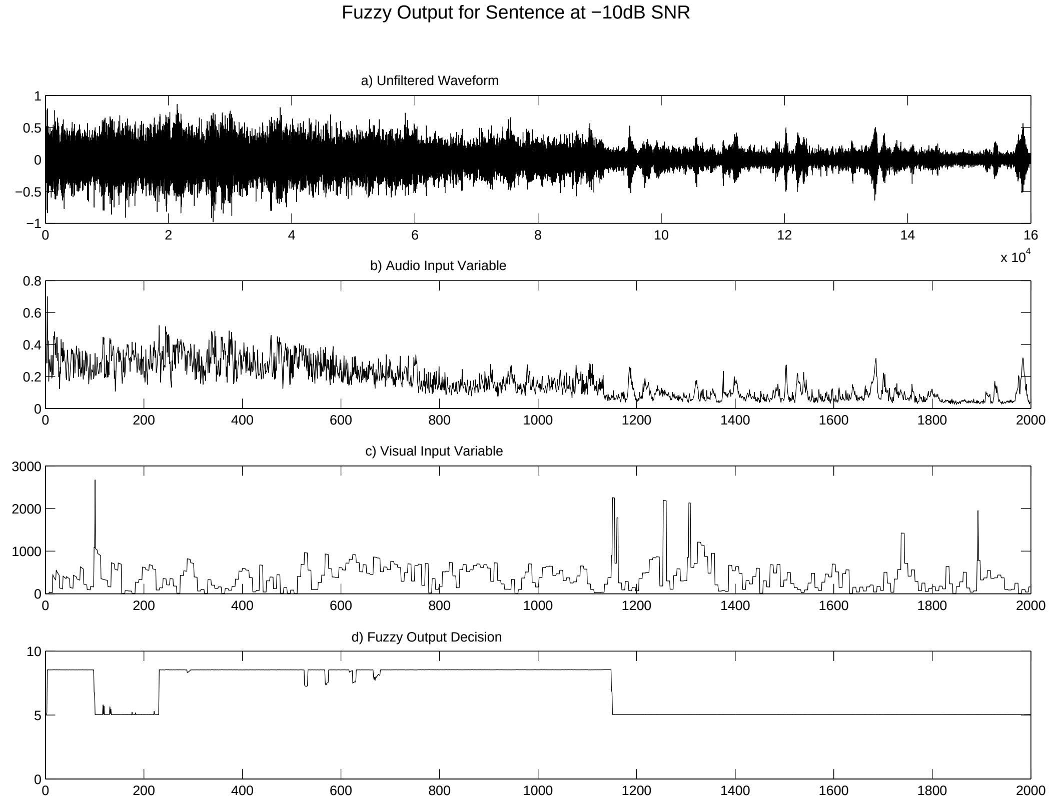


Figure 111: fuzzy logic output decision depending on SNR level. (a) shows the input audio waveform, with speech and noise mixed at a SNR of -10dB. (b) shows the audio input variable. (c) shows the visual input variable, with a small number of frames considered to be of low quality. Finally, (d) shows the fuzzy output decision.

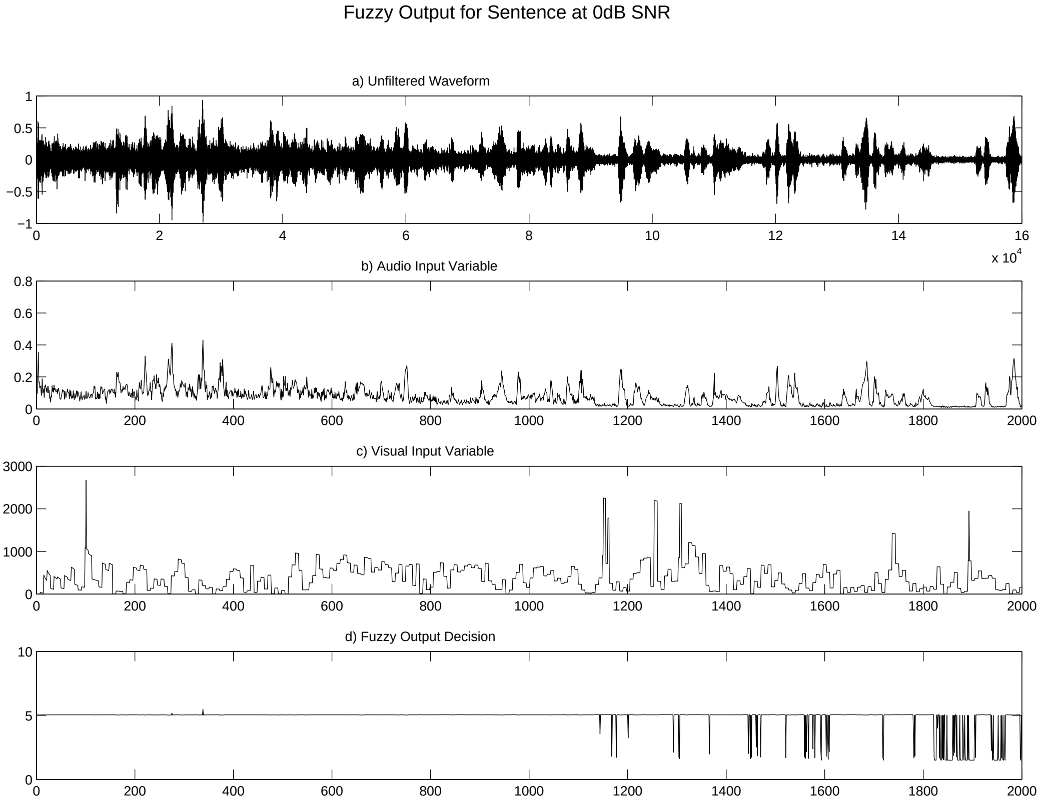


Figure 112: fuzzy logic output decision depending on SNR level. (a) shows the input audio waveform, with speech and noise mixed at a SNR of odB. (b) shows the audio input variable. (c) shows the visual input variable, with a small number of frames considered to be of low quality. Finally, (d) shows the fuzzy output decision.

Firstly, in figure 111, it can be seen in (a) that the speech is more visible in the waveform, which is a reflection on the increased SNR level. It can be seen in (d) that as the input level variable decreases, the fuzzy logic system chooses the audio-only option for much of the second part of the sentence, which is very different from previous examples of the same sentence with the same noise but a lower SNR. Figure 112 shows that at a SNR of odB, the audiovisual option is no longer chosen in any frames, and most frames make use of audio-only processing, with a small number of frames making use of the unprocessed option (reflected in the lower fuzzy output value).

Finally, figure 113 shows that at a SNR of +10dB there are a much greater number of examples of the fuzzy logic system choosing to not filter the frame of speech. Overall, it is shown that the system will adapt to changing audio input levels, with an example of the same

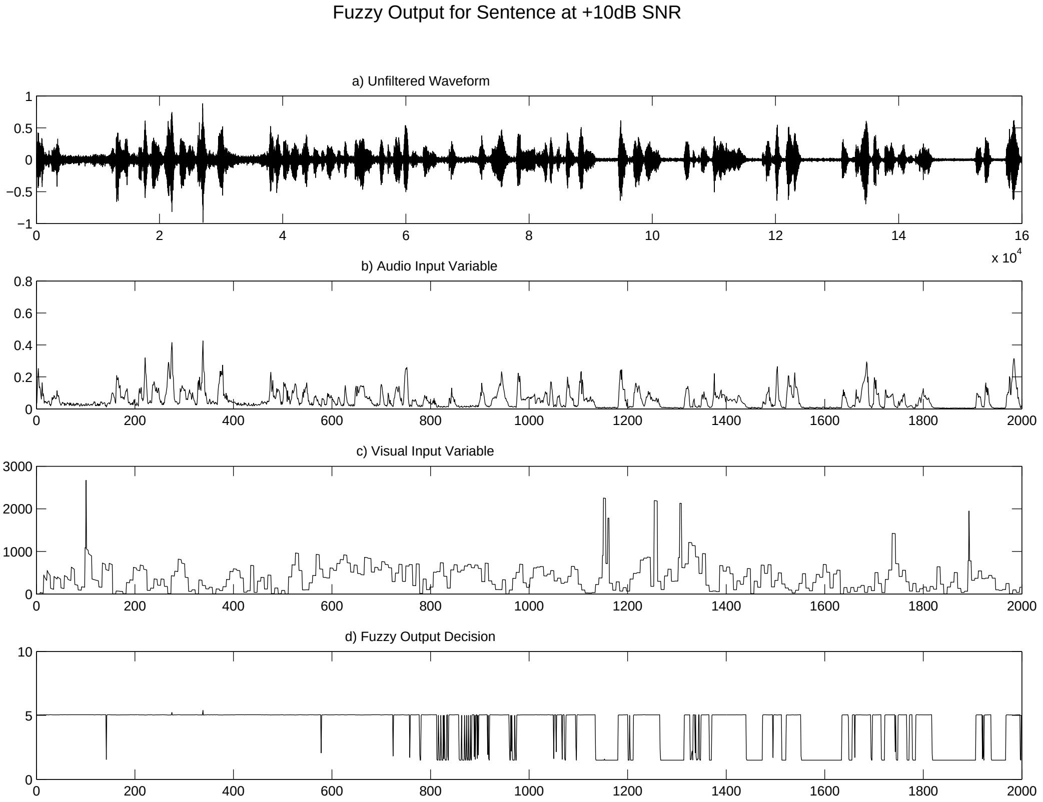


Figure 113: fuzzy logic output decision depending on SNR level. (a) shows the input audio waveform, with speech and noise mixed at a SNR of +10dB. (b) shows the audio input variable. (c) shows the visual input variable, with a small number of frames considered to be of low quality. Finally, (d) shows the fuzzy output decision.

sentence, with the same visual input variable, and the same type of noise source, producing a different decision from frame-to-frame, depending on the SNR, and therefore the level of noise.

### *Fuzzy Switching Conclusions*

In conclusion, it can be seen that although there are limitations with the specific audio-only and audiovisual techniques, as identified in previous sections, this section has demonstrated that the fuzzy logic switching system is functioning as expected in a range of different conditions. It has been shown to perform differently when the same speech sentence has been mixed with different types of noise, demonstrating that it is capable of adapting to different types of noise. It has also been shown to function as expected with a number of different sentences, with different visual quality input values. This shows that the system is versatile enough to adapt to different speakers and content, and also only makes use of audiovisual processing when there is considered to be good quality visual information available. Finally, an inspection of the same sentence with the same noise, but at different SNR levels shows that as the SNR changes, the processing output decision also changes; again demonstrating the adaptability of this fuzzy logic based switching system.

## 7.8 DISCUSSION OF RESULTS

Firstly, the evaluation performed in this chapter confirmed that the fuzzy-based system performs as expected. The system switches between processing options when considered appropriate, as confirmed by the results in section 7.7.7. It can be seen that at a very low SNR, the system will switch to making use of audiovisual information, but only if that visual information is available. Section 7.6 presented an evaluation of the input variables, and it was concluded that the initial visual fuzzy input variable can successfully be used to classify visual information. It was shown with a range of challenging conditions and widely varying conversation snippets from different speakers that the method correctly identified the quality of visual information in the majority of cases. Tracking errors due to animated movement of the speakers was generally correctly identified. This section justified the use of fuzzy variables by showing that different speech sentences had different input values, matching the manually estimated predictions, and the values chosen for the fuzzy thresholds were suitable to cover a wide range of potential input data.

However, there are further improvements that could be made to this approach. There are occasions when data that was manually identified as being of poor quality, was classified as being correct by the detector. To improve the accuracy of the fuzzy input variable, it could be possible to create an improved input variable using a machine learning technique discussed in the previous chapter, such as a HMM or ANNs, to improve the accuracy of the input variable.

Section 7.6 also discussed the use of the previous fuzzy controller output value as an input into the system for the subsequent frame. The aim of making use of this system is to reduce rapid switching between processing options on a frame-by-frame basis. The potential benefit of using the single previous frame or a floating mean of 3, 5, and 10 previous outputs was investigated. The results showed that although using a floating mean smoothed the input variable on a frame-to-frame basis, it made very little difference to the fuzzy output value, justifying the use of a single frame. This section also evaluated the effect of using the fuzzy variable the rapid switching of processing options from frame-to-frame. Switching processing options rapidly can cause potential listener discomfort due to the different audio output from different processing options, and while it is expected that the processing option will change in response to environmental conditions, rapid oscillation should be prevented where possible.

To investigate this, the fuzzy rules pertaining to the previous input variable were disabled and the system was run with a number of sentences at different SNR levels. The results, when compared to running the system with the rules enabled demonstrated that using the previous variable fulfilled the requirement of reducing the oscillation from frame-to-frame. This, combined with the positive evaluation result of the visual fuzzy variable, demonstrated that the input fuzzy variables were successfully used to provide inputs into the fuzzy logic based system.

However, there are a number of ways in which these inputs could be improved. As discussed above, a model could be trained to accurately identify the quality of an image. Also, in addition to the relatively basic audio power input, additional detectors such as a VAD could be used to positively identify the presence or absence of speech. This would serve as an additional input into the fuzzy-based system (and so would require the writing of additional rules), as used in some current commercial hearing-aids. This could also include specific front-back or wind detectors, to add versatility to the system. These detectors were discussed in more depth in chapter 3.

With regard to the audio output of the system, it can be seen from the evaluation that the results are of limited value. When making use of washing machine noise, the audiovisual filtering produces a significantly worse result than using beamforming. This was discussed in chapter 5, where the result was found to be significantly worse when used with data not similar to that which the system had not previously been trained with. Therefore, poor results were expected with novel data. In this scenario, the audio-only approach produced significantly stronger results than any others, again, as expected. The simulated room environment is specifically designed for demonstrating the performance of the system, therefore, the audio-only approach is the best performer, as shown by the objective tests in section 7.7.4. This was also confirmed by the listening tests presented in section 7.7.5, where the beamforming approach resulted in the highest MOS.

The fuzzy results are of interest because they demonstrate that the fuzzy-based system performs as expected. At a very low SNR, the system makes use of the audiovisual processing option, and at a high SNR, the system predominantly makes use of the audio-only approach as expected. However, at even the lowest SNR, the objective and subjective scores are not identical to the audiovisual scores. This is because the fuzzy-based approach makes use of different processing options, depending on the fuzzy input variables, and so there is a difference in scores. A similar pattern can be seen at a higher SNR, when the audio-only approach is predominantly used, but again, it is not used in all cases, and is dependent on the input fuzzy variables. However, the score is again rated as lower than the beamforming approach.

This would initially suggest that the beamforming approach is always better; however, this result has to be interpreted with a degree of caution. Section 7.7.6 discussed the results of objective tests when using an inconsistent clapping noise with transients and silences, designed to be extremely challenging for a beamformer. In this scenario, it was found that the audio-only approach produced no results of value, as shown by the extremely low objective scores. However, the audiovisual approach also performed poorly, due to the limitations discussed previously. Accordingly, although the fuzzy-based approach performed as expected, the limitations identified with the speech processing techniques also show that the system is currently only suitable for testing in specialised environments, and needs further development before being suitable for more general purpose use.

Overall, when testing the system with more challenging data, in terms of audio output, there are significant limitations with the evaluation of the system in its current condition. When tested in more general scenarios outside of specific conditions, the techniques proposed in this work for processing speech are in need of improvement. This was a limitation discussed in more depth in chapter 5, when it was concluded that the audiovisual approach was limited with regard to data that it had not been trained with.

Despite the limitations identified above, an investigation of the performance of the fuzzy switching system has shown that the system switches between inputs as expected. In noisy environments with a high SNR, the system automatically selects a different form of processing (in this case audiovisual), but only when there is only suitable associated visual information. Section 7.7 uses an example of a noise that is gradually decreasing. It can be seen that at some SNR levels, as the noise decreases, the processing option accurately switches to an alternative processing mode, as expected. This demonstrates that the system is capable of adapting to a range of different audiovisual environments, and is capable of solving the problem of lack of availability of visual information. The use of the fuzzy input variable for a previous frame was also shown to reduce the rate of oscillation between frames, and so despite the limited audio output results, the system was shown to perform as expected with regard to the fuzzy switching. It should be emphasised that this is a preliminary system, and future work specifically with regard to this aspect of the system could involve adding additional detectors and rules to perform more sophisticated analysis of the input data. Future work would also investigate the processing cost of using such a system, and potential performance savings to be gained from using different processing options.

Overall, this preliminary system demonstrates that there are limitations with the system in its current form, primarily in the limitations with the specific processing options, as outlined in chapter 5. In order to improve this system, the most important aspect is that the audiovisual filtering approach needs to be improved and refined. The results show that there is considerable scope for improvement when using data that the system has not been trained with. This limitation was tested and explains the limited objective and subjective results. Another significant improvement needed is to further develop the system to enable more accurate evaluation. The results showed that the beamforming results were good with the appropriate type of noise, but extremely limited with an unsuitable noise, and so therefore had to be treated with caution. Future work would involve the development of this system to be able to use a true multi-microphone environment rather than a simulated room, to fully and accurately evaluate the system. This would involve further refinement, but would also require the acquisition of improved hardware to use for testing. This improved hardware would allow for improved data synchronisation, correct acquisition of impulse responses and directional information, and would allow for noise to be added during recording rather than afterwards, taking more account of the Lombard Effect. However, the fuzzy-system was shown in this chapter to successfully switch between processing options based on input variables. A number of improvements could be made, such as the addition of further detectors such as a VAD. Overall, this chapter demonstrated that a fuzzy logic system could be used to accurately switch between processing options depending on fuzzy input variables.

## 7.9 SUMMARY

In addition to the two-stage filtering system discussed in chapter 4 and tested in chapter 5, this thesis also explores the concept of utilising fuzzy logic as part of an autonomous, adaptive, and context aware system. The limitations of the two-stage system were discussed in chapter 6, with a preliminary fuzzy logic system also presented. This fuzzy logic system used a number of fuzzy inputs and rules to determine the most suitable method of processing each frame of speech, either by audio-only filtering, two-stage audiovisual speech enhancement, or simply leaving the frame unprocessed. In this chapter, the results of initial testing of the fuzzy logic system were presented. Firstly, the requirements of these tests were discussed, including the need for visual data of varying quality, more challenging speech sentences, and novel data not previously tested with the system. Some issues with the system in its current preliminary state of implementation with regard to recording and testing of challenging real data were also described. With these limitations taken into account, along with the requirements for testing a system with challenging data, the recording of a novel corpus containing this challenging data was discussed in section 7.5. This corpus contained examples of reading tasks, more animated speech, made use of longer conversation snippets rather than simple two second sentences, and a varying quality of visual data. This corpus was then used to present an evaluation of the preliminary system.

Firstly, the fuzzy input variables were evaluated, demonstrating that the visual input variable was a reasonable estimation of the input data, and could accurately represent varying input data over a range of sentences. The use of a previous frame for input was also justified, showing that this presented a smoother output with less switching from frame-to-frame. However, the audio output results were of less significance, as discussed in section 7.7, the performance is very much dependent on the audio-only beamforming, which varied widely in performance depending on the noise that was used. The audiovisual method also produced a poor quality output, as expected. Therefore, in order to evaluate the system more fully, the speech processing techniques should be improved in future work. Additional hardware resources are also required in order to fully test the beamforming performance. The results also demonstrated that the proposed system performed as expected, with the fuzzy logic controller adjusting the output depending on the input variables. This showed that a fuzzy logic approach can be applied to speech processing, and that it is context aware and capable of adapting to environmental conditions.

The final chapter concludes the thesis as a whole, providing an overview of the research and the original contributions presented in this work, and outlining some proposed future research directions.

# 8

# CONCLUSIONS AND FUTURE WORK

## 8.1 CONCLUSIONS

The speech enhancement research presented in this thesis was motivated by several factors. Firstly, the development in recent years of audio-only hearing aids that utilise sophisticated decision rules to determine the appropriate level of speech processing served as an inspiration. A second motivational factor was the exploitation of the established cognitive relationship between audio and visual elements of speech to produce multimodal speech filtering systems. Another motivation was the desire to utilise audiovisual speech filtering to extend the concept of audio-only speech processing to become multimodal, from the perspective of potential application to hearing aids. Based on these motivations, the goal of the work reported in this thesis was primarily to develop a flexible two-stage multimodal speech enhancement system, working towards the development of a fuzzy logic based speech enhancement framework that is autonomous, adaptive, and context aware. The novel proof of concept framework presented in this thesis makes use of audio-only beamforming, visually derived Wiener filtering, state-of-the-art lip tracking with Viola-Jones ROI detection, and a fuzzy logic controller, to present a novel speech enhancement framework.

This thesis presented two review chapters. The first, chapter 2, presented a brief background to the research domain. This summarised the background research into the relationship between the audio and visual speech modalities with respect to speech production and perception. This chapter also briefly provided a definition of some speech phenomena of relevance to this thesis, summarising the Cocktail Party Problem, the McGurk Effect, and the Lombard Effect. Previous research into audio and visual correlation by others was also described, and one original research contribution was presented, an investigation into audiovisual correlation and the change in multimodal correlation when beamforming is applied to noisy speech. The second review chapter provided a detailed description of the specific research context to the work presented in this thesis. Firstly, a review of commercial hearing aid technology was presented, describing speech filtering techniques utilised in modern hearing aids. This included techniques such as directional microphones, noise cancelling algorithms, and decision rules. This chapter also provided a review of state-of-the-art audiovisual speech enhancement techniques in the literature. The historical development of this field was described, and a number of recent research developments were then examined, such as audiovisual source separation, multimodal fragment decoding, and visually derived Wiener filtering, and also a review of ROI detection technology. Finally, the chapter also evaluated a number of audiovisual speech databases, including GRID, VidTIMIT and others, in order to determine the most suitable corpora to use for developing and testing the framework subsequently presented in this thesis.

The key original contributions of this thesis were presented in chapters 4, 5,6, and then chapter 7. Chapter 4 provided a detailed description of the novel two-stage speech enhancement system. The feature extraction process for audio and visual features was described. The automated lip tracking approach used in this thesis was also summarised. The two-stage audiovisual speech filtering framework was then discussed in this chapter. Noise free speech estimation using GMM-GMR (to the knowledge of the author, this was a novel application of this technique to this research domain) with the use of 2D-DCT features as an input was described, along with visually derived filtering and audio-only beamforming, and the two filtering techniques used as part of this two-stage system. In chapter 5, this system was then evaluated using objective and subjective testing. Experiments in a range of challenging scenarios confirmed that this system is capable of delivering encouraging results, and the strengths of this system were identified. The results presented in this chapter, to the best knowledge of the author, represent benchmark results, with no pre-existing two-stage multimodal speech enhancement system to use for comparison. However, some limitations of this system were also discussed, with issues such as the introduction of distortion at a high SNR due to limitations of the two-stage approach, particularly the chosen GMM-GMR approach, and also the poor performance of the two-stage system when tested with a corpus it was not trained with.

Chapter 6 refined and expanded the system first presented in chapter 4 to present a preliminary, novel, fuzzy logic based, multimodal, two-stage speech enhancement framework. The limitations of the initial system tested in chapter 5 were discussed in depth, and a justification of the decision to use a fuzzy logic system was presented, along with a review of other potential techniques such as HMMs. This proof of concept framework used the same fundamental components as discussed in the previous chapters, but introduced a number of fuzzy input variables to determine the most suitable speech processing option to apply, depending on the audio and visual input data. For each frame, there is a choice of applying audio-only speech processing, leaving the frame unprocessed, or applying audiovisual two-stage processing. Finally, chapter 7 presented an evaluation of this preliminary concept, evaluating the performance of the fuzzy input variables, and then presenting subjective and objective testing of the fuzzy-based system. Initial evaluation results, concluded that although there is potential with the system, significant further refinements are needed in order to improve on the initial limited results in terms of audio output quality. However, the fuzzy switching system performed as expected, switching between processing options depending on the fuzzy input variables.

Based on the initial investigation of speech correlation, the experiments carried out with the two-stage system developed in this thesis, and then the examination of the novel preliminary fuzzy-logic based multimodal speech enhancement framework, the following conclusions can be drawn:

- The thesis explores the relationship between audio and visual elements of speech in the literature, and presents work published by the author into the effect of noise on audiovisual correlation. Performing an investigation of audiovisual correlation using the MLR technique, and adding noise to speech and comparing results to speech with the noise filtered concluded that, as expected, multimodal correlation can be improved with the use of filtering. This corroborated work in the literature and served as a test of some of the initial components used in this thesis.

- The work presented in this thesis showed that relevant individual components from different research domains, specifically modelling techniques designed for robot arm training and image tracking, can be successfully integrated into a novel multimodal speech enhancement system. The successful use of fuzzy logic as part of a multimodal speech enhancement framework was also novel (as far as the author has been able to ascertain), and the integration of GMM-GMR for speech estimation was also novel (again, to the best knowledge of the author). The use of these components as part of the same framework represented an original contribution of this work.

- The relationship between audio and visual aspects of speech production can be effectively used to successfully develop a multimodal speech enhancement system, combining both audio-only and audiovisual speech filtering elements. This thesis presented a two-stage audiovisual speech enhancement system, and a thorough evaluation confirmed that this system performed well in very noisy environments, when the SNR is extremely low, and speech is very difficult to identify from the noisy speech mixture without filtering. However, in less noisy environments, it was found that using visually derived filtering could add distortion to the speech and produce poor results, so there are both strengths and weaknesses to using this two-stage system. Overall, the discussion of these results concluded that while the system described in this chapter had limitations, it was capable of producing results in extremely noisy environments, and the extension of audio-only ideas to the multimodal domain added flexibility and functionality when compared to single modality speech enhancement systems.

- The novel fuzzy logic based multimodal speech enhancement framework presented in this thesis has been shown to successfully perform speech enhancement in limited experimental scenarios. The chosen fuzzy input variables have been shown to correctly determine the most appropriate technique for processing a noisy speech frame, depending on the quality of the input data from both speech production domains. The fuzzy-logic system was found to produce less conclusive results due to the limitations identified with the audiovisual technique and the recording environment. However, it was demonstrated that the system was able to successfully switch between different processing techniques depending on the input variables, with the oscillation limited. This framework extends ideas found in hearing aids in the audio modality, with the use of audio and visual speech input, and also the use of fuzzy logic. Although the framework has been found to be autonomous, adaptive, and context aware in terms of being able to switch between processing decisions on a frame by frame basis, further development is needed in order to improve performance.

Overall, this thesis presents a novel audiovisual speech enhancement system. The initial system combined a number of state-of-the-art techniques such as lip tracking, beamforming, and audiovisual Wiener filtering (using GMM-GMR), to develop a two-stage speech filtering system. The strengths and limitations of this system were thoroughly examined, and from this, a preliminary proof of concept novel fuzzy based multimodal two-stage speech enhancement framework was demonstrated.

## 8.2 FUTURE WORK RECOMMENDATIONS

### 8.2.1 *Improvement of Individual Speech Processing Components*

As discussed in chapter 5, limitations have been identified with some of the individual speech processing components presented in this system which could be improved. One significant example of this is the Wiener filtering approach used in this thesis. The current implementation is fairly basic, utilising GMM-GMR to provide an estimation of the noise free speech signal in the filterbank domain and interpolating this. A single GMM is also used for speech estimation. However, this has limitations due to the relative simplicity of its implementation. The GMM-GMR approach was originally devised by Calinon *et al.* [33] to calculate efficient robot arm movement. Although this thesis experimented with the novel application of this technique to the domain of speech filtering, the results suggest that this technique is ultimately not as accurate as the MAP GMM approach utilised by Almajai & Milner [11]. Furthermore, the speech modelling technique used does not make use of some of the most recent developments in speech enhancement, which may improve results. So for example, Almajai & Milner [11] make use of phoneme specific GMMs that attempt to identify the phoneme spoken, and then apply a specific GMM to this portion of speech.

Other state-of-the-art beamforming techniques could be investigated to consider for integration within this framework, and alternatives to using GMMs, such as reservoir computing (Maass *et al.* [124]), an area which has been recently applied to the signal processing domain for tasks such as multimodal laughter detection and music classification (Scherer *et al.* [162], Newton & Smith [134] can also be considered, to improve on the visually derived filtering approach used in this thesis and improve results.

### 8.2.2 *Extension of Overall Speech Filtering Framework*

One outcome of the work presented in this thesis is the initial development of a novel, scalable, speech processing framework that extends from feature extraction to speech filtering, with the use of a fuzzy logic controller. However, there is still much potential for extension of this framework. In addition to future work to upgrade the existing components of the system and investigate new speech enhancement techniques, it is also possible to add additional components to the framework. Some examples include the possibility of adding additional inputs such as spike trains Maass [123], Smith & Fraser [165] to potentially improve the filtering process. Other speech processing research (for example, by Sargin *et al.* [160] and also the author in Abel *et al.* [2]) has found that asynchrony can also result in an improved audiovisual speech relationship, and this could be exploited in future work.

Another way the framework can be extended is to include a number of more sophisticated input detectors, such as wind and front-back detectors, as discussed in chapter 3. Work by Almajai & Milner [12] has resulted in a speech enhancement system that uses a VAD to identify areas of speech and non speech in the input signal. This additional detector has precedent for being used in the literature, and may improve the fuzzy logic aspect of this system greatly. If the system was to be extended to successfully process real world data, then some of the other detectors discussed by Chung [44] such as wind detectors and front back detectors could also be integrated into the system, all of which would add sophistication and feasibility to the framework presented in this thesis.

### 8.2.3 *Further Development of Fuzzy Logic Based Switching Controller*

The fuzzy logic controller presented as part of this novel speech enhancement framework in chapter 6 is a very basic implementation, demonstrating that this framework could be developed further. Although it has been demonstrated that the fuzzy-based system is capable of responding to environmental conditions as expected, the results of running tests with real data have to be treated with some caution, due to the limitations of the test environment, the preliminary nature of the system (in that it is not implemented in real time), and the limitations with the filtering techniques identified in chapter 5. Although tests have been carried out using more challenging data, in order to test the system still further hardware based tests using multiple microphones and more real data is needed.

Additionally, the range and quality of input variables and fuzzy sets could be improved. As stated in chapter 6, the three variables used, audio frame power, visual DCT detail level, and previous frame selection, represent fairly simple detectors to use as fuzzy inputs. Although these are sufficient for demonstrating the novel framework presented in this thesis, an extension of this would naturally investigate the use of the detectors mentioned earlier, such as modulation and wind detectors, as inputs to the fuzzy switching system. These could then be used to develop the rules further. Although the current rules are adequate for demonstrating proof of concept, there are potential areas of improvement such as tweaking the weighting of the rules to give priority as suitable, rewriting the rules to cope with potential new inputs, and considering other aspects such as engaging/adaption/attack time when it comes to selection of the processing option. As discussed in chapter 7, additional refinement could also be carried out with regard to the visual input variable. Although the initial implementation was found to function well, it could be improved further by using a machine learning technique such as a trained HMM or an **ANN!** (ANN!) to classify input lip images.

It is also important that any future work carries out further testing of the refined framework. As stated previously, the current proof of concept framework has only undergone limited evaluation with it being concluded that further refinement is needed. Listener comfort is of particular importance. The system is designed with the considerations of users with hearing loss in mind, and is designed to automatically switch between processing options as needed, but it is important that this is done in a manner that does not cause irritation to the listener. Although the fuzzy inputs were shown to minimise frame to frame oscillation, this could be investigated and evaluated further in future work.

### 8.2.4 *Practical Implementation of System*

The system is currently purely implemented through software and simulations. MATLAB has been used for development, and testing has been carried out using a pre-recorded corpus, mixed with noise using a simulated room. Future development of this system would be to extend this initial software implementation (and the proposed refinements discussed previously in this section) and work towards the development of an initial hardware prototype. This would implement the improved fuzzy logic based speech enhancement framework physically, and would be expected to function with live data and real world noise, rather than simply with pre-recorded corpora. An example of a potential implementation strategy would be to make use of Field-Programmable Gate Arrays (FPGAs). These are semiconductor devices that can be programmed after manufacturing and thus allow for rapid prototyping and debugging. For this reason, they are commonly used in initial hardware development of technology. The evaluation process could also be improved by hardware implementation, in that it would be possible to carry out listening tests in a truly noisy environment, taking full account of the Lombard Effect, room impulse responses, and data synchrony, providing a full evaluation of a real time system able to function with a wide range of challenging data.

# REFERENCES

[1] Abdullah, Rudwan A, Hussain, Amir, & Polycarpou, Marios M. 2007. Fuzzy logic based switching and tuning supervisor for a multi-variable multiple controller. *Pages 1–6 of: Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International*. IEEE. (Cited on pages 175 and 176.)

[2] Abel, A., Hussain, A., Nguyen, Q.D., Ringeval, F., Chetouani, M., & Milgram, M. 2009. Maximising Audiovisual Correlation with Automatic Lip Tracking and Vowel Based Segmentation. *Pages 65–72 of: Biometric ID Management and Multimodal Communication: Joint COST 2101 and 2102 International Conference, BioID_MultiComm 2009, Madrid, Spain, September 16-18, 2009, Proceedings*, vol. 5707. Springer-Verlag. (Cited on pages 3, 20, 21, 69, 73, 91, 138, and 290.)

[3] Acero, A., & Stern, R.M. 2002. Environmental robustness in automatic speech recognition. *Pages 849–852 of: Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE. (Cited on page 46.)

[4] Adrian, E.D. 1928. The basis of sensation. (Cited on page 181.)

[5] Agui, T., Kokubo, Y., Nagahashi, H., & Nagao, T. 1992. Extraction of face regions from monochromatic photographs using neural networks. *Proceedings o] International Conferer˜ ce on Robotics*. (Cited on page 71.)

[6] Ahmed, N., Natarajan, T., & Rao, KR. 1974. Discrete cosine transfom. *Computers, IEEE Transactions on*, **100**(1), 90–93. (Cited on pages 24 and 98.)

[7] Alcántara, J.I., Moore, B.C.J., Kühnel, V., & Launer, S. 2003. Evaluation of the noise reduction system in a commercial digital hearing aid: Evaluación del sistema de reducción de ruido en un auxiliar auditivo digital comercial. *International Journal of Audiology*, **42**(1), 34–42. (Cited on page 44.)

[8] Allen, J.B., Trevino, A., & Han, W. 2012. Speech perception in impaired ears. (Cited on pages 172 and 179.)

[9] Almajai, I. 2009. *Audiovisual Speech Enhancement*. Ph.D. thesis, University of East Anglia. (Cited on pages 16, 17, and 144.)

[10] Almajai, I., & Milner, B. 2007. Maximising audio-visual speech correlation. *In: Proc. AVSP*. (Cited on pages 3, 8, 9, 14, 16, 19, 20, 21, 22, 23, 26, 27, 30, 31, 59, 97, 107, and 205.)

[11] Almajai, I., & Milner, B. 2009a. Effective visually-derived Wiener filtering for audio-visual speech processing. *In: Proc. Interspeech, Brighton, UK*. (Cited on pages xiv, 5, 6, 59, 60, 64, 81, 99, 100, 107, 143, 144, 149, 188, 203, 289, and 290.)

[12] Almajai, I., & Milner, B. 2009b. Enhancing Audio Speech using Visual Speech Features. *In: Proc. Interspeech, Brighton, UK*. (Cited on pages 46, 59, 62, 105, 107, 108, 110, 138, 143, 144, 149, 153, 165, 206, and 290.)

[13] Almajai, I., Milner, B., Darch, J., & Vaseghi, S. 2007. Visually-derived Wiener filters for speech enhancement. *Pages 585–588 of: IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*, vol. 4. (Cited on pages 1, 3, 8, 48, 59, 85, 89, 98, 99, and 107.)

[14] Avci, Engin, & Akpolat, Zuhtu Hakan. 2006. Speech recognition using a wavelet packet adaptive network based fuzzy inference system. *Expert Systems with Applications*, **31**(3), 495–503. (Cited on page 175.)

[15] Bagis, A. 2003. Determining fuzzy membership functions with tabu search–an application to control. *Fuzzy sets and systems*, **139**(1), 209–225. (Cited on page 185.)

[16] Bailly-Bailliere, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., *et al.* 2003. The BANCA database and evaluation protocol. *Pages 1057–1057 of: Audio-and Video-Based Biometric Person Authentication*. Springer. (Cited on pages xiv, 74, and 75.)

[17] Bansal, P., Kant, A., Kumar, S., Sharda, A., & Gupta, S. 2008. Improved Hybrid Model of HMM/GMM For Speech Recognition. *Intelligent Technologies and Applications*, 69–74. (Cited on pages 178 and 180.)

[18] Barker, J., & Shao, X. 2007. Audio-visual speech fragment decoding. *Pages 37–42 of: Proc. Int. Conf. on Auditory-Visual Speech Processing*. (Cited on pages xiii, 47, 48, 54, 55, 57, and 81.)

[19] Barker, J., & Shao, X. 2009. Energetic and informational masking effects in an audiovisual speech recognition system. *Audio, Speech, and Language Processing, IEEE Transactions on*, **17**(3), 446–458. (Cited on page 55.)

[20] Barker, J., Coy, A., Ma, N., & Cooke, M. 2006. Recent advances in speech fragment decoding techniques. *Pages 85–88 of: Proceedings of Interspeech*, vol. 2006. Citeseer. (Cited on page 54.)

[21] Barker, JP, & Berthommier, F. 1999a. Estimation of speech acoustics from visual speech features: A comparison of linear and non-linear models. *In: AVSP'99-International Conference on Auditory-Visual Speech Processing*. (Cited on pages 16 and 19.)

[22] Barker, J.P., & Berthommier, F. 1999b. Evidence of correlation between acoustic and visual features of speech. *Ohala et al*, 199–202. (Cited on pages 3 and 8.)

[23] Barker, JP, Cooke, MP, & Ellis, D.P.W. 2005. Decoding speech in the presence of other sources. *speech communication*, **45**(1), 5–25. (Cited on page 54.)

[24] Bentler, R.A., Palmer, C., & Dittberner, A.B. 2004. Hearing-in-noise: Comparison of listeners with normal and (aided) impaired hearing. *Journal of the American Academy of Audiology*, **15**(3), 216–225. (Cited on page 39.)

[25] Bingham, E., & Hyvarinen, A. 2000. A fast fixed-point algorithm for independent component analysis of complex valued signals. *International Journal of Neural Systems*, **10**(1), 1–8. (Cited on pages 51 and 52.)

[26] Bishop, C., & Viola, P. 2003. Learning and vision: Discriminative methods. *ICCV Course on Learning and Vision*, **2**(7), 11. (Cited on page 72.)

[27] Bolia, R.S., Nelson, W.T., Ericson, M.A., & Simpson, B.D. 2000. A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, **107**, 1065. (Cited on page 79.)

[28] Boymans, M., Dreschler, W.A., Schoneveld, P., & Verschuure, H. 1999. Clinical evaluation of a full-digital in-the-ear hearing instrument. *International Journal of Audiology*, **38**(2), 99–108. (Cited on page 44.)

[29] Bradski, G. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*. (Cited on pages 71, 73, 93, and 119.)

[30] Bregman, A.S. 1990. *Auditory scene analysis: The perceptual organization of sound*. The MIT Press. (Cited on page 54.)

[31] Bregman, A.S. 1993. Auditory scene analysis: Hearing in complex environments. (Cited on page 54.)

[32] Brubaker, S.C., Wu, J., Sun, J., Mullin, M.D., & Rehg, J.M. 2008. On the design of cascades of boosted ensembles for face detection. *International Journal of Computer Vision*, **77**(1), 65–86. (Cited on page 72.)

[33] Calinon, S., Guenter, F., & Billard, A. 2007. On learning, representing, and generalizing a task in a humanoid robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, **37**(2), 286–298. (Cited on pages 8, 10, 100, 101, 165, and 289.)

[34] Cauwenberghs, G., & Poggio, T. 2001. Incremental and decremental support vector machine learning. *Pages 409–415 of: Advances in neural information processing systems 13: proceedings of the 2000 conference*. The MIT Press. (Cited on page 95.)

[35] Chang, C.Y., & Shyu, K.K. 2002. A self-tuning fuzzy filtered-U algorithm for the application of active noise cancellation. *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, **49**(9), 1325–1333. (Cited on pages 173 and 175.)

[36] Chatterjee, S., & Hadi, A.S. 2006. *Regression analysis by example, 4th edition*. John Wiley and Sons. (Cited on page 23.)

[37] Chen, W.H., & Pratt, W. 1984. Scene adaptive coder. *Communications, IEEE Transactions on*, **32**(3), 225–232. (Cited on pages 24 and 98.)

[38] Cherry, E.C. 1953. Some experiments on the recognition of speech, with one and with two ears. *Journal of the acoustical society of America*, **25**(5), 975–979. (Cited on pages 17, 47, and 49.)

[39] Cheung, Y., Liu, X., & You, X. 2012. A local region based approach to lip tracking. *Pattern Recognition*, **45**, 3336–3347. (Cited on pages 65, 66, 67, 68, 69, and 92.)

[40] Chibelushi, CC, Deravi, F., & Mason, JSD. 1996. Survey of audio visual speech databases. *Speech and Image Processing Research Group, Department of Electrical and Electronic Engineering, University of Wales Swansea*. (Cited on page 74.)

[41] Chiou, G.I., & Hwang, J.N. 1997. Lipreading from color video. *Image Processing, IEEE Transactions on*, **6**(8), 1192–1195. (Cited on page 68.)

[42] Choi, Young Sik, & Krishnapuram, Raghu. 1997. A robust approach to image enhancement based on fuzzy logic. *Image Processing, IEEE Transactions on*, **6**(6), 808–825. (Cited on page 175.)

[43] Christensen, L.A., Helmink, D., Soede, W., & Killion, M.C. 2002. Complaints about hearing in noise: a new answer. *Hear Rev*, **9**(6), 34–36. (Cited on page 41.)

[44] Chung, K. 2004. Challenges and recent developments in hearing aids. Part I. Speech understanding in noise, microphone technologies and noise reduction algorithms. *Trends Amplif*, **8**(3), 83–124. (Cited on pages 6, 35, 36, 38, 39, 41, 42, 43, 44, 80, 169, 170, 172, 177, 189, 190, and 291.)

[45] Cifani, S., Abel, A., Hussain, A., Squartini, S., & Piazza, F. 2009. An Investigation into Audiovisual Speech Correlation in Reverberant Noisy Environments. *Pages 331–343 of: Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions: COST Action 2102 International Conference Prague, Czech Republic, October 15-18, 2008 Revised Selected and Invited Papers*, vol. 5641. Springer-Verlag. (Cited on pages 3, 8, 14, 19, 21, and 31.)

[46] Cooke, M., Barker, J., Cunningham, S., & Shao, X. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, **120**(5 Pt 1), 2421–2424. (Cited on pages 56, 78, 81, 115, 125, 197, and 202.)

[47] Cootes, TF, Edwards, GJ, & Taylor, CJ. 1998. Active appearance models. *Computer Vision–ECCV 98*, 484–498. (Cited on pages 20, 67, 70, and 92.)

[48] Cord, M.T., Surr, R.K., Walden, B.E., & Olson, L. 2002. Performance of directional microphone hearing aids in everyday life. *Journal of the American Academy of Audiology*, **13**(6), 295–307. (Cited on pages 37 and 171.)

[49] Cord, M.T., Surr, R.K., Walden, B.E., & Dyrlund, O. 2004. Relationship between laboratory measures of directional advantage and everyday success with directional microphone hearing aids. *Journal of the American Academy of Audiology*, **15**(5), 353–364. (Cited on pages 37 and 171.)

[50] Crow, F.C. 1984. Summed-area tables for texture mapping. *Computer Graphics*, **18**(3), 207–212. (Cited on pages 71 and 92.)

[51] Das, A., & Ghoshal, D. 2012. Extraction of time invariant lips based on Morphological Operation and Corner Detection Method. *International Journal of Computer Applications*, **48**(21), 7–11. (Cited on pages 65 and 66.)

[52] Deligne, S., Potamianos, G., & Neti, C. 2003. Audio-visual speech enhancement with AVCDCN (audio-visual codebook dependent cepstral normalization). *Pages 68–71 of: Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002*. IEEE. (Cited on pages 4 and 46.)

[53] Deng, L., Acero, A., Jiang, L., Droppo, J., & Huang, X. 2002. High-performance robust speech recognition using stereo training data. *Pages 301–304 of: Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 1. IEEE. (Cited on page 46.)

[54] El-emary, I.M.M., Fezari, M., & Attoui, H. 2011. Hidden Markov model/Gaussian mixture models (HMM/GMM) based voice command system: A way to improve the control of

remotely operated robot arm TR45. *Scientific Research and Essays*, **6**(2), 341–350. (Cited on page 178.)

[55] El-Wakdy, Mohamed, El-Sehely, Ehab, El-Tokhy, Mostafa, El-Hennawy, Adel, Mastorakis, NE, Mladenov, V, Bojkovic, Z, Simian, D, Kartalopoulos, S, Varonides, A, *et al.* 2008. Speech Recognition Using a Wavelet Transform to Establish Fuzzy Inference System Through Subtractive Clustering and Neural Network(ANFIS). *In: WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering*. WSEAS. (Cited on page 176.)

[56] Elberling, C. 2002. About the VoiceFinder. *News from Oticon*. (Cited on page 44.)

[57] Esposito, A., Ezin, E., & Reyes-Garcia, C. 2000. Designing a fast neuro-fuzzy system for speech noise cancellation. *MICAI 2000: Advances in Artificial Intelligence*, 482–492. (Cited on pages 6, 11, 166, 170, 176, 179, and 182.)

[58] Eveno, N., Caplier, A., & Coulon, P.Y. 2001. New color transformation for lips segmentation. *Pages 3–8 of: Multimedia Signal Processing, 2001 IEEE Fourth Workshop on*. IEEE. (Cited on page 66.)

[59] Eveno, N., Caplier, A., & Coulon, P.Y. 2002. Key points based segmentation of lips. *Pages 125–128 of: Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, vol. 2. IEEE. (Cited on page 66.)

[60] Flynn, M.C. 2004. Maximizing the voice-to-noise ratio (VNR) via voice priority processing. *Hearing Review*, **11**(4), 54–59. (Cited on page 39.)

[61] Flynn, MC, & Lunner, T. 2004. Clinical evidence for the benefits of Oticon Syncro. *News from Oticon*, 1–10. (Cited on page 39.)

[62] Freedman, D., & Brandstein, M.S. 2000. Contour tracking in clutter: a subset approach. *International Journal of Computer Vision*, **38**(2), 173–186. (Cited on page 66.)

[63] Freund, Y., & Schapire, R. 1995. A deciaion-theoretic generalization of on-line learning and an application to boosting. *Pages 23–37 of: Computational learning theory*. Springer. (Cited on page 72.)

[64] Friedman, J., Hastie, T., & Tibshirani, R. 2000. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The annals of statistics*, **28**(2), 337–407. (Cited on page 72.)

[65] Fritsch, FN, & Carlson, RE. 1980. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, **17**(2), 238–246. (Cited on pages 100 and 110.)

[66] Gader, P.D., Mohamed, M., & Chiang, J.H. 1997. Handwritten word recognition with character and inter-character neural networks. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **27**(1), 158–164. (Cited on page 182.)

[67] Gannot, S., Burshtein, D., & Weinstein, E. 2001. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, **49**(8), 1614–1626. (Cited on pages 2, 34, and 103.)

[68] Ghahramani, Z. 2001. An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, **15**(01), 9–42. (Cited on pages 173, 178, and 180.)

[69] Girin, L., Schwartz, J.L., & Feng, G. 2001. Audio-visual enhancement of speech in noise. *The Journal of the Acoustical Society of America*, **109**, 3007. (Cited on pages 3, 5, 45, and 46.)

[70] Girin, L., Feng, G., & Schwartz, JL. 2002. Fusion of auditory and visual information for noisy speech enhancement: A preliminary study of vowel transitions. *Pages 1005–1008 of: Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2. IEEE. (Cited on page 49.)

[71] Goecke, R., Potamianos, G., & Neti, C. 2002. Noisy audio feature enhancement using audio-visual speech data. *Pages 2025–2028 of: Acoustics, Speech, and Signal Processing, 2002. Proceedings.(ICASSP'02). IEEE International Conference on*, vol. 2. IEEE. (Cited on pages 3, 5, 6, and 46.)

[72] Golub, G.H., & Van Loan, C.F. 1996. *Matrix computations*. Johns Hopkins Univ Pr. (Cited on pages 95 and 96.)

[73] Griffiths, L., & Jim, C. 1982. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, **30**(1), 27–34. (Cited on pages 2 and 34.)

[74] Halavati, Ramin, Shouraki, Saeed Bagheri, & Zadeh, Saman Harati. 2007. Recognition of human speech phonemes using a novel fuzzy approach. *Applied Soft Computing*, **7**(3), 828–839. (Cited on page 175.)

[75] Hammer, B., & Steil, J.J. 2002. Tutorial: Perspectives on learning with rnns. *Pages 357–368 of: Proc. ESANN*. (Cited on page 181.)

[76] Hansen, J.H.L., & Pellom, B. 1998. An effective quality evaluation protocol for speech enhancement algorithms. *Pages 2819–2822 of: ICSLP, Sydney, Australia*. Citeseer. (Cited on pages 109, 111, and 113.)

[77] Harris, C., & Stephens, M. 1988. A combined corner and edge detector. *Page 50 of: Alvey vision conference*, vol. 15. Manchester, UK. (Cited on page 66.)

[78] Haykin, S. 1998. Neural networks: a comprehensive foundation. (Cited on pages 180 and 181.)

[79] Haykin, S., & Chen, Z. 2005. The cocktail party problem. *Neural computation*, **17**(9), 1875–1902. (Cited on page 18.)

[80] Hellmann, M. 2001. Fuzzy logic introduction. *Université de Rennes*. (Cited on pages 173 and 174.)

[81] Herault, J., Jutten, C., & Ans, B. 1985 (May). Detection De Grandeurs Primitives Dans Un Message Composite Par Une Architecture De Calcul Neuromimetrique En Apprentissage Non Supervise. *Pages 1017–1020 of: Actes du Xeme colloque GRETSI*, vol. 2. (Cited on page 47.)

[82] Hershey, J., & Casey, M. 2001. Audio-visual sound separation via hidden markov models. *Advances in Neural Information Processing Systems*, **14**, 1173–1180. (Cited on pages 178 and 180.)

[83] Hiller, AD, & Chin, RT. 1990. Iterative Wiener filters for image restoration. *Pages 1901–1904 of: Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. (Cited on pages 59 and 98.)

[84] Hotelling, H. 1936. Relations between two sets of variates. *Biometrika*, **28**(3/4), 321–377. (Cited on pages 1, 19, and 20.)

[85] Houtgast, T., & Steeneken, H.J.M. 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America*, **77**, 1069. (Cited on page 38.)

[86] Hu, Y., & Loizou, P.C. 2006. Evaluation of objective measures for speech enhancement. *Pages 1447–1450 of: Proc. Interspeech*. Citeseer. (Cited on pages 109, 111, 113, and 114.)

[87] Hu, Y., & Loizou, P.C. 2008. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing*, **16**(1), 229–238. (Cited on pages 109, 110, 111, 112, 113, 114, 236, and 237.)

[88] Huang, X., Li, S.Z., & Wang, Y. 2005. Jensen-shannon boosting learning for object recognition. *Pages 144–149 of: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE. (Cited on page 72.)

[89] Hussain, A., & Campbell, D.R. 1998. Binaural sub-band adaptive speech enhancement using artificial neural networks. *Speech communication*, **25**(1), 177–186. (Cited on page 182.)

[90] Hussain, A., & Campbell, DR. 2001. Intelligibility improvements using binaural diverse sub-band processing applied to speech corrupted with automobile noise. *Pages 127–132 of: Vision, Image and Signal Processing, IEE Proceedings-*, vol. 148. IET. (Cited on pages 2, 34, and 110.)

[91] Hussain, A., Cifani, S., Squartini, S., Piazza, F., & Durrani, T. 2007. A Novel Psychoacoustically Motivated Multichannel Speech Enhancement System. *Verbal and Nonverbal Communication Behaviours*, 190–199. (Cited on pages 2, 34, and 86.)

[92] Hyvarinen, A., Karhunen, J., & Oja, E. 2001. *Independent component analysis*. Vol. 26. Wiley-Interscience. (Cited on page 51.)

[93] Iyengar, G., Potamianos, G., Neti, C., Faruquie, T., & Verma, A. 2001. Robust detection of visual ROI for automatic speechreading. *Pages 79–84 of: Multimedia Signal Processing, 2001 IEEE Fourth Workshop on*. IEEE. (Cited on page 92.)

[94] Jaeger, H. 2001. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *Tecnical report GMD report*, **148**. (Cited on page 181.)

[95] Ji, Z., Su, Y., Wang, J., & Hua, R. 2009. Robust sea-sky-line detection based on horizontal projection and hough transformation. *Pages 1–4 of: Image and Signal Processing, 2009. CISP'09. 2nd International Congress on*. IEEE. (Cited on page 66.)

[96] Jiang, J., Alwan, A., Keating, P.A., Auer, E.T., & Bernstein, L.E. 2002. On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP Journal on Applied Signal Processing*, **11**, 1174–1188. (Cited on page 19.)

[97] Joshi, P., & Maass, W. 2005. Movement generation with circuits of spiking neurons. *Neural Computation*, **17**(8), 1715–1738. (Cited on page 181.)

[98] Jutten, C., & Herault, J. 1991. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, **24**(1), 1–10. (Cited on pages 47 and 49.)

[99] Kass, Michael, Witkin, Andrew, & Terzopoulos, Demetri. 1988. Snakes: Active contour models. *International Journal of Computer Vision*, **1**, 321–331. (Cited on pages 65, 66, and 70.)

[100] Kjeldsen, R., & Kender, J. 1996. Finding skin in color images. *Pages 312–317 of: Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*. IEEE. (Cited on page 70.)

[101] Klatt, D. 1982. Prediction of perceived phonetic distance from critical-band spectra: a first step. *Pages 1278–1281 of: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, vol. 7. IEEE. (Cited on page 112.)

[102] Kohonen, T. 1989. *Self-Organisation and Associative Memory (Springer-Verlag)*. (Cited on page 70.)

[103] Kotropoulos, C., & Pitas, I. 1997. Rule-based face detection in frontal views. *Pages 2537–2540 of: Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 4. IEEE. (Cited on page 70.)

[104] Kroon, D-J. 2010. *Viola Jones Object Detection*. `http://www.mathworks.com/matlabcentral/fileexchange/29437-viola-jones-object-detection`. (Cited on pages 92 and 119.)

[105] Kuk, F., Ludvigsen, C., & Paludan-Müller, C. 2002. Improving hearing aid performance in noise: Challenges and strategies. *The Hearing Journal*, **55**(4), 34. (Cited on page 42.)

[106] Kuk, F., Keenan, D., Lau, C.C., & Ludvigsen, C. 2005. Performance of a fully adaptive directional microphone to signals presented from various azimuths. *Journal of the American Academy of Audiology*, **16**(6), 333–347. (Cited on page 36.)

[107] Lane, H., & Tranel, B. 1971. The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, **14**(4), 677. (Cited on pages 64 and 77.)

[108] Lankton, S., & Tannenbaum, A. 2008. Localizing region-based active contours. *Image Processing, IEEE Transactions on*, **17**(11), 2029–2039. (Cited on page 68.)

[109] Laugesen, S., & Schmidtke, T. 2004. Improving on the speech-in-noise problem with wireless array technology. *News from Oticon*, 3–23. (Cited on page 41.)

[110] Lee, B., Hasegawa-johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., & Huang, T. 2004. AVICAR: Audio-Visual Speech Corpus in a Car Environment. *Pages 2489–2492 of: Proc. Conf. Spoken Language, Jeju, Korea.* Citeseer. (Cited on pages 19, 64, 76, and 77.)

[111] Levey, A., & Lindenbaum, M. 2000. Sequential Karhunen-Loeve basis extraction and its application to images. *IEEE Transactions on Image processing*, **9**(8), 1371–1374. (Cited on page 95.)

[112] Levitt, H. 2001. Noise reduction in hearing aids: An overview. *Journal of rehabilitation research and development*, **38**(1). (Cited on page 44.)

[113] Li, J., Sakamoto, S., Hongo, S., Akagi, M., Suzuki, Y., *et al.* 2008. A two-stage binaural speech enhancement approach for hearing aids with preserving binaural benefits in noisy environments. *Journal of the Acoustical Society of America*, **123**(5), 3012–3012. (Cited on pages 2, 6, and 34.)

[114] Li, J., Sakamoto, S., Hongo, S., Akagi, M., & Suzuki, Y. 2010. Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication. *Speech Communication*. (Cited on pages 9 and 34.)

[115] Li, S., Zhu, L., Zhang, Z.Q., Blake, A., Zhang, H.J., & Shum, H. 2006. Statistical learning of multi-view face detection. *Computer Vision, ECCV 2002*, 117–121. (Cited on pages 72, 119, and 124.)

[116] Lienhart, R., & Maydt, J. 2002. An extended set of haar-like features for rapid object detection. *Pages I–900 of: Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1. IEEE. (Cited on page 72.)

[117] Liew, A.W.C., Leung, S.H., & Lau, W.H. 2002. Lip contour extraction from color images using a deformable model. *Pattern Recognition*, **35**(12), 2949–2962. (Cited on pages 65 and 68.)

[118] Lisboa, P.J., & Taktak, A.F.G. 2006. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural networks*, **19**(4), 408–415. (Cited on page 182.)

[119] Loizou, Philipos C. 2007. *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*. 1 edn. CRC. (Cited on pages 111 and 113.)

[120] Lombard, E. 1911. Le signe de l'elevation de la voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, **37**(101-119), 25. (Cited on pages 18 and 197.)

[121] Lu, Y., & Loizou, P.C. 2008. A geometric approach to spectral subtraction. *Speech communication*, **50**(6), 453–466. (Cited on pages 126, 127, and 146.)

[122] Luettin, J., Thacker, N.A., & Beet, S.W. 1996. Visual speech recognition using active shape models and hidden Markov models. *Pages 817–820 of: Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2. IEEE. (Cited on pages 67, 70, and 180.)

[123] Maass, W. 1997. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, **10**(9), 1659–1671. (Cited on pages 181 and 290.)

[124] Maass, W., Natschläger, T., & Markram, H. 2002. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, **14**(11), 2531–2560. (Cited on pages 181 and 290.)

[125] Malden Electronics Ltd. 2004. *Speech Quality Assessment Background Information for DSLA and MultiDSLA Users*. Malden Electronics Ltd. (Cited on pages xv and 112.)

[126] McGurk, H., & MacDonald, J. 1976. Hearing lips and seeing voices. *Nature*, **264**, 746–748. (Cited on pages 3, 18, and 107.)

[127] Meir, R., & Rätsch, G. 2003. An introduction to boosting and leveraging. *Advanced lectures on machine learning*, 118–183. (Cited on page 72.)

[128] Mens, L.H.M. 2011. Speech understanding in noise with an eyeglass hearing aid: Asymmetric fitting and the head shadow benefit of anterior microphones. *International Journal of Audiology*, **50**(1), 27–33. (Cited on pages 40 and 84.)

[129] Messer, K., Matas, J., Kittler, J., Luettin, J., & Maitre, G. 1999. XM2VTSDB: The extended M2VTS database. *Pages 965–966 of: Second International Conference on Audio and Video-based Biometric Person Authentication*, vol. 964. Citeseer. (Cited on pages xiv and 76.)

[130] Milner, B., & Almajai, I. Noisy audio speech enhancement using Wiener filters derived from visual speech. *In: Proc. International Workshop on Auditory-Visual Speech Processing (AVSP)*. (Cited on pages 5, 21, 59, 81, 144, and 168.)

[131] Mita, T., Kaneko, T., & Hori, O. 2005. Joint haar-like features for face detection. *Pages 1619–1626 of: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2. IEEE. (Cited on pages 72 and 124.)

[132] Moore, TJ. 1981. Voice communication jamming research. *In: AGARD Conference Proceedings 331 Aural Communication in Aviation*. Advisory Group for Aerospace Research and Development (AGARD) Conference Proceedings No. 311, Aural Communication in Aviation, CP311, 2-1-2-6. (Cited on page 79.)

[133] Naqvi, S.M., Yu, M., & Chambers, J.A. 2010. A multimodal approach to blind source separation of moving sources. *Selected Topics in Signal Processing, IEEE Journal of*, **4**(5), 895–910. (Cited on pages xiii, 49, 50, 51, 52, and 203.)

[134] Newton, M.J., & Smith, L.S. 2012. A neurally inspired musical instrument classification system based upon the sound onset. *The Journal of the Acoustical Society of America*, 4785–4798. (Cited on pages 181, 182, and 290.)

[135] Nguyen, Q.D., & Milgram, M. 2009. Semi Adaptive Appearance Models for lip tracking. *Pages 2437–2440 of: ICIP09*. (Cited on pages 4, 8, 10, 65, 69, 70, 73, 91, and 92.)

[136] Nguyen, Q.D., Milgram, M., & Nguyen, T.H.L. 2008. Multi features models for robust lip tracking. *Pages 1333–1337 of: Control, Automation, Robotics and Vision, 2008. ICARCV 2008. 10th International Conference on*. IEEE. (Cited on page 67.)

[137] Owens, FJ, & Lynn, P.A. 1993. Signal Processing of Speech (Macmillan New Electronics). (Cited on page 16.)

[138] P.835, ITU-T. 2003. Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. (Cited on pages 110 and 114.)

[139] P.862, ITU-T. 2001. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. (Cited on page 111.)

[140] Parra, L.C., & Alvino, C.V. 2002. Geometric source separation: Merging convolutive source separation with geometric beamforming. *Speech and Audio Processing, IEEE Transactions on*, **10**(6), 352–362. (Cited on page 51.)

[141] Patterson, E., Gurbuz, S., Tufekci, Z., & Gowdy, JN. 2002. CUAVE: A new audio-visual database for multimodal human-computer interface research. *Pages II–II of: Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2. IEEE. (Cited on pages 74 and 80.)

[142] Potamianos, G., Neti, C., & Deligne, S. 2003. Joint Audio-Visual Speech Processing for Recognition and Enhancement. *Pages 95–104 of: AVSP 2003-International Conference on Auditory-Visual Speech Processing*. (Cited on pages 4 and 46.)

[143] Rabiner, L.R., & Schafer, R.W. 1978. *Digital processing of speech signals*. Vol. 100. Prentice-hall Englewood Cliffs, NJ. (Cited on page 17.)

[144] Ricketts, T., & Henry, P. 2002. Evaluation of an adaptive, directional-microphone hearing aid: Evaluación de un auxiliar auditivo de micrófono direccional adaptable. *International Journal of Audiology*, **41**(2), 100–112. (Cited on page 39.)

[145] Ricketts, T., & Mueller, H.G. 1999. Making sense of directional microphone hearing aids. *American Journal of Audiology*, **8**(2), 117. (Cited on page 36.)

[146] Ricketts, TA, & Dahr, S. 1999. Aided benefit across directional and omni-directional hearing aid microphones for behind-the-ear hearing aids. *Journal of the American Academy of Audiology*, **10**(4), 180–189. (Cited on page 44.)

[147] Rivet, B., Girin, L., & Jutten, C. 2007a. Log-rayleigh distribution: A simple and efficient statistical representation of log-spectral coefficients. *IEEE transactions on audio, speech, and language processing*, **15**(3), 796–802. (Cited on pages 3, 46, 47, 49, and 86.)

[148] Rivet, B., Girin, L., & Jutten, C. 2007b. Mixing Audiovisual Speech Processing and Blind Source Separation for the Extraction of Speech Signals From Convolutive Mixtures. *IEEE Transactions on Audio, Speech and Language Processing*, **15**(1), 96–108. (Cited on pages 46, 47, 49, and 205.)

[149] Rivet, B., Girin, L., Serviere, C., Pham, Dinh-Tuan, & Jutten, C. 2007c. Using a Visual Voice Activity Detector to Regularize the Permutations in Blind Separation of Convolutive Speech Mixtures. *Pages 223 –226 of: Digital Signal Processing, 2007 15th International Conference on*. (Cited on pages 46, 47, 49, and 169.)

[150] Rivet, B., Girin, L., & Jutten, C. 2007d. Visual voice activity detection as a help for speech source separation from convolutive mixtures. *Speech Communication*, **49**(7-8), 667–677. (Cited on pages 46, 47, and 49.)

[151] Rivet, Bertrand. 2009. Blind Non-stationnary Sources Separation by Sparsity in a Linear Instantaneous Mixture. *Pages 314–321 of:* Adali, Tulay, Jutten, Christian, Romano, Joao, & Barros, Allan (eds), *Independent Component Analysis and Signal Separation*. Lecture Notes in Computer Science, vol. 5441. Springer Berlin / Heidelberg. (Cited on pages 18, 81, and 105.)

[152] Rivet, Bertrand, & Chambers, Jonathon. 2010. Multimodal Speech Separation. *Pages 1–11 of:* Sole-Casals, Jordi, & Zaiats, Vladimir (eds), *Advances in Nonlinear Speech Processing*. Lecture Notes in Computer Science, vol. 5933. Springer Berlin / Heidelberg. (Cited on pages 5, 46, 47, and 49.)

[153] Rix, A.W., Beerends, J.G., Hollier, M.P., & Hekstra, A.P. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. *Pages 749–752 of: Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE. (Cited on pages 109 and 111.)

[154] Rosen, S. 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions: Biological Sciences*, 367–373. (Cited on pages 38 and 41.)

[155] Rosistem. *Speech Production*. `http://www.barcode.ro/tutorials/biometrics/voice.html`. (Cited on pages xiii and 16.)

[156] Rumelhart, D.E., Hintont, G.E., & Williams, R.J. 1986. Learning representations by back-propagating errors. *Nature*, **323**(6088), 533–536. (Cited on pages 180 and 181.)

[157] Sanderson, C. 2008. *Biometric person recognition: Face, speech and fusion*. VDM Verlag Dr. Muller. (Cited on pages 14, 22, 77, 81, 144, 197, and 202.)

[158] Sanderson, C., & Paliwal, K.K. 2002. Polynomial features for robust face authentication. *Pages 997–1000 of: Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 3. IEEE. (Cited on page 77.)

[159] Sargın, ME, Erzin, E., Yemez, Y., & Tekalp, AM. 2005. Lip feature extraction based on audio-visual correlation. *In: Proc. EUSIPCO*, vol. 2005. (Cited on pages 19, 20, and 97.)

[160] Sargin, ME, Yemez, Y., Erzin, E., & Tekalp, AM. 2007. Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Transactions on Multimedia*, **9**(7), 1396–1403. (Cited on pages 1, 3, 8, 20, 99, and 290.)

[161] Schapire, R.E., & Singer, Y. 1999. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, **37**(3), 297–336. (Cited on page 72.)

[162] Scherer, Stefan, Glodek, Michael, Schwenker, Friedhelm, Campbell, Nick, & Palm, Gunther. 2012. Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, **2**(1), 4. (Cited on pages 119, 124, 181, and 290.)

[163] Schrauwen, B., Verstraeten, D., & Van Campenhout, J. 2007. An overview of reservoir computing: theory, applications and implementations. *In: Proceedings of the 15th European Symposium on Artificial Neural Networks*. Citeseer. (Cited on page 181.)

[164] Schum, D.J. 2003. Noise-reduction circuitry in hearing aids:(2) Goals and current strategies. *The Hearing Journal*, **56**(6), 32. (Cited on page 44.)

[165] Smith, Leslie S, & Fraser, Dagmar S. 2004. Robust sound onset detection using leaky integrate-and-fire neurons with depressing synapses. *Neural Networks, IEEE Transactions on*, **15**(5), 1125–1134. (Cited on page 290.)

[166] Sodoyer, D., Schwartz, J.L., Girin, L., Klinkisch, J., & Jutten, C. 2002. Separation of audio-visual speech sources: A new approach exploiting the audio-visual coherence of speech stimuli. *EURASIP Journal on Applied Signal Processing*, **2002**(1), 1165–1173. (Cited on pages 3 and 49.)

[167] Sodoyer, D., Girin, L., Jutten, C., & Schwartz, J.L. 2004. Developing an audio-visual speech source separation algorithm. *Speech Communication*, **44**(1-4), 113–125. (Cited on pages 3 and 49.)

[168] Sumby, W.H., & Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**(2), 212–215. (Cited on pages 3 and 107.)

[169] Sung, K.K. 1996. Learning and example selection for object and pattern detection. (Cited on page 71.)

[170] Tanaka, K., Iwasaki, M., & Wang, H.O. 2001. Switching control of an R/C hovercraft: stabilization and smooth switching. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **31**(6), 853–863. (Cited on pages 173, 175, and 192.)

[171] Tellier, N., Arndt, H., & Luo, H. 2003. Speech or noise? Using signal detection and noise reduction. *Hearing Review*, **10**(6), 48–51. (Cited on pages xxvii, 42, 43, and 170.)

[172] Valente, M. 2000. Use of microphone technology to improve user performance in noise. *Textbook of hearing aid amplification*, 247. (Cited on page 36.)

[173] Van den Bogaert, T., Doclo, S., Wouters, J., & Moonen, M. 2009. Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids. *The Journal of the Acoustical Society of America*, **125**, 360–371. (Cited on pages 2, 6, 34, and 83.)

[174] Varibel-Innovations. *Varibel - De bril die hoort*. `http://www.varibel.nl/`. (Cited on pages xiii, 40, and 84.)

[175] Viola, P., & Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. *Pages 511–518 of: Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 1. IEEE Comput. Soc. (Cited on pages 50, 65, 71, 72, 73, 92, 93, 119, 124, and 187.)

[176] Wakasugi, T., Nishiura, M., & Fukui, K. 2004. Robust lip contour extraction using separability of multi-dimensional distributions. *Pages 415–420 of: Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*. IEEE. (Cited on pages 65 and 68.)

[177] Wang, S., & Abdel-Dayem, A. 2012. Improved viola-jones face detector. *Pages 321–328 of: Proceedings of the 1st Taibah University International Conference on Computing and Information Technology, ICCIT '12*. (Cited on pages 70 and 119.)

[178] Wang, SL, Lau, WH, & Leung, SH. 2004. Automatic lip contour extraction from color images. *Pattern recognition*, **37**(12), 2375–2387. (Cited on page 68.)

[179] Wark, T., & Sridharan, S. 1998. A syntactic approach to automatic lip feature extraction for speaker identification. *Pages 3693–3696 of: Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 6. IEEE. (Cited on page 92.)

[180] Wiener, N. 1949. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. The MIT Press. (Cited on pages 2, 34, 59, and 98.)

[181] XM2VTS. *XM2VTS Website*. `http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/`. (Cited on pages xiv and 76.)

[182] Yang, G., & Huang, T.S. 1994. Human face detection in a complex background. *Pattern recognition*, **27**(1), 53–63. (Cited on page 70.)

[183] Yang, M.H., Kriegman, D.J., & Ahuja, N. 2002. Detecting faces in images: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **24**(1), 34–58. (Cited on pages 69 and 70.)

[184] Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. 1998. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, **26**(1-2), 23–43. (Cited on pages 16 and 19.)

[185] Yow, K.C., & Cipolla, R. 1996. A probabilistic framework for perceptual grouping of features for human face detection. *Pages 16–21 of: Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*. IEEE. (Cited on page 70.)

[186] Yuille, A.L., Hallinan, P.W., & Cohen, D.S. 1992. Feature extraction from faces using deformable templates. *International journal of computer vision*, **8**(2), 99–111. (Cited on page 68.)

[187] Zadeh, L.A. 1965. Fuzzy sets*. *Information and control*, **8**(3), 338–353. (Cited on pages 171 and 172.)

[188] Zayed, A.S., Hussain, A., & Abdullah, R.A. 2006. A novel multiple-controller incorporating a radial basis function neural network based generalized learning model. *Neurocomputing*, **69**(16), 1868–1881. (Cited on page 180.)

[189] Zelinski, R. 1988. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. *Pages 2578–2581 of: Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. (Cited on pages 2, 9, 34, 59, 83, and 98.)

[190] Zeng, Z., Tu, J., Pianfetti, B., Liu, M., Zhang, T., Zhang, Z., Huang, T.S., & Levinson, S. 2005. Audio-visual affect recognition through multi-stream fused HMM for HCI. *Pages 967–972 of: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE. (Cited on pages 178 and 180.)

[191] Zhang, C., & Zhang, Z. 2010. A survey of recent advances in face detection. *Microsoft Research, June*. (Cited on pages 72 and 73.)

[192] Zhang, X., & Mersereau, R.M. 2000. Lip feature extraction towards an automatic speechreading system. *Pages 226–229 of: Image Processing, 2000. Proceedings. 2000 International Conference on*, vol. 3. IEEE. (Cited on page 66.)

[193] Zurada, J.M. 1992. *Introduction to artificial neural systems*. West St. Paul, Minn. (Cited on page 180.)