Research Program in Speech, Audio, Image and Video Technology
School of Engineering Systems

ROBUST SPEECH RECOGNITION USING SPEECH ENHANCEMENT

Tristan Friedrich Kleinschmidt
BEng(Elec&CompEng)(Hons), MEngSc(Comp&Communications Eng), GCRC

SUBMITTED AS A REQUIREMENT OF
THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
QUEENSLAND UNIVERSITY OF TECHNOLOGY
BRISBANE, QUEENSLAND
1 MARCH 2010

Keywords

speech processing, automatic speech recognition, robust, adverse environments, speech enhancement, phase spectrum, phase estimation, optimisation, likelihood maximisation, automotive

Abstract

Automatic Speech Recognition (ASR) has matured into a technology which is becoming more common in our everyday lives, and is emerging as a necessity to minimise driver distraction when operating in-car systems such as navigation and infotainment. In "noise-free" environments, word recognition performance of these systems has been shown to approach 100%, however this performance degrades rapidly as the level of background noise is increased.

Speech enhancement is a popular method for making ASR systems more robust. Single-channel spectral subtraction was originally designed to improve human speech intelligibility and many attempts have been made to optimise this algorithm in terms of signal-based metrics such as maximised Signal-to-Noise Ratio (SNR) or minimised speech distortion. Such metrics are used to assess enhancement performance for intelligibility *not* speech recognition, therefore making them sub-optimal ASR applications.

This research investigates two methods for closely coupling subtractive-type enhancement algorithms with ASR: (a) a computationally-efficient Mel-filterbank noise subtraction technique based on likelihood-maximisation (LIMA), and (b) introducing phase spectrum information to enable spectral subtraction in the complex frequency domain.

Likelihood-maximisation uses gradient-descent to optimise parameters of the enhancement algorithm to best fit the acoustic speech model given a word sequence known *a priori*. Whilst this technique is shown to improve the ASR word accuracy performance, it is also identified to be particularly sensitive to non-noise mismatches between the training and testing data.

Phase information has long been ignored in spectral subtraction as it is deemed to have little effect on human intelligibility. In this work it is shown that phase

information is important in obtaining highly accurate estimates of clean speech magnitudes which are typically used in ASR feature extraction. Phase Estimation via Delay Projection is proposed based on the stationarity of sinusoidal signals, and demonstrates the potential to produce improvements in ASR word accuracy in a wide range of SNR.

Throughout the dissertation, consideration is given to practical implementation in vehicular environments which resulted in two novel contributions – a LIMA framework which takes advantage of the grounding procedure common to speech dialogue systems, and a resource-saving formulation of frequency-domain spectral subtraction for realisation in field-programmable gate array hardware.

The techniques proposed in this dissertation were evaluated using the Australian English In-Car Speech Corpus which was collected as part of this work. This database is the first of its kind within Australia and captures real in-car speech of 50 native Australian speakers in seven driving conditions common to Australian environments.

Contents

List of Tables

List of Figures

Acronyms & Abbreviations

ADC	Analogue-to-digital converter
AEICS	Australian English In-Car Speech corpus
ASR	Automatic speech recognition
AWGN	Additive white Gaussian noise
BSS	Blind source separation
CMS	Cepstral mean subtraction
CSS	Complex spectrum subtraction
DCT	Discrete cosine transform
DFT	Discrete Fourier transform
DSP	Digital signal processing
EM	Expectation maximisation
FPGA	Field programmable gate arrays
GMM	Gaussian mixture model
GSC	Generalised sidelobe canceller
HMM	Hidden Markov model
HVAC	Heating, ventilation and air-conditioning
IFT	Inverse Fourier transform
LIMA	Likelihood-maximisation
LM	Language model
LPC	Linear predictive coding
LSS	Linear spectral subtraction
MAP	Maximum *a posteriori*
MBSS	Multi-band spectral subtraction
MFCC	Mel-frequency cepstral coefficients

MFNS	Mel-filterbank noise subtraction
MLLR	Maximum-likelihood linear regression
MMSE	Minimum mean-sqare error
MS	Minimum statistics
PEDEP	Phase estimation via delay projection
PEF	Phase-error filtering
PLP	Perceptual linear prediction
PMC	Parallel model combination
PSC	Phase spectrum compensation
RASTA	Relative spectra
SAD	Speech activity detection
SDS	Speech dialogue system
SNR	Signal-to-noise ratio
TRA	Time-recursive averaging
XSG	Xilinx System Generator^{TM}

Authorship

The work contained in this thesis has not been previously submitted for a degree or diploma at any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signed: _____

Date: _____

Acknowledgements

No Ph.D. could be completed without the support provided by the supervisory team, fellow students, family and friends. This program was no exception!

A very big thank-you to the supervisory team – Prof. Sridha Sridharan and Dr Michael Mason – for all their hard work over the past four years. To Sridha – thank-you for providing the opportunity to enhance my research career by undertaking a Ph.D., as well as the chance to work on other research projects. The latter opened my eyes to benefits of research collaboration along with all the associated challenges. Thank-you also for the top-class research environment and resources you provide to all your students – these make a huge difference to the research experience and successful candidature completion.

To Michael, thank-you in particular for attending to all my technical concerns, for helping formulate responses to the numerous challenges along the way, and most of all for your approachability and friendship during this time. I hope I have been able to give something back through all the hours of tutoring and marking!

A number of other people have provided invaluable technical assistance along the way. To Darren Moore – a former student of the Speech, Audio, Image and Video Technology (SAIVT) laboratory – thanks for taking the effort to meet for early morning coffees and technical discussions. These discussions provided valuable feedback and opened new research directions some of which are mentioned in these pages, others which I hope to pursue. Also, thanks to various members of SAIVT – in particular Dr Eddie Wong, Dr Robbie Vogt, Ivan Himawan, and Chris Lustri – for your speech- and math-related technical assistance. A big thank-you must also go to Gleb Sechenov and Mark Cox for your computing support, particularly whilst I was overseas on internship. Without you all, experimentation would have taken much more time than it did!

To the rest of the members of SAIVT – thanks for your friendship, and those lunch-time card games which helped break the monotony of research and to reset the mind for an afternoon of hard work. I wish you all well in completing your Ph.D. research (to those whom it applies), and your future research endeavours.

I am very thankful for the opportunity given by Prof. John H. L. Hansen to undertake an internship at the Center for Robust Speech Systems (CRSS) at the University of Texas at Dallas during mid-2009. Whilst the timing may not have been ideal for completing this thesis, it opened my eyes to a much broader range of research challenges and methods. My full appreciation goes to Dr Pinar Boyraz and Dr Hynek Bořil for your collaboration which I hope will continue in the future. To the rest of the members of CRSS – thank-you for your friendship which was a remedy for the constant homesickness during this time!

It is also necessary to acknowledge the various sources of funding which provided support during this Ph.D. – without it the research would not have been possible. In particular, I would like to thank the Australian government for providing the Australian Postgraduate Award, Queensland University of Technology and the Vice Chancellor, as well as the Co-operative Research Centre for Advanced Automotive Technology (AutoCRC) and Prof. Sridharan for financial assistance, particularly during the latter stages of this research. To the AutoCRC, thank-you for also providing opportunities to engage with other Ph.D. students across Australia, and for giving us access to resources not available to most Ph.D. students. My only regret was not being able to take part in a number of activities because of the need to travel to Melbourne to attend.

Last, but definitely not least, a thank-you to my family and friends for your continued support throughout the entire candidature. In particular, thanks to my parents Fred & Cherie for the Saturday morning chats, for listening to all my concerns, and for your affirmation that I would eventually reach this point. And finally to my wonderful wife Colleen for your uplifting words and smiles at the end of a long week, and for your full support whilst I undertook a 3 month internship on the other side of the world. Your generosity was unsurpassed, and I hope now that this dissertation is finished, I can return to a normal life!

Chapter 1

Introduction

1.1 Motivation

At its core, Automatic Speech Recognition (ASR) is the process of determining a sequence of words spoken by a human using machines. From its earliest beginnings in vowel and isolated digit recognition in the 1950s and then connected speech recognition in the 1970s [117], ASR systems have matured to the point of widespread deployment in an ever-increasing range of applications. Small vocabulary systems (typically less than 100 words) are used in command and control applications such as computer-based games and entertainment, voice dialling on mobile phones, and for providing instructions for navigating menus on portable devices such as GPS and personal music players. Large vocabulary systems (greater than 1,000-2,000 words) are required for office dictation applications and captioning of television programs.

One of the emerging target applications for ASR is that of human-machine interfaces for automotive environments. As consumers become more accustomed to the use of hand-held devices such as mobile phones, navigation systems and music players in their everyday lives, there is an increasing demand for these devices to be integrated with their vehicles for use whilst driving. Research has shown however, that during the primary driving task, visual cognitive resources can easily reach the point of overload [89] without the introduction of secondary visual stimuli. Therefore, adding visual displays only further increases driver

cognitive workload in this perceptual channel which will likely result in unsafe driving behaviour. Speech-based interfaces on the other hand, utilise the auditory perception channel which is generally rarely used during the primary driving task [96]. As a result, speech recognition is viewed as a key enabler to interfacing portable devices and in-car informational systems whilst minimising the effects on driver behaviour and overall road safety.

At the turn of the 21^{st} century, state-of-the-art speech recognition systems operating in "noise-free" environments were able to produce word recognition accuracies exceeding 95% accuracy on well-defined large vocabulary tasks and approaching 100% on small vocabulary tasks [58]. Clearly, speech recognition technology is no longer a fantasy confined to the realms of science fiction – it has already reached the levels of user expectation in these "noise-free" environments.

Despite this success, the performance of the same ASR systems degrades significantly in the presence of even moderate levels of environmental noise. This phenomenon is particularly problematic in automotive applications where the level and type of noise continually changes as the driver negotiates their desired route. To counteract this rapid decrease in word recognition accuracy, focus in speech recognition research in the past 15 years has centred on making systems robust in the noisy scenarios typical of crowded airports, street, and restaurants, as well as in-car environments and airplane cockpits [29].

A number of approaches have been proposed in order to increase the robustness of ASR systems ([27, 46, 81] provide comprehensive reviews). One of these approaches – speech enhancement – can be ported between a range of different environments and used with a wide range of speech recogniser configurations with little to no modifications; this makes it a widely effective solution. Recent advances in speech enhancement techniques have come in the form of multiple microphones which enable spatial filtering and improved enhancement performance. The automotive industry necessitates low-cost manufacturing, therefore single-channel solutions are preferred over multi-channel systems which are still too expensive at this point in time.

The term speech enhancement is often used interchangeably with noise reduction since algorithms are typically designed to improve the intelligibility of noisy

speech signals as perceived by humans through either emphasising speech components or removing/reducing the noise components. Given the initial design for improving human perception, implementing these techniques for computer-based ASR requires different signal processing and changes the evaluation criteria. Despite showing small improvements in speech recognition accuracy, enhancement techniques designed primarily for human intelligibility are not ideal for ASR application. This mismatch between design and application is slowly being acknowledged by the research community, and researchers are beginning to divert their attention to designing enhancement techniques specifically for use in ASR front-end processing [10, 127, 130]. This small change of focus in the field of robust ASR has motivated the research contained in this dissertation.

1.2 Aims and Objectives

Speech enhancement was originally designed for improving human speech intelligibility, however research has typically considered the recognition system and speech enhancement algorithm as separate entities [127]. An example of this traditional approach is the European Standard ES 202 050 [33] which stipulates an advanced speech recognition front-end in which enhanced waveforms – not speech features – are generated based on signal-level criteria rather than criteria related to ASR. The focus on signal-level criteria has been identified as a major problem of current speech enhancement approaches for ASR [127].

In the literature to date, only a few examples of speech enhancement for robust ASR have explicitly taken into account the operation of the recognition system [10, 127, 130]. This dissertation aims to extend these existing approaches and discover new methods by which speech enhancement can be designed for more effective use in robust ASR.

The general aims of this thesis are:

1. To demonstrate that speech enhancement techniques optimised for human intelligibility are sub-optimal for integration with state of the art speech recognition systems.

2. To propose novel techniques which improve current speech enhancement algorithms when used as part of the front-end processing for in-car ASR.

3. To consider the implementation of speech enhancement algorithms within the constraints of the automotive environment.

In order to achieve these aims, the specific research objectives are:

1. To quantify the word accuracy performance of ASR systems when speech is collected in real car environments and is therefore corrupted by a wide range of in-vehicle noise conditions.

2. To analyse the effectiveness of traditional speech enhancement and model adaptation techniques for increasing noise-robustness of ASR systems.

3. To analyse how subtractive-type enhancement algorithms have been previously used for speech enhancement and ASR and identify the shortfalls of these approaches in terms of the resulting ASR performance.

4. To propose novel speech enhancement algorithms which directly improve the performance of the underlying speech recognition engine.

5. To design frameworks which are suitable for integration with existing in-car speech systems, and where possible, are designed with computational and hardware requirements in mind.

6. To assess each of the proposed techniques and report their performance based on speech recognition accuracy and computational requirements.

1.3 Scope of Research

Robust automatic speech recognition is a very broad area of research which encompasses the fundamentals of speech recognition such as acoustic and language model representation, pattern recognition and signal processing, as well as all the potential methods for increasing robustness. It is therefore very important to fully define the scope of this thesis in order to ensure the attainment of the research aims. The scope has been defined as:

Single-channel speech enhancement. Despite showing the potential to provide superior speech enhancement performance, systems employing multiple microphones are still some way from widespread deployment in the automotive industry due to their increased manufacturing costs. In order to ensure the techniques developed in this dissertation are suitable for improving the *current* state of the art in-vehicle speech systems, detailed analysis and novel contributions are focused solely on single-channel speech enhancement techniques.

Spectral subtraction. This widely used single-channel speech enhancement technique has undergone a wide range of alterations in the past 30 years in order to provide better performance under different experimental configurations. Spectral subtraction is chosen as it is a computationally simple solution to the additive noise problem yet provides sufficient levels of noise reduction. This cost-performance trade-off is a very important factor when considering application in automotive environments.

Small-to-medium vocabulary speech recognition. In order to evaluate the improvements in speech recognition performance that can be achieved using the proposed speech enhancement techniques, small and medium vocabulary tasks were chosen as both are common in automotive applications. Small vocabulary tasks are used for command and control of non-critical functions such as adjusting the air-conditioning or entertainment systems. Medium vocabulary systems find their applications in address entry for context-aware navigation systems as well as communication with information retrieval services via the internet – both of these applications are becoming common in luxury vehicles.

Real in-car noise environments. Whilst the techniques developed in this research could be easily applied in other noisy environments, evaluations are focused solely on in-car speech data. Importantly, data collected in a moving vehicle is generally preferred over artificially generated data as this "real-world" data will exhibit real-time variations in noise conditions and also incorporate (to some extent) the effects of driver stress on speech production.

1.4 Outline of Dissertation

The remainder of this dissertation is organised as follows:

Chapter 2 presents the fundamental theory behind automatic speech recognition including speech parameterisation, acoustic modeling and decoding. Whilst ASR is not a new field of research, it is imperative to understand the recognition process in order to analyse and develop techniques aimed at improving recognition performance in the case of train-test mismatches. Three common methods for improving speech recognition performance in noisy environments are discussed with reference to challenges facing implementation of ASR in automotive applications. These techniques are speech enhancement, model adaptation, and robust feature extraction and recognition algorithms.

Chapter 3 reviews state of the art techniques for both single- and multi-microphone speech enhancement citing examples of implementations in automotive environments. Theory and implementation of various forms of spectral subtraction are presented in detail to provide reference for the novel contributions of this research. This literature review highlights the traditional focus of speech enhancement algorithms on optimising signal-based criteria; research directions focusing on optimising spectral subtractive algorithms specifically for ASR applications are proposed in order to guide the remainder of the research.

Chapter 4 describes two in-car speech databases used for ASR evaluation throughout this thesis – the AVICAR database and the Australian English In-Car Speech (AEICS) corpus which was collected as part of this research. Baseline ASR performance of each dataset is obtained and compared with that of model adaptation and three different implementations of spectral subtractive speech enhancement.

Chapter 5 introduces LIkelihood-MAximising (LIMA) speech enhancement designed specifically for robust ASR. These techniques optimise enhancement parameters based on maximising the speech recognition likelihood as opposed to signal-level criteria. A review of previous LIMA studies leads to the application of this approach to Mel-Filterbank Noise Subtraction (MFNS) in order to reduce the computational requirements of other single-channel implementations.

The inclusion of cepstral liftering is proposed to overcome differences in dynamic range between cepstral coefficients which can bias gradient-descent optimisation towards components with larger magnitudes. ASR and computational performance of LIMA-based MFNS is obtained with respect to optimising each of the enhancement parameters, application of cepstral liftering, and using enhancement with noise-adapted acoustic models.

Chapter 6 proposes a dialogue-based LIMA framework specifically for implementation in automotive applications. This framework utilises the grounding procedure which is common to all speech dialogue systems. ASR metrics are used to compare this proposed framework with existing frameworks based on one-time adaptation. This chapter also considers the trade-off between processing time and ASR performance (which is very important in the automotive industry) by analysing the effects of different levels of optimisation.

Chapter 7 investigates the use of phase information for improving frequency-domain spectral subtraction by performing the enhancement in the complex domain as opposed to the magnitude domain. Potential ASR improvements using Complex Spectral Subtraction (CSS) are demonstrated using oracle-type experimentation. A method for estimating the phase information is proposed based upon the stationarity of sinusoidal signals which enables a reference phase to be projected forward knowing the time between two observations. Further oracle-type experiments are used to evaluate the necessary frame advances required in order to exploit phase stationarity. Evaluation is completed by incorporating soft-decision speech activity detection (SAD) to project the noise phase estimate through periods of speech, making the proposed technique viable in real-world scenarios. All experiments in this chapter demonstrate the effectiveness of the proposed phase estimation procedure and CSS for improving speech recognition performance in noisy environments.

Chapter 8 describes simplifications to the traditional frequency-domain spectral subtraction algorithm specifically for cost-effective, real-time implementation in Field Programmable Gate Array (FPGA) hardware. The resulting ASR performance of this implementation is comparable to that of an equivalent floating-point model and uses minimal amounts of FPGA resources, allowing integration with

other in-car processing modules.

Chapter 9 summarises the major outcomes of this research with particular reference made to the novel contributions outlined in Section 1.5. Directions for future research which extend the work in this dissertation are also suggested.

1.5 Original Contributions

1.5.1 Major Contributions

1. *The collection and validation of the first in-car speech database recorded with native Australian speakers in Australian driving conditions.* Unlike other in-car speech corpora of this size, driver speech data was collected rather than passenger speech. Multi-channel recordings of navigation addresses and menu commands were obtained for 50 speakers in seven different noise conditions. Being the first of its kind, this database will impact both the Australian speech and natural language research community and the automotive industry.

2. *Application of Mel-filterbank noise subtraction to a likelihood-maximising speech enhancement framework specifically for in-car speech recognition.* Traditionally, enhancement techniques of this nature have considered speech enhancement and ASR as separate systems; the LIMA approach considers both as one entity. In this dissertation, MFNS was mathematically derived for use in this framework, and this approach was also shown to provide a computationally efficient solution compared to frequency-domain spectral subtractive speech enhancement. The LIMA-based MFNS system was integrated within a newly proposed dialogue-based optimisation framework specifically for use in car environments, and its ASR performance in a range of scenarios was evaluated and compared with traditional calibrated frameworks. The analysis showed significant improvements in ASR performance using the proposed framework, and led to a number of recommendations for use in vehicular environments to ensure optimal ASR performance and satisfy other requirements implicitly imposed by the automotive sector.

3. *The use of the short-time phase spectrum to improve the ASR performance of frequency-domain spectral subtraction.* In this work it is shown that ignoring the phase information in frequency-domain spectral subtraction results in unavoidable errors in the cleaned magnitude spectrum which is used in common feature representations. Spectral subtraction is subsequently reformulated to be performed in the complex frequency domain making use of a pioneering short-time phase estimation procedure called Phase Estimation via Delay Propagation. Experiments demonstrate the importance of phase information for robust ASR using spectral subtraction, and verify the effectiveness of the PEDEP algorithm in a range of signal-to-noise ratios.

1.5.2 Other Contributions

1. *A speaker-independent, continuous ASR evaluation protocol for the AVICAR database.* This protocol extends those released with the database, enabling adaptation, development and evaluation testing on continuous speech tasks, whilst ensuring that reliable comparisons can be made between single- and multi-microphone speech enhancement techniques on the same data set; such comparisons are not always possible with other corpora.

2. *Evaluation of LIMA-based enhancement on test data incorporating multiple layers of acoustic mismatch.* No previous studies on LIMA-based enhancement have used test data consisting of acoustic mismatches other than background noise. In this dissertation, a second level of mismatch (speaker dialects) is introduced using the Australian English In-Car Speech corpus. The ASR performance of the LIMA framework is demonstrated to be highly sensitive to this second level of mismatch.

3. *Simplification of the frequency-domain spectral subtraction algorithm for cost-effective, real-time implementation in FPGA hardware.* This work lead to a minimal resource solution – which closely matched the ASR performance of a floating-point model – being ported to an automotive-grade FPGA.

1.6 Publications Resulting from Research

The following fully-refereed publications have been produced during the course of this PhD research:

1. **T. Kleinschmidt**, S. Sridharan, M. Mason, "A Modified LIMA Framework for Spectral Subtraction Applied to In-Car Speech Recognition," in *Proceedings of 1st International Conference on Signal Processing and Communication Systems*, (Gold Coast, Australia), pp. 335-338, December 2007.

2. **T. Kleinschmidt**, D. Dean, S. Sridharan, M. Mason, "A Continuous Speech Recognition Protocol for the AVICAR Database," in *Proceedings of 1st International Conference on Signal Processing and Communication Systems*, (Gold Coast, Australia), pp. 339-344, December 2007.

3. **T. Kleinschmidt**, M. Mason, E. Wong, S. Sridharan, "The Australian English Speech Corpus for In-Car Speech Processing," in *Proceedings of 34th IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Taipei, Taiwan), pp. 4177-4180, April 2009.

4. **T. Kleinschmidt**, S. Sridharan, M. Mason, "Likelihood-Maximising Frameworks for Enhanced In-Car Speech Recognition," in *Proceedings of 4th Biennial Workshop on DSP for In-Vehicle Systems and Safety*, (Dallas, TX, USA), paper DSP08, pp. 1-8, June 2009.

5. **T. Kleinschmidt**, P. Boyraz, H. Bořil, S. Sridharan, J. H. L. Hansen, "Assessment of Speech Dialog Systems using Multi-Modal Cognitive Load Analysis and Driving Performance Metrics," to be presented at *2009 IEEE International Conference on Vehicular Electronics and Safety*, (Pune, India), pp. 167-172, November 2009.

6. J. Whittington, K. Deo, **T. Kleinschmidt**, M. Mason, "FPGA Implementation of Spectral Subtraction for In-Car Speech Enhancement and Recognition," in *Proceedings of 2nd International Conference on Signal Processing and Communication Systems*, (Gold Coast, Australia), December 2008.

7. J. Whittington, K. Deo, **T. Kleinschmidt**, M. Mason, "FPGA Implementation of Spectral Subtraction for Automotive Speech Recognition," in *Proceedings of IEEE Workshop on Computational Intelligence in Vehicles and Vehicular Systems*, (Nashville, TN, USA), pp. 72-79, March-April 2009.

8. H. Ye, J. Whittington, I. Himawan, **T. Kleinschmidt**, M. Mason, "FPGA Implementation of Dual-Microphone Delay-and-Sum Beamforming for In-Car Speech Enhancement and Recognition," in *Proceedings AutoCRC Conference*, (Melbourne, Australia) March 2009.

1.7 Research at CRSS, University of Texas at Dallas

During the latter stages of this candidature, research was conducted as part of a three month internship at the Center for Robust Speech Systems (CRSS) at the University of Texas at Dallas (UTD). Due to the timing of this visit, and the nature of the work performed during this time, much of the research undertaken was not directly applicable to the work contained in this dissertation. Where possible, passing references are made to demonstrate the research undertaken, to assist the discussion, and to also demonstrate progress towards some of the proposed future research directions. These references can be found in footnotes in the appropriate sections of this dissertation.

Chapter 2

Automatic Speech Recognition

2.1 Introduction

Automatic speech recognition (ASR) is the process of converting a sequence of words contained in an acoustic signal into a textual representation. ASR is used in a range of applications including dictation, device command-and-control, audio-based keyword searching, and in security systems. This chapter provides a summary of the ASR fundamentals relevant to the work contained in this dissertation – a complete review of ASR technology is beyond the scope of this work.

The typical components of an ASR system are shown in Fig. 2.1. Speech feature extraction consisting of signal acquisition and parameterisation (Section 2.2) is required to reduce the dimensionality of the pattern recognition system whilst emphasising the distinguishing characteristics of speech. Acoustic models based on the chosen feature set are trained using data from a specific operating environment, and these models are used with pronunciation dictionaries and language models in the decoding process. Recognition fundamentals using acoustic and language models are detailed in Section 2.3.

Current state of the art ASR systems perform remarkably well in controlled conditions. Under more adverse conditions such as noisy or reverberant environments, speech recognition performance decreases dramatically. There are a number of different approaches to making ASR systems more robust – these methods are described in Section 2.4.

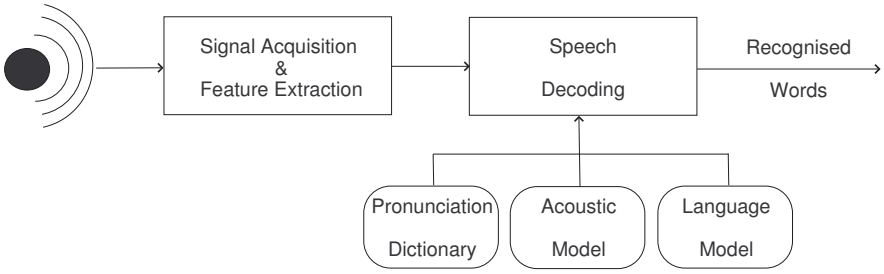


Figure 2.1: Block diagram depicting the major components of an ASR system.

2.2 Speech Feature Extraction

The purpose of the feature extraction module of Fig. 2.1 is to reduce the data rate of the incoming acoustic signal, to condition the signal in order to remove background noise and obtain the features of speech which are most useful for speech recognition. An expanded view of the feature extractor is shown in Fig. 2.2. The following sections describe the signal acquisition and preparation stage, as well as discussing some common speech parameterisation techniques.

2.2.1 Signal Acquisition and Preparation

An acoustic signal containing the word(s) to be recognised is received by a microphone or an array of microphones. Before the digital signal processing elements of the feature extractor can operate on the signal, the analogue waveform is sampled by an analogue-to-digital converter (ADC) to create a digital signal. For speech recognition, the common sampling rates used by the ADC are 8 kHz and 16 kHz. The higher sampling rate produces the best recognition performance since the majority of the useful information in the speech signal lies within the 8 kHz bandwidth [58]; a 16 kHz sampling rate ensures the Nyquist sampling criterion is satisfied.

A well known characteristic of audio signals is the tendency for high frequencies to have less energy than low frequencies – a phenomenon referred to as spectral slope [36]. Speech recognition systems utilise information from the entire frequency spectrum, therefore it is necessary to equalise the dynamic range to offset the spectral slope. A pre-emphasis filter boosts the signal energy of the higher frequencies which contain the majority of the speech information. Pre-emphasis

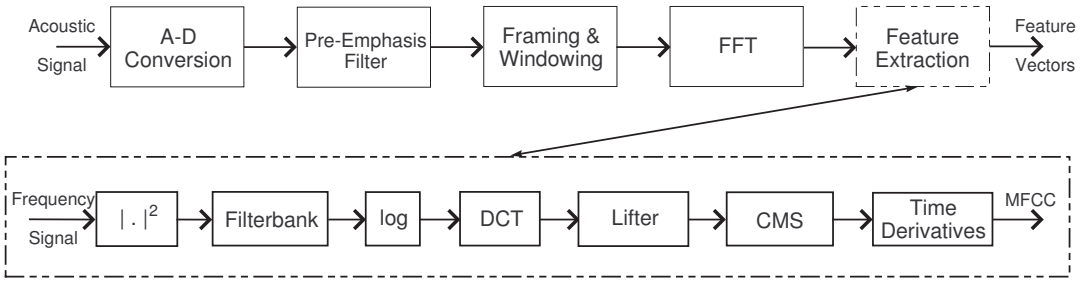


Figure 2.2: Signal acquisition, preparation and feature extraction using Mel-Frequency Cepstral Coefficients.

filtering is achieved using a first-order finite impulse response filter:

$$y[n] = x[n] - ax[n-1] \tag{2.1}$$

where the value of the filter coefficient is typically $0.9 < a < 1.0$ for speech processing applications. In this research, a value of $a = 0.97$ is used.

Spectral analysis methods used in speech parameterisation are based on short-time analysis of speech signals which assumes the time segment is short enough that the signal exhibits short-time stationarity. Such analysis requires the pre-emphasised speech signal to be divided into a series of overlapping frames. For speech processing, frames are typically 20-40 ms in length with 10-20 ms advances between adjacent frames.

Spectral leakage is caused by discontinuities introduced at both ends of every frame. The distortion caused by spectral leakage is reduced by applying a window function to each frame which causes the frame samples to be tapered towards the frame boundaries. A commonly used window in speech processing which provides a suitable tradeoff between spectral leakage and resolution is the Hamming window:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N}\right), \qquad n = 0, 1, ..., N-1 \tag{2.2}$$

where N is the length of the frame, and n is the sample index within the frame.

For feature extraction methods including Mel-Frequency Cepstral Coefficients (MFCC), each frame of speech is converted to the frequency domain using a Discrete Fourier Transform (DFT):

$$Y(k) = \frac{1}{N}\sum_{n=0}^{N-1} y(n)w(n)e^{-j\frac{2\pi kn}{N}} \tag{2.3}$$

where $y(n)$ is the pre-emphasised signal and $w(n)$ is the Hamming window. The resulting frequency domain signal $Y(k)$ is defined only at discrete frequencies (through a binning process) and is a complex-valued representation which contains the magnitude $|Y(k)|$ and phase spectra $e^{j\theta(k)}$ of the signal:

$$Y(k) = |Y(k)|e^{j\angle Y(k)}. \tag{2.4}$$

2.2.2 Speech Parameterisation

As mentioned previously, parameterisation techniques are required to reduce the dimensionality of the pattern recognition problem encountered in ASR. Speech representations used for feature extraction must be compact whilst ensuring the distinctive characteristics of speech are preserved. Speech signal representations used in ASR can be classified broadly into two types: (a) methods which model the speech production process, and (b) those which model speech perception.

Representations using Speech Production Models

Linear Predictive Coding (LPC) [8] utilises an all-pole filter to approximate the vocal tract. LPC estimates the current speech sample given p previous samples by minimising the error between the predicted and actual sample value. Error minimisation is achieved by utilising the autocorrelation method to calculate the all-pole filter coefficients.

While this model is effective for voiced sounds, LPC doesn't perform well for unvoiced sounds which introduce zeros into the speech model. Additive noise also introduces zeros, and therefore the LPC representation doesn't perform as well in additive noise as representations derived from the Fourier transform magnitude spectrum. As a result, representations based on the Fourier transform are generally favoured for ASR systems.

Perceptually Motivated Representations

Perceptually motivated speech representations capitalise on knowledge of the human auditory system. The two most common feature extraction methods incorporating perception models are MFCC and Perceptual Linear Prediction (PLP).

The MFCC representation is used extensively in this research, and is described separately in the next section.

Perceptual linear prediction [54] uses the Levinson-Durbin algorithm to perform linear predictive analysis as per LPC. Before the LPC analysis, Fourier transform analysis (as per Eq. (2.3)) is required to determine the power spectrum of the signal. Using a series of filterbanks, the power spectrum is warped to a frequency axis known as the Bark scale which approximates the known human hearing filters [54]. Equal-loudness filtering and intensity-loudness non-linear compression are applied to the filterbank outputs to produce a perceptually motivated power spectrum.

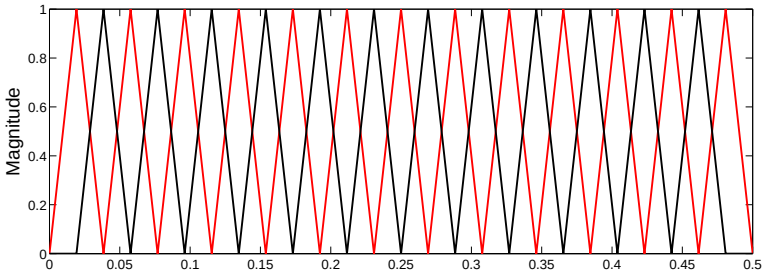
The LPC coefficients are calculated using the Inverse Fourier Transform (IFT) since the autocorrelation function is the IFT of the power spectrum. The resulting LPC coefficients are typically transformed to the LPC-cepstrum, where the cepstrum is the logarithm of the all-pole filter (more on the cepstrum in the following section). Whilst LPC produces a finite number of coefficients, the cepstral transform results in an infinite number of cepstral coefficients; research has shown that 12-20 coefficients are sufficient for ASR applications [54, 114].

Mel-Frequency Cepstral Coefficients

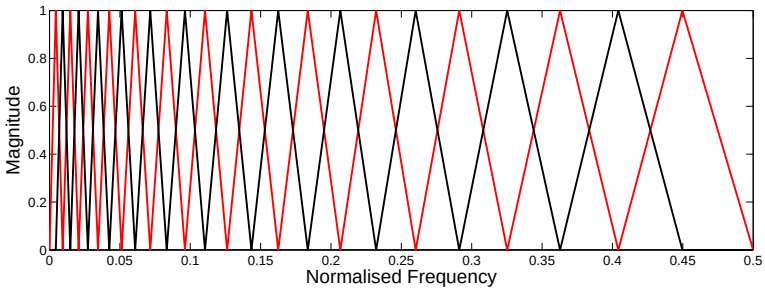
Mel-frequency cepstral coefficients are a perceptually motivated speech representation based on Fourier transform and filterbank analysis (as shown in Fig. 2.2). It was proposed in 1980 by Davis and Mermelstein [28] and has regularly been shown to be superior to other feature representations for ASR on clean speech. The MFCC representation has a particular advantage over LPC in that it is more robust to background noise; this was demonstrated for car environments in [85].

MFCCs are based on the Mel-frequency scale which describes the behaviour of the human auditory system whereby a *perceived* halving (or doubling) of the frequency is a *true* halving (or doubling) in the Mel-scale [138]. The Mel-scale is approximately linear below 1 kHz and logarithmic for all frequencies greater than 1 kHz. This frequency warping is approximated by:

$$W(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.5}$$



(a)



(b)

Figure 2.3: Linear filterbanks in (a) Hertz, and (b) Mel-frequency scale.

where f is the linear frequency scale (in Hertz), and $W(f)$ is the perceived frequency (in Mel). This frequency-warping procedure is similar to that used in PLP where the Mel-frequency scale is used in place of the Bark scale.

The Mel-warping is applied through filterbank analysis which produces a weighted sum of band-limited power spectrum values $|Y(k)|^2$. An example of linear-frequency filterbanks in the Mel scale is shown in Fig. 2.3. Filterbank energies are used to generate the cepstrum as they are a more robust representation of speech than the power spectrum which exhibits fine spectral harmonics at multiples of the fundamental frequency. Despite this, a large number of filterbanks are required for high-performance ASR and therefore cepstral representations are more effective.

The conversion of filterbank energies to the cepstrum is performed using logarithmic compression and a Discrete Cosine Transform (DCT):

$$C_l = \sum_{n=0}^{N-1} \log_{10}(M_l) \cos\left(\frac{\pi l(n+\frac{1}{2})}{N}\right) \tag{2.6}$$

where M_l is the filterbank energy of the l^{th} filterbank, N is the number of cepstral coefficients, and C_l are the cepstral coefficients. The coefficient C_0 provides a measure of the energy in the signal and is often included in the feature vector as phonemes tend to have differing energy levels. The DCT is used to decorrelate the data (since filterbank energies are highly correlated due to the overlap of adjacent filters), and to also remove one-sided distributions which would be present due to filterbank energies always being positive. Both of these characteristics are important for ASR systems which rely on Gaussian-like distributions.

The DCT also causes the signal energy to be contained in the lower frequencies which enables reductions in dimensionality. Compacting the energy into the lower frequencies however, results in a variation in the dynamic range of the cepstral coefficients. As a result, a cepstral lifter can be incorporated (i.e. a filter applied to the cepstrum) [61]:

$$C_l' = \left(1 + \frac{L}{2}\sin\left(\frac{\pi l}{L}\right)\right) C_l \tag{2.7}$$

where L is the order of the lifter, and C_l' are the cepstrally liftered coefficients. Cepstral liftering will be assessed in Chapter 5 as a method to overcome one of the limitations of likelihood-maximisation on MFCC.

The transform to the cepstrum also provides the ability to remove the effect of the channel response. The speech signal in the time-domain $s(n)$ propagates through the communication channel which has an impulse response $h(n)$. This is represented as a convolution in the time-domain:

$$x(n) = s(n) * h(n). \tag{2.8}$$

Due to the properties of the Fourier transform, time-domain convolution becomes multiplication in the frequency-domain. Application of the logarithm turns the frequency-domain multiplication into an addition in the cepstrum. If the channel response is assumed to be constant over long periods of time, techniques such as Cepstral Mean Subtraction (CMS) [37] can be used to remove the channel response. The mean cepstrum is typically calculated over an entire recording of speech, but can operate in real-time through recursive-averaging techniques. Throughout this research, the mean calculation is taken over the entire recording.

Speech recognition systems using Hidden Markov Models (HMM) assume that each frame of speech is independent, therefore no temporal information is used which indicates how individual speech sounds evolve. First- and second-order temporal derivatives (also referred to as delta and acceleration coefficients) calculated over short periods (e.g. 5 frames) can be included in the feature vector to provide complementary information to the HMM speech recogniser. These features were first proposed by Furui [38] and have been shown to be particularly robust as they are less sensitive to slowly varying noise than the static cepstral coefficients from which they are calculated.

Temporal derivatives are calculated using linear regression over successive frames. Given frame j as the current frame of reference (and therefore central frame in the regression calculation), first-order derivative features can be calculated as follows:

$$\Delta C_l^j = \frac{\sum_{d=1}^{D} d(C_l^{j+d} - C_l^{j-d})}{2\sum_{d=1}^{D} d^2} \tag{2.9}$$

where D is the order of regression. Second-order derivatives are calculated as the linear regression of the first-order coefficients as per Eq. (2.9). A full MFCC feature vector typically includes 39 features consisting of 13 cepstral coefficients (including C_0), 13 first-order derivatives, and 13 second-order derivatives.

2.3 Speech Recognition Fundamentals

Humans communicate worded messages by converting them into a sequence of speech sounds or acoustic events. The role of the automatic speech recogniser is to reverse engineer the underlying message given a sequence of acoustic observations. The effectiveness of ASR systems can be attributed to the choice of acoustic (speaker-dependent versus speaker-independent) and language models, and particulars of the application.

As an example, consider a small vocabulary system for number entry on a mobile phone. Speaker-dependent ASR is a perfect choice for this situation as model training on a speaker-by-speaker basis can be performed very quickly. Dictation on the other hand, is a large vocabulary task, and therefore speaker-independent ASR is more appropriate in order to train effective acoustic models

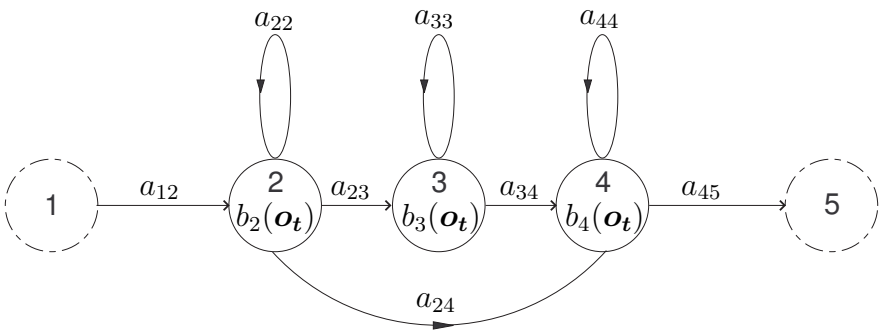


Figure 2.4: Hidden Markov Model typically used for ASR.

and deploy to large populations. Further, in terms of language models, the mobile phone task requires a very simple digit loop, however the dictation task requires a far more complex model as the probability of word combinations must be taken into account. The scope of the discussion and experiments in this dissertation is limited to small and medium vocabulary speaker-independent ASR.

2.3.1 Acoustic Modeling using Hidden Markov Models

The feature extraction methods outlined in Section 2.2.2 generate observation sequences $\boldsymbol{O}$ defined as:

$$\boldsymbol{O} = \boldsymbol{o_1}, \boldsymbol{o_2}, \boldsymbol{o_3}, ..., \boldsymbol{o_T} \tag{2.10}$$

where $\boldsymbol{o_t}$ is the observation at time t. The observed sequence of feature vectors is assumed to be generated by a finite state machine known as a Markov model which enables non-stationary speech signals to be transformed into piecewise stationary states. For speech recognition, only the acoustic feature vectors from the speech signal of interest are known; therefore hidden Markov models are used to determine the unknown (i.e. hidden) state sequence s which generated the observed sequence of feature vectors $\boldsymbol{O}$. HMM-based recognition systems are used throughout this research, and are the basis for the likelihood-maximisation technique discussed in Chapter 5 and Chapter 6.

The simple five state left-to-right HMM shown in Fig. 2.4 consists of three emitting and two non-emitting (entry and exit) states. The non-emitting states are required to chain together multiple HMM for continuous sub-word speech

recognition. The emitting states (i.e. states 2, 3 and 4 in this example) provide all the statistical information required for training acoustic models and recognising speech signals.

There are two important parameters contained within the model; the first is the observation probability density $b_j(\boldsymbol{o_t})$ which determines the probability of generating observation $\boldsymbol{o_t}$ from the emitting model state j. In ASR applications, probability densities are commonly represented by Gaussian Mixture Models (GMM). A GMM is comprised of multivariate Gaussian probability density functions represented by:

$$\aleph(\boldsymbol{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{o}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{o}-\boldsymbol{\mu})} \tag{2.11}$$

where $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Sigma}$ is the covariance matrix and n is the dimensionality of the feature vector. Given a number of mixture components M, the observation probability density becomes:

$$b_j(\boldsymbol{o_t}) = \sum_{m=1}^{M} \gamma_{jm} \aleph(\boldsymbol{o_t}; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}) \tag{2.12}$$

where γ_m is the weight of the m^{th} mixture component.

In determining the likelihood of a given state sequence, it is also necessary to include the transition probabilities between states. In Fig. 2.4, transition probabilities a_{ij} are defined for all allowable state transitions. For each model, the sum of all transition probabilities will be 1. It should be noted that each state of the HMM can generate consecutive acoustic observations as transitions within the same state a_{jj} are permitted.

A key consideration for HMM-based speech recognition are the acoustic units which are represented by each HMM. For small vocabulary and isolated word recognition tasks such as digit recognition, a popular approach is to represent each word by a single HMM [115]. This approach does not scale well to medium and large vocabulary tasks which contain several hundred to several thousand words. As a result, sub-word units based on phonemes are used for large vocabulary continuous speech recognition. Using sub-word representations requires a pronunciation dictionary which maps all words in the recognition vocabulary into corresponding sub-word sequences.

Monophone models consist of an HMM for each basic phonetic unit for a particular language (e.g. there are approximately 43 phonemes used in the English language). Triphone models incorporate context-dependency in terms of the phones which occur before and after each phoneme [124]; this is sometimes referred to as left-and-right context-dependency. Whilst the use of triphone models requires more training data and results in a larger acoustic model, it does also provide better speech recognition performance [74]. Triphone models are used for all acoustic modeling in this thesis.

2.3.2 Recognition using Viterbi Decoding

The recognition problem can be viewed as determining the sequence of words $\hat{w}$ with the maximum likelihood of all possible word sequences W:

$$\hat{w} = \arg\max_{w \epsilon W} P(w|\boldsymbol{O}) = \arg\max_{w \epsilon W} \frac{P(\boldsymbol{O}|w)P(w)}{P(\boldsymbol{O})} \tag{2.13}$$

which is determined using Bayes' Rule shown on the right side of Eq. (2.13). In this equation, $P(\boldsymbol{O}|w)$ is the *acoustic score* representing the probability that the observation sequence $\boldsymbol{O}$ was generated by the word sequence w, and $P(w)$ is the *language model score* which is explained in the following section. The term $P(\boldsymbol{O})$ can be ignored since it represents a fixed sequence of observations and will be the same for all possible word sequences. This results in the recognition hypothesis:

$$\hat{w} = \arg\max_{w \epsilon W} P(\boldsymbol{O}|w)P(w). \tag{2.14}$$

Since continuous speech recognition takes place on sub-word units, Eq. (2.14) can be represented as the sum of all states in the state sequence $\hat{s}$:

$$\hat{s} = \arg\max_{s \epsilon S} \sum_{s} \left(\prod_{i} P(\boldsymbol{o_i}|s_i)P(s_i|s_{i-1}, w) \right). \tag{2.15}$$

The Viterbi algorithm [142] is used to determine the best possible state sequence $\hat{s}$ for the given observation sequence $\boldsymbol{O}$. The forward log-likelihood of occupying state j at time t is calculated as:

$$\psi_j(t) = \max_{i} \{\psi_i(t-1) + \log(a_{ij})\} + \log(b_j(\boldsymbol{o_t})) \tag{2.16}$$

where $\psi_i(t)$ is the log-likelihood of the previous model state. The recognition log-likelihood is therefore dependent upon the log-likelihood of all previous states in the sequence.

In order to maintain all possible paths to time t during the decoding process, a token passing model [156] is employed. Using this model, each state i of the HMM at time $t-1$ contains a token which consists of (among other information) the current log-likelihood $\psi_i(t-1)$ which is passed to all connecting states j in the network at time t. In doing so, the log-likelihood for the partial path token is continuously updated and the best possible paths are maintained. After all tokens have been passed to the next state, they are examined, and the least likely tokens are discarded. The route taken by the most likely token to the end of the observation sequence represents the recognition hypothesis.

2.3.3 HMM Parameter Estimation

Prior to recognition, the transition probabilities and output probability densities must be estimated through training. Using a well labeled set of training data large enough to contain sufficient examples for each triphone, parameter estimation can be performed using an algorithm such as Baum-Welch re-estimation [11] or the Expectation Maximization (EM) algorithm [59]. All acoustic models in this dissertation have been trained using Baum-Welch re-estimation, and so all discussion is based on this method of training.

In Baum-Welch re-estimation, an initial estimate of the parameters is obtained by distributing all training observations between all the required model states. The re-estimation procedure then assigns each observation vector to each state in proportion to the probability of the model being in that particular state when the feature vector was observed. In other words, feature vectors which are most likely to have been generated by a particular model state contribute more to the final parameter values than other less likely feature vectors.

In order to calculate the proportionalities for model training, the probability of state occupation must be calculated using the Forward-Backward algorithm. The forward probability is the joint probability described by Eq. (2.16); that is, the likelihood of observing all previous feature vectors and being in state j at

time t. The backward probability is the conditional probability of observing all subsequent feature vectors given the model was in state j at time t. The overall state occupation probability is calculated as the multiplication of the forward and backward probabilities.

The forward and backward probabilities are calculated for each state and time for each example of the model token and all model parameters (including the transition probabilities) are updated. The process continues until the observation likelihood converges for each model state. This procedure is very time consuming, however models can be trained in parallel using the above procedure for each triphone.

2.3.4 Task Grammars and Language Modeling

Task grammars are used to specify legal word sequences for a particular ASR application. For example, in a system which recognises phone numbers, the grammar would dictate that eight (or ten) digits be spoken consecutively. An alternative to a task grammar is to use an open word loop which means any in-vocabulary word can be spoken at any time with no restrictions. Both open word loops and well-defined task grammars have been used in the evaluation protocols defined in Chapter 4.

Word loop grammars are typically combined with a Language Model (LM) which estimates the probability of a particular word sequence w. The language model score shown in Eq. (2.14) can be calculated as the product of conditional probabilities:

$$P(w_1, w_2, ..., w_N) = \prod_{i=1}^{N} P(w_i | w_1, ..., w_{i-1}) \tag{2.17}$$

where each word probability is conditional upon the previous words in the sequence. In a bi-gram LM, the probability of each word is only dependent on the previous word; therefore the LM predicts the next word in the sequence. Language model scores are added to the acoustic scores described in the previous section each time the recognised sequence moves from the end of one word to the start of the next word. In-car speech recognition systems are often based on well-defined command words and digit sequences [6, 67]; therefore it is possible

to use a grammar to restrict the legal word sequences rather than incorporate a LM. This dissertation adheres to this use of well defined task grammars rather than language models.

Insertion penalties and grammar scales are closely related to the concepts of language modeling. Insertion penalties are added every time a word is inserted to the recognition hypothesis, and are typically used to limit the insertion of short words. The grammar scale is used to place more emphasis on the language model score with respect to acoustic scores. The values of both insertion penalties and grammar scales are often determined empirically using a development data set.

2.4 Noise-Robust Speech Recognition

Speech recognition systems are susceptible to a wide range of mismatches between training and testing which cause severe decreases in speech recognition accuracy. Whilst the word accuracy performance of speech recognition systems in "noise-free" environments has approached 100%, performance in real-world environments such as automobiles is still failing to meet user expectation [6, 53, 86]. In these environments, the train-test mismatch can be attributed to the addition of background noise as well as the variation in speech production which results from humans communicating in these environments [60]. The latter cause is often referred to as the Lombard effect [49, 88, 113]. The work contained in this dissertation focuses solely on combating the effects of additive background noise.

To counteract additive noise, three key approaches for making ASR systems robust have been proposed (see [3, 46] for some reviews on this field): speech enhancement, robust acoustic modeling, and the use of robust features or recognition algorithms. None of these methods are specifically designed to be used in isolation; further ASR improvements can be obtained through implementing combinations of these techniques (e.g. [22, 151]). The following sections provide a brief discussion of each of these broad classes of robust ASR techniques.

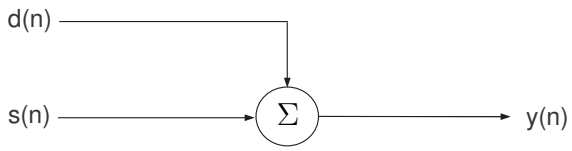


Figure 2.5: Additive background noise model commonly used in speech enhancement.

2.4.1 Speech Enhancement

Background noise $d(n)$ is typically considered to be added acoustically to the speech signal of interest $s(n)$ as shown in Fig. 2.5. Speech enhancement techniques aim to remove the additive noise from the signal and thereby recover a clean speech representation from the noisy speech signal. In most cases, little to no prior knowledge of the noise environment is needed, however estimation of statistical properties of the noise are often required. The clean speech representation allows acoustic models trained in noise-free environments to be utilised in ASR; this is particularly important since there is a requirement to collect significantly large amounts of data on which to train accurate acoustic models [53, 75]. In automotive environments, collecting enough data to model all variations in noise conditions is a very expensive and time consuming process, and therefore enhancement techniques allow clean speech data (of which there is an abundance) to be used for model training.

Speech enhancement techniques can be broadly divided into single- or multiple-microphone techniques, and signal- or feature-space techniques. Signal-space techniques operate directly on the signal in time or frequency domains, and are generally aimed at improving signal-to-noise ratios (SNR) of the incoming speech signal. Feature-space algorithms operate on data as part of the feature extraction process (see Section 2.2.2); this helps minimise extra computational resources required to incorporate into existing ASR systems. A literature review of state of the art speech enhancement techniques is presented in Chapter 3.

Enhancement techniques (even for ASR purposes) have been typically designed to satisfy signal-level criteria (e.g. maximising SNR) and not designed specifically for speech recognition accuracy [46, 127]. Whilst this is the case, improvements in ASR accuracy can still be observed when traditional signal-space

speech enhancement algorithms are used in the pre-processing stage. Novel contributions of the research contained in this dissertation modify the criteria of enhancement algorithms with the sole aim of optimising for speech recognition applications.

2.4.2 Robust Speech Modeling

Robust speech modeling techniques provide an almost opposite approach to that of speech enhancement. These techniques aim to incorporate noise into the acoustic model rather than remove noise from the signal of interest. There are four main approaches to noise-robust acoustic modeling: (a) model re-training, (b) multi-style model training, (c) model adaptation, and (d) Parallel Model Combination (PMC). The first two techniques relate to the way the acoustic model is trained, whilst the latter methods transform existing (typically "clean") acoustic models to the noisy environment.

Acoustic model re-training continually trains an environment-dependent acoustic model when data from new test environments becomes available [3]. In this way, it is possible to have matched conditions for a wide range of noise conditions. This method requires *a priori* knowledge of the environment characteristics which is not always available, as well as large amounts of data for each noise condition which makes the collection and transcription process very demanding. This approach is not suitable for automotive environments where there is the need to consider a very large range of noise conditions which arise due to different vehicle and engine types, road materials and driving conditions.

Multi-style training results in an environment-independent model where training data from a wide range of acoustic environments is available [84] but does not need to be relevant to the application environment. The limitations of data collection make the ability to obtain sufficient data from a large enough range of environmental conditions difficult, resulting in a acoustic models which are not truly environmental-independent.

Model adaptation schemes are used to transform reference speech models into the noisy application environment. These methods are very sensitive to variations

in noise conditions present in the adaptation data which may restrict improvements in ASR performance. Nevertheless, model adaptation techniques require considerably less data than that required for acoustic model training, making it a more viable approach for in-car speech recognition applications. In Chapter 4, model adaptation is used as part of the evaluation protocols developed within that chapter. Two common model adaptation techniques are described here: Maximum *A Posteriori* (MAP) adaptation and Maximum-Likelihood Linear Regression (MLLR).

MAP adaptation (sometimes referred to as Bayesian adaptation) [75] uses an existing speech model to provide prior knowledge of the model parameter distributions. Having knowledge of the distributions, adaptation of the j^{th} HMM state means μ can be obtained for each mixture component m using:

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm}+\tau}\bar{\mu}_{jm} + \frac{\tau}{N_{jm}+\tau}\mu_{jm} \tag{2.18}$$

where τ is used to weight the influence of the prior acoustic model, N is the occupation likelihood of the adaptation data, and $\bar{\mu}_{jm}$ is the observed mean of the adaptation data. More details on the calculation of the occupation likelihood and observed mean of the adaptation data can be found in [75]. In order to place more emphasis on the prior model, higher values of τ should be used. MAP adaptation ensures that only mean components of the state models observed in the adaptation are updated – therefore sufficient data is required to provide coverage of all state models. MAP adaptation is used to assess the database evaluation protocols defined in Chapter 4 and is shown to provide widespread word accuracy improvements for in-car ASR.

MLLR adaptation [77] uses the EM algorithm to produce a set of linear transformations for the mean and variance parameters in a GMM-based HMM acoustic model. This transformation shifts the mixture component means and changes the variances so that each state in the HMM system has a greater likelihood of generating the observed adaptation data. MLLR can be used to generate a global transformation for all Gaussian components in the presence of small amounts of adaptation data, or to produce more specific transformations based on regression classes as more data becomes available [155]. Since MLLR produces

transformations which cover all Gaussian components, less adaptation data is required than for MAP adaptation. In the case where large amounts of adaptation data is available, MAP outperforms MLLR because it adapts details about each Gaussian component rather than using a grouped regression class approach [58]. The results of adaptation can be improved further by combining both MAP and MLLR [133, 155].

Parallel model combination [39] is another method of adaptation whereby acoustic models are trained for clean speech and environmental noise separately, and these are added together during system operation. In car environments, this addition can be based on the SNR as well as the noise condition [121] and can considerably improve recognition rates in vehicles traveling at speed. As there is a requirement to model (and subsequently store) a wide range of environmental conditions, PMC has limited practical application in the automotive industry which emphasises low-cost software and hardware solutions.

2.4.3 Robust Speech Parameterisation and Recognition Algorithms

Robust speech parameterisation techniques seek representations of speech which are immune to the effects of environmental noise and are therefore effective for both clean and noisy speech recognition. In this way, the effect of the noise is not removed directly but is reduced in the feature extraction process. The major advantage of these techniques is that only very weak assumptions about the noise are made and explicit estimation of statistical parameters are not required [46]. The latter is also a limitation in some environments as the techniques are not specifically tuned to particular characteristics of the noise signal.

RelAtive SpecTrAl (RASTA) processing is one method for improving the robustness of common feature extraction techniques [55, 129]. The idea behind RASTA is to use a band-pass filter which suppresses slowly and quickly varying components of speech, and in the process emphasises important parts of the speech that are most robust against noise [129]. RASTA processing has been applied in various forms, and has shown positive results when integrated in feature

extraction processing in car environments [134].

Robustness can be incorporated into speech decoding algorithms to account for the masking effect of noise on the speech signal, as achieved using the missing features paradigm [24]. This approach is based on the notion that when some speech features (and sometimes entire frames) are masked by particular types and levels of noise, they become unreliable for use in the decoding stage. In doing so, emphasis is placed on speech characteristics which are robust to noise. Missing feature methods determine a mask (e.g. using auditory scene analysis [17]) and this mask can be used to reconstruct the unreliable features [116] or modify the decoding stage [25]. Whilst this paradigm has been shown to be effective for additive car noise [25], the major drawback of this method is that uncertainty mask estimation relies on estimation of the noise signal, which inherently assumes that the noise is stationary [126], and therefore performance is limited in environments subject to non-stationary noise sources.

Another advance in robust speech recognition algorithms is to use a Recogniser Output Voting Error Reduction (ROVER) paradigm [35] which has resulted from an increase in available computing resources. In a system following the ROVER framework, multiple speech recognisers are run in parallel, each employing a different robust technique. For example, one recogniser might use RASTA-PLP or MFCC feature extraction, whilst another uses spectral subtraction speech enhancement. This paradigm operates under the assumption that the parallel recognisers will exhibit different types of errors under different noise conditions. In vehicular environments, the increase in required computing resources to employ the ROVER paradigm make it unsuitable for the near real-time operation demanded by such an application [3].

2.5 Summary

In this chapter, the fundamentals of automatic speech recognition including speech parameterisation, acoustic modeling and methods for making ASR systems more robust in adverse environments have been presented. The common Mel-frequency cepstral coefficients feature extraction algorithm was detailed in full including

the use of cepstral mean subtraction, cepstral liftering and temporal derivatives which are used throughout this research. Acoustic modeling using hidden Markov models was explained with reference to both model training and speech decoding. Important aspects of this discussion will be made clear when discussing the novel contributions of this dissertation.

A range of common approaches to making ASR systems more robust in the presence of additive background noise including speech enhancement, acoustic model training and adaptation, and robust feature extraction and recognition algorithms were discussed. Particular reference to implementation in in-car speech recognition applications was made for each of these methods. Speech enhancement techniques which overcome most of the limitations of data requirements, processing requirements, and stationarity assumptions are deemed to be most suitable for in-car speech recognition, and will be discussed in further detail in Chapter 3.

Chapter 3

Speech Enhancement

3.1 Introduction

In "real-world" environments where the levels of background noise cannot be considered insignificant, the recognition accuracy of ASR systems degrades significantly. Speech enhancement is a popular approach to improving the robustness of ASR through the removal of additive noise from recorded speech signals. Robustness for ASR applications is only one motivation for the use of speech enhancement – discussion of the motivations for enhancement in general speech processing applications is made in Section 3.2.

Speech enhancement techniques can be broadly classified by the number of microphones used. Single-channel techniques (Section 3.3) are well suited to a number of applications (e.g. in-car ASR) where hardware costs are a key factor. Multi-channel speech enhancement techniques, whilst increasing hardware requirements, have been shown to provide superior enhancement performance through the use of spatial filtering. Common multi-microphone techniques are reviewed in Section 3.5. In these sections, particular reference is made to examples of speech enhancement used for in-car speech recognition.

A number of these single- and multi-channel enhancement techniques require the estimation of statistical characteristics of the background noise. Common techniques for noise estimation are detailed in Section 3.4.

The chapter concludes by discussing the research directions investigated in this

dissertation. In particular, shortfalls of the traditional optimisation criteria for single-channel speech enhancement used as a front-end for ASR are highlighted. This discussion leads to the definition of the scope of the research.

3.2 Motivations for Speech Enhancement

In "real-world" environments, ASR has so far failed to live up to consumer expectations. For automotive applications, the low performance of ASR is due primarily to the large number of noise sources both within the car (e.g. air-conditioning fans, infotainment systems, other passengers), and also external to the cabin (e.g. engine noise, road noise, wind noise, other vehicles). As briefly mentioned in Chapter 2, the perception of these noise sources can also lead drivers to change their vocal effort; this is commonly referred to as the Lombard effect [88]. In this research, emphasis is placed only on the effects of additive noise on the speech signal (i.e. the Lombard effect is assumed to be absent).

Since the performance of ASR systems in adverse environments is generally unsatisfactory, it is required to make them more robust. Three common approaches are to use speech enhancement, robust acoustic modeling or robust speech parameterisation and recognition algorithms. All of these approaches were analysed in Chapter 2, with speech enhancement suggested as being most appropriate to improve the robustness of in-car ASR systems.

The motivation for using speech enhancement in automotive environments arises primarily from the time and expense required to collect significant amounts of data on which to train or adapt acoustic models for a wide range of noise conditions. Speech enhancement techniques can handle the constantly changing noise conditions, allowing noisy speech signals to be transformed into a clean speech representation (either waveform or feature vectors) which enables the use of well-trained clean speech acoustic models for ASR.

Another motivation for the use of speech enhancement is to improve the quality of speech communication in noisy environments. In this application, the aim is to satisfy signal-level criteria such as maximising signal-to-noise ratio, minimising the signal error, or improving human perceptual quality [127]. These criteria are

used for both single- and multi-microphone enhancement techniques [46, 127].

Most speech enhancement techniques were originally designed for speech intelligibility rather than for ASR [46]. Whilst there have been numerous examples where enhancement techniques have been shown to be a successful pre-processing stage for robust speech recognition, some enhancement techniques distort the speech signal in ways which can cause ASR performance to decrease. Any solutions derived from signal-level criteria which produce improvements in word accuracy are therefore typically sub-optimal for ASR. Redesigning enhancement techniques to optimise for speech recognition has gained increased interest in recent years [9, 127, 130], and is the focus of the research in this dissertation.

3.3 Single-Channel Techniques

3.3.1 Spectral Subtraction

Spectral subtraction was first proposed by Steven Boll in 1979 [14]. It has become one of the most widely used single-channel noise reduction techniques, and is commonly used as a baseline for comparing novel speech enhancement techniques. Approaches for optimising spectral subtraction specifically for automatic speech recognition applications are the focus of this thesis.

The aim of spectral subtraction is to estimate the spectrum of the clean speech signal by subtracting an estimate of the noise spectrum from that of the noise-corrupted speech signal. Subtraction typically takes place in the magnitude or power spectrum, but may also take place on filterbank energies as will be described in this section.

The basis for many speech enhancement algorithms (including spectral subtraction) is the assumption that the noise and speech signals are statistically independent [13]. In this instance, noise can be regarded as being added acoustically to the clean speech signal as was shown in Fig. 2.5. In the time domain this addition is represented as:

$$y(n) = s(n) + d(n) \tag{3.1}$$

where $s(n)$, $d(n)$, and $y(n)$ are the clean speech, additive background noise and

noisy speech signals respectively. During speech segments, the noise signal is assumed to remain stationary, and if the noise environment changes between two consecutive speech segments, there should be sufficient time in which to accurately re-estimate the noise characteristics [14]. Methods of estimating noise characteristics are explained in Section 3.4.

As per MFCC feature extraction detailed in Chapter 2, the noisy speech signal is broken up into frames and transformed to the complex discrete frequency domain using Eq. (2.3) to produce:

$$Y^i(k) = S^i(k) + D^i(k) \tag{3.2}$$

where i is the frame index. The generalised frequency-domain spectral subtraction rule derived from the early works of Boll [14] and Berouti *et al.* [13] is defined by:

$$|\hat{S}^i(k)|^\gamma = \begin{cases} |Y^i(k)|^\gamma - \alpha^i(k)|\hat{D}^i(k)|^\gamma & |Y^i(k)|^\gamma - \alpha^i(k)|\hat{D}^i(k)|^\gamma > \beta|\hat{D}^i(k)|^\gamma \\ \beta|\hat{D}^i(k)|^\gamma & \text{otherwise} \end{cases} \tag{3.3}$$

where $|\hat{D}^i(k)|$ is an estimate of the noise magnitude spectrum obtained during periods of non-speech, and $|\hat{S}^i(k)|$ is the resulting estimate of the clean speech signal. The parameter γ determines the spectrum the subtraction takes place in; for example this may be the magnitude spectrum ($\gamma = 1$) [14], or the power spectrum ($\gamma = 2$) [13, 34, 85]. Whilst these values of γ produce spectra which have theoretical relevance, there is actually no limit to the values that this parameter can take [26, 68, 79].

Since the noise spectrum is obtained through estimation, time- and frequency-dependent subtraction factors, $\alpha^i(k)$, are introduced to compensate for under-estimating or overestimating the potentially non-stationary instantaneous noise spectrum. Optimisation of $\alpha^i(k)$ has been the subject of much of the spectral subtraction research to date. An SNR-weighted subtraction factor was first introduced in [13], but numerous methods for determining the subtraction factors have since been proposed [63, 79], including examples where in-car speech recognition was the target application [86, 122, 150]. A considerable amount of this research has been aimed at improving speech intelligibility by reducing the levels of musical noise present in the enhanced signal [45]. Musical noise is an artefact

of spectral subtraction resulting from both over-estimation and under-estimation of the noise spectrum [13].

Since the instantaneous signal spectrum can be smaller than the estimated noise spectrum, it is possible for $|\hat{S}^i(k)|^\gamma$ to become negative; therefore, the spectral flooring factor, β, is used to reduce the effects of over-subtraction. The value for the flooring factor is typically $0 < \beta \ll 1$ [13, 91, 105]. Most implementations set a constant spectral floor factor, however some studies have determined the noise floor factor dynamically [122]. The spectral floor factor can be applied to the noise spectral estimate [13, 68, 146], and also to the instantaneous noisy speech signal [34, 91, 131] to enforce a maximum level of signal attenuation. In this research, the spectrum to which spectral flooring is applied varies based upon ASR word accuracy performance; the relevant spectra are outlined when explaining the individual experimentation configurations.

Rather than using a subtraction factor for each DFT frequency bin, more recent research has looked at reducing the number of subtraction parameters by using pre-determined frequency bands [10, 63, 132]. These methods are referred to as Multi-Band Spectral Subtraction (MBSS) techniques. For example, the spectral subtraction rule used by Kamath and Loizou [64] is:

$$|\hat{S}_b^i(k)|^\gamma = \begin{cases} |Y_b^i(k)|^\gamma - \alpha_b^i \delta_b(k) |\hat{D}_b^i(k)|^\gamma & |Y_b^i(k)|^\gamma - \alpha_b^i \delta_b(k) |\hat{D}_b^i(k)|^\gamma > 0 \\ \beta |Y^i(k)|^\gamma & \text{otherwise} \end{cases} \tag{3.4}$$

where the values for α_b^i are determined by the local SNR in the b^{th} sub-band. The three critical bands used in MBSS are specified by:

$$\delta_b = \begin{cases} 1 & f_{U,b} \leq 1 \text{ kHz} \\ 2.5 & 1 \text{ kHz} < f_{U,b} \leq \frac{f_s}{2} - 2 \text{ kHz} \\ 1.5 & f_{U,b} > \frac{f_s}{2} - 2 \text{ kHz} \end{cases} \tag{3.5}$$

where f_s is the signal sampling frequency and $f_{U,b}$ is the upper frequency of the b^{th} sub-band. The motivation for using the representation detailed in Eqs. (3.4)-(3.5) is to reduce the levels of speech distortion in critical speech frequency bands [87].

Performing MBSS also considerably reduces the number of required subtraction parameters; the above implementation uses only 3 frequency bands, whilst

Singh and Sridharan [132] use 18 critical frequency bands, and Chen *et al.* [19] and BabaAli *et al.* [10] use 24 and 25 Mel-scaled frequency bands respectively. These implementations are all in stark contrast to the 257 parameters needed for each DFT frequency when a 16 kHz signal is analysed with 32 ms windows.

At this point, it should also be noted that it is assumed that spectral subtraction is only required on the magnitude spectrum, leaving the phase spectrum $e^{j\angle Y^i(k)}$ unchanged. This assumption is generally appropriate as phase has been regarded as unimportant for human perception [109, 143], although this has been challenged in recent times [110]. The noisy speech phase spectrum is re-combined with the modified magnitude spectrum $|\hat{S}^i(k)|$ for synthesis to a time-domain signal via an IFT and overlap-add reconstruction [4].

All examples of spectral subtraction described thus far have performed the subtraction in the frequency-domain. Whilst this is appropriate for both speech intelligibility and ASR applications, noise subtraction can also be performed on the Mel-filterbank energies which are a part of MFCC feature extraction [103, 108]. Throughout this dissertation, this method will be referred to as Mel-Filterbank Noise Subtraction (MFNS). If the frequency bands k are split into M sub-bands based on the Mel-scale, MFNS can be defined as:

$$E_Y^i(m) = \int_{f_{L,m}}^{f_{U,m}} |Y^i(k)|dk$$

$$E_{\hat{D}}^i(m) = \int_{f_{L,m}}^{f_{U,m}} |\hat{D}^i(k)|dk$$

$$\hat{E}_S^i(m) = \begin{cases} E_Y^i(m) - \alpha^i(m)E_{\hat{D}}^i(m) & E_Y^i(m) - \alpha^i(m)E_{\hat{D}}^i(m) > \beta E_Y^i(m) \\ \beta E_Y^i(m) & \text{otherwise} \end{cases} \tag{3.6}$$

where $E_Y^i(m)$, $E_{\hat{D}}^i(m)$ and $\hat{E}_S^i(m)$ are the energies of the m^{th} Mel-filterbank of the noisy speech, noise estimate and the estimate of the clean speech filterbank energy respectively. The scaling factor β provides a maximum level of signal energy attenuation and ensures output filterbank energies remain positive as per the frequency-domain formulation. In this instance, the subtraction factors $\alpha^i(m)$ are filterbank-dependent rather than frequency-dependent. In [103], the value of $\beta = 0.1$ was used with constant values of $\alpha^i(m) = 1$ across all filterbanks. Implementing noise subtraction in this domain provides close coupling with the

MFCC feature extraction process and also provides robustness against large, impulsive spectral magnitudes.

It was previously stated that much of the research in noise subtraction techniques has been based around optimising the oversubtraction parameters. Typically, this has been to improve speech intelligibility, but some reports do show improvements in ASR accuracy using signal-level criteria such as SNR. Only one research group have looked at methods for optimising these parameters specifically for speech recognition applications. BabaAli *et al.* [10] utilise multi-band spectral subtraction to enhance the noisy speech, and apply a likelihood-maximising (LIMA) framework [127] to optimise the subtraction factors. Much of their work was performed in parallel with the work contained in this dissertation, which uses Mel-filterbank noise subtraction as the enhancement technique in a LIMA framework. Further discussion and analysis of LIMA-based noise subtraction techniques can be found in Chapters 5 and 6.

3.3.2 Wiener Filtering

The spectral subtraction techniques discussed in Section 3.3.1 were not derived using well-defined mathematical error criterion; they simply assume that additive noise can be subtracted from the noisy speech signal. Wiener filters – whilst still assuming that noise is additive – reduce noise levels by minimising the mean-square error between the estimated and desired signals [148]. In deriving the Wiener filter, it is also assumed that the signals under analysis are stationary, which is not always the case. Kalman filters (which have been used for speech enhancement for both intelligibility and ASR applications [40, 44, 85, 111]) are an extension of the Wiener filter which enables handling of non-stationary noise.

According to Wiener filter theory, Eq. (3.1) is altered such that the noise and speech signals are passed through a linear system with impulse response, $h(n)$:

$$y(n) = h(n) * [s(n) + d(n)] \tag{3.7}$$

The goal of the Wiener filter approach is to determine the optimal impulse (or frequency) response of the linear filter $h(n)$. Assuming clean speech and noise signals are uncorrelated, the parametric frequency-domain Wiener filter response

can be derived as [82]:

$$H^i(k) = \left(\frac{S^i(k)^2}{S^i(k)^2 + a^i D^i(k)^2} \right)^\beta = \left(\frac{\xi^i(k)}{a^i + \xi^i(k)} \right)^\beta \tag{3.8}$$

where a^i and β are used to alter the signal attenuation for each frame i [87], and $\xi^i(k)$ is the *a priori* SNR in frequency k. The traditional Wiener filter (i.e. $a = \beta = 1$) attenuates noise at each frequency in proportion to the *a priori* SNR in much the same way that frequency-dependent spectral subtraction does. Deriving the frequency-response of spectral subtraction filter in [13] will lead to the same response [87, 101].

The major drawback with this derivation of the Wiener filter is the requirement to have *a priori* knowledge of the power spectrum of the clean speech signal. Since this is the desired result of enhancement, numerous methods have been proposed to overcome this limitation, including iterative Wiener filtering with [51, 112] and without constraints [82]. In these implementations, the clean speech signal is continually estimated (after initialisation as the noisy speech signal) using an updated Wiener filter.

Application of Wiener filtering to ASR has been less prevalent than spectral subtraction. This is due to the sub-optimality of the Wiener filter in non-Gaussian noise environments, and also the computational requirements of iterative Wiener filtering. Specific examples of Wiener filtering in ASR front-ends include [2, 51], with some application to in-car speech recognition [7, 20, 41, 98].

3.3.3 MMSE-Based Spectral Enhancement

Research has typically shown that only the magnitude spectrum is important for human speech intelligibility [109, 143], therefore the optimal complex spectrum estimator – the Wiener filter – is not the optimal magnitude spectrum estimator in the Minimum Mean-Square Error (MMSE) sense. Unlike the Wiener filter, the MMSE estimator does not assume a linear relationship between the observed spectral information and the estimator; it does, however, make assumptions about and required knowledge of the statistical distributions of the speech and noise magnitude spectra [31].

The MMSE uses a Bayesian probability approach to determine the clean speech amplitudes assuming Gaussian distributions for the speech and noise magnitudes [31]. Derivation of the spectral estimator results in the following spectral gain function:

$$G(\xi(k),\gamma(k)) = \frac{\sqrt{\pi}}{2}\frac{\sqrt{v(k)}}{\gamma(k)}\exp\left(-\frac{v(k)}{2}\right)\left[(1+v(k))I_0\left(\frac{v(k)}{2}\right)+v(k)I_1\left(\frac{v(k)}{2}\right)\right] \tag{3.9}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ are zeroth and first order Bessel functions, $v(k)$ is defined by:

$$v(k) = \frac{\xi(k)}{\xi(k)+1}\gamma(k) \tag{3.10}$$

and $\xi(k)$ and $\gamma(k)$ are the *a priori* and *a posteriori* SNRs respectively. The dependence on the current frame i has been dropped for simplicity. Even though the influence on overall attenuation is less than the *a priori* SNR, the MMSE estimator also relies on the *a posteriori* SNR which is useful for reducing the levels of musical noise [87]. It should also be noted that when the *a priori* SNR is large, the MMSE estimator behaves similarly to the Wiener filter.

Attempts to derive an MMSE estimator for the phase spectrum were also described in [31], however the derivation led either to the magnitude estimator becoming sub-optimal, or the noisy speech phase being the optimal phase spectrum. As a result, only the magnitude MMSE is used in speech enhancement applications.

A key issue with the MMSE estimator described in Eqs. (3.9)-(3.10) is the reliance on estimating both the variance of the noise magnitude spectrum, and the *a priori* SNR. The noise variance is easily calculated during non-speech periods using the noise estimation techniques described in Section 3.4. Solutions for estimating the *a priori* SNR include the maximum-likelihood approach [31], as well as decision-directed approaches [23, 31, 52].

Derivatives of the original MMSE technique have also been proposed in the literature, including the log-MMSE estimator [32], and the p^{th}-power magnitude estimator [154]. Alternatives to using Gaussian-distributed spectral values such as Gamma [93] and Laplacian distributions [18] have also been reported.

Examples of MMSE-based speech enhancement as a front-end for speech enhancement include [43, 94]. In highly mismatched noise conditions [43], a modified

MMSE estimator showed ASR word accuracy improvements over Wiener filtering and the original MMSE estimator [31], whilst Matassoni *et al.* [94] showed that a highly optimised spectral subtractor (based on [13]) can outperform both MMSE and log-MMSE estimators in automotive environments.

3.3.4 Phase Spectrum Compensation

The speech enhancement methods discussed thus far have only operated on the magnitude spectrum of the incoming signal. Recently, researchers at Griffith University, Australia, have been exploiting the phase spectrum for speech enhancement [135, 137, 149]. These have been some of the first attempts to use the phase spectrum for speech enhancement applications.

Phase Spectrum Compensation (PSC) utilises the synthesis procedure (i.e. IFT and overlap-add reconstruction) commonly used in speech enhancement where an enhanced waveform is required for playback. Since the incoming speech signal is real-valued, the DFT coefficients are conjugate symmetric. PSC controls the amount of reinforcement or cancellation that occurs during synthesis by adding a noise-weighted anti-symmetry function $\Lambda^i(k)$ to the noisy speech signal in the complex frequency domain [137]:

$$Y^i_\Lambda(k) = Y^i(k) + \Lambda^i(k). \tag{3.11}$$

For frequencies with low noise magnitudes, the anti-symmetry function causes little change to the original signal. For high noise components however, the anti-symmetry function causes the conjugate pairs to cancel during the synthesis stage. The reader is directed to [137, 149] for detailed descriptions of this technique.

Whilst PSC has shown promising improvements in human intelligibility, it has not yet been used in a speech recognition application. Chapter 7 provides discussion of another method of incorporating phase information into speech enhancement techniques for ASR applications.

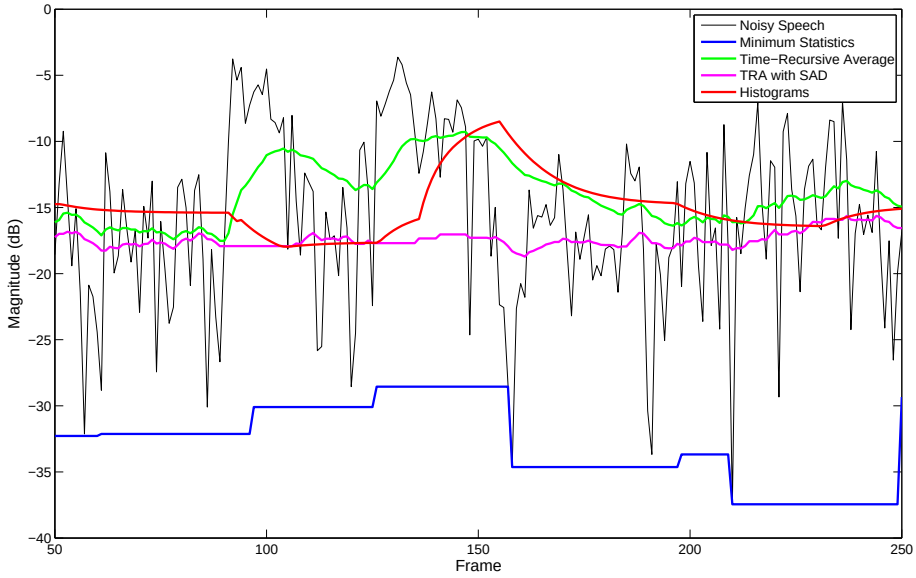


Figure 3.1: Tracking performance of noise estimation techniques on a 2 second segment of a noisy speech signal.

3.4 Single-Channel Noise Estimation

The single-channel speech enhancement algorithms discussed in Section 3.3 all assume that estimates of the noise characteristics are available. Noise estimation techniques therefore play an important role in the effectiveness of these enhancement algorithms. For example, if the noise estimate is too low, the resulting clean speech estimate will still contain significant levels of noise. Alternatively, if the noise estimate of too high, there is potential for speech distortion. Another important consideration for applications such as in-car speech recognition is the ability to track continually changing noise conditions (see Fig. 3.1). This section explains the concept of Speech Activity Detection (SAD) and analyses three common approaches to noise estimation with particular reference to Fig. 3.1.

3.4.1 Speech Activity Detection

Speech activity detection is the process of determinating the presence of either speech or silence in a segment of speech. A number of features can be used for determining SAD including short-time energy, zero-crossings [62], cepstral features [48] or periodicity [139], and output a binary decision of whether the segment contains speech or silence. SAD detects silence periods not only at the

beginning or end of an utterance, but also in the middle of sentences. In low-SNR and non-stationary noise environments, these algorithms typically perform poorly, and are often left out of noise estimation methods in such environments.

3.4.2 Minimum Statistics

Minimum Statistics (MS) noise estimation utilises the premise that the spectrum of the noise signal generally exhibits lower magnitudes than the underlying speech signal. As a result, noise estimates can be derived by tracking the minimum spectral magnitudes over finite time windows for each frequency [91]. In order to reduce the effects of outlying spectral values, the minimum statistics algorithm typically smoothes the signal spectrum prior to calculating the noise estimate.

The blue line in Fig. 3.1 shows the estimation performance of the original implementation of the MS algorithm [91]. Examination of this figure shows two mains drawbacks of MS: (a) the noise level is consistently lower than the true noise level, and (b) the algorithm fails to respond rapidly to increases in the noise spectrum. To counteract (a) in spectral subtractive speech enhancement, the subtraction values are typically $\alpha^i(k) > 1$. Methods for reducing this bias have also been proposed in the literature [92].

The slow response to increase in noise levels is attributed to the use of finite time windows. To overcome this limitation, methods for continually updating the noise spectrum have also been proposed [30] and shown to improve the performance of the original implementation [99].

3.4.3 Time-Recursive Averaging

Time-Recursive Averaging (TRA) techniques are used regularly for noise estimation as they can be implemented efficiently, and can track increases and decreases in the noise signal effectively. The recursive nature of this technique comes from the fact that the noise estimate $|\hat{D}^i(k)|^\gamma$ is updated according to the noise estimate of the previous frame $|\hat{D}^{i-1}(k)|^\gamma$. The TRA algorithm has the general

form:

$$|\hat{D}^i(k)|^\gamma = \begin{cases} \eta|\hat{D}^{i-1}(k)|^\gamma + (1-\eta)|Y^i(k)|^\gamma & |Y^i(k)|^\gamma > \lambda|\hat{D}^{i-1}(k)|^\gamma \\ |\hat{D}^{i-1}(k)|^\gamma & \text{otherwise} \end{cases} \tag{3.12}$$

where η is the smoothing factor which can be frame- and/or frequency-dependent. The value for η can be determined using estimated SNR [83], or probabilities of speech presence [90, 136]. In Eq. (3.12), soft-decision SAD is used to determine which frequencies of the noise estimate are updated based on the instantaneous signal spectrum, previous noise estimate and scaling factor λ. Soft-decision SAD can also be left out of this technique by setting $\lambda = 0$. This dissertation uses a static η and λ for all frames and frequencies.

The magenta and green lines in Fig. 3.1 shows the performance of TRA with and without the use of soft-decision SAD respectively. It can be seen that TRA without SAD is able to respond quickly to both increases and decreases in the noise spectrum, making it more effective than MS and histogram-based methods in noise tracking. Without the use of SAD, this tracking ability can easily lead to noise over-estimation since the TRA estimator will also track speech components (which generally exhibit higher spectral magnitudes); this can lead to speech distortion when using subtractive-type enhancement algorithms. The use of soft-decision SAD is able to reduce this response sensitivity, providing a lower noise estimate in cases where the noise spectrum increases significantly which avoids tracking speech components. Using high smoothing factors (typically $0.9 < \eta < 1$) can also be used to reduce the response sensitivity. Finding an appropriate combination of λ and η is important for the effectiveness of the TRA noise estimation technique.

3.4.4 Histogram-Based Techniques

Histogram-based noise estimation resulted from the observation that the most frequent spectral values correspond to the level of the noise spectrum [56, 87, 118]. Finite time windows are used to construct a histogram of past spectral values, from which the maximum value is taken as the noise estimate. This technique

is generally robust to scenarios where the histogram has two modes – one representing speech and the other noise. In most cases the noise mode dominates, and therefore histogram estimation accurately extracts the noise level. Similar to MS estimation, the noisy signal spectrum is smoothed prior to constructing the histogram to reduce extreme spectral values.

The red line in Fig. 3.1 shows the noise tracking performance of the basic histogram noise estimator [56] with a time window of 40 frames. The histogram method is seen to track the true signal level more effectively than MS estimation, although – compared to the TRA algorithm – it does show noticeable delays in increasing (frames 80-130) and decreasing (frames 160-195) the noise estimate when the spectral magnitude changes rapidly. Whilst this makes it robust to the large spectral magnitudes present during speech segments, it can result in speech distortion due to noise over-estimation (from delays in reducing the estimate) or increased residual noise levels due to under-estimation (due to delays in increasing the estimate). Having determined an appropriate trade-off between distortion and residual noise, the response time can be increased/decreased by changing the finite time window used in the histogram construction [87].

3.5 Multi-Channel Speech Enhancement

Speech enhancement with multiple microphones has become a popular and effective approach for speech enhancement, particularly in ASR systems. Speech signals are captured simultaneously by all microphones in the system, and this multi-sensor information is then used to filter the signal to produce an estimate of the clean speech signal. As a result of the use of multiple signal channels, multi-microphone methods have shown more impressive results than single-channel techniques for ASR applications in car environments [80]. Despite these improvements, multi-channel techniques still have limited application in some domains (particularly automotive) due to the increased hardware requirements (and consequently increased costs) and extra processing required for each of the channels.

In this section, common approaches for multi-channel speech enhancement are briefly described in order to provide a complete review of state-of-the-art speech

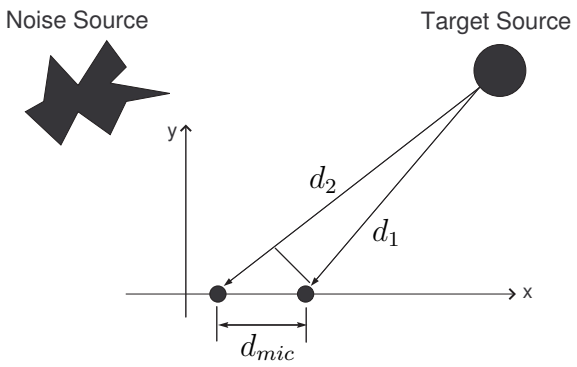


Figure 3.2: Near-field spatial information used by multi-channel beamforming algorithms.

enhancement techniques. Multi-channel techniques are not further explored in this dissertation; readers are encouraged to refer to the cited publications for more technical details.

3.5.1 Beamforming

Multi-channel beamforming combines the acoustic signals from all microphones to perform filtering which differentiates the signal of interest from the background noise based on physical locations. If the geometry of the microphones with respect to the target source is known *a priori* (as shown by the near-field scenario in Fig. 3.2), a beam can be formed which includes the target source but excludes the noise source which could be any form of noise field, or a point source as shown in the diagram. Under a near-field assumption – which is more appropriate than the far-field for in-car applications – the beam is created by compensating the respective propagation delays between the source and each microphone. Practically, this operation results in each microphone being aligned in the time axis.

Having compensated the delays, the microphone channels are individually weighted and combined in order to reinforce the speech signal; for this reason the technique is referred to as filter-and-sum beamforming. This operation causes cancellation of noise and other signal sources outside the target direction as they are assumed to be uncorrelated in each microphone. In the frequency domain,

the filter-and-sum beamformer is represented as:

$$S(k) = \frac{1}{N}\sum_{n=1}^{N} G_n(k)Y_n(k)\exp^{-j2\pi k\Delta_n} \tag{3.13}$$

where N is the number of microphones, $Y_n(k)$ is the signal received at the n^{th} microphone, $G_n(k)$ are the filter coefficients, and the exponential term is compensation for the delay Δ_n.

A number of fixed and adaptive beamforming techniques to determine the filter weights $G_n(k)$ have been proposed in the literature. A common fixed beamformer used as a baseline system for comparison with novel beamformer algorithms [102] is the delay-and-sum beamformer, in which $G_n(k) = 1$ [16]. This beamformer was shown to be effective for dual-microphone hardware implementations in a range of noise conditions in car environments [152]. Many other filter-and-sum beamformers have been proposed to optimise the filter weights $G_n(k)$ for particular noise fields and conditions [95]. Whilst the majority of these beamformers maximise signal level criteria, a likelihood-maximisation framework to optimise the filter coefficients specifically for ASR in noisy environments has also been proposed [127].

The most commonly used beamformer for automotive applications is the adaptive Generalised Sidelobe Canceller (GSC) based on the Griffiths-Jim beamformer [47, 106, 107]. In [107], the GSC configuration was shown to be more effective than delay-and-sum beamforming in reducing word error rates for both city and highway driving.

3.5.2 Blind Source Separation

Blind Source Separation (BSS) techniques aim to distinguish a set of signals $(s_1, ..., s_N)$ which have been mixed with some unknown model (see Fig. 3.3) [73, 153]. Examples of mixing scenarios include adding background noise to clean speech, or two speakers talking at the same time (i.e. a simplified version of the "cocktail-party" problem [21]). Therefore, whilst BSS is not explicitly a speech enhancement technique, it can be used for such applications where the target speech has been corrupted by background noise. In order to perform the separation, it is generally required to have at least the same number of channels

Figure 3.3: Signal and mixing model assumed in blind source separation.

as there are sources to separate (i.e. for the simple speech in noise problem, a minimum of two microphones are required).

Blind source separation (in similarity with the general speech enhancement model) assumes that the signals that have been mixed are uncorrelated, and can therefore be considered statistically independent. As such, no assumption is made about the spatial locations of the sources, which is in contrast to the beamforming techniques described in the previous section. Examples of BSS which incorporate beamforming concepts can also be found in the literature for improved separation performance [120].

BSS has been applied as a front-end for speech recognition in high-noise car environments with around 30% improvement in word recognition accuracy [12].

3.5.3 Phase-Error Filtering

Phase-Error Filtering (PEF) is a recent and unique development in multi-channel speech enhancement. It was originally proposed for dual-channel speech enhancement [1], but was extended for multiple microphones in [72]. This technique has been included in this review as it incorporates two increasingly-popular methods for speech enhancement and recognition which are the emphasis of this dissertation: enhancement using phase information, and integration with LIMA frameworks [130]. Despite this influence, PEF is not considered in detail elsewhere in this dissertation since it is a multi-channel enhancement technique and this research is concerned only with single-channel enhancement.

The delay compensation previously discussed theoretically time-aligns the speech components in all microphones; therefore, the values of the phase spectrum should be the same (i.e. have a phase error of zero). Due to the presence of noise however, the true phase error is non-zero, and the mean square phase error increases as the level of noise increases [1]. Using this observation, the phase-error

filter applied to the noisy speech signal is defined as:

$$G(k) = \frac{Y(k)}{1 + \gamma\theta^2(k)} \tag{3.14}$$

where $\theta(k)$ is the phase difference between two channels, and γ controls the aggressiveness of the filter which can be optimised using a LIMA framework. The likelihood-maximised phase-error filter showed consistent word error rate reductions compared to delay-and-sum beamforming across a wide range of SNRs [130], proving its validity as a speech enhancement technique for ASR applications.

3.6 Research Directions

In-car speech recognition has been chosen as the application environment for design and evaluation of speech enhancement techniques in this dissertation. As will be shown in Chapter 4, considerable improvements in speech recognition accuracy are still required in order perform at levels which meet driver expectation and will ultimately make these systems commercially viable on a large scale.

At the present time, multi-microphone systems are deemed too expensive for widespread adoption in the highly competitive automotive industry. As a result, this research has chosen to investigate only single-channel enhancement techniques designed specifically for improved speech recognition accuracy.

Throughout the review in this chapter, numerous examples of speech enhancement used in front-end processing for ASR have been cited. In the majority of these techniques, the optimisation criteria are based on signal level measures which are most suited to speech enhancement for speech quality and human intelligibility. Whilst some techniques have shown improvements in word accuracy rates, the results are purely by-products of the enhancement process, and are therefore sub-optimal for ASR applications. Approaches such as the LIMA framework briefly introduced in this chapter can further optimise the performance of these techniques specifically for ASR. Progress has been made to applying LIMA frameworks in real-world environments using single-channel speech enhancement, but there are still limitations to current implementations which are evaluated in Chapter 5 along with potential solutions. Consideration of specific application to automotive environments is the focus of Chapter 6.

Finally, careful analysis of single-channel enhancement reveals the preference to use only information about the magnitude (or power) spectrum of the speech and noise signals. In these instances, the noisy phase spectrum is left unaltered, and is used for synthesis to the time-domain if required. For frequency-domain spectral subtractive techniques in particular, estimates of the phase spectrum could be used to reduce the errors in the overall magnitude estimate by performing the subtraction in the complex frequency-domain as opposed to the magnitude spectrum. Chapter 7 proves this concept, and proposes methods for obtaining phase spectrum estimates of the noise or speech signal.

Practical application of these enhancement techniques inherently introduce a resource versus performance trade-off which must be carefully considered. Simplification of frequency-domain spectral subtraction for a cost-effective hardware implementation which also ensures optimal speech recognition performance in vehicular environments is considered in Chapter 8.

3.7 Summary

This chapter has outlined state of the art speech enhancement using both single- and multi-microphone approaches. Single-channel techniques discussed include spectral subtraction, Wiener filtering, MMSE-based methods, and the recent development of phase spectrum compensation. These algorithms rely heavily on estimation of the noise spectrum – three solutions to this problem were discussed and compared.

Brief descriptions of three multi-microphone speech enhancement techniques completed the review of the state of the art technology in this field. Despite showing better enhancement performance than single-channel techniques, multi-microphone techniques were still deemed too expensive for the application domain of interest in this research – in-car ASR. As a result, the remainder of this work concentrates solely on single-channel speech enhancement.

The optimisation of speech enhancement techniques for intelligibility rather than automatic speech recognition has been identified as the major shortcoming

of the existing approaches. Research directions aimed at specifically optimising common single-channel speech enhancement techniques for ASR applications have been discussed and comprise the novel contributions of this dissertation. Two particular approaches which integrate the enhancement technique and the ASR system more closely are of interest in this work – likelihood-maximisation frameworks (Chapters 5-6), and the incorporation of phase spectrum information (Chapter 7).

Chapter 4

ASR Evaluation Databases

4.1 Introduction

Since most ASR systems are trained for use in controlled environments, they fail to produce satisfactory performance under more adverse conditions such as those encountered in automotive environments. One of the major limitations in making ASR systems more robust is the inability to collect sufficient amounts of data on which to train acoustic models and perform meaningful evaluations. The collection of data requires hundreds of hours of work in recording data as well as transcribing it for training and evaluation purposes. Once acoustic models have been trained, experimental evaluation requires a significant amount of test data – which must be different from the training data – in order to obtain statistically relevant descriptions of the performance of any robust speech recognition technique.

Two in-car speech databases – including the Australian English In-Car Speech corpus (AEICS) collected as part of this research – and their corresponding evaluation protocols are outlined in Section 4.2. In defining the evaluation protocol, it is important to also declare the parameters used in the development of the baseline ASR system (Section 4.3). This declaration includes all parameters relating to acoustic model training, grammar definition, performance measures, and parameters used in speech enhancement and model adaptation. Using these

well-defined evaluation protocols, the performance of subtractive-type speech enhancement and MAP adaptation on both databases are analysed in Section 4.4.

4.2 In-Car Speech Databases

Much of the previous research on in-car ASR has typically been performed with small amounts of data collected for a particular study (e.g. [86, 107]), or data which was artificially generated [12, 20, 41, 98, 122]. The former approach typically results in limited amounts of evaluation data, which not only puts doubt over the statistical significance of the experiment, but also makes comparisons between techniques difficult since all researchers need access to that data.

To alleviate the limited data scenario, the Aurora experimental framework was introduced [57]. Although this database has been used extensively to report and compare experimental results [7, 43, 71], the framework has two very important limitations. Since noisy speech data was created by adding various noise sources (including car noise) to a large speaker-independent isolated digit database [78], no alteration was made to the speech waveform. As a result, it fails to reflect changes in speech production which are due to the Lombard effect or other types of stress. Further, state of the art speech enhancement techniques (see Chapter 3) use multiple microphones; therefore the Aurora framework is unsuitable for evaluating these methods. Under this framework, the comparison of both single- and multi-microphone enhancement techniques is impossible.

In order to overcome these limitations, a number of large in-car speech databases have been collected [50, 119]. These collections contain recorded speech from a large number of speakers under an extensive range of real noise conditions. Unfortunately, datasets of this size have seen limited use because they are either not publicly available or very expensive to acquire. The AVICAR ("Audio-Visual speech In a CAR") database collected at the University of Illinois [76] is an exception to this rule as it is freely available; this dissertation has developed an evaluation protocol [66] to enable widespread use for in-car ASR evaluations (see Section 4.2.1).

Whilst the cited databases (including AVICAR) provide significant resources

Table 4.1: AVICAR database in-car noise conditions.

Noise	Description
IDL	Engine running, car stopped, windows up
35U	Car travelling at 35mph, windows up
35D	Car travelling at 35mph, windows down
55U	Car travelling at 55mph, windows up
55D	Car travelling at 55mph, windows down

for in-car ASR with American and European speakers, no data previously existed for Australian speakers driving under Australian road conditions. Noticeable differences exist between Australian English and both American and British English [141]; therefore data is required to optimise in-car ASR systems for Australian environments. For this reason, the Australian English In-Car Speech corpus [67] was collected to assist this research (Section 4.2.2). This database was a major outcome of this Ph.D. research program, as well as the research collaboration between QUT, LaTrobe University, Melbourne, and General Motors Holden as part of a project under the Co-operative Research Centre for Advanced Automotive Technology (AutoCRC).

4.2.1 AVICAR

The AVICAR database contains multi-channel audio and video speech recordings in 5 different driving conditions (see Table 4.1). The microphone and camera arrays were placed on the sun-visor and dash in front of the front-seat passenger while the driver ensured the desired noise conditions were being met. Since the passenger's speech is recorded, this collection only *simulates* driver speech, and therefore is unsuitable for analysing the effects of speaking on driver distraction and vice versa. For such a purpose, a data collection such as UTDrive [6] would be more appropriate. Despite this limitation, AVICAR is suitable for evaluating low-SNR *neutral* speech recognition through combining multi-channel audio and visual speech recognition.

Data was recorded under four distinct tasks – isolated digits, isolated letters, phone numbers and TIMIT sentences. The isolated digits task closely resembles

command and control applications, whilst the isolated letters task mimics spelling which may be required in navigation systems. The other two tasks constitute continuous speech recognition tasks – phone numbers represent small vocabulary systems, whilst the sentences match medium vocabulary tasks. Further information about the utterance scripts used in the collection can be found in [76]. A recognition framework for the isolated digit and letter tasks is provided with the database, however this research is primarily interested in continuous speech recognition, and therefore an evaluation protocol was developed as part of this dissertation to enable model adaptation, development testing and evaluation [66].

More detailed information about the full recording setup can be found in [76]. It should be noted that the released portion of the AVICAR database contains less data than documented in [76]. It includes audio and video for 87 and 86 speakers respectively. This reduced amount of data was taken into account in the evaluation protocol development outlined in the following section.

Evaluation Protocol

Whilst all recorded speech is English, approximately 40% of the speakers are from Latin America, Europe, East or South Asia. Of these speakers, fifty-five suitable American English speakers were chosen to create a k-fold leave-one-out experimental procedure, with a smaller set chosen to analyse the performance of non-native speakers. The American English speakers were randomly divided into groups I-V as shown in Table 4.2. Group VI (comprising 11 randomly selected non-native speakers) is designed purely to characterise the expected decreases in ASR performance for non-native speakers – there is insufficient non-native data to make adaptation useful, and the variation of nationalities is likely to make system tuning problematic. In all instances, an effort was made to distribute male and female speakers evenly, as well as distribute the utterance scripts to limit text-dependency. For each speaker group, 160 utterances were randomly chosen for each noise condition, giving a total of 800 utterances per group.

The five native English groups were split into a series of experimental folds comprising 60% of the data for model adaptation, and 20% each for development

Table 4.2: AVICAR database protocol speaker groups.

Group	Speakers
I	AM4, BM4, CF5, DF1, EF4, EM1, FF2, GM4, HF2, HM3, IF1
II	AM3, BF5, BM1, CM1, DM2, EM4, FF5, GF2, HF3, IM5, JM2
III	AM2, BM3, CF1, DF4, EF1, EM2, FM2, GF1, GM1, HF5, JF1
IV	AM5, BF1, CF2, DF2, EF5, FM5, GF4, GM3, HM1, IM4, JF4
V	AF2, BF2, DF3, EF3, EM3, FM4, GF5, GM5, HF1, HM4, JF5
VI	AF3, BM2, CF4, CM3, DM3, FM3, GF3, HF4, IF3, JF2, JM4

Table 4.3: Protocol groups for k-fold leave-one-out ASR experiments.

Fold	Adapt.	Dev. Test	Eval. Test	Fold	Adapt.	Dev. Test	Eval. Test
1	I, II, III	IV	V	6	II, III, V	IV	I
2	III, IV, V	I	II	7	I, III, IV	V	II
3	I, II, V	III	IV	8	II, IV, V	I	III
4	II, III, IV	V	I	9	I, III, V	II	IV
5	I, IV, V	II	III	10	I, II, IV	III	V

testing and evaluation testing. These groupings are shown in Table 4.3. Averaging results over a number of folds enables more indicative speaker-independent recognition results since individual groups may be affected by poor (or very good) performance of one or two speakers. The first 5 folds are used for all experimentation in this dissertation as this ensures each speaker is used once in the evaluation.

The evaluation protocol also stipulates the use of the centrally located array microphone (M4) for all single-microphone experiments. Multi-microphone experiments can use whichever combination of microphones is required for the particular enhancement technique.

In order to reflect command and control applications in car environments, task grammars are chosen to be unconstrained word loops. This type of grammar produces worst-case recognition results (i.e. it is a true baseline system); this also allows improvements in ASR performance to be shown through the use of language models or grammar constraints. For the phone numbers and TIMIT sentences tasks, the number of words in the grammar are 11 and 773 respectively therefore constituting the small and medium vocabulary tasks. Throughout this research, only the phone numbers task is used as the performance of the sentences

Figure 4.1: Location of 8-microphone array used in collecting the Australian English In-Car Speech corpus.

task under such a task grammar is comparatively very low (see [66]).

More specific details on how this evaluation protocol was developed can be found in [66]. A copy of the file lists used in this evaluation has been made publicly available.

4.2.2 Australian English In-Car Speech Database

To collect speech data from the driver, a linear microphone array consisting of 8 high-quality omni-directional elements was fitted to the central roof console of a 2008 VE Commodore as shown in Fig. 4.1. This location is an industry-favoured position due to the ease of integration with existing electronics whilst still providing good signal-to-noise ratios [119]. The microphones were spaced symmetrically around the midline of the vehicle with 2 cm spacing between each adjacent microphone. The average location of the driver's mouth was estimated (with reference to the microphone closest to the driver) to be 35 cm to the right, 25 cm below, and 17.5 cm behind this reference microphone.

A total of 50 native-English speakers were collected for this corpus consisting of 20 female and 30 male drivers, all who had lived in Australia for at least 5 years to allow for naturalisation to the Australian English dialect. Female speakers were aged between 21 and 53 years; male speakers between 20 and 67 years old.

A command and control grammar (shown in Extended Backus-Nauer form in

Table 4.4: Extended Backus-Nauer form grammar used in the collection of the Australian In-Car Speech Corpus.

```
$Numbers     = [ NUMBER ] ( $Single_Digit |$Two_Digits |$Three_Digits |$Four_Digits );
$Street      = [ $Street_Prefix ] $Street_Name $Street_Type;
$In_Suburb   = [ ( AT |IN ) ] $Suburb;
$Corner      = ( CORNER |JUNCTION |INTERSECTION ) OF $Street AND $Street [ $In_Suburb ] ;
$Address     = ( [ $Numbers ] $Street $Suburb_List ) |( $Suburb_List $Street [ $Numbers ] ) |$Corner;
$Addr_Cmd    = ENTER ( ADDRESS |DESTINATION ) $Address;
$Other_Cmd   = RECALL DESTINATION $Single_Digit |( START |STOP ) NAVIGATION |RETURN |BACK |MAIN MENU;
$Cmd         = $Addr_Cmd |$Other_Cmd;
```

Table 4.5: Seven in-car noise conditions in the AEICS database.

Condition	Description
C0	Car idle, sealed cabin, no HVAC
C1	Medium speed (50-60 km/h), sealed cabin, no HVAC
C2	Medium speed (50-60 km/h), sealed cabin, HVAC on full fan
C3	Medium speed (50-60 km/h), driver window open, no HVAC
C4	High speed (90-100 km/h), sealed cabin, no HVAC
C5	High speed (90-100 km/h), sealed cabin, HVAC on full fan
C6	Car idle, sealed cabin, HVAC on full fan

Table 4.4) was formulated to generate a large number of consistent utterances for drivers to say in a variety of driving conditions based on a mock navigation task. The lists of 20 suburbs, 1931 street names, 16 prefixes and 37 street types were extracted from the Ausway index database. The task-oriented grammar provides the potential to investigate language processing techniques which may aid medium and large vocabulary command and control applications.

Seven different driving conditions were used as general audio scenes for utterance recordings. These conditions were chosen to capture variety in general noise types and levels present in the cabin of a vehicle whilst also representing likely driving scenarios in Australia at that time. Table 4.5 lists these recording conditions, where HVAC stands for the Heating, Ventilation, and Air Conditioning system.

Each speaker was recorded speaking a series of utterances in these driving conditions. For each driving condition, the speaker recorded 6 utterances consisting of one common, two repeated and three unique utterances. Common utterances were a set of utterances which each participant recorded, one associated with each specific driving condition. Repeated utterances occur more than once in the entire database (and may occur in the same noise condition), though never occur twice by the same speaker. Unique utterances occur only once across the entire database. This procedure was chosen to collect some data which can be regarded as speaker dependent whilst minimising the effect of text-dependency.

Utterances consisted of two different types of information – an address-style utterance, or a chain of six commands (never in the same order for unique utterances) used in a navigation system. It should be noted that the command chains were separated into individual commands after collection. Examples of both of these types of utterances are shown below.

Navigation: ENTER ADDRESS TWO GEORGE STREET BRISBANE *or* ENTER DESTINATION JUNCTION OF COLLINS WAY AND CORDOVA STREET WEST END

Command: MAIN MENU *or* START NAVIGATION *or* RETURN *or* RECALL DESTINATION THREE.

More details on the data collection procedure can be found in [67].

Evaluation Protocol

The AEICS database is suitable for use in a number of speech processing fields such as speech enhancement and speech recognition. The multiple channel recording process ensures investigations into current beamforming techniques are possible. For single-channel experiments, microphone 0 is chosen as it is closest to the driver and generates the highest ASR accuracies based on preliminary experiments. Multi-channel techniques can utilise any combination of channels as required by the individual technique.

Like the evaluation protocol for the AVICAR database, the 50 speakers are divided into 5 groups of 10 speakers to enable model adaptation, development

Table 4.6: Speaker groupings used in the Australian In-Car Speech Database evaluation protocol.

Group	Speakers	# Utterances
I	P04, P05, P11, P14, P16, P17, P21, P26, P35, P42	714
II	P08, P09, P12, P15, P22, P27, P30, P34, P47, P49	840
III	P02, P07, P18, P23, P38, P39, P43, P46, P54, P55	790
IV	P10, P19, P24, P25, P31, P32, P36, P45, P52, P53	720
V	P03, P06, P13, P20, P28, P29, P33, P37, P41, P51	749

and evaluation testing through the use of k-fold leave-one-out testing. The groups (shown in Table 4.6) were randomly generated with some gender balancing as per the previous evaluation protocol. Again, 60% of the data is made available for adaptation, with 20% set aside for both development and evaluation testing. The experimental folds are the same as those shown in Table 4.3. The first 5 folds are used for all experimentation in this dissertation as this ensures all 50 speakers are used in the evaluation.

Unlike the AVICAR evaluation protocol, the constrained grammar used for utterance generation (see Table 4.4) is also used for ASR. The constrained grammar is necessary since there are two types of potential input (i.e. commands or navigation addresses), and using an open-word loop grammar could lead to recognition hypotheses which make no sense for this application.

4.3 Experimental Configuration

4.3.1 Baseline Speech Recogniser

Throughout this research, the Hidden Markov Model Toolkit (HTK) [155] is used for acoustic model training and utterance decoding. Context-dependent 3-state left-to-right triphone HMMs were trained using the speaker-independent Wall Street Journal 1 training dataset which consists of almost 70,000 utterances. Unless otherwise noted, the full procedure outlined in Section 2.2.2 including CMS and cepstral liftering was used to generate 39-D MFCC feature vectors – 13 MFCC (including C_0) along with delta and acceleration coefficients – for each

32 ms frame with 10 ms advance between frames. HMM states were represented using a 16-component GMM for speech components, and a 48-component GMM for silence models.

4.3.2 Speech Recognition Performance Measure

Evaluating the performance of ASR systems requires a performance metric which enables simple comparisons of results. Two common measures used in such evaluations are the *word recognition rate* and the *word error rate.* In this dissertation, only the word recognition rate is referred to. The word recognition rate of a speech recognition system is defined as:

$$Accuracy = \frac{N - D - S - I}{N} \times 100\% \tag{4.1}$$

where N represents the total number of words in the experiment, D the number of deletions, S the number of substitutions, and I the number of insertions. A deletion occurs when a word from the known sequence is removed from the hypothesised word sequence. A substitution occurs when a word from the known sequence is replaced by a different word in the hypothesised word sequence. An insertion occurs when a word is added in the hypothesised word sequence. The hypothesised word sequence is that which is determined by the speech decoder as the sequence with the greatest likelihood, whilst the known word sequence is the true sequence of words. These two sequences must first be dynamically aligned in order to correctly determine the word accuracy.

As an example of calculating the word accuracy, the following dynamically aligned sequences comprise a known and a hypothesised word sequence from the phone numbers task of the AVICAR database [76]. Substitutions are shown in **bold**, deletions by ** and insertions are underlined. The overall word accuracy of this sequence is 70% since there is one substitution, one deletion and one insertion.

True:	Six three zero seven one nine five eight seven three.
Hypothesised:	Six three zero seven one nine **nine** eight oh seven **.

Table 4.7: Parameters used in evaluating the various techniques.

Technique	Parameters	Spectral Floor
LSS	$\gamma = 1$, $\alpha = 1$, $\beta = 0.45$	Noise Estimate
MBSS	$\gamma = 2$, $\beta = 0.35$	Noisy Speech Signal
MFNS	$\alpha = 1$, $\beta = 0.45$	Noisy Speech Signal

4.4 Baseline ASR Evaluation

A baseline ASR evaluation was conducted on both databases outlined in Section 4.2 to assess the performance of three spectral subtractive speech enhancement techniques, as well as model adaptation and task grammars in the case of the AVICAR database. The enhancement techniques evaluated include frequency-domain linear spectral subtraction (Eq. (3.3)), Kamath's multi-band spectral subtraction (Eq. (3.4)), and Mel-filterbank noise subtraction (Eq. (3.6)). The subtraction and flooring parameters used for each of these techniques are shown in Table 4.7. For all enhancement techniques, time recursive averaging with soft-decision SAD (Eq. (3.12)) was used to estimate the noise with $\lambda = 5$ and $\eta = 0.97$ determined empirically based on preliminary ASR word accuracy performance.

Using data from the AVICAR database, the mean and variances of the original acoustic models were adjusted using MAP adaptation with $\tau = 16$ used for weighting the prior speech model. In [66], it was observed that variance and mean adaptation was most successful in adapting Gaussian mixtures to in-car noise conditions. The high emphasis placed on the prior acoustic model can be attributed to ensuring the models remain speaker-independent since there are only 33 speakers in each adaptation fold. The experimental results in [66] have not been included here in order to maintain focus on the evaluation of the enhancement techniques.

Based on the empirically determined value of $\tau = 16$ for the AVICAR database, a value of $\tau = 8$ was used for the AEICS corpus since there is a need to also put emphasis on the Australian English dialect in the adapted acoustic models and not just adapt to the noise conditions.

Finally, to demonstrate the performance difference between constrained and unconstrained grammars, the open-word loop specified for the AVICAR phone

Table 4.8: ASR baseline evaluation results for phone numbers task of the AVICAR database.

	Grammar	ASR Word Accuracy(%)				
		IDL	35U	35D	55U	55D
Baseline	Open	71.6	49.6	37.2	42.9	24.7
Baseline	Constrained	78.4	53.9	37.8	43.9	22.9
LSS	Open	75.0	54.4	41.0	50.6	31.0
MBSS	Open	74.2	50.7	37.6	47.3	29.6
MFNS	Open	74.2	49.7	36.9	46.8	28.6
MAP Adaptation	Open	82.8	77.4	69.4	76.2	59.2
MAP + MFNS	Open	80.6	73.8	66.2	75.4	61.0

numbers task was altered to ensure decoding always produced sequences of 10 digits. As a result, the possible types of errors was reduced to deletions and substitutions.

4.4.1 Experimental Results & Discussion

The speech recognition results for the AVICAR phone numbers task and the AEICS corpus are shown in Tables 4.8 and 4.9 respectively. Throughout this dissertation, references to "Baseline" results are those obtained by decoding the original noisy speech signals. It should also be noted that results presented in [67] for the AEICS corpus were a combined average of the command and navigation tasks presented here.

Discussion

General Data Trends: Analysing the baseline system results for both datasets, a number of observations related to in-car noise conditions can be made. Comparing the results for all car speeds with either windows up or down (AVICAR) or HVAC on or off (AEICS), it can be seen that an increase in vehicle speed – which increases noise levels due to wind and road friction – causes degradation in recognition accuracy. The decrease in performance is particularly noticeable in the navigation task of the AEICS corpus when the air-conditioning system is

Table 4.9: ASR baseline evaluation results on the AEICS corpus.

	Task	ASR Word Accuracy(%)						
		C0	C6	C1	C2	C3	C4	C5
Baseline	Commands	92.0	55.8	84.6	49.0	81.2	81.8	43.2
LSS	Commands	94.0	72.6	88.1	68.3	88.6	86.3	67.7
MBSS	Commands	94.2	72.3	86.6	69.6	88.4	85.9	67.6
MFNS	Commands	93.8	71.0	85.4	69.6	87.0	85.5	66.3
MAP Adaptation	Commands	99.1	96.9	98.7	95.5	98.6	98.5	94.8
MAP + MFNS	Commands	98.7	98.4	98.5	98.3	99.2	98.3	96.6
Baseline	Addresses	83.3	36.8	67.9	30.8	46.4	47.0	27.2
LSS	Addresses	84.9	48.6	74.2	42.7	53.5	55.2	39.3
MBSS	Addresses	85.8	48.1	74.7	41.5	56.4	53.3	37.7
MFNS	Addresses	85.6	48.1	73.1	41.4	55.1	52.5	37.5
MAP Adaptation	Addresses	91.3	69.4	86.2	71.2	81.6	81.5	68.6
MAP + MFNS	Addresses	90.1	76.9	88.7	80.2	85.5	85.0	79.6

off (C0, C1, C4) – accuracies are 83.3%, 67.9% and 47.0% for idle, 50-60 km/h and 90-100 km/h respectively.

In the AVICAR results, having the windows open appears to have more affect on the recognition accuracy than simply increasing vehicle speed. This is demonstrated through recognition accuracies showing better performance for the car traveling at 55 mph with windows up (55U) compared to 35 mph with windows down (35D). This result is in accordance with the findings of Zhang and Hansen [157] who determined that road and wind noise dominate the noise field when the windows are open. Compared with having the windows closed, this scenario leads to greater decreases in accuracy as vehicle speed increases. This is due to the fact that a sealed cabin acts as a filter and rejects some of the external noise; opening the window subjects the cabin to amplified levels of road and wind friction as vehicle speed increases.

For AEICS, having the air-conditioning system on appears to have the greatest effect on ASR accuracy. In the idle case (C0 and C6), the performance difference is approximately 46.5% for the navigation task. Further, having the window

down (C3) doesn't degrade the performance anywhere near as drastically as air-conditioning (C2). This effect is attributed to the location of the air-conditioning vents which are directly beneath the microphone array; therefore fan noise is recorded by the microphone at considerably higher amplitudes than noise coming from the driver's side window. This observation (along with that of the previous paragraph) demonstrates the need for careful consideration of microphone placement in the vehicle, as well as an understanding of ASR performance limitations that may result from a particular choice of location.

Grammar: Using a constrained grammar on the AVICAR phone numbers task improves recognition performance in most noise conditions, although the amount of improvement reduces as the noise level increases, and leads to inferior word accuracies in the noisiest condition (55D). Two factors are involved which lead to these results. Firstly, the open word loop grammar was found to be particularly susceptible to deletion errors, which is a direct result of the background noise. Words are deleted in this instance as the noise completely masks the speech, and the silence model is determined to be more likely than any speech component. This is particularly true for words beginning with unvoiced components like /f/ as in "FOUR" or "FIVE" (which exhibited the highest deletion occurrences out of the 11 digits). Unvoiced components are commonly regarded as having characteristics similar to white Gaussian noise, and therefore tend to be easily masked by background noise.

In the case of a constrained grammar, deletion and insertion errors occur together since every utterance must contain 10 digits. This grammar exhibited less combined occurrences of deletions and insertion errors than the open word loop grammar, but substitutions occurred more regularly. This was particularly true in the 55 mph with windows down noise condition where the number of substitutions was more than twice the number for the open word loop grammar. Inspection of the confusion matrix showed most substitutions resulted in either "OH" or "EIGHT" being recognised. These two words respectively accounted for 48% and 33% of the substitutions in this noise condition. As observed previously, words which start with an unvoiced component are masked until a voiced

component is present. Since there must now be 10 digits in the hypothesised utterance, these words are substituted with words beginning with voiced phonemes ("OH" or "EIGHT") rather than deleted as was the case with the open word loop grammar. Whilst there are other examples of words beginning with voiced speech in this vocabulary ("ONE", "NINE" and "ZERO"), these two words are phonemically shorter which will make them more likely.

Model Adaptation: The application of MAP adaptation shows global improvements over the baseline results in all noise conditions for both corpora. For the AVICAR database, since the adaptation set consists of native American English speakers, adaptation adjusts the models solely to the noise levels in the vehicle. The improvement in word accuracy for all noise conditions excepting idle (which is significantly more quiet) is more than 25%, proving the effectiveness of MAP adaptation for this application. The scale of these improvements can be attributed to task similarity between adaptation and test sets which both include a large number of digit samples from the small vocabulary phone numbers task. This ensures the adaptation process accurately transforms the active triphone models to the new in-car environment since there are many examples in the adaptation set.

For the AEICS corpus, the increases in ASR performance can be attributed to the two-fold adaptation to both the noisy in-car environment and the Australian English speakers; although exact contributions of each factor are difficult to ascertain and not considered important. The average 7.5% improvement in word accuracy across the two tasks in the very low noise idle condition (C0) demonstrates the contribution of adaptation to the Australian accent. All other noise conditions exhibit improvements in excess of 14% for the command task, and 18% for the navigation task due to dialect and noise adaptation. These results show that the AEICS corpus is more than suitable for adapting well-trained American English acoustic models for in-car speech applications in Australian environments.

Speech Enhancement: All three single-channel spectral subtractive enhancement techniques provide word accuracy improvements over the baseline system. This is true for all noise conditions with very few exceptions including idle in

both datasets. The absolute improvements are seen to be greater in the noisier conditions where the background noise considerably hinders speech recognition accuracy. These results prove the value of using spectral subtractive speech enhancement in the front-end of a speech recogniser.

Comparing the performance of the three techniques shows that frequency-domain LSS provides better overall performance than the other two techniques. Despite using static values for the subtraction parameter ($\alpha = 1$), these results verify the belief in [10, 127] that optimising enhancement parameters based on signal-level criteria (such as SNR in Kamath & Loizou's MBSS method) does not lead to optimal ASR performance. This result strengthens the motivation to use likelihood-maximising techniques to optimise speech enhancement techniques used in the ASR front-end; this is explored in Chapters 5 and 6.

Despite showing consistent improvements over the baseline system, MFNS fails to match the ASR performance of frequency-domain LSS. Given that noise components overlap with speech components in the frequency range below 1000 Hz (a phenomenon which can be seen in Fig. 4.2), to explain these results we consider the extreme case where two adjacent frequencies, f_1 and f_2, contain only noise and only speech respectively. In this instance, the superior performance of LSS can be attributed to the fine-grained nature of the subtraction process which would remove most of the noise from f_1 but not subtract anything from f_2 ensuring the speech signal is fully preserved. This is because the noise estimate at f_1 will be close to the instantaneous magnitude, while the estimate at f_2 will be zero.

The spectral averaging prior to MFNS on the other hand, causes the noise and speech components to be combined into the same filterbank which puts greater emphasis on the noise estimation procedure and also the subtraction factors α (which have been kept constant in this experiment and comparison). If noise is oversubtracted from this filterbank, the speech component of the signal will likely be distorted resulting in decreased ASR word accuracy.

From this analysis, it can be seen that the performance of MFNS is susceptible to instances where noise and speech are present in adjacent frequencies which are averaged into the same Mel-filterbank energy. This scenario is likely to occur

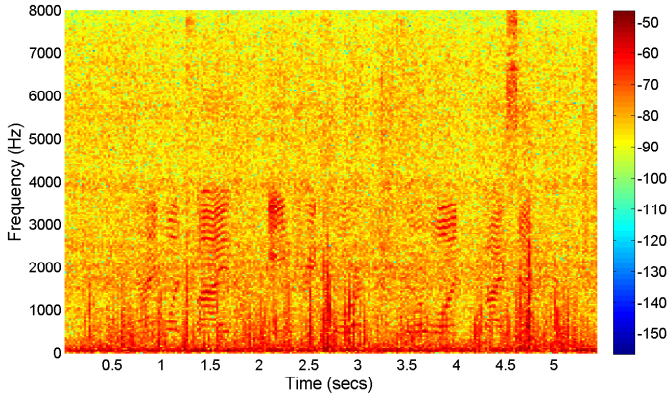


Figure 4.2: Example spectrogram of speech recorded at 55 mph with windows down.

in low-frequencies where the noise and speech signals overlap as seen for the in-car noise case shown in Fig. 4.2. It should be noted however, that there is no advantage in using LSS over MFNS when speech and noise signals exist at the *same* frequency f as both methods have the potential of inaccurately subtracting the noise estimate from the signal.

Speech Enhancement and Model Adaptation: Combining model adaptation and MFNS speech enhancement yields results which are always better than MFNS alone, but in the case of AVICAR are inferior to model adaptation in most noise conditions. For the AVICAR database, the decreased ASR performance is due to an objective conflict between the two techniques. MAP adaptation transforms the clean speech models to some level of background noise based on the adaptation data. Speech enhancement on the other hand, aims to transform the noisy speech back to a clean speech estimate.

An example of this conflict for the AVICAR database is shown in Fig. 4.3 where the 'blue mismatch' represents the difference between adapted models and noisy speech, and the 'orange mismatch' is the difference between adapted models and enhanced speech for the 55U noise condition. It can be seen that by performing speech enhancement and then ASR using adapted models can actually increase the mismatch between the data and the models compared to the use of model adaptation only. This explains the majority of the results seen in Table 4.8, the only exception being the 55 mph with windows down condition which

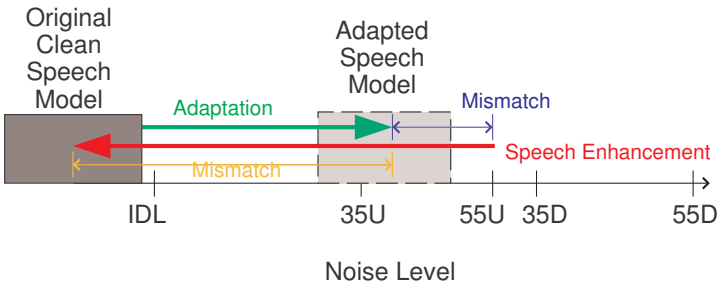


Figure 4.3: Example of the adaptation-enhancement conflict on the AVICAR database.

improves by 1.7%. In this case, the very high level of noise in the original recordings increases the mismatch between noisy speech and adapted acoustic models (i.e. the 'blue mismatch'); therefore speech enhancement is able to make the orange 'mismatch' smaller (albeit by enhancing the speech to contain less noise than that present in the adapted models).

The results for the combined adaptation and enhancement scenario on the AEICS corpus (Table 4.9) exhibits a different performance trend. Only the idle noise condition (C0) shows a degradation in word accuracy which follows the adaptation-enhancement conflict previously discussed. The improvement in ASR accuracy observed in all other noise conditions suggests some imbalance between adapting to both noise and dialect mismatches. For example, if adaptation focuses on dialect more than noise, the overall noise levels in the acoustic model will be lower than experienced in the AVICAR case. In this case, speech enhancement will be able to further reduce the noise mismatch between the test data and the adapted models.

Since Australian English speakers are used throughout the adaptation and test sets, the improvement due to dialect adaptation should be constant across all noise conditions, whilst improvements due to noise adaptation will vary due to the remaining 'blue mismatch' in Fig. 4.3. Since the idle case (C0) has little to noise background noise, it should benefit primarily from dialect adaptation, and may even suffer from noise adaptation since the resulting noise levels in the model will be greater than the idle condition. Looking at the results in Table 4.9 more closely, the relative improvement in word accuracy performance for idle on

the address task is approximately 48%. In the case of traveling at 50 km/hr with window down (C3), the relative improvement is increased to 65%, which suggests the noise adaptation on this condition only contributes around 17% improvement which is considerably less than the improvement due to dialect adaptation. As a result, there will likely be noticeable noise mismatch, and so speech enhancement is able to produce improvements over the system using only MAP adaptation.

This analysis of the speech recognition results provides some insight into the processes occurring in both MAP adaptation and speech enhancement. Further experimentation could be performed to examine this phenomenon more closely, however a number of factors including the amount of training data and the emphasis to be placed on the original acoustic model will need to be considered carefully. As a result, it was deemed unnecessary to further investigate this effect in this dissertation.

Given the results on the accent-matched AVICAR database, it should be noted that ASR performance could be improved with the joint adaptation-enhancement approach if the enhancement was used as part of the feature extraction procedure used in model training. Another approach to improve performance would be to alter the enhancement parameters ($\alpha \neq 1$) to better suit the new acoustic model. Within the aims of this dissertation, the latter approach is preferred as it enables the enhancement technique to be used with any acoustic model making the approach independent of the data used for training and adaptation.

4.5 Research Directions

The experimental evaluation in Section 4.4 showed the ASR performance characteristics of single-channel subtraction-type speech enhancement techniques. In summary, these techniques were able to improve word accuracies when using clean speech models, but failed to provide any performance gains when using dialect-matched adapted acoustic models. Further, the best word accuracy performance in very noisy conditions was around 60% which is unlikely to meet user expectations.

The research directions proposed in Chapter 3 were specifically directed at designing speech enhancement techniques with improvements in ASR performance in mind. The experimental results presented in this chapter have solidified this position and also added the consideration of making these techniques suitable for use with any acoustic model regardless of the nature of the data used in the training process. Chapters 5 and 6 consider a likelihood-maximisation technique which satisfies both these objectives, whilst Chapters 7 and 8 look solely at improving recognition accuracy when only clean speech acoustic models are available.

It would also be beneficial to examine the effects of combining model adaptation and speech enhancement specifically in scenarios where there are environmental mismatches (e.g. office versus in-car) as well as mismatches due to speech production (e.g. accent/dialect or stressed speech). The results of such a study would enable system designers to fine-tune acoustic models based on a wider range of user and environment requirements, and would be particularly useful in creating effective in-car speech recognition systems for the Australian automotive industry where there is currently limited speech resources.

4.6 Summary

This chapter has presented two in-car speech databases which are used throughout this dissertation for ASR evaluation of speech enhancement techniques. The first in-car speech database including Australian speakers (AEICS) was collected as part of this thesis and was described in Section 4.2.2. The AEICS database was shown to be suitable for developing in-car ASR applications in Australian environments given existing systems trained on American English data.

The baseline recognition system and evaluation protocols for each database used in this research were also outlined. Initial evaluation of the performance of three different single-channel speech enhancement techniques was performed, and showed each technique capable of improving the performance of a baseline ASR system. The performance of these techniques failed to outperform acoustic model adaptation, and combining both enhancement and adaptation showed different performance characteristics depending on the levels of mismatch. For

a system with matched dialects in adaptation and test phases, decreased overall recognition accuracy was observed because of the mismatch incurred due to the adaptation-enhancement conflict. A system in which there were mismatches between dialects as well as noise conditions produced an acoustic model which favoured dialect adaptation, enabling speech enhancement to improve the overall ASR word accuracy when the two systems were combined.

The research directions proposed in Chapter 3 regarding the design of speech enhancement algorithms which specifically improve speech recognition accuracy were supported with experimental evidence which showed poor performance of ASR in high noise environments and a need to carefully consider the data used for acoustic modeling. In Chapter 5, a method called likelihood-maximisation is applied to MFNS in order to optimise the enhancement parameters for improved speech recognition performance; this approach is suitable for use with clean speech and noise-adapted acoustic models.

Chapter 5

Likelihood-Maximising Speech Enhancement for Robust ASR

5.1 Introduction

Many of the single- and multi-channel speech enhancement algorithms presented in the literature review in Chapter 3 were designed primarily to produce improvements in the human intelligibility of speech signals. In doing so, most of these techniques optimise enhancement parameters based on signal-level criteria such as maximising signal-to-noise ratio, minimising speech distortion or minimising the mean-squared signal error. Automatic speech recognition systems however, hypothesise the most likely sequence of statistical acoustic models produced by the observed feature vectors. Since ASR is a computer pattern recognition problem, traditional optimisation of speech enhancement algorithms based on waveform criteria do not translate into optimal improvements in ASR word accuracy.

The likelihood-maximisation (LIMA) framework was designed to overcome this incompatibility of optimisation criteria in traditional enhancement techniques. In Section 5.2, likelihood-maximisation is derived in general form, and the limited studies which have employed this approach are reviewed with particular reference made specifically to studies employing spectral subtractive enhancement. The scope of research in this chapter is confined only to spectral subtractive techniques as they are a common approach for single-channel speech

enhancement, are computationally simple and provide sufficient levels of noise reduction. Such considerations are very important for applications in automotive environments as was outlined in the Scope of Research (Section 1.3). In this dissertation, it is proposed to apply LIMA to Mel-Filterbank Noise Subtraction (MFNS), and a direct theoretical comparison with the existing application employing frequency-domain Multi-Band Spectral Subtraction (MFNS) is made in Section 5.3.

Whilst these studies have shown the promise of LIMA frameworks with a range of enhancement techniques, some limitations in the approaches and experiment procedures have yet to be addressed. Section 5.4 introduces potential methods for overcoming some of these limitations and proposes extensions to the previous application of LIMA to subtractive-type algorithms.

Having proposed the application of the LIMA framework to Mel-filterbank noise subtraction, validation experiments are performed in Section 5.5. These experiments also assess simplifications and extensions to the existing parameter set, along with the performance of this technique using noise-adapted acoustic models.

5.2 Likelihood-Maximising Speech Enhancement

5.2.1 Development of LIMA Framework

In Chapter 2, ASR was presented as a statistical pattern recognition problem rather than a signal processing problem. For example, in this research acoustic events are modeled as mixtures of Gaussian probability distributions; therefore the goal of ASR is to determine the sequence of acoustic models which most likely correspond to the observed feature vectors. Since speech enhancement is incorporated as part of the feature extraction process, LIMA aims to determine the set of enhancement parameters which maximises the likelihood of the correct sequence of acoustic events being output from the ASR system. This section demonstrates how this can be achieved for *any* speech enhancement technique

with a set of P enhancement parameters defined by:

$$\boldsymbol{\xi} = \{\xi_1, \xi_2, ..., \xi_P\}. \tag{5.1}$$

In order to derive the LIMA approach for optimisation, first recall from Section 2.3.2 the optimal Bayes' classifier used for ASR decoding:

$$\hat{w} = \arg\max_{w \epsilon W} P(\boldsymbol{O}|w)P(w) \tag{5.2}$$

where $\boldsymbol{O}$ is the sequence of observed features described by Eq. (2.10), and $P(\boldsymbol{O}|w)$ and $P(w)$ are the acoustic and language scores respectively. The observed feature vectors are a function of both the input speech and the feature extraction procedure; therefore, if speech enhancement is part of feature extraction, the observed features are a function of the enhancement parameters:

$$\hat{w} = \arg\max_{w \epsilon W} P(\boldsymbol{O}(\boldsymbol{\xi})|w)P(w). \tag{5.3}$$

In Eq. (5.3) it can be seen that the language score is not dependent upon the observed feature vectors (and therefore $\boldsymbol{\xi}$), and can be ignored [127]. The optimal set of enhancement parameters $\boldsymbol{\xi}$ is calculated such that it maximises the acoustic likelihood $P(\boldsymbol{O}(\boldsymbol{\xi})|w)$ given a transcription w_C which is assumed to be known *a priori*:

$$\hat{\boldsymbol{\xi}} = \arg\max_{\boldsymbol{\xi}} P(\boldsymbol{O}(\boldsymbol{\xi})|w_C). \tag{5.4}$$

Practicalities of obtaining and utilising the correct word transcription w_C in vehicular environments are the focus of Chapter 6, and as such will not be discussed further in this chapter.

For HMM-based speech recognition systems, there are many possible state sequences which generate the correct word transcription w_C – these sequences all contribute to the overall acoustic likelihood. Amongst the collection of correct state sequences S_C is the most likely state sequence s_i for all frames i – it is assumed that this particular sequence contributes the most to the total acoustic likelihood (since many sequences are highly unlikely). This assumption considerably reduces the computational complexity of the LIMA approach. The maximum-likelihood estimate of the enhancement parameters $\boldsymbol{\xi}$ which optimises

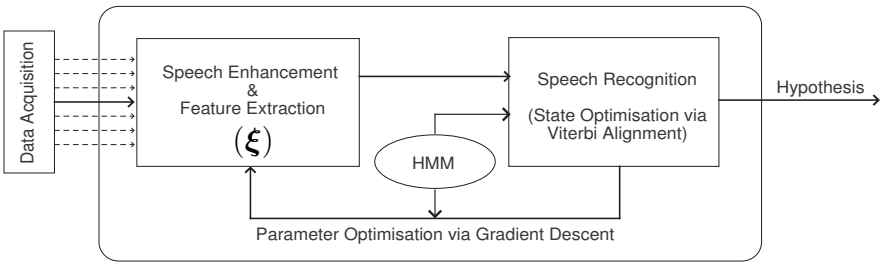


Figure 5.1: Generalised ASR likelihood-maximising framework for speech enhancement.

the log-likelihood of the acoustic state sequence s_i is therefore determined as [125]:

$$\hat{\boldsymbol{\xi}} = \arg \max_{\boldsymbol{\xi}, s \in S_C} \left\{ \sum_i \log(P(\boldsymbol{o_i}(\boldsymbol{\xi})|s_i)) + \sum_i \log(P(s_i|s_{i-1}, w_C)) \right\}. \tag{5.5}$$

This equation contains two terms which demonstrate the joint optimisation problem encountered in this framework. The first term, $\sum_i \log(P(\boldsymbol{o_i}(\boldsymbol{\xi})|s_i))$, determines the optimal set of enhancement parameters given a correct and constant state sequence s_i. The second term, $\sum_i \log(P(s_i|s_{i-1}, w_C))$, determines the optimal state s_i given the correct word transcription w_C and the previous state s_{i-1} (N.B. the enhancement parameters are constant). Thus, the state sequence s_i and the set of enhancement parameters $\boldsymbol{\xi}$ are jointly optimised, a process shown graphically in the generalised LIMA framework in Fig. 5.1. Optimisation of the recognised state sequence is achieved through the use of Viterbi alignment which is similar to Viterbi decoding described in Chapter 2 but uses the known transcription w_C to generate a frame-by-frame alignment of model states.

Since the second part of Eq. (5.5) optimises only the state alignment, the optimisation of the set of enhancement parameters for an HMM-based speech recognition system is defined as:

$$\hat{\boldsymbol{\xi}} = \arg \max_{\boldsymbol{\xi}} \log(P(\boldsymbol{o_i}(\boldsymbol{\xi})|s_i)). \tag{5.6}$$

A closed form solution to this optimisation problem does not exist due to the complex signal processing involved in the feature extraction and speech enhancement processes. Therefore, non-linear optimisation approaches such as gradient-descent methods are required to solve this problem. In order to use gradient-descent optimisation, it is required to determine the gradient of the likelihood

function $L(\boldsymbol{\xi})$:

$$L(\boldsymbol{\xi}) = \sum_{i} \log(P(\boldsymbol{o_i}(\boldsymbol{\xi})|s_i)). \tag{5.7}$$

Assuming the GMM defined by Eqs. (2.11)-(2.12), calculation of the gradient and appropriate simplifications leads to the gradient function with respect to each of the enhancement parameters (see [125] for full derivation) :

$$\nabla_{\boldsymbol{\xi}} L(\boldsymbol{\xi}) = -\sum_{i} \sum_{m=1}^{M} \chi_{im}(\boldsymbol{\xi}) \frac{\partial \boldsymbol{o_i}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} \Sigma_{im}^{-1} (\boldsymbol{o_i}(\boldsymbol{\xi}) - \mu_{im}) \tag{5.8}$$

where $\chi_{im}(\boldsymbol{\xi})$ is the *a posteriori* probability of the m^{th} mixture component in state s_i given the observed feature vector $\boldsymbol{o_i}(\boldsymbol{\xi})$ (for more details, refer to [125]). The mean vector μ and covariance matrix Σ are required for each state i and mixture component m in order to calculate the gradient. The remaining term in Eq. (5.8) is the Jacobian matrix, $\frac{\partial \boldsymbol{o_i}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}$, which consists of the partial derivatives of each feature vector element at frame i with respect to each of the parameters of the speech enhancement technique. The derivation of the Jacobian matrix is unique to each speech enhancement technique – examples of Jacobian matrices can be found in the references cited in Section 5.2.2.

Having obtained the Jacobian elements for the speech enhancement technique of interest, the optimal set of enhancement parameters can be obtained using the method of conjugate gradients [104]. This has been the most popular approach to solving this non-linear optimisation problem in previous studies [10, 127], and will be used in all experimentation involving the LIMA framework in this dissertation.

The joint optimisation process continues until both the enhancement parameters and the state sequence converge. Given variations in speakers, noise conditions and utterance lengths, the number of optimisation iterations is not bounded, and could be very time-consuming in order to guarantee convergence. This particular consideration is explored in more detail for in-car applications in Chapter 6.

5.2.2 Previous LIMA Speech Enhancement Studies

Despite the potential of LIMA-based speech enhancement for improved robust ASR, very few studies have yet to apply this approach. The first major study and derivation was performed by Michael Seltzer and colleagues at Carnegie Mellon

University [125, 127]. In their research, the LIMA approach was originally applied to a time-domain multi-microphone filter-and-sum beamformer; they termed this technique LIMABEAM. Their research also considered practical use of the LIMA approach – the proposed methods will be discussed in detail in Chapter 6.

Optimisation in LIMABEAM was used to determine 140 filter weights in the log-Mel spectral domain as opposed to the Mel-frequency cepstral domain, however ASR was still performed on MFCC features using parallel acoustic models. Performance evaluation in moderately noisy and reverberant office environments showed an average 17.5% relative improvement over both conventional delay-and-sum beamforming and a filter-and-sum beamformer with post-filtering. Further improvements in ASR performance were obtained by applying other robust ASR techniques including MLLR adaptation after LIMABEAM. This collection of experiments demonstrated the potential of LIMA frameworks for robust ASR using speech enhancement. The authors successfully extended the system to sub-band LIMABEAM designed specifically to improve the performance of LIMABEAM in highly reverberant environments [128].

A similar framework was used by Shi *et al.* [130] for optimising the single parameter which determines the aggressiveness of the dual-channel phase-error filter discussed in Section 3.5.3. This parameter was optimised using a generalised EM algorithm in the MFCC feature space, however the acoustic models were trained with phase-error filtered clean speech data rather than unprocessed clean speech data. This system outperformed the static phase-error filter as well as a dual-channel delay-and-sum beamformer with post-filtering under a wide range of input SNR.

More recently, LIMA was applied to a multi-band spectral subtraction technique [9, 10]. In this research, 25 Mel-spaced sub-bands were used to perform spectral subtraction in the frequency-domain. The expression for the Jacobian elements, $\frac{\partial \boldsymbol{o_i}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}$, was also extended to account for the inclusion of CMS in the feature extraction process:

$$\frac{\partial \boldsymbol{o_i}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \frac{\partial \boldsymbol{o_i}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} - \frac{1}{T}\sum_{i=1}^{T} \frac{\partial \boldsymbol{o_i}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} \tag{5.9}$$

where T is the total number of frames in the utterance. Details of this technique

are provided in Section 5.3.1 and compared with the proposed Mel-filterbank noise subtraction implementation in Section 5.3.2.

In [10], the multi-band spectral subtraction approach was accompanied by an extensive experimental evaluation. Phoneme recognition was performed on two clean speech datasets which were mixed with a range of noises to satisfy a range of SNR. The LIMA approach showed consistent accuracy improvements over a modified version of the multi-band spectral subtraction proposed by Kamath and Loizou [64] as well as a system with no speech enhancement. Other important outcomes from this study include:

1. On clean speech data, the LIMA approach recovers decreases in ASR performance which are incurred by traditional speech enhancement techniques; the resulting recognition accuracies closely match those obtained on clean speech.

2. In scenarios where noise levels change during a recording, LIMA speech enhancement still provides improved speech recognition accuracy.

3. The use of a single enhancement parameter in the LIMA framework can outperform the spectral subtraction method proposed by Berouti *et al.* [13].

4. LIMA speech enhancement can be used with acoustic models which are trained on noisy speech data. The LIMA optimisation process adapts the enhancement parameters to best fit the statistics of the acoustic model; in other words, it minimises any train-test mismatch which results from the operating environment. As such, the LIMA approach can be seen to be independent of the data used for acoustic model training.

5. The inclusion of CMS in the feature extraction process (and subsequently LIMA speech enhancement) results in over 30% absolute word error reduction on data recorded in a 15 dB office environment.

Despite these promising outcomes and the size of the evaluation, there are a few shortfalls which will be discussed in Section 5.4.

All the studies discussed in this section optimised the enhancement parameters based upon the state sequence for an entire utterance. As a result, the

enhancement parameters are the best match for a wide range of HMM states, but not necessarily optimal for a specific acoustic event. The standard LIMABEAM algorithm was extended to derive a separate set of filter weights during calibration for each phone observed in the calibration utterance [70]. Their experiments showed a small number of filter-taps is best in order to avoid over-fitting to the small amount of calibration data for each phone. Using this setup, it was possible to achieve a 12.2% relative improvement over the utterance-based LIMABEAM method used by Seltzer *et al.* [127] in low SNR conditions. Despite the improvements in ASR performance, this method is likely to incur considerable overhead during decoding of an unknown utterance as each input frame must be processed by the filters for all phones. At this point in time, the authors have not proposed a method to reduce this overhead.

5.3 LIMA Applied to Spectral Subtractive Speech Enhancement

5.3.1 Multi-Band Spectral Subtraction

BabaAli *et al.* [9, 10] propose the use of overlapping Mel-spaced sub-bands to perform multi-band spectral subtraction in the frequency domain. The advantage of using MBSS as opposed to applying a subtraction factor α for each frequency band is that the size of the parameter space to be optimised is considerably reduced. Using MFCC feature extraction as defined in Chapter 2, the number of parameters P can be reduced from 257 (i.e. $(N/2)+1$ where N is the length of the DFT) to 26. In their implementation, a subtraction factor α_l is used for each Mel-spaced filter l, which is defined in the frequency domain as:

$$\hat{\alpha}_l(k) = \begin{cases} \alpha_l & f_L^l \leq k < f_U^l \\ 0 & \text{otherwise} \end{cases} \tag{5.10}$$

where f_L^l and f_U^l are the lower and upper frequencies of the l^{th} Mel-filter. Due to the overlapping nature of these filters, it is also necessary to introduce a vector $\boldsymbol{B}$ which accounts for this overlap (N.B. in frequencies with overlapping filters

$B(k) = 2$, otherwise $B(k) = 1$). The resulting spectral subtraction rule used in the MBSS implementation [10] is defined as:

$$|\hat{S}^i(k)|^2 = \left(|Y^i(k)|^2 - \sum_{l=1}^{L} \frac{\hat{\alpha}_l(k)}{B(k)} |\hat{D}^i(k)|^2\right) \times U\left(|Y^i(k)|^2 - \sum_{l=1}^{L} \frac{\hat{\alpha}_l(k)}{B(k)} |\hat{D}^i(k)|^2\right) \tag{5.11}$$

where U is the Heaviside step function, and i is the frame index. From this equation they derive the gradient of $|\hat{S}(k)|^2$ with respect to each of the subtraction factors α_l as:

$$\frac{\partial |\hat{S}(k)|^2}{\partial \alpha_l} = \begin{cases} \frac{-|\hat{D}(k)|^2}{B(k)} & f_L^l \leq k < f_U^l \\ 0 & \text{otherwise.} \end{cases} \tag{5.12}$$

In implementing Eqs. (5.11)-(5.12) the authors removed the spectral flooring condition $\beta|Y^i(k)|^2$ by setting $\beta = 0$; this was due to their belief that the subtraction factor α is the sole important parameter in spectral subtraction for ASR. It should be noted that this formulation differs from all implementations of spectral subtraction since the original work by Boll [14] which remove the hard spectral floor imposed by half-wave rectification. This implementation was confirmed with the authors, however we were unable to replicate their results as attempts to implement their approach have failed to converge. Consequently, we have not been able to perform an experimental comparison between their technique and the proposed MFNS approach. Please refer to Section 5.3.2 for a comparison of computational complexity.

The Jacobian element for the c^{th} element of the observed feature vector $\boldsymbol{o}$ in frame i w.r.t. the l^{th} subtraction factor is determined as:

$$\frac{\partial o_c^i}{\partial \alpha_l} = -\sum_{l=0}^{L-1} \frac{\Phi_{cl}}{M_l^i} \sum_{k=0}^{N/2} v_l(k) \frac{\partial |\hat{S}^i(k)|^2}{\partial \alpha_l} \tag{5.13}$$

where $v_l(k)$ is the coefficient of the l^{th} Mel-filterbank for the k^{th} frequency component, N is the length of the DFT, M_l is the energy of the l^{th} Mel-filterbank, and Φ is the $C \times L$ DCT matrix.

It should also be noted that BabaAli *et al.* apply no constraints on the estimated parameters, therefore it is possible for the subtraction parameters to become negative in situations where the test data is less noisy than the training

data. Whilst this may be suitable for that particular scenario, it may cause problems if the next speech recording comes from a noisier environment. For example, in an automotive application where training data may come from a range of noise conditions, optimisation performed on speech from the idle noise condition would be the most likely to result in negative enhancement parameters. This outcome is potentially problematic if the next speech recording comes from a significantly more noisy environment such as driving at 90-100 km/h with the HVAC on. The effect on ASR performance of applying constraints to the optimised parameters will be investigated in Section 5.5.

5.3.2 Mel-Filterbank Noise Subtraction

In this research, Mel-filterbank noise subtraction is considered for application in the LIMA framework. In this case, the number of enhancement parameters in the set $\boldsymbol{\xi}$ is the same as that used in the MBSS approach described in Section 5.3.1, however the subtraction is performed on Mel-filterbank energies rather than power spectra. For this application, the form of MFNS presented in Section 3.3.1 is used:

$$E_Y^i(l) = \int_{f_L^l}^{f_U^l} |Y^i(k)|dk$$

$$E_{\hat{D}}^i(l) = \int_{f_L^l}^{f_U^l} |\hat{D}^i(k)|dk$$

$$\hat{E}_S^i(l) = \begin{cases} E_Y^i(l) - \alpha_l E_{\hat{D}}^i(l) & E_Y^i(l) - \alpha_l E_{\hat{D}}^i(l) > \beta E_Y^i(l) \\ \beta E_Y^i(l) & \text{otherwise} \end{cases} \tag{5.14}$$

where $E_Y^i(l)$, $E_{\hat{D}}^i(l)$ and $\hat{E}_S^i(l)$ are the energies of the l^{th} Mel-filterbank of the noisy speech, noise estimate and the estimate of the clean speech filterbank energy respectively, and β is the spectral flooring factor. It should be noted that the clean speech filterbank energy estimate $\hat{E}_S^i(l)$ is equivalent to M_l in Eq. (5.13). In the case of MFNS, the Jacobian elements are:

$$\frac{\partial o_c^i}{\partial \alpha_l} = \sum_{l=0}^{L-1} \frac{\Phi_{cl}}{E_{\hat{S}}^i(l)} \frac{\partial \hat{E}_S^i(l)}{\partial \alpha_l} \tag{5.15}$$

The expression for $\frac{\partial \hat{E}_S^i(l)}{\partial \alpha_l}$ is fully derived in Appendix A. Substituting in the

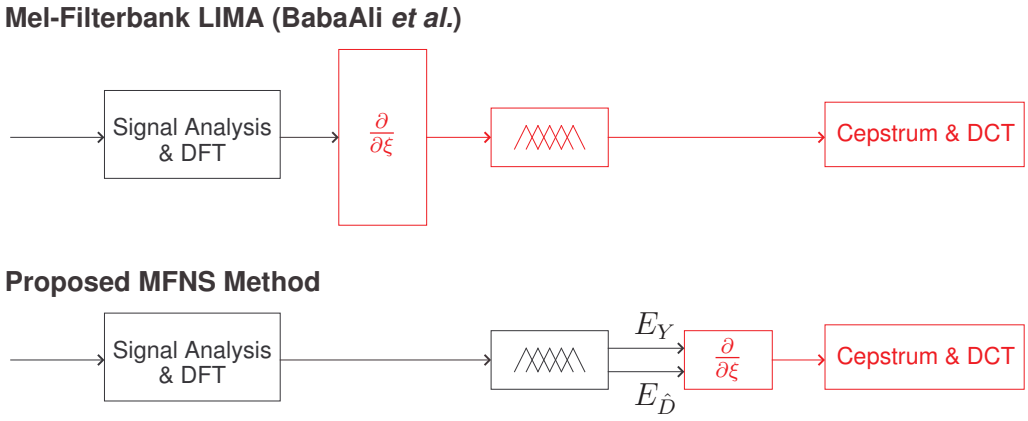


Figure 5.2: Comparison of the computational requirements for each iteration of the method by BabaAli *et al.* [10] and the proposed MFNS-based method.

result to Eq. (5.15), the full expression for the Jacobian elements is expressed as:

$$\frac{\partial o_c^i}{\partial \alpha_l} = -\frac{1}{2}\sum_{l=0}^{L-1} \frac{\Phi_{cl} E_{\hat{D}}^i(l)}{\hat{E}_S^i(l)} \left(1 + \text{sign}\left\{E_Y^i(l)(1-\beta) - \alpha_l E_{\hat{D}}^i(l)\right\}\right) \tag{5.16}$$

where $\text{sign}\{X\} = X/|X|$.

The major differences between the proposed method and that used by BabaAli *et al.* are shown in Fig. 5.2. The two approaches have a number of blocks in common – signal analysis and DFT, application of the Mel-filterbank, and the cepstral transform and DCT. Common processes used on an iteration-by-iteration basis will not be considered in the following computational complexity comparison since they contribute identically to both techniques. The main difference between the two approaches is the domain in which the partial derivatives $\frac{\partial}{\partial \xi}$ are calculated (i.e. their domain of operation).

In optimisation problems, the aspect which contributes most to the overall computational complexity is the number of iterations required for convergence. Although we have been unable to quantify this experimentally, there is no reason to believe that one of these methods will converge faster than the other (N.B. this assumption will be revisited). Making this assumption, it is therefore necessary to compare the complexity of these two methods in terms of the processing required during each iteration.

The iteration-by-iteration processing for each approach is highlighted in red in Fig. 5.2. This figure shows that the proposed method avoids the need to apply

the Mel-filterbank to the derivative $\frac{\partial}{\partial \xi}$, a process with O(K) complexity where $K = N/2 + 1$.

The other point of deviation between the two methods is the derivative calculation. Ignoring all information which can be calculated once and stored, both derivative calculations involve a single multiplication for each element of the relevant spectrum. For the proposed MFNS-based method, the complexity of the derivative calculation is O(P) (where P is the number of filterbanks/enhancement parameters), whereas the complexity of the original method – since it operates in the frequency domain – is O(K).

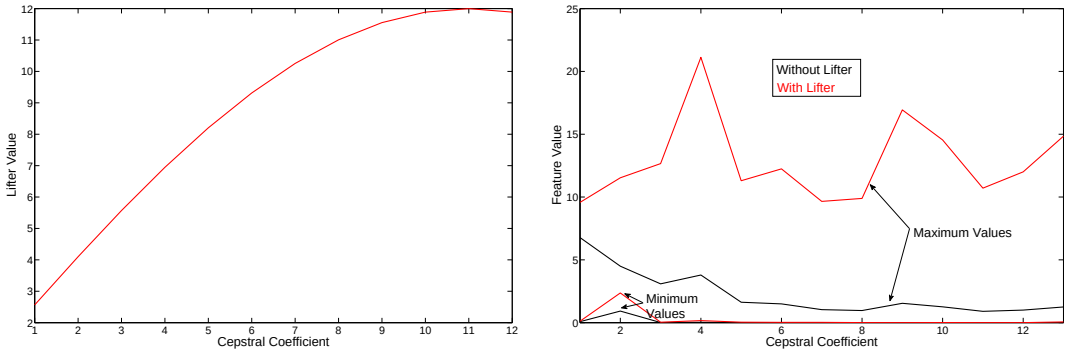
The relevant computational complexity of the proposed algorithm is therefore O(P) compared to the original method which is O(K)+O(K). Since $P \propto \log(K)$, it can be seen that the original method is exponentially more expensive than the proposed MFNS-based method on an iteration-by-iteration basis. Therefore, if the original method did converge faster, it would need to do so exponentially in order to counteract the extra complexity in each iteration.

Another consequence of the logarithmic relationship between P and K is the effect of scaling on these algorithms. For example, if the signal sampling rate was doubled, and the analysis window size maintained, the resulting signal spectrum would become $2K$, therefore doubling the computational complexity of the method proposed by BabaAli *et al.* [10]. For the proposed method however, the increase in complexity is considerably less since the increase in Mel-filterbanks is proportional to $\log(2)$.

5.4 Extensions to Existing LIMA Research

5.4.1 Optimisation on MFCC Features

In the initial evaluation of likelihood-maximisation, Seltzer *et al.* [127] optimised the filter-and-sum beamformer filter weights using features from the log-Mel spectral domain rather than the cepstral domain despite using MFCC features for ASR. Log-Mel spectral coefficients were chosen since the magnitudes of all coefficients are of a similar dynamic range whereas the magnitude of cepstral features



(a) (b)

Figure 5.3: (a) Cepstral lifter and (b) output of the cepstral lifter on a single frame of speech.

decrease as the order increases. In gradient-descent optimisation, if dynamic ranges vary, components with larger magnitudes dominate the resulting objective function [127] which makes optimising cepstral coefficients problematic. Previous research has failed to propose methods for normalising the cepstral coefficient dynamic range so optimisation can take place effectively on MFCC features, yet some studies have still applied optimisation to these features [10].

The cepstral lifter – which was introduced in Section 2.2.2 – is designed to increase the dynamic range of the higher-order coefficients. The cepstral lifter Ψ used throughout the experiments in this dissertation has the form:

$$o_c'^i = \Psi_c o_c^i = \left(1 + \frac{22}{2}\sin\left(\frac{\pi c}{22}\right)\right) o_c^i \tag{5.17}$$

and is shown graphically in Fig. 5.3(a). It can be seen that as the order of the cepstral coefficient is increased, the lifter response also increases. The effect on the speech signal is shown in the example in Fig. 5.3(b) which shows the maximum values of each cepstral coefficient are better matched across the full range of coefficients than was the case without liftering. This reduction in variation across the cepstral coefficients is the property of the lifter which has the potential to reduce the effect of components dominating the objective function used in gradient-based optimisation.

To integrate cepstral liftering into the gradient function, each partial derivative in Eq. (5.15) is multiplied by the lifter coefficients Ψ:

$$\frac{\partial o_c^i}{\partial \alpha_l} = \Psi_c \sum_{l=0}^{L-1} \frac{\Phi_{cl}}{E_{\hat{S}}^i(l)} \frac{\partial \hat{E}_S^i(l)}{\partial \alpha_l} \tag{5.18}$$

5.4.2 Evaluation of Spectral Subtractive LIMA

Despite the comprehensive evaluation performed in [10], there are still some gaps in the knowledge acquired from their experimentation. Below are proposed evaluation procedures to fill these knowledge gaps.

1. The effect of the spectral flooring factor β on the optimisation has yet to be determined. To examine these effects, it is proposed to perform optimisation on β alone (i.e. keeping α constant), and also in combination with α.

2. No comparison between the use of a single subtraction factor for all filterbanks and a separate factor for each filterbank has been made. If the single subtraction factor produces similar performance, it will considerably reduce the processing time required to perform the optimisation.

3. All speech recognition experiments performed were either isolated word or phone recognition tasks; no reference was made to continuous speech recognition. This research will evaluate the performance of spectral-subtractive LIMA on continuous speech recognition tasks in the AVICAR database and the AEICS corpus (navigation address task).

4. LIMA-based MBSS was shown to provide further improvements to an ASR system using noise-adapted speech models, however the adapted models didn't consider the performance under other mismatches such as differing dialects or stressed speech. Using the Australian English data in the AEICS corpus, this research can evaluate the word accuracy performance of the LIMA framework when a second mismatch is present in the test data.[1]

[1] A study on the performance under Lombard speech was initiated during an internship at the University of Texas at Dallas, however results of this study were incomplete at the time of submission of this dissertation.

5.5 Experiments & Discussion

A series of experiments were conducted primarily to address the limitations of previous evaluations (Section 5.4.2). In particular, these experiments aim to provide:

- A comparison of constrained and unconstrained optimisation of the over-subtraction factor α;
- An evaluation of the effectiveness of the spectral flooring factor β as a parameter for optimisation;
- An analysis of the performance of difference combinations of enhancement parameters used in optimisation;
- An evaluation of the use of cepstral liftering to reduce the variation in dynamic range of the cepstral features and therefore provide even weightings for optimising each element;
- An evaluation of LIMA frameworks on test data which consists of two mismatches – one due to noise and one due to speech production variabilities (e.g. the Australian English dialect using American English acoustic models); [2]
- An indication of the general importance of model adaptation and the resulting choice of enhancement parameters.

The experiments reported in this section utilise the baseline speech recogniser and evaluation protocols described in Section 4.3. An alteration to the protocol for the AVICAR database was required in order to ensure utterances were available for each speaker in each in-car noise condition. This led to a subset of 38 speakers for these experiments.

For experiments reported here which use only CMS (i.e. no cepstral lifter), a separate acoustic model was trained using the same Wall Street Journal 1

[2]Dialect – as opposed to accent – is used throughout this dissertation to represent the difference between American English and Australian English. The term accent can be used to refer to the realisation of English by non-native speakers; as a result, dialect is chosen since only data from native English speakers are used in these evaluations.

data. Data for model adaptation was made available according to the evaluation protocols defined in Section 4.3.

Being an optimisation problem, the number of iterations used in the joint optimisation process is an important part of the experimental set-up. In this chapter, only one joint optimisation iteration was performed (i.e. only one decode pass). This level of optimisation was based on the results presented in Chapter 6 which show that care must be taken to avoid over-optimising the enhancement parameters when LIMA-based speech enhancement is employed on in-car speech data. The number of gradient-descent iterations differ depending on the acoustic model used and the set of enhancement parameters $\boldsymbol{\xi}$ being optimised – they are chosen to provide the best improved performance across all noise conditions. These values will be noted on an experiment-by-experiment basis.

The LIMA framework applied in this chapter is speaker- and noise-dependent calibration whereby the first utterance for each speaker in each different noise condition is used to optimise the enhancement parameters. This particular framework was used since it best matches the experimental procedures used in previous research [10]. More information on this implementation of the LIMA framework can be found in Chapter 6. Limitations on the number gradient-descent iterations have been applied throughout this chapter given the ease of over-optimisation and its detrimental effects on ASR performance – effects which are highlighted in Chapter 6.

The ASR performance of various LIMA-based MFNS configurations is compared with both a baseline ASR system with no speech enhancement, and a version of MFNS which uses static parameters as per the experimentation in Chapter 4. For the experiments without cepstral liftering, $\beta = 0.4$, whereas $\beta = 0.45$ for the cases with cepstral liftering. The static oversubtraction factor $\alpha_l = 1$ was used for all experiments.[3] These parameter values were also used as the initial values for the optimisation process.

Computational factors were obtained by running a number of the configurations on a set of 35 utterances. In all cases, the code was optimised to ensure

[3]The notation α_l refers to a separate parameter for each filterbank as opposed to a global parameter denoted as α. The same notation is used for β in Sections 5.5.3-5.5.4.

minimal processing was required during each iteration. The results quoted in this section were obtained by comparing CPU times for each experiment and normalising to a reference experiment (i.e. the reference experiment has a computational factor equal to 1). Computational analysis was performed only on the AVICAR dataset because the length of all utterances are approximately the same, whereas utterances in the AEICS corpus consist of both short commands and long navigation addresses which makes comparison of computation times misleading. Computation factors for the baseline systems are not included as they do not provide a fair comparison with the optimisation framework.

In this chapter, only results for the navigation address task of the AEICS corpus are quoted since this thesis is most interested in continuous speech recognition. Appendix B contains the results for the commands task.

5.5.1 Constrained Optimisation

In the work by BabaAli *et al.* [10], they suggested removing constraints on the oversubtraction factors α in order to reinforce some frequencies; this approach may be particularly useful if the general level of noise in the acoustic model is greater than the noise in the test data. This did, however, raise concern over the generality of the optimised parameters (as discussed in Section 5.3.1). In order to determine if there is potential to lose generality by relaxing the parameter constraints, the ASR performance with and without parameter constraints was compared. For both cases, it was found that the best ASR performance across all noise conditions occurred using 4 gradient-descent iterations. These results are shown in Tables 5.1-5.2.

Given that the best ASR performance for both constrained and unconstrained optimisation was obtained at the same number of iterations, the results in these tables can be used to directly compare both the computation time and ASR performance of these approaches. From a computational perspective it can be seen that the application of constraints incurs only 2.6% processing overhead. This small increase in processing is attributed to the fact that for all optimisation iterations prior to convergence, constraints were applied in less than 0.5% of cases.

In terms of speech recognition accuracy, it can be seen for both datasets that

Table 5.1: Comparison of constrained and unconstrained optimisation on the AVICAR phone numbers task.

Experiment	GD Iter.	ASR Word Accuracy (%)					Computation Factor
		IDL	35U	35D	55U	55D	
Baseline	NA	70.4	48.8	36.2	41.8	23.5	NA
Static MFNS	NA	73.8	48.3	37.9	44.8	27.2	NA
Constrained	4	74.1	50.2	38.5	45.3	27.2	1.0
Unconstrained	4	74.0	50.8	38.1	45.4	27.0	0.974

Table 5.2: Comparison of constrained and unconstrained optimisation on the AEICS navigation address task.

Experiment	GD. Iter	ASR Word Accuracy (%)						
		C0	C6	C1	C2	C3	C4	C5
Baseline	NA	83.4	37.8	67.5	30.6	47.8	48.0	26.5
Static MFNS	NA	85.4	48.3	74.4	43.0	56.9	55.4	38.5
Constrained	4	85.2	45.4	73.6	40.6	56.6	54.3	37.1
Unconstrained	4	84.9	46.0	74.0	41.0	56.7	54.1	36.6

neither approach produces global improvements over the other, with word accuracies varying by at most 0.6%, but generally by less than 0.3%. This similarity in ASR performance is attributed to the fact that constraints are only applied to approximately 20% of the calibration utterances, resulting in the majority of the optimised parameters being common to both constrained and unconstrained optimisation.

The similarity in computation and ASR performance between constrained and unconstrained optimisation shows that relaxing the constraints on α_l does not result in the final parameter values losing generality, which justifies the unconstrained optimisation approach chosen by BabaAli *et al.* [10]. Despite this justification, it is chosen to utilise constrained optimisation throughout the remainder of this research, namely for the improvements in performance that are shown for the noisier conditions of the AVICAR database (i.e. the two windows down conditions) which are the worst performing conditions across both evaluation datasets.

Before moving on to assessing cepstral liftering, it is important to note the

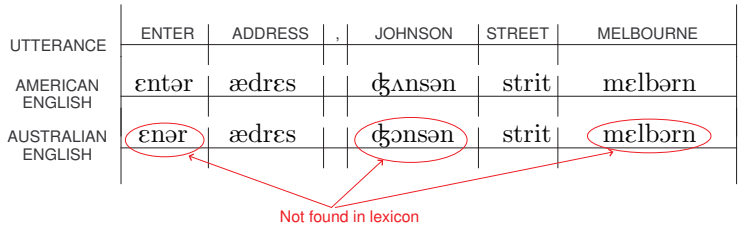
overall performance characteristics of LIMA-based MFNS on the two datasets. For the AVICAR database, there are small (but ever-present) improvements in ASR performance over the static speech enhancement case. The same can not be said of the AEICS corpus where the closest the proposed system comes to the static enhancement case is 0.2%, with the performance dropping by as much as 2.9% below baseline enhancement. This observation is attributed to the sensitivity of the LIMA approach to a second level of acoustic model mismatch.

As explained in Sections 5.2-5.3, LIMA-based speech enhancement is designed to reduce the background noise present in the test data given the acoustic model used for ASR. If the acoustic model is trained or adapted using noisy speech data, the LIMA approach will match the level of noise in the incoming speech recording to the acoustic model. As such, the LIMA approach only deals with the mismatch due to noise. In the case of the AEICS corpus however, there is a second mismatch which is due to dialectal differences between American (training data) and Australian English (test data). This dialectal difference was shown to be quite significant given the improvements in ASR performance in Chapter 4 when the American English acoustic model was adapted with a subset of the AEICS corpus prior to testing.

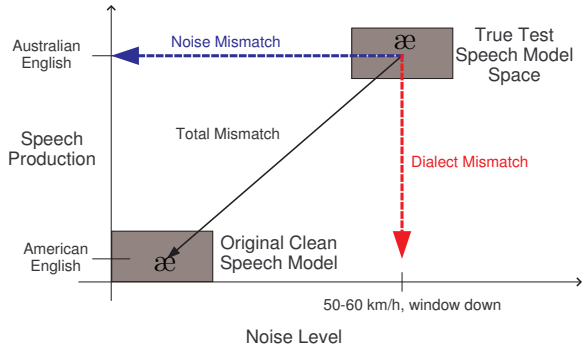
Differences between accents and dialects are typically divided into four main categories [141, 145]. Of particular interest to the scenario discussed here are differences between the lexical realisations of words, and differences in the acoustic realisations of the two dialects. Examples of these two differences are shown in Fig. 5.4.

Lexical differences occur due to the pronunciations of words in two dialects – phonemes can be deleted, substituted, or inserted. In the example in Fig. 5.4(a), the phoneme /t/ is deleted in the Australian English realisation of “ENTER”, whilst both “JOHNSON” and “MELBOURNE” contain phoneme substitutions. Lexical differences are particularly problematic for the forced alignment process that occurs in the LIMA framework.

Forced alignment converts the word sequence (which is known *a priori*) into its phone-level transcription; this lower-level transcription is used to match each observation vector to a HMM state. Thus, if the lexicon fails to provide a true



(a)



(b)

Figure 5.4: Two examples of multiple levels of acoustic mismatch due to differences in the (a) lexical realisations, and (b) acoustic realisations of two dialects.

realisation of a particular word in the Australian dialect, the sequence of HMM states chosen during forced alignment may be incorrect. In the case of the deletion seen in Fig. 5.4(a), the phoneme /t/ will be forced to be a part of the state sequence, even if the speaker used the alternative pronunciation. These states will therefore be unreliable and are likely to throw off the values of the optimised enhancement parameters since the wrong models are being used in the optimisation. In this dissertation, some effort was made to convert the American English lexicon to an Australian English version, so the effect of lexical differences should be minimal, although some inaccuracies will still exist as a number of words have multiple pronunciations.

Differences in the acoustic realisation of phonemes are therefore regarded as the major factor in the performance characteristic observed in Table 5.2. In comparison to both British and American English, Australian vowels exhibit a number of differences in the formant space [141]. In particular, some Australian

vowels exhibit a rise in formant frequency, some have more open realisations, and some vowels are even closer together than in American English. This means, for words pronounced with the same phonemes, the observed acoustic events will be considerably mismatched, as can be seen by the example for phoneme /æ/ in Fig. 5.4(b) which shows both noise and dialect mismatches. In this instance, any phoneme that is realised acoustically different to the American English acoustic model will be unreliable, which could be a significant number of the force-aligned states.

Despite removing the noise mismatch, the LIMA framework appears susceptible to the second layer of mismatch present in the AEICS corpus. This could be counteracted by adapting the original acoustic models to incorporate the Australian English dialect; this approach is evaluated in Section 5.5.4. The evaluation of cepstral liftering and parameter combinations in the next two sections utilise the original clean speech models; therefore only the AVICAR database is used to avoid the problems observed regarding secondary acoustic mismatches.

5.5.2 Cepstral Liftering

In Section 5.4 cepstral liftering was introduced to the LIMA framework in order to make the optimisation more effective on Mel-frequency cepstral coefficients which previously suffered from the effects of varying dynamic ranges across the coefficients. The ASR performance using LIMA-based MFNS with CMS and cepstral liftering on the AVICAR database is compared with that of a framework using only CMS in Table 5.3. Like the CMS-only system evaluated in the previous section, the liftered version also performed best using constrained optimisation with 4 gradient-descent iterations.

Despite incurring very minor increases in computation (0.1%), cepstral liftering appears to provide no advantages over the CMS-only system when the LIMA framework is applied. Analysing the baseline systems (with and without enhancement) it can be seen that cepstral liftering fails to provide a solution which improves ASR performance in all noise conditions. This tends to suggest that the cepstral lifter employed in this research is not suitable for improving

Table 5.3: Comparative performance evaluation of LIMA framework on the AVICAR phone numbers task with and without cepstral liftering.

Experiment	GD. Iter	ASR Word Accuracy (%)					Computational Factor
		IDL	35U	35D	55U	55D	
Baseline	NA	70.4	48.8	36.2	41.8	23.5	NA
Static MFNS	NA	73.8	48.3	37.9	44.8	27.2	NA
CMS Only	4	74.1	50.2	38.5	45.3	27.2	1.0
Baseline	NA	71.5	47.8	35.6	40.2	24.0	NA
Static MFNS	NA	74.5	49.2	36.6	44.0	28.1	NA
CMS + Liftering	4	74.2	49.4	37.5	43.9	27.6	1.001

the performance of in-car speech recognition, in spite of the dynamic range normalisation that occurs as shown in Fig. 5.3(a). If this lifter *is* inappropriate for this type of test data, the comparison of the two LIMA-based enhancement systems in this section will be unreliable. Therefore, future research is required to determine an appropriate lifter prior to evaluating the effect of liftering in the likelihood-maximisation approach.

5.5.3 Parameter Combinations

Previous work in LIMA-based spectral subtraction failed to quantify the effects of optimising the oversubtraction factors α as well as the spectral flooring factor β. In this section, various combinations of parameters have been evaluated on the AVICAR database. These combinations include single parameters for all Mel-filterbanks (α, β), parameters for each Mel-filterbank (α_l, β_l), as well as combinations of α_l and β.

Prior to performing ASR experiments, the number of iterations required for convergence of each set of parameters was determined. The mean and median number of iterations for each set of parameters can be found in Table 5.4. In general, the number of iterations required for convergence increases as the number of parameters increase, however an exception to this rule are the combined parameter sets which converge faster than using a single oversubtraction factor. The reason behind this may be due to the incorporation of β which converges

Table 5.4: Comparison of number of iterations for convergence for each of the parameter combinations.

	$1 \times \alpha$	$1 \times \beta$	$26 \times \alpha$	$26 \times \beta$	$[26 \times \alpha] + [1 \times \beta]$	$[26 \times \alpha] + [26 \times \beta]$
Mean Iterations	12.5	5.7	17.5	9.1	11.0	11.7
Median Iterations	10	5	17	8	11	11

in half the number of iterations required for the same number of oversubtraction factors. If β is converging quickly in the cases where both parameters are optimised, the oversubtraction factors may be forced to search for local maxima in the vicinity of β. Given the fact that over-optimisation is a serious issue for this framework (as shown in Chapter 6), convergence caused by local maxima should be avoided since less iterations than required for convergence will be best suited to ASR.

The reason why the two parameters converge at different rates is likely due to differences in the range of values each parameter can take which is dictated by the constraints. Constraints were always applied to ensure $0 < \beta_l \leq 1$, however the only constraint on α_l was that it be positive. The constraints applied to β_l potentially makes the initial guess ($\beta_l = 0.4$) much closer to the final value than that for α_l, resulting in less iterations to achieve convergence.

Given the concerns surrounding over-optimisation, the number of gradient-descent iterations used in the ASR experiments were determined using preliminary experiments which are not reported here. In general, the lower the number of optimised parameters, the lower the number of iterations. The exception to this rule is the combination of all oversubtraction factors with a single spectral flooring factor – a low number of iterations was used to account for the fast convergence rate of the single spectral flooring factor. The number of iterations used for each set of parameters are shown along with the recognition results and normalised processing times in Table 5.5.

Despite requiring less than 25% of the processing for optimising parameters for each filterbank, the use of a single oversubtraction factor or a single spectral flooring factor fails to produce word accuracy improvements over the static

Table 5.5: Performance evaluation of LIMA framework on the AVICAR phone numbers task for different parameter sets.

Experiment	GD. Iter	ASR Word Accuracy (%)					Comp.
		IDL	35U	35D	55U	55D	Factor
Baseline	NA	70.4	48.8	36.2	41.8	23.5	NA
Static MFNS	NA	73.8	48.3	37.9	44.8	27.2	NA
$1 \times \alpha$	1	73.7	49.4	38.2	44.4	27.1	0.237
$26 \times \alpha$	4	74.1	50.2	38.5	45.3	27.2	1.0
$1 \times \beta$	1	73.3	50.6	38.8	45.3	26.5	0.238
$26 \times \beta$	4	74.1	50.1	37.5	44.3	26.3	0.954
$[26 \times \alpha] + [1 \times \beta]$	1	73.4	49.2	37.8	44.4	26.4	0.551
$[26 \times \alpha] + [26 \times \beta]$	5	74.3	51.7	39.3	45.9	26.9	2.160

enhancement system across the full range of noise conditions. Whilst some conditions show consistent improvements with single parameters (e.g. 35 mph with windows up), the idle condition performance decreases in both cases, and no performance improvements can be seen in the noisiest condition (55D). Since the enhancement parameters are applied across all Mel filterbanks, they fail to provide enough resolution to allow for sufficient attenuation of the energies of the lower filterbanks (which are most affected by noise) whilst preserving the energies of the higher filterbanks which exhibit mostly speech.

By using a parameter for each of the Mel-filterbanks, more consistent performance across the range of noise conditions can be observed, particularly when optimising α. In all conditions, the optimised oversubtraction factors improve (or at least maintain) ASR performance, with the largest relative improvement being 3.7% for the 35U noise condition. For the spectral flooring factors, similar improvements in word accuracy can be observed in the two least noisy conditions (IDL and 35U), but they fail to improve performance in the noisier conditions.

Considering these two results, it appears that optimising the oversubtraction factor(s) is more important than optimising the spectral flooring factor(s). This observation confirms the thoughts of BabaAli *et al.* who removed the spectral flooring factor from their multi-band spectral subtraction implementation as they believed it to be least influential on overall ASR performance [10]. The

results here however show that some improvement in performance is possible by optimising this parameter, and therefore further improvements may be possible when combining both sets of parameters.

Two combinations of parameters are shown in the bottom two rows of Table 5.5. It can be seen that adding a single spectral floor factor to the full set of oversubtraction factors degrades speech recognition accuracy; although not documented here, this decrease was more prevalent as the number of gradient-descent iterations were increased. From this result and the previous experiment using a single β parameter, it is clear that using a global value is not ideal for use in MFNS-based LIMA.

The combination of 52 parameters shown at the bottom of Table 5.5 produces the best overall word accuracy performance, with four of the five noise conditions improving by at least 2% relative to the performance of the static enhancement system with a maximum improvement of 6.6% in the 35U condition. For the remaining condition (55D), a 0.3% improvement was obtained using only two iterations of gradient-descent optimisation (not shown in the table) which enabled it to outperform a system which only optimises the oversubtraction factors ($26 \times \alpha$). This experiment demonstrates that by increasing the degree of freedom in the enhancement algorithm (i.e. increasing the number of enhancement parameters which are optimised), the greater the likelihood can be maximised which ultimately results in better ASR performance. The major drawback of using more parameters is that more than double the processing time is required. For practical implementation, a trade-off between computational cost and improvement in ASR performance must be made, and it may therefore be decided to optimise just the oversubtraction factors.

5.5.4 Acoustic Model Adaptation

Previous studies on LIMA-based speech enhancement highlighted the ability for the framework to be used with both clean speech and noise-adapted acoustic models. To verify the proposed MFNS-based LIMA implementation behaves in the same way, the original clean speech acoustic models were adapted with in-car speech data as per the evaluation protocols outlined in Chapter 4. This was

Table 5.6: Performance evaluation of LIMA framework on the AVICAR phone numbers task with MAP adaptation.

Experiment	Iter.	ASR Word Accuracy (%)				
		IDL	35U	35D	55U	55D
Baseline	NA	82.5	76.0	68.3	74.7	58.3
Static MFNS	NA	80.0	71.7	64.5	73.0	58.4
$26 \times \alpha$	∞	81.2	75.5	67.3	74.5	59.9
$26 \times \beta$	∞	82.6	76.2	68.5	75.2	59.6
$[26 \times \alpha] + [26 \times \beta]$	∞	82.4	76.0	68.1	75.1	59.6

Table 5.7: Performance evaluation of LIMA framework on the AEICS navigation address task with MAP adaptation.

Experiment	GD. Iter	ASR Word Accuracy (%)						
		C0	C6	C1	C2	C3	C4	C5
Baseline	NA	90.5	69.4	86.2	71.4	81.7	82.4	69.3
Static MFNS	NA	89.1	77.0	88.2	80.6	86.2	85.7	79.1
$26 \times \alpha$	1	88.5	75.2	88.4	77.6	84.6	84.9	77.5
$26 \times \beta$	1	89.6	77.9	88.1	81.6	86.5	85.5	81.3
$[26 \times \alpha] + [26 \times \beta]$	1	89.0	73.8	87.7	77.7	84.2	84.8	76.1

done for both the AVICAR and AEICS datasets and the results can be found in Tables 5.6 and 5.7 respectively.

Before analysing the speech recognition results, it is important to note the change in the number of gradient-descent iterations (compared to Sections 5.5.1 and 5.5.3) used to generate these results. An analysis of the number of iterations required for convergence on the AVICAR database revealed that more iterations were required than when using clean speech models (e.g. median 20 iterations for $[26 \times \alpha]$).

Considering this effect more closely, frames which are aligned incorrectly (and therefore can be considered unreliable) are more prevalent when the clean speech acoustic models are used for forced alignment since the overall ASR performance of these models is lower compared to the MAP-adapted models (refer to the experiments in Chapter 4). During optimisation, these unreliable frames act as a source of noise which limits the amount of change that can be made to the overall

acoustic likelihood which in turn leads to faster convergence to the maximum likelihood. By using noise-adapted speech models, the number of unreliable frames (and therefore noise) is reduced, which reduces the limits on the overall maximised likelihood. In short, the more accurate the model alignment, the greater the maximum likelihood that can be obtained.

Despite the greater number of iterations required to achieve convergence, the two corpora exhibited considerably different optimal ASR behaviour w.r.t. the number of gradient descent iterations. The AEICS corpus (Table 5.7) experienced continual over-optimisation as the parameters converged, and as a result the best ASR performance was obtained with minimal gradient descent iterations. Limiting the number of gradient descent iterations however failed to produce a solution which was able to *consistently* outperform even the static enhancement system, regardless of the combination of parameters used.[4]This behaviour matches that of LIMA-based enhancement on the original acoustic model observed in Section 5.5.1.

This consistency in behaviour suggests that despite model adaptation introducing Australian dialect information into the original American English acoustic models[5], the amount of adaptation is insufficient to fully counteract the effects of this second layer of acoustic mismatch. Given the behaviour on both original and adapted acoustic models, it is inferred that the LIMA framework is highly sensitive to the data used for model training and that encountered during testing. This observation is an important outcome for the research community and practical implementations of LIMA-based systems, and therefore warrants considerable future research in order to overcome these issues; discussion on this is provided in Section 5.6.

For the AVICAR database, the best overall word accuracy was achieved (in

[4]By examining the results in Table 5.7, it can be seen that the system optimising only β parameters improves on the static MFNS in 5 out of the 7 noise conditions. Whilst this result demonstrates promise in the LIMA-based enhancement on the AEICS corpus, the inconsistency still suggests that the model adaptation employed here is not fully effective in counteracting the second layer of acoustic mismatch.

[5]In Chapter 4 – by examining the performance of the idle noise condition – it was observed that the adaptation process was capable of improving ASR performance by introducing Australian dialect information into the original American English acoustic models. This can be further verified by comparing the baseline results in Tables 5.2 and 5.7.

most noise conditions) when the system allowed for convergence of the parameters which is the expected behaviour of any optimisation problem. This was not true however for the very noisy conditions which produced maximum recognition accuracies in the first few iterations, and then gradually decreased with further iterations owing to over-optimisation. Only the results generated by convergence have been quoted here since they generally outperformed the baseline system with and without enhancement.

On closer examination of Table 5.6, the baseline enhancement system is identified to behave as it did in Chapter 4 where the conflict between speech enhancement and model adaptation caused the overall ASR accuracy to drop for the quieter noise conditions. It is important to note that in all cases however, the LIMA framework is able to recover at least some of these performance losses; this is true for any combination of optimised parameters.

For the case where only the spectral flooring factors β_l are optimised, the LIMA approach is able to provide at least minor improvements in word recognition accuracy for all noise conditions; relative improvements range from 0.6% for idle through to 3.1% for 55 mph with the windows down. These improvements can be attributed to the need for a different noise floor, since the noise level present in the new adapted models is greater than that of the original clean speech models. This suggests that optimising the spectral floor factor is more important than the oversubtraction factors when noisy adaptation data is made available to the acoustic model.

For both of the previous observations, the initial force-aligned state sequence was generated using the static enhancement case even though it produces lower word accuracies than the baseline system for the quieter noise conditions. As a result, the state alignment used for optimisation will be less reliable than that possible using the baseline enhancement system. This lower accuracy transcription is likely to be the reason why the LIMA system is only able to (at best) marginally improve on the baseline recognition system. There is potential to improve this performance by using the baseline transcription to optimise the enhancement parameters.

5.6 Research Directions

Over-fitting the enhancement parameters was found to be a significant problem surrounding the application of the LIMA framework. The core of this problem relates to the dependency on the data used for calibration. One perspective of this data-dependency is the amount of speech required to provide a good fit for *all* states in the acoustic model, and not just those present during optimisation. This issue was briefly considered by Seltzer *et al.* [127] who suggested that if the utterance length is kept constant, over-fitting is more likely to occur as the number of parameters are increased. It could also be hypothesised that as the length of the adaptation utterance is decreased (therefore less models used for optimisation), the more reliant the optimised parameters become on the states present. This problem is very similar to that of training speaker-independent acoustic models – it is required to have enough speakers in order to reduce the reliance on any one speaker.

Considering the evaluation datasets used in this thesis, the AVICAR phone numbers task can potentially hide this reliance since the ten digits in each utterance are likely to utilise at least half of the "active" model states (i.e. those states present given the task grammar and vocabulary) which makes the optimised parameters suitable for *most* successive utterances. However, if the frame alignments are highly reliable (e.g. when using noise-adapted acoustic models), there is potential to overfit the parameters to those model states observed in the calibration utterance, and this effect will be emphasised when digits are repeated.

Likewise for the AEICS corpus, if optimisation takes place on a command phrase, it may overfit the parameters for the small set of models present; these parameters may not be applicable for a navigation address which will be more phonetically balanced. This reliance on the model coverage provided by the adaptation data is likely to be the primary reason behind BabaAli *et al.* using phonetically balanced sentences for the large part of their evaluation [10].

It is therefore of practical importance to fully understand the dependency of the LIMA framework on the calibration data, particularly in scenarios (like that in the AEICS corpus) where the same speech system is used for a range of task

grammars and vocabulary sizes. Using the AEICS corpus, it is intended to extend the current research to analyse the effect of different lengths of adaptation data on the overall performance of the LIMA framework.

Looking at the model coverage problem from a different perspective, perhaps the triphone acoustic models used throughout this research are not the best choice for the LIMA framework. Given the same length of available calibration data, the model coverage on a triphone acoustic model would be considerably less than that using a phone-based model; this may explain the smaller improvements in ASR performance demonstrated in this chapter (around 2% relative). Small model coverage will also favour over-fitting in the same way as described above for short adaptation utterances. Another side-effect of using triphone models is the potential to make state alignments less accurate due to increased confusability between models with the same base phone but different contexts. It is therefore seen as very important to assess the influence of the model unit used within this framework by extending the work contained in this dissertation to compare the use of triphone, phone and broad phone class models for both the forced alignment and final recognition stages.

Experiments using the AEICS corpus highlighted the sensitivity of the LIMA framework to secondary acoustic mismatches between training and testing – even performing model adaptation was unable to remove the second layer of acoustic match which was due to speech production rather than background noise. There are a number of approaches which could be taken to alleviate this sensitivity. For example, greater amounts of data could be utilised for dialect-adaptation of the acoustic model. In the evaluation protocol utilised throughout this dissertation, only limited data has been made available for model adaptation, which is unlikely to cover all states in the acoustic model adequately. In this instance, any unadapted states observed during calibration will act as a source of noise to the likelihood-maximisation algorithm, resulting in sub-optimal enhancement parameters. In order to provide sufficient amounts of data, a speech database such as the Australian National Database of Spoken Language (ANDOSL) [100] may be suitable.

Another approach to counteract this sensitivity would be to first analyse the

influence of particular broad phone classes on the optimisation outcome in order to establish whether particular phones are more problematic than others. This suggestion is based on the fact that different accents of English differ more in the realisation of vowels than other phonetic classes [141, 145]. Once the influence of each phone class has been established, it may be possible to apply a reward-punish scheme which emphasises or attenuates the influence of phones based on their likeness or differences between dialects.

For in-car speech applications, the speaker- and noise- dependent framework employed for the experimentation in this chapter and in previous research is highly impractical. In particular, this framework requires not only a calibration utterance for every speaker (which is not restrictive) but also in every conceivable noise condition. For example, if a calibration utterance is required for every combination of speed (for instance in 10 km/hr intervals), air-conditioning status or window position, the result would be a scenario in which the driver is continually asked for calibration utterances and not their desired set of commands. This problem could be reduced by clustering noise conditions, however considerable thought must be placed on how to best group these conditions to ensure the system covers the full range of noise conditions yet isn't over-generalised. This practical limitation of existing LIMA frameworks motivated the research of a dialogue-based solution which is presented in Chapter 6.

Despite deriving the theoretical computational advantage over the original frequency-domain method proposed by BabaAli *et al.* [10], this research was unable to compare the two approaches experimentally. This limitation was due to an inability to replicate the system used in [10], despite many attempts to communicate with the authors. In the future, the aim is to resolve the issues with the current implementation from whence it will be possible to truly verify the computational savings of the proposed LIMA-based spectral subtraction in Mel-filterbank domain and also compare the ASR performance of the two approaches.

In Section 5.4, cepstral liftering was proposed to correct the problem surrounding varying dynamic ranges in the cepstral coefficients which causes the optimisation process to favour elements with greater magnitudes. In evaluating this solution, liftering failed to provide consistent improvements in ASR word

accuracy, even on test data without speech enhancement applied. Despite this performance characteristic, it is still firmly believed that cepstral liftering is a computationally efficient solution to the problem, however further research is required to pinpoint whether the lifter implementation is inappropriate or whether the phenomenon seen in this chapter is confined to in-car noise environments and not applicable to other adverse environments.

Finally, very little attention was paid to characteristics of the optimisation technique and their possible effects on the rate of convergence in this chapter (and the next). Whilst this was not necessary for demonstrating the concepts of the LIMA approach, this lack of focus may have led to slower rates of convergence than possible. For example, it was noticed that β parameters converged faster than α parameters; this could be related to either the initial guess (i.e. the initial guess of β was more accurate than α) or a considerable difference in the range of possible values each could take. Whilst analysis of the effects of these two optimisation variables (among others) is not sufficiently significant to contribute to the body of scientific research in this field, it may be of importance when porting the LIMA approach to commercial applications in which fast convergence rates are essential.

5.7 Summary

In this chapter the generalised framework of likelihood-maximising speech enhancement for robust ASR was introduced. An analysis of previous LIMA implementations of single-channel spectral subtractive enhancement schemes was presented, and it was shown that this type of framework was suitable for use with a wide range of test data and acoustic models. It was highlighted that the only previous use of LIMA-based enhancement for spectral subtractive-type algorithms was computationally expensive, and therefore it was proposed to reduce this computational burden by using Mel-filterbank noise subtraction rather than frequency-domain spectral subtraction.

As well as the general formulation of MFNS-based LIMA, it was proposed to optimise the spectral flooring factor β as well as the oversubtraction factor α.

Cepstral liftering was also incorporated in order to reduce the effects of the varying dynamic ranges of cepstral coefficients which are present in the MFCC representation – in these scenarios, the cepstral coefficients with the greatest magnitude dominate the gradients calculated during the optimisation.

Evaluation of the proposed system began by analysing the effects of constraining the oversubtraction factors to positive values. It was found that the constraints were applied infrequently and this led to only minor increases in processing time with comparable performance to unconstrained optimisation.

Despite incurring minimal increases in computational times, cepstral liftering was shown to provide no improvements in speech recognition performance regardless of its use in baseline systems or the LIMA framework. It was hypothesised that cepstral lifter used in this dissertation is not effect for in-car speech data – this hypothesis will be tested in future studies.

An evaluation was conducted which compared various combinations of enhancement parameters for optimisation. It was seen that optimising the spectral flooring factor(s) causes faster convergence rates, but fails to provide any improvements in word recognition accuracy over optimising the oversubtraction factors. By optimising both the oversubtraction factor and spectral flooring factor for *each* Mel-filterbank, the overall word accuracy was increased from the static enhancement system by over 2% in most noise conditions and outperformed optimisation of only the oversubtraction factors. Despite the superior ASR performance, optimising both sets of parameters incurs considerable processing overhead compared to only optimising the oversubtraction factors.

The LIMA framework was finally applied to noise-adapted speech models on both the AVICAR and AEICS datasets. Results on the AEICS corpus showed that model adaptation was unable to fully remove the detrimental effects of a second level of acoustic mismatch (due to the Australian English dialect). This was consistent with the use of the original clean speech models in which the performance of the LIMA-based enhancement failed to match a system with static enhancement parameters. For both databases, it was observed that optimising the spectral flooring factor β was more important than the oversubtraction factors α; this was necessary in order to produce a noise floor which better matched noise

levels in the adapted acoustic models.

A number of future research directions were proposed to contribute to the existing body of scientific knowledge about LIMA-based systems and to also speed up convergence rates of practical LIMA-based systems. Of particular future interest is the need to determine a solution to the problem of model sensitivity when multiple layers of acoustic mismatch are present between training and testing. Proposed avenues of investigating this particular problem and its solutions include increasing the amounts of adaptation data, as well as a thorough analysis of the effects different phonemes have on the optimisation process.

Throughout the evaluations in this chapter, a one-time calibration framework was used for each speaker in each in-car noise condition. This particular framework is impractical for in-car applications due to the requirement for a large number of calibration utterances. To overcome this impracticality, Chapter 6 proposes a framework which couples LIMA-based speech enhancement with speech dialogue systems.

Chapter 6

LIMA Frameworks for In-Car Speech Recognition

6.1 Introduction

In the previous chapter, likelihood-maximisation applied to Mel-filterbank noise subtraction was proposed and evaluated by optimising enhancement parameters using a single utterance for each noise condition for each speaker. This particular application of LIMA was referred to as a Calibrated LIMA framework [127]. Given the number of noise sources (engine, wind, air-conditioning etc.) and within-source variations due to speed, traffic and external weather conditions, there is an almost endless number of noise characteristics possible inside a vehicle. Calibrated frameworks in the form used in Chapter 5 are therefore impractical for in-car applications as they require calibration utterances (and associated storage of optimised parameters) to cover every possible noise condition.

Other forms of LIMA frameworks have been proposed in the literature, however these also have shortfalls for in-car applications. The discussion of these frameworks (Section 6.2) leads to the formulation of a dialogue-based LIMA framework which is suitable for in-car environments in Section 6.3.

To evaluate the LIMA frameworks discussed in this chapter, the trade-off between ASR performance and processing requirements is examined by altering the number of iterations used in parameter optimisation (Section 6.4.1). Using

the results obtained from this investigation, each of the LIMA frameworks are evaluated, and recommendations for the use of LIMA-based MFNS for automotive speech recognition are made in Section 6.4.2.

In order to make the proposed dialogue-based framework better suited to deployment in automotive environments, the computation time of the optimisation process needs to be further reduced. Section 6.5 discusses research directions which could be pursued in order to make LIMA frameworks realisable for application in future generations of vehicles.

6.2 Review of Practical LIMA Frameworks

In vehicular environments, the goal is to create human-machine interfaces (HMI) to in-vehicle infotainment, navigation and command and control systems via voice. Speech-based interfaces allow drivers to keep their eyes on the road and hands on the steering wheel whilst accessing these services instead of, for example, looking away and manually changing radio stations, cabin temperature or entering navigation addresses. Therefore, speech-based interfaces are an important development in improving road safety.

In this application, the human-machine interface is a speech dialogue system (SDS). Figure 6.1 shows the general architecture of the speech dialogue system; it can be seen that speech recognition is only one component of a much larger architecture required for effective dialogue systems. Although the purpose of LIMA-based speech enhancement is to make ASR systems more robust to environmental noise, when analysing the strengths and weaknesses of its practical implementation it is also important to consider how the ASR component is incorporated in a SDS. Throughout the following sections, practical LIMA frameworks will be considered with reference to deployment in SDS rather than standalone ASR systems.

6.2.1 Calibration

The simplest and most common approach for optimising the enhancement parameters in a LIMA-based framework is to use an adaptation session in which a

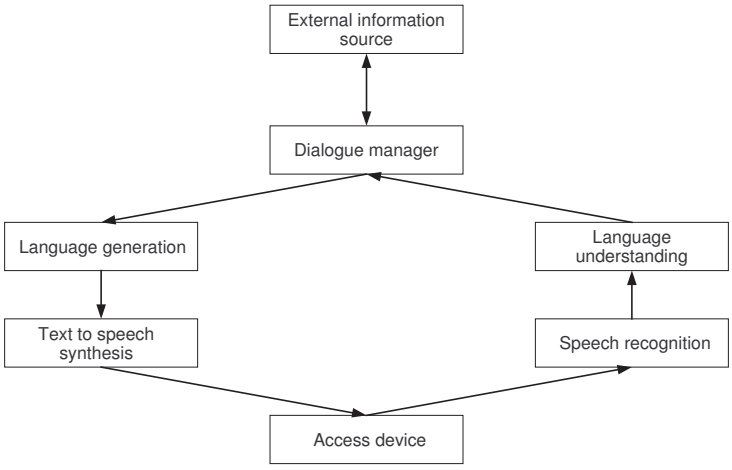


Figure 6.1: Architecture of a spoken dialogue system (taken from [97]).

single utterance with a known transcription w_C is used to determine the optimal set of enhancement parameters. Following the adaptation session, the optimised enhancement parameters are kept constant for all other utterances. In this approach, the user is aware of the adaptation process as they will be requested to say a particular phrase which is not likely to be part of the desired dialogue transaction. The "calibration" approach was used in previous studies [10, 127] and was used for all experiments in Chapter 5 where an adaptation utterance was used for *each* speaker in *each* noise condition. Each of these studies have shown that calibration-type frameworks produce satisfactory improvements in word recognition accuracy.

Whilst the use of a known word transcription in the calibration framework ensures that optimisation takes place on a state sequence which is correct (excluding alignment errors), this type of framework inherently assumes that the background noise conditions do not change between the calibration and testing sessions. This is a major challenge for in-car speech recognition since vehicular environments are subjected to continually changing noise levels and conditions – this approach would require a calibration utterance every time noise conditions change significantly from the previous optimisation. Two simple solutions to this problem are:

1. The optimised enhancement parameters could be stored for each common noise condition, however this still requires an initial calibration utterance for

each of these conditions. Since there is a wide range of noise conditions, the user would be continually asked to repeat the adaptation utterance in order to obtain the optimal set of parameters. This operation is an unnecessary annoyance for the driver, and is likely to lead drivers to become frustrated with the SDS; such emotions could lead to further repercussions on ASR and driving performance.[1]

2. Enhancement parameter calibration could be performed once only for each driving session; for example, a common startup utterance such as "Start dialogue" could be used for adaptation. Whilst this removes the need for regular adaptation sessions (and reduces user awareness of the adaptation process), it introduces the risk of inferior recognition in noise conditions significantly different from those present during calibration.

The results in Chapter 5 led to the hypothesis that small coverage of the acoustic model space leads to ASR performance which (at best) only marginally improves on an enhancement technique with fixed parameters. Since the calibration framework is reliant on the words in the adaptation utterance, it is therefore necessary for this utterance to be phonetically balanced and sufficiently long enough to provide as much model coverage as possible in order to generalise the optimised enhancement parameters. This is in conflict with the majority of SDS which promote simpler linguistic structures than human conversation and are therefore unlikely to be phonetically balanced. Thus, a separate utterance unrelated to the dialogue transaction is required which is likely to be seen by the user as an inconvenience, and is therefore another reason why calibration frameworks are impractical for use in this particular application.

Despite the impracticalities described in this section, calibrated frameworks are the most simple LIMA framework to implement. In Section 6.4, calibration is performed on a speaker-by-speaker basis, noise-by-noise basis and also a combination of the two (i.e. as per the experimental procedure in Chapter 5). Calibrating on a speaker-by-speaker basis cuts down the number of adaptation sessions per

[1] These phenomena were studied during an internship at the University of Texas at Dallas and results are presented in [65].

speaker, however it does lead to mismatches in the noise conditions used for adaptation and operation. Performing calibration for each noise condition avoids the problems of noise mismatch, but fails to incorporate speaker variabilities in speech production in these environments. Combining speaker and noise calibration captures both speaker and noise variation and is therefore expected to provide the best ASR performance, but is the least practical approach.

6.2.2 Unsupervised

An unsupervised LIMA framework was also proposed in [127] whereby online optimisation takes place on an utterance-by-utterance basis using the hypothesised transcription w as opposed to the true transcription w_C (i.e. there is no *a priori* knowledge of the transcription). Whilst this approach removes the restriction of a calibration session and makes the adaptation process transparent to the user, it is highly reliant on the accuracy of the state sequence generated by Viterbi alignment since the word transcription is unknown. In other words, the framework is reliant on the effectiveness of the underlying acoustic models and speech recogniser.

Since the true transcription w_C is unknown, it is possible that states in the hypothesised transcription w are incorrect due to misrecognition and frame alignment errors (N.B. frame alignment errors will occur even when the transcription is known *a priori*, but should be limited to only a few). These inaccurate states are likely to lead to the resulting enhancement parameters being sub-optimal since optimisation is performed on the wrong state model. In turn, sub-optimal enhancement parameters could lead to further decreases in accuracy in the subsequent decoding stage. This effect is particularly likely when the number of incorrectly labeled frames is greater than the correctly labeled frames.

Although there is not a 1:1 relationship between word accuracy and state accuracy, an ASR system which performs poorly in terms of word accuracy will also generate highly inaccurate state alignments. In Chapter 4, it was shown that word accuracy performance was as low as 31% even after speech enhancement was applied. It is for this reason that the unsupervised LIMA framework is

not assessed in this dissertation as the overall performance of the speech recogniser is low (less than 50% average word accuracy across all noise conditions on the AVICAR database), which will make the hypothesised transcriptions – and therefore the optimised parameters – highly unreliable.

6.3 Dialogue-Based LIMA Framework for In-Car Applications

Having identified the problems with both the calibrated and unsupervised LIMA frameworks in Section 6.2, it is proposed to drive the optimisation process by exploiting the need for user confirmation in a dialogue system. A block diagram of the proposed framework within the dialogue exchange is shown Fig. 6.2. The proposed system mimics the calibrated and unsupervised frameworks by performing an initial decode using default enhancement parameter values in the feature extraction stage. Instead of immediately performing optimisation however, the hypothesised word sequence is first verified through the grounding process which is required in SDS in order to detect any misrecognition errors which need to be corrected prior to executing a desired action such as determining route navigation.

Since it is cumbersome for the dialogue manager to request confirmation from the user after each response, grounding often occurs once the dialogue system has gathered a number of pieces of information, for example the suburb, street name and number of a destination address. In the case where the user states the information is incorrect, the dialogue manager will attempt to recover from these errors by either asking for corrections to specific information, or restarting the dialogue transaction altogether. The former method is preferred in modern-day systems as it reduces the total dialogue transaction time. In this instance, the enhancement parameters are left unaltered.

When the user confirms the information to be correct, this affirmation is fed back to the dialogue manager for further processing (e.g. a call to an external information source such as the navigation system), but also triggers the optimisation of the enhancement parameters. In order to interface the optimisation process with the grounding procedure, it is required to store the user responses as

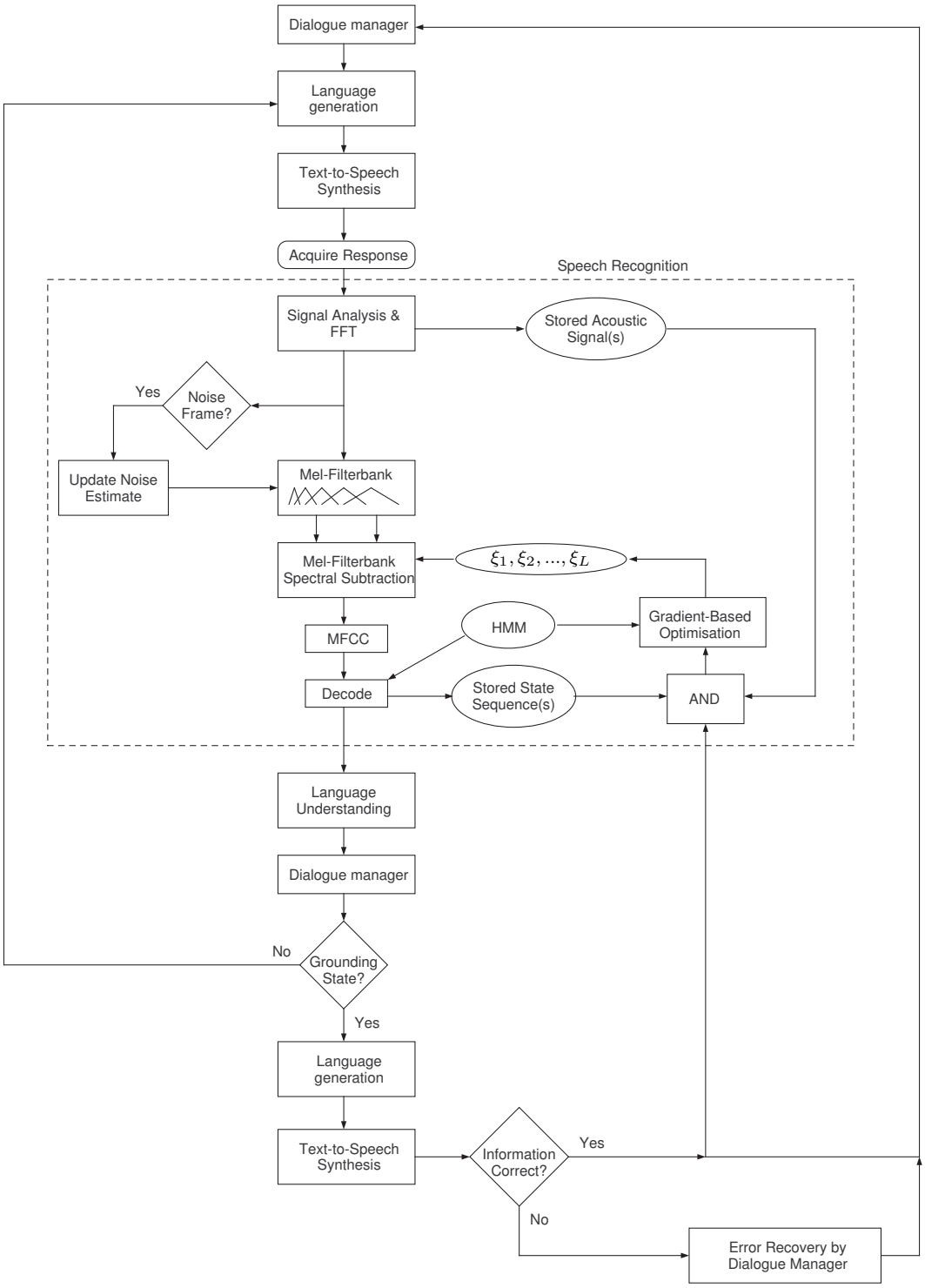


Figure 6.2: Proposed LIMA speech enhancement framework for in-car speech dialogue systems.

well as the hypothesised state sequences – this is shown in Fig. 6.2. On confirmation, this stored information is used in the optimisation process; if rejected, the stored state sequence is therefore unreliable, and so the memory can be cleared in preparation for responses in the error recovery stage.

The primary advantage of this proposed dialogue-based LIMA framework is that optimisation never takes place on inaccurate transcription hypotheses, which overcomes the limitation of the unsupervised framework described in Section 6.2.2. The other advantage of this framework is the ability to continually update the enhancement parameters as the noise conditions inside the vehicle change. This is achieved by maintaining the previous enhancement parameters until the next successful dialogue transaction, by which time the noise conditions may have changed. As a result, the dialogue-based system is able to overcome the need for matched noise conditions required for calibrated operation to be fully effective.

It should be noted that this framework is similar to the supervised LIMA framework proposed in [10] which was published at the same time as the publication which arose from this research [69]. The discussion in this dissertation is directed towards full practical application of this framework and is therefore more thorough in this regard than the description made by BabaAli *et al.* [10].

6.4 Experiments & Discussion

Following the findings of Chapter 5, experiments in this chapter are focused solely on the AVICAR database which showed consistent improvements in word accuracy when using clean speech acoustic models. This database enables analysis of LIMA frameworks based on speaker or noise calibration, as well as a combination of both. Therefore, in this evaluation the following LIMA frameworks were tested:

1. Calibrated LIMA framework using optimisation on a noise-by-noise basis;
2. Calibrated LIMA framework using optimisation on a speaker-by-speaker basis under a single, randomly chosen noise condition (i.e. mismatched conditions between calibration and testing);

3. Calibrated LIMA framework using optimisation for each speaker in each noise condition (i.e. matched conditions);

4. Proposed dialogue-based LIMA framework without initial calibration;

5. Proposed dialogue-based LIMA framework with a single calibration utterance in a randomly chosen noise condition; and

6. Proposed dialogue-based LIMA framework with a single calibration utterance in the idle noise condition.

As stated in Section 6.2.2, the unsupervised LIMA framework was not evaluated in this research due to the relatively low overall performance of the baseline speech recognition system.

Each calibrated LIMA framework used a single, randomly generated utterance treated as the adaptation session. For the noise-only calibration framework, a random utterance from a random speaker was chosen for each experimental fold in the evaluation protocol. For speaker-based calibration (applied in both calibrated and dialogue frameworks), a single utterance from a random noise condition was used for each speaker, with the remaining utterances ordered randomly to simulate continually changing noise conditions in the vehicle.

The proposed dialogue-based system was run using no prior calibration and optimisation occurred every time the decoder correctly recognised *all* 10 digits in the phone number. Utterances which occur prior to the first optimisation exhibit the same performance as the static enhancement system (i.e. $\alpha_l = 1$) and are therefore ignored in the final evaluation (N.B. this is the reason why the baseline results differ throughout this section).

In order to also simulate *a priori* knowledge relating to previously optimised enhancement parameters, the dialogue-based framework was also tested using an initial adaptation utterance which was either randomly chosen, or from the idle condition. The idle condition was chosen as this is a likely scenario for users to first communicate with the in-car SDS – for instance, for entering a destination address before setting off on the journey. Again, all utterances which occurred prior to the first subsequent optimisation (excluding calibration) were ignored in the evaluation.

Since LIMA is an optimisation problem, over-optimisation of the enhancement parameters to a specific noise condition, speaker or subset of the acoustic model is highly possible and should be avoided. Over-optimisation to a subset of acoustic models was one of the reasons suggested to explain the limited improvements in ASR accuracy for the triphone-based recognition system used in this research (refer to Chapter 5). The potential for over-optimisation suggests the number of optimisation iterations should be limited in order to maintain generality, but insufficient iterations may result in the LIMA framework operating less effectively than a standard enhancement system. Considering the need to also limit the total processing time (which is another important consideration for in-car ASR) also suggests a limit on the number of iterations.

To address this issue, two experiments were designed to determine a suitable balance between ASR performance and minimal processing delays using the noise-only calibration framework prior to comparing the performance of the six frameworks listed in this section. This particular framework was used as the belief was that noise conditions have a greater effect on the resulting enhancement parameters than individual speakers since this research uses speaker-independent acoustic models (refer to Section 4.3). In the first experiment, the number of gradient-descent iterations was varied whilst using a single joint optimisation iteration (i.e. one full recognition and parameter optimisation cycle). The second experiment varied the number of joint optimisation iterations whilst the gradient-descent iterations (determined from the former experiment) were kept constant. The combined outcomes of these experiments dictated the level of optimisation used for assessing all six frameworks.

Optimisation was performed only on the oversubtraction factors α_l using cepstral mean subtraction in the feature extraction; cepstral liftering was not used as it was shown in the previous chapter not to further improve word accuracy. It was also decided not to use the combined α_l and β_l optimisation which showed the best performance in Section 5.5.3 as this incurs considerable processing overheads with little improvements in word accuracy – these characteristics make the simplified system used throughout these experiments more suitable for the in-car application.

Table 6.1: ASR accuracies for increasing gradient-descent iterations used in parameter optimisation.

# Iter.	ASR Word Accuracy (%)				
	IDL	35U	35D	55U	55D
Baseline	70.4	48.8	36.2	41.8	23.5
Static MFNS	73.3	47.8	36.8	44.5	26.1
1	73.9	48.7	37.9	44.8	**26.4**
2	**74.2**	49.3	37.7	44.8	**26.4**
3	74.1	49.1	38.1	**45.1**	**26.4**
4	**74.2**	49.5	37.8	**45.1**	26.1
5	74.1	49.6	**38.2**	45.0	25.9
10	**74.2**	49.7	37.7	44.6	26.1
15	**74.2**	49.8	37.5	44.8	25.6
20	**74.2**	**49.9**	37.6	44.7	25.7
25	**74.2**	**49.9**	37.6	44.7	25.7

As per Chapter 5, the enhancement parameters were initialised to $\alpha_l = 1$ for all 26 Mel-filterbanks. The noise estimation procedure used in the MFNS technique in this chapter was a simple average of the initial silence consisting of N frames. Supporting results in Appendix B show that this method has similar word accuracy performance to time-recursive averaging with and without SAD based on a soft decision, but has the added advantage of reducing the processing delay of the technique which is important for in-car applications.

6.4.1 Optimisation Iterations

Gradient-Descent Iterations: The effect on ASR word accuracy as the number of gradient-descent iterations increase is shown in Table 6.1. Maximum recognition accuracies for each noise condition have been highlighted in boldface font for clarity. Recognition results with no enhancement (Baseline) and MFNS with static subtraction factors are also shown for comparison.

Analysis of these results shows the optimal number of gradient-descent iterations is considerably different for each noise condition. For the more quiet conditions (idle and 35 mph with windows up), best performance is obtained

with more than 20 iterations of gradient-descent optimisation. For the noisier conditions, less than 5 optimisation iterations provide the best performance (particularly for the 55 mph with windows down noise condition). These three noise conditions also show clear trends of decreasing word accuracy as the number of iterations is increased above 5. Since the noise conditions are approximately ordered by increasing levels of noise, it can be concluded that as the noise levels in the vehicle increase (i.e. higher speeds or open windows), the level of gradient-descent optimisations needs to be reduced in order to avoid over-optimisation of the enhancement parameters.

The best overall performance across all 5 noise conditions can be seen to be at 3 iterations. At this level of optimisation, the 55 mph conditions both exhibit maximum performance, with two other noise conditions (IDL and 35D) being only 0.1% below their maximum word accuracies. The 35 mph with windows up condition is the only condition which is well below its best performance (0.8% absolute), but still provides improvements over the baseline and static enhancement systems. As a result of these observations, 3 gradient-descent iterations were used for the remainder of the experiments in this chapter.

Joint Optimisation Iterations: Having established the most effective number of gradient-descent iterations, the number of joint optimisation iterations was analysed. Table 6.2 shows these results with the best performance across all noise conditions again highlighted in boldface for clarity.

Apart from the 35 mph with windows up noise condition, the results clearly indicate that only one joint optimisation iteration is required for in-car speech recognition. This result indicates that only minor changes are made to the decoded state sequences and therefore there appears to be no advantage in performing more than one joint optimisation iteration. Relating this observation to the results of the gradient-descent iterations, if the state sequence did not change at all, the parameter optimisation would continue from exactly the same position that it finished previously, and therefore over-optimisation is likely to occur as the number of joint optimisation iterations increased.

This result combined with that of the analysis of gradient-descent iterations indicate that over-optimisation is a serious issue for LIMA frameworks operating

Table 6.2: ASR accuracies for increasing joint optimisation iterations.

# Iter.	ASR Word Accuracy (%)				
	IDL	35U	35D	55U	55D
Baseline	70.4	48.8	36.2	41.8	23.5
Static MFNS	73.3	47.8	36.8	44.5	26.1
1	**74.1**	**49.1**	**38.1**	**45.1**	**26.4**
2	74.1	49.4	37.7	44.8	26.1
3	73.9	49.9	37.2	44.8	26.0
4	74.0	50.1	37.2	44.5	26.3
5	74.0	50.3	37.1	44.4	26.1
10	74.1	50.2	37.5	44.1	25.9

in vehicular environments. It is therefore suggested that optimisation iterations be kept to a minimum in order to keep the enhancement parameters generalised for future driver responses. The practical advantage of these findings is the ability to achieve improved ASR performance using LIMA frameworks whilst creating minimal processing delays due to the need for only a few optimisation iterations.

6.4.2 Evaluation of LIMA Frameworks

The six LIMA frameworks listed at the beginning of Section 6.4 were tested using the results obtained in Section 6.4.1. Table 6.3 presents the ASR results for all three calibrated frameworks. The best results for all the frameworks are again highlighted in boldface for clarity. Regardless of the calibration method used, the results show a global improvement over an enhancement system which does not utilise a LIMA framework.

Using matched conditions for speaker-based optimisation (i.e. employing calibration for each speaker in each noise condition) produces the best results in all cases in Table 6.3 except idle. Whilst the idle noise condition shows a 0.5% absolute decrease in word accuracy in its matched condition as opposed to optimising in 55U, the word accuracy performance is still an improvement over the baseline enhancement case (73.7% versus 73.3%). As a result, this outlier is not regarded as a significant issue.

Table 6.3: ASR results for the calibrated LIMA frameworks.

Calibration Framework	ASR Word Accuracy (%)				
	IDL	35U	35D	55U	55D
Baseline	70.4	48.8	36.2	41.8	23.5
Static MFNS	73.3	47.8	36.8	44.5	26.1
Noise	74.1	49.1	38.1	45.1	26.4
Speaker (Random Noise)	73.6	49.5	38.2	44.9	26.5
Speaker (IDL)	73.7	49.3	37.8	44.6	26.8
Speaker (35U)	73.8	**49.9**	38.6	45.0	27.0
Speaker (35D)	73.0	49.4	**39.2**	45.1	26.7
Speaker (55U)	**74.2**	49.7	37.9	**45.5**	26.8
Speaker (55D)	73.1	49.1	38.2	44.7	**27.1**

In order to assess the effectiveness of the proposed dialogue-based LIMA framework, all utterances occurring prior to the first optimisation (or first optimisation after calibration) for each speaker were ignored. This approach was required since the proposed technique requires 100% word accuracy in order to trigger optimisation, a result which was achieved on only 3% of all utterances and mostly in the idle noise condition. This low number of optimisation instances is due to the relatively low performance of the ASR system and the nature of the recognition task which requires all 10 digits to be recognised correctly.

The results of this final evaluation are summarised in Table 6.4. It should be noted that word accuracies in this table are better than previous results because the analysis procedure removed a lot of utterances which exhibited poor ASR performance.

Almost all comparisons that could be made on the results in Table 6.4 show that the proposed dialogue-based LIMA framework provides improved performance over the baseline enhancement system. Applying this framework can also recover losses in word accuracy incurred when using standard Mel-filterbank noise subtraction (e.g. in the two 35 mph noise conditions).

The results of this evaluation also prove the effectiveness of the proposed dialogue-based framework when used with or without explicit calibration even though there is a very low number of optimisation instances. For the case without

Table 6.4: ASR results for the calibrated LIMA frameworks.

Framework	ASR Word Accuracy (%)				
	IDL	35U	35D	55U	55D
Baseline	79.1	55.8	42.1	49.8	27.6
Static MFNS	81.8	53.9	41.6	51.7	30.1
Dialogue-Based System	82.6	55.9	42.3	53.1	31.1
Baseline	80.7	55.5	43.3	49.5	28.6
Static MFNS	81.4	53.3	45.3	50.0	33.6
Speaker Calibration (Random Noise)	82.5	55.7	46.4	52.5	33.3
Dialogue-Based System	82.3	57.7	45.5	52.7	32.3
Baseline	80.4	57.7	44.7	53.3	28.4
Static MFNS	82.2	52.5	42.9	53.9	30.3
Speaker Calibration (IDL)	82.4	55.4	44.6	54.9	31.0
Dialogue-Based System	82.9	55.9	46.0	55.5	30.9

calibration – which is the ideal operational behaviour of such a framework since the user would be completely unaware of the adaptation – global improvements over both baseline systems can be observed, with the best relative performance improvement over a system without enhancement being 16.7% in the idle condition. This particular result demonstrates the true potential of the framework to improve ASR accuracy, since utterances spoken during idle are most likely to trigger the optimisation process. In comparison to the baseline enhancement system, the proposed framework shows relative improvements of between 1.2% and 4.4% in this mode of operation.

There are also noticeable improvements over the calibration-only LIMA framework, particularly one performing calibration during idle. In this case, the relative improvements range from 1.2% to 2.8% (excluding the marginal decrease in performance in the 55D noise condition). Given that most users will first speak to the in-car dialogue system when entering their vehicle, this result verifies the potential of the proposed framework to be incorporated with a calibration session to produce further improvements in system performance.

Considering the operation of the proposed dialogue-system, there is potential for a loss of generality if a particular noise condition is consecutively optimised;

this phenomenon was observed when assessing the number of optimisation iterations in Section 6.4.1. The consistent improvements in Table 6.4 indicate however that this is not an issue within the proposed framework as regular changes in noise conditions allow the optimisation process to effectively track the noise conditions inside the cabin and reset the enhancement parameters appropriately.

6.5 Research Directions

The major shortfall of the experimental evaluation in this chapter was the nature of the test data. In this instance, it was required that all 10 digits of the phone number were recognised correctly before optimisation was allowed to take place by the proposed framework. This constraint led to a very low number of optimisation instances (3%); nevertheless, it did demonstrate the capability of the framework for this application. The most obvious scope for future research is to compare the frameworks with data which constitutes a much simpler recognition task; performance on such a task is expected to exceed that shown in this chapter due to increased regularity of optimisation occurrences. Once the issues identified in Chapter 5 regarding dialect mismatch have been completely resolved, the commands task of the AEICS corpus would be ideal for this evaluation. Another advantage of using a simpler recognition task is the increased word accuracy of the ASR system, which would make comparison against an unsupervised LIMA framework meaningful.

Despite the proposed dialogue-based framework showing the potential to avoid parameter over-optimisation, there are situations where this problem may still occur. For example, consider driving on the highway for an extended period of time. Since the noise conditions will remain relatively stationary since the speed is kept constant (barring changes to the internal environment such as opening the window), successive utterances will lead to further optimisation for the same speaker and noise condition. To overcome this problem when the state of the environment changes significantly (e.g. slowing from 100 km/h to 50 km/h), it would be useful to incorporate knowledge relating to these states in order to appropriately reset the enhancement parameters. Such information could also be used

to create a system with adaptive numbers of gradient-descent iterations for each noise condition as it was seen both here and in Chapter 5 that noisier conditions necessitate less iterations in order to avoid over-optimisation. It was also seen in Chapter 5 that the baseline performance for the quieter conditions was better than a static set of enhancement parameters; this information could therefore be used to change the initial set of enhancement parameters if noise-adapted acoustic models being used. Given the connectivity of modules in modern-day vehicles, knowledge about the vehicle speed, climate control status, window status etc. can be extracted from the already available CAN-bus signals [6]. How best to separate noise conditions using CAN-bus signals in order to assist speech systems is a research topic which should gain increased interest in the coming years.

Whilst the proposed dialogue-based LIMA framework has shown the ability to deal with considerable changes in noise conditions *between* successive optimisations, data limitations restricted the analysis to constant noise conditions *during* the utterance. BabaAli *et al.* [10] attempted to evaluate within-recording condition changes by altering the SNR of white noise and periodic alarm noise; this procedure performs frequency-independent scaling of the magnitude spectrum, but in car environments when speed is changed or the air-conditioning system is turned on, the changes would be frequency-dependent. Analysing the performance of speech systems in a full range of in-car conditions will indicate the scenarios which are most problematic and will assist in developing in-car speech systems which are effective in all possible driving conditions. In order to conduct future evaluation of LIMA frameworks in a full range of realistic in-car environments, data to complement the existing AVICAR database and AEICS corpus was collected shortly before completion of this dissertation.[2]

A general drawback of the LIMA framework is the extensive computing required for gradient-descent optimisation. The dialogue-based system does not need to run in real-time since optimisation occurs *after* confirmation of spoken responses, however it is still practically beneficial to reduce overall processing costs. Improvements in this respect could be achieved in a number of simple

[2]The data was collected in collaboration with the UTDrive project [6] during an internship at the University of Texas at Dallas. Details of this collection can be found in Appendix C.

ways such as reducing the parameter space, simplifying the acoustic model, or removing frames deemed unreliable and potentially harmful to the final result. Parameter reduction through examination of optimised parameter values may reveal correlation between parameters or identify parameters which only produce minor changes to the final result. Reducing the acoustic classes on which the optimisation occurs (e.g. from triphone to phoneme) will reduce the model space and may also lead to additional robustness and improved performance. Finally, frame masks like those in missing feature techniques (see Section 2.4.3) could reduce the number of frames on which to optimise by removing unreliable segments altogether which may also increase the accuracy of the optimisation. Suitable masks could be based on proximity to model boundaries (i.e. to counteract alignment errors) or the effect of speech production mismatches such as stress on specific HMMs.

6.6 Summary

This chapter discussed different practical implementations of LIMA frameworks specifically for in-car speech dialogue systems and identified a number of drawbacks with the previously proposed calibration and unsupervised frameworks. In order to overcome these limitations, a new LIMA framework which exploits the grounding process used in speech dialogue systems was proposed. This framework permits optimisation to occur only when the user has confirmed that the speech dialogue system has correctly recognised their spoken responses. The advantages of this framework include the ability to deal with continually changing in-car noise conditions, as well as ensuring recognised state sequences are reliable for use in the optimisation process.

An analysis of the number of gradient-descent and joint optimisation iterations revealed that minimal optimisation is required for the best average speech recognition performance in this application. This observation enables processing delays to be reduced whilst also providing word accuracy improvements over a static Mel-filterbank noise subtraction speech enhancement technique.

The proposed dialogue-based framework was evaluated against a calibrated

LIMA framework operating under a number of different adaptation scenarios. Experimental results showed the proposed system provides improved recognition performance over baseline systems with and without enhancement as well as the calibration-only framework, particularly when it is assumed that the initial calibration occurs when the vehicle is idling. Despite the low number of optimisation instances, this framework is particularly suited to improve the speech recognition performance of in-car speech dialogue systems. Future improvements to the proposed LIMA framework are focussed on reducing the overall processing time whilst maintaining the speech recognition accuracy which is an important factor in pushing this technique into the automotive industry.

Both Chapter 5 and Chapter 6 have provided a comprehensive review of likelihood-maximising speech enhancement for robust ASR. This has included the practical application of this framework to optimise various parameters of the enhancement algorithm. In the next chapter, a novel method is presented which potentially removes the need for parameter optimisation through the use of complex spectrum subtraction in the frequency domain.

Chapter 7

The Use of Phase in Spectral Subtraction

7.1 Introduction

One of the primary aims of this dissertation is to demonstrate that speech enhancement techniques designed for human intelligibility are not necessarily optimal for use with ASR systems. Spectral subtraction is no exception – it was initially designed to improve human intelligibility through noise reduction, and subsequent research has shown that the phase used for signal reconstruction in the time domain has little effect on human intelligibility. As a result, only the magnitude spectrum has been used within the subtraction process.

Following the lead of some recent speech enhancement studies which utilise some form of phase spectrum information (Section 7.2), this chapter examines the use of the short-time phase spectrum in frequency-domain spectral subtraction. It is shown in Section 7.3 that the only way to obtain accurate estimates of the clean speech magnitude (which is used in MFCC feature extraction) is to utilise the phase spectrum as part of spectral subtraction in the complex frequency domain.

Obtaining estimates of the phase spectrum however, is non-trivial compared to the techniques used for magnitude spectrum estimation, and has been therefore been overlooked in the past. A novel phase spectrum estimation procedure is presented in Section 7.4 which exploits the assumption of phase stationarity on

sinusoidal waveforms. The effectiveness of this proposed phase estimation procedure as part of Complex Spectrum Subtraction (CSS) is examined for estimating either clean speech or noise phase spectra (Section 7.5).

The novel contributions contained in this chapter are highly exploratory; they have been assessed using oracle-type experimentation which allows access to all spectral information. Therefore, prior to summarising the work contained in this chapter, Section 7.6 places considerable thought into the research directions required to make the use of the phase spectrum in spectral subtraction suitable for integration with state of the art speech recognition systems.

7.2 Phase Spectrum and Speech Enhancement

Despite the use of phase for other speech processing applications (such as features for ASR [123, 158] or speaker identification [144]), phase information has only been used in two speech enhancement examples to date. These two techniques – phase spectrum compensation [135, 137, 149] and multi-channel phase-error filtering [1, 72] – were briefly described in the speech enhancement literature review in Chapter 3. For the scope of the work contained in this dissertation however, neither of these approaches are deemed appropriate; PSC requires reconstruction to the time domain (which is unnecessary and sometimes undesirable for ASR applications), and PEF is a multi-microphone solution.

Attention is therefore shifted to the use of phase (or rather, lack thereof) in frequency-domain spectral subtraction. As described throughout this dissertation, the core of spectral subtraction is to subtract an estimate of the noise *magnitude* spectrum from the noisy speech *magnitude* spectrum. The noisy speech phase spectrum is left unaltered, and is utilised only when reconstructing the enhanced spectrum back to the time domain. This ignorance of the phase spectrum is the direct result of studies undertaken in the 1980s which showed that the phase spectrum provided no perceptual difference to the enhanced signals [109, 143]. In recent times these claims have been challenged [110] which may result in a shift of attention in speech enhancement research in the coming years.

Whilst the early experiments showed that phase was unimportant for perception, it was duly noted that if "using the phase estimate to further improve the magnitude spectrum, then a more accurate estimation of phase may be important" [143]. This statement has direct relevance for speech recognition applications since features for speech recognition are typically derived from the magnitude of the incoming speech signal (e.g. in MFCCs). Therefore, speech recognition performance with spectral subtraction could be improved by estimating phase spectrum information and using that information as part of the algorithm. More recently it has been shown that in order to obtain accurate estimates of the clean speech magnitude spectrum, estimates of the true phase of the clean speech are also required [87].

This observation was supported by a recent investigation into the limitations of spectral subtraction for speech recognition applications [34]. This study specifically looked at the effect on recognition accuracy of three sources of error when the algorithm is implemented in the power spectrum (i.e. $\gamma = 2$). Two of these errors are directly related to the use of phase – reconstruction phase and the phase difference between clean speech and noise which is encapsulated in the spectral cross-terms [87]. It should be noted however, that cross-terms are typically ignored under the assumptions that speech and noise signals are uncorrelated or that their phasor representations are colinear (i.e. have the same phase).

Figure 7.1 shows the results of this particular study using data from the Aurora database [57]. It can be seen that the effect of reconstruction phase errors (red) only produce minor decreases in word recognition accuracy. Errors due to the phase between spectral cross-terms (green) become significant at low SNRs where the decrease in recognition accuracy can be as much as 15%. Whilst the overall effect of these two errors is small compared to errors in the magnitude spectrum (compare the blue and purple lines in Fig. 7.1), these results confirm that neglecting information contained in the phase spectrum can lead to noticeable losses in speech recognition performance.

Despite the original statement made by Wang & Lim in 1982 [143], and the recent findings in both [34] and [87], research has so far failed to develop an appropriate method for estimating either the noise or speech phase spectrum.

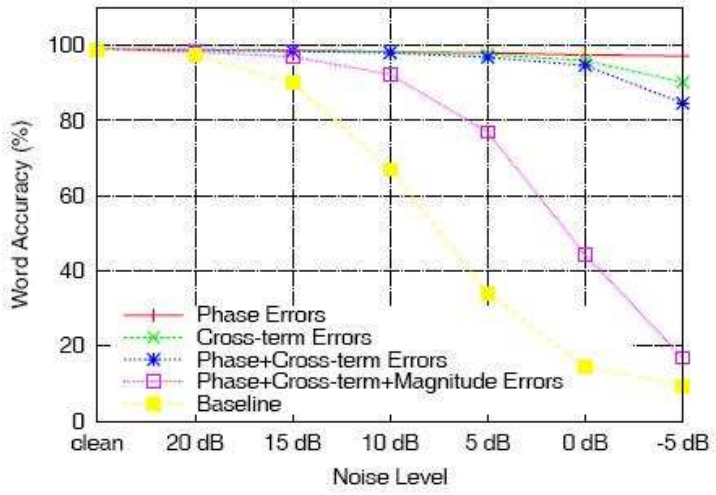


Figure 7.1: ASR word accuracy for the Aurora database with different sources of errors in spectral subtraction (from [34]).

Motivated by the lack of a suitable solution, the novel contributions contained in this chapter include a demonstration of the importance of performing spectral subtraction in the complex frequency domain, and the subsequent realisation of CSS using an original phase estimation algorithm referred to as Phase Estimation via DElay Projection; this method can be used to estimate either noise or speech spectra.

7.3 Incorporating Phase Information into Spectral Subtraction

7.3.1 The Effect of Phase on ASR

Throughout this dissertation, Mel-frequency cepstral coefficients have been used for speech feature extraction. As described in Chapter 2, MFCCs are calculated by passing the magnitude spectrum of the signal to be recognised through a series of filterbanks, taking the log spectrum, and then decorrelating the cepstrum. Whilst it is intuitive that phase information is not explicitly used in this representation, it can be shown that phase information is necessary in deriving accurate estimates of the magnitude spectrum [87].

Traditional magnitude spectral subtraction (i.e. when $\gamma = 1$) assumes that

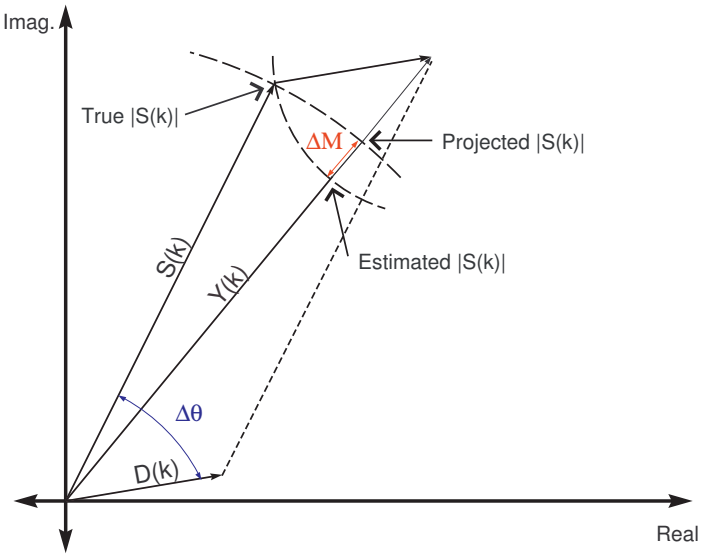


Figure 7.2: Single-frequency phasor diagram showing the effect on clean speech magnitude estimates when assuming colinearity of noise and noisy speech signals.

the noisy speech and noise (and subsequently the clean speech) have the same phase; that is, they are colinear. The shortfall of this assumption is demonstrated by the single-frequency phasor diagram shown in Fig. 7.2. This figure represents – in vector form – the spectral subtraction operation which takes place on a frequency-by-frequency basis.

Under the common assumption of additive background noise, the vectors for the noise and speech signals ($D(k)$ and $S(k)$ respectively) are added to produce the observed noisy speech vector $Y(k)$. Assuming that the noise magnitude estimate accurately represents the instantaneous noise magnitude, the use of conventional magnitude subtraction produces the estimated clean speech magnitude denoted by "Estimated $|S(k)|$". To compare this estimate to the true clean speech magnitude, the "True $|S(k)|$" label is rotated about the origin onto the noisy speech vector to the point "Projected $|S(k)|$". It can be seen from this projection that there is an error between the true magnitude and the clean speech magnitude estimate (denoted ΔM). Since this example has assumed that there is no error in the noise magnitude estimate, the error in the resulting clean speech magnitude is purely due to the phase difference between the clean speech and noise (denoted $\Delta\theta$).

In this example, the noise signal is added to the clean speech signal to produce

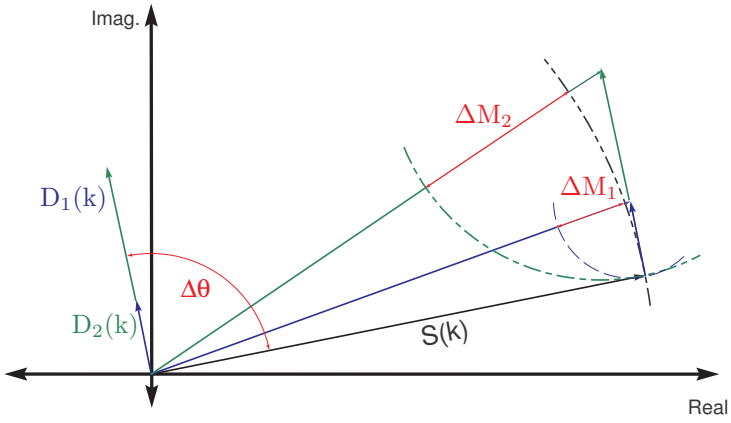


Figure 7.3: The effect on the magnitude error of decreasing the SNR.

the noisy signal. As the noise magnitude is increased (assuming the clean speech signal remains constant and the phase of the noise signal remains unchanged), the overall SNR decreases and the influence of the noise signal on the final noisy speech becomes greater. This effect (shown in Fig. 7.3) results in an increase in the resulting clean speech magnitude error (i.e. ΔM) as the SNR decreases. This phenomenon is particularly problematic for in-car environments which can exhibit SNRs less than 0 dB when using distant microphones [15].

The clean speech magnitude error is also magnified as the difference in phase between the noise signal and the clean speech is increased. A visualisation of this type of error is shown on the x-axis of Fig. 7.4, where the negative scale on the colourmap indicates the magnitude is always underestimated. It can be seen that for SNRs greater than 0 dB, the magnitude error increases as the noise and speech signals become increasingly out of phase up to π radians (180^o).

It can also be seen that at high SNR (e.g. $>$ 30 dB), the effect of phase error is much less than around 5 dB where the error reaches its maximum – $-(1-\beta)|Y(k)|$. For SNR less than 5 dB, large clean speech magnitude errors (as shown by the blue regions of Fig. 7.4) are spread over a much wider range of phase differences, with these maxima moving asymptotically towards $\frac{\pi}{2}$ and $\frac{3\pi}{2}$.

An interesting phenomena occurs when the SNR is less than 0 dB – the magnitude error gets smaller as the noise and speech signals move towards π radians out of phase. In this instance, the additive assumption results in the noisy speech signal having phase more closely related to the background noise

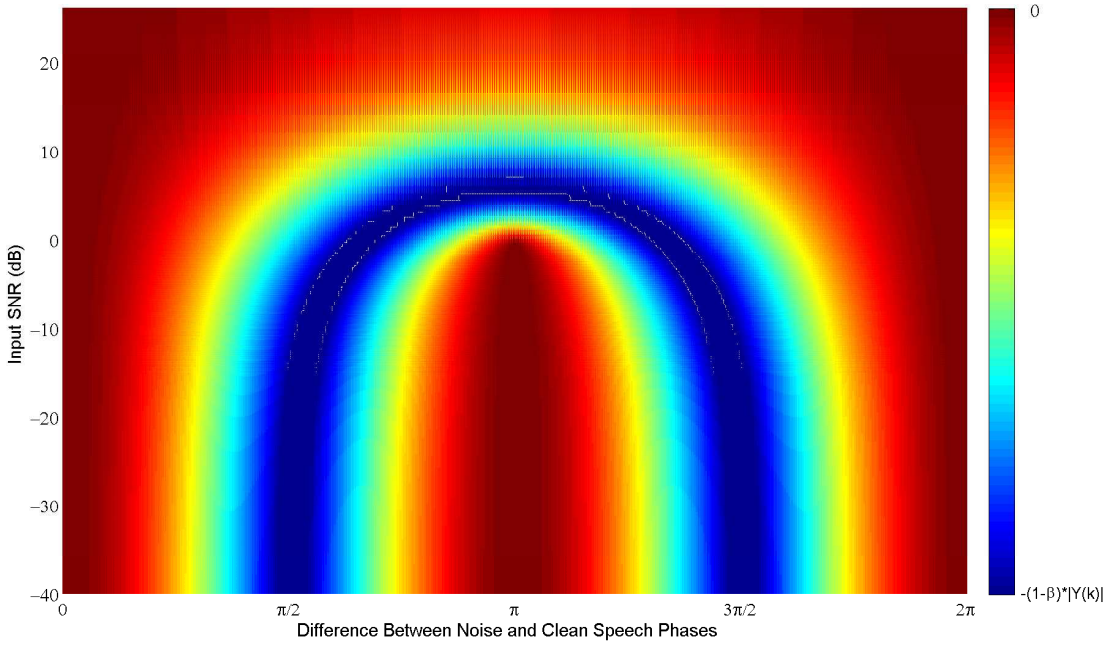


Figure 7.4: Visualisation of the effects of both SNR and difference between noise and speech phases on the output clean speech magnitude estimate.

than the speech signal; therefore the assumption of colinearity used in traditional magnitude spectral subtraction is restored.

7.3.2 Complex Spectrum Subtraction

The discussion in Section 7.3.1 highlighted the need to incorporate phase information into the spectral subtraction algorithm in order to obtain accurate clean speech estimates. This section demonstrates how to perform the subtraction on the complex frequency spectrum by incorporating phase information.

Recall the representation of the noisy speech signal in the complex frequency spectrum:

$$Y^i(k) = S^i(k) + D^i(k) \tag{7.1}$$

where i is the frame index and k is the index of the discrete frequency bin. Rearranging Eq. (7.1) yields the subtraction rule in the complex spectrum:

$$S^i(k) = Y^i(k) - D^i(k). \tag{7.2}$$

If Eq. (7.2) is expanded into individual subtraction rules for both the real and

imaginary components:

$$|S^i(k)|\cos(\theta_S) = |Y^i(k)|\cos(\theta_Y) - |D^i(k)|\cos(\theta_D)$$
$$|S^i(k)|\sin(\theta_S) = |Y^i(k)|\sin(\theta_Y) - |D^i(k)|\sin(\theta_D) \tag{7.3}$$

it can be seen that a number of pieces of information are required in order to accurately determine the clean speech magnitude. This information consists of the magnitude and phase spectrum of *both* the noisy speech and noise signals. The output of the DFT contains the magnitude and phase information about the noisy speech spectrum $Y^i(k)$, however it is not possible to know the instantaneous noise spectrum $D^i(k)$ exactly, therefore it must be estimated. The magnitude of the noise $|D^i(k)|$ can be estimated by any of the techniques described in Section 3.4 (and many more); in this dissertation magnitude estimation uses a time-recursive averaging estimation with soft-decision SAD.

Whilst magnitude estimation is a relatively simple task, estimating the phase spectrum is not as trivial, particularly due to phase wrapping and other signal processing problems [5]. An averaging process similar to that used for magnitude estimation is particularly inappropriate as the phase spectrum is circular, existing in the range $-\pi < \theta \leq \pi$. In this case, a large number of phase samples would result in a mean of zero – this provides no useful information.

These challenges in finding an appropriate representation of the phase spectrum is another reason why spectral subtraction is traditionally performed only on the magnitude (or power) spectrum. Despite these challenges, it was shown in Section 7.3.1 that in order to derive the true clean speech magnitude, it is necessary to also include information about the phase spectrum. In Section 7.4, a novel method for estimating the phase spectrum is presented and used to perform complex spectrum subtraction as per Eqs. (7.2) and (7.3).

It should be noted that using complex spectrum subtraction removes all reliance on the flooring factor β, the oversubtraction factors α, and also the spectrum in which subtraction takes place (i.e. γ). Optimal operation of conventional spectral subtraction techniques requires data-dependent tuning of each of these parameters [68]. Therefore, CSS is an attractive alternative to magnitude-based spectral subtraction as it is able to provide an enhancement solution without the

requirement to tune algorithm parameters.

7.4 Phase Spectrum Estimation

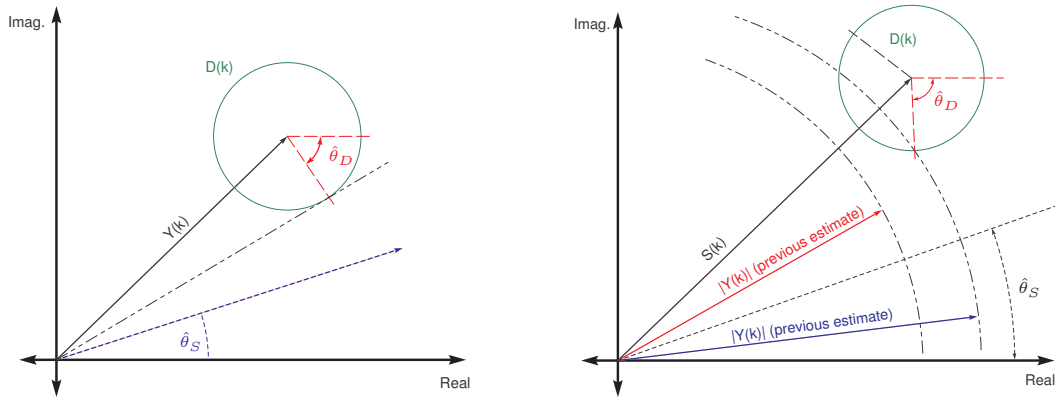
7.4.1 Estimation Domains

Complex spectrum subtraction as proposed in the previous section requires estimation of the noise phase spectrum in order to utilise Eqs. (7.2) and (7.3). There are two ways in which this could be achieved – direct estimation of the noise phase spectrum, or estimation of the clean speech phase and interpolating to deduce the noise phase spectrum (or complex spectral subtraction result) using geometrical relationships. In this section a new method is proposed to estimate either noise or speech phase.

To interpolate the noise phase from the clean speech phase estimate, there are two possible methods which are referred to as the *tangent* and *intersection* methods. These two methods are shown graphically in Fig. 7.5. For both methods it is acknowledged that the full range of possible outputs from the subtraction process is represented by a circle with radius equal to the noise magnitude (i.e. $D(k)$) and centred on the complex representation of the noisy speech signal. In the case where the noise magnitude is perfectly accurate, one of the points on this circle will constitute the original clean speech signal.

The tangent method (Fig. 7.5(a)) assumes that the clean speech phase estimate is accurate, and derives the clean speech spectrum using tangents to the circle from a line drawn from the origin. Since there are two tangents, the final result is chosen as the tangent point which most closely matches the clean speech phase estimate $\hat{\theta}_S$. If the clean speech phase estimate lies between the two tangent points, the subtraction result is taken as the point with the smallest magnitude where a line drawn from the origin at that phase intersects the circle.

The intersection method (Fig. 7.5(b)) on the other hand, assumes that the previous estimate of the clean speech magnitude is more accurate than the new phase estimate $\hat{\theta}_S$. In this instance, the clean speech magnitude estimate from the previous frame (blue) is rotated in order to intersect the circle representing



(a) (b)

Figure 7.5: Two methods for interpolating noise phase from the clean speech phase – (a) tangent method, and (b) intersection method.

all possible subtraction results (green). The true intersection point (as again there are two) is taken as the point most closely matching the clean speech phase estimate in the current frame. In the case that the clean speech magnitude from the previous frame does not intersect the circle (red), colinear complex spectral subtraction is performed – i.e. the interpolation reverts to the complex equivalent of the traditional spectral subtraction implementation.

7.4.2 Estimation Based on Stationarity

Noise magnitude estimates are typically calculated during non-speech periods and are assumed to remain stationary during speech periods. Here the concept of stationarity applied to single-frequency sinusoids is utilised to explore the possibility of deriving phase estimates. The overall aim is to project both the noise magnitude and phase spectra through periods of speech.

Consider the single sinusoid case shown in Fig. 7.6 which is divided into 32 ms frames with 10 ms advances between adjacent frames (i.e. common speech processing frame rates). At the beginning of "Frame 1", the sinusoid has a phase of 0 radians, but at "Frame 2" and "Frame 3" this phase has changed. If it is assumed that this sinusoid is stationary between frames, the expected phase at the start of each of these successive frames can be inferred if the frequency of the sinusoid

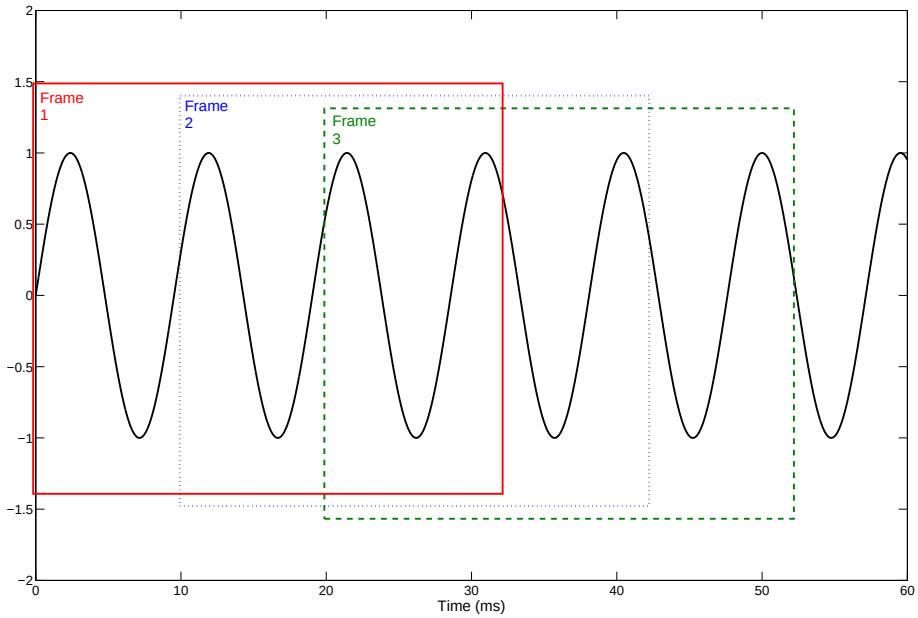


Figure 7.6: Demonstration of phase changes due to advancing frames on a single-frequency sinusoid.

is known along with the time delay between adjacent frames τ. Therefore, the expected phase of frequency f at frame i can be determined as:

$$\phi^i = \phi^{i-n} + 2\pi n \tau f \tag{7.4}$$

where n is the number of frames prior to the current frame in which the reference phase was taken, and τ is the time advance of each frame in seconds. This estimation approach is termed Phase Estimation via DElay Projection (PEDEP).

PEDEP using Eq. (7.4) is straight forward when only one sinusoid is present, however speech signals contain a mixture of sinusoids each with a different frequency. A few assumptions are therefore required in order to explore PEDEP as a useful estimation technique.

The DFT accumulates sinusoidal components into discrete frequency bands with centre frequencies which are determined by the sampling rate f_s and the length of the analysis window. Sinusoidal components contribute largely to the frequency band which encompasses their true frequency, however it is highly unlikely that the true frequency is exactly equal to a particular DFT centre frequency. The consequence of this deviation is a smearing effect across a wider range of frequencies. This smearing effect is demonstrated in Fig. 7.7 which shows the magnitude spectrum of two sinusoids – one at 1000 Hz which matches

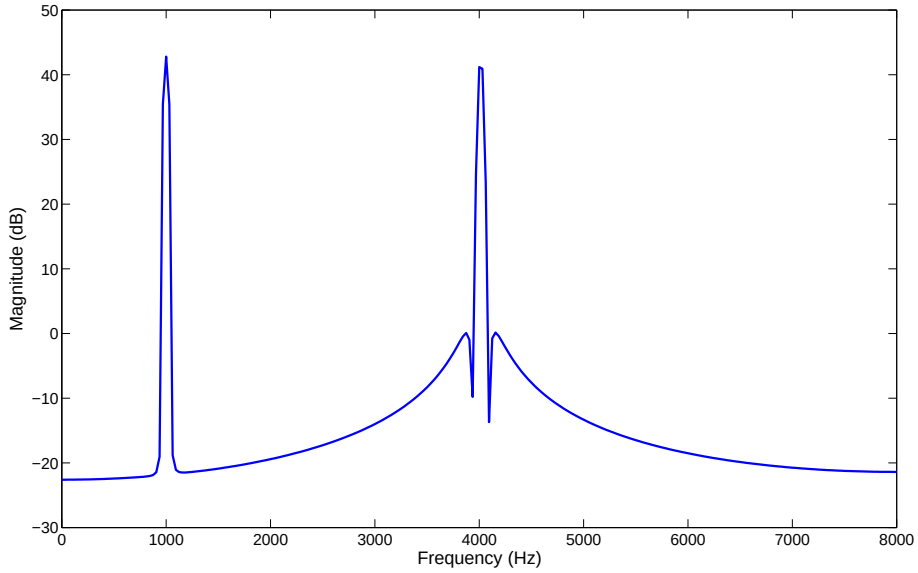


Figure 7.7: Demonstration of the effect of spectral smearing when sinusoidal frequencies differ from DFT frequencies.

a DFT frequency, and one at 4015 Hz which deviates from the nearest centre frequency by 15 Hz. Both signals cause smearing across adjacent frequencies, however the higher frequency signal has a greater effect across a much wider range of frequencies. Despite this smearing effect, PEDEP assumes that this effect is minimal on the final frequency representation, and therefore frequencies can still be analysed independently.

PEDEP also assumes that the signal in each independent frequency band remains stationary (or very close to) from one sampling frame to the next. In this way, it is inherently assumed that no other sinusoidal components are added or removed between these instances in time.

7.5 Experiments & Discussion

7.5.1 Investigation

To determine the validity of the PEDEP approach, it was required to study the stationarity behaviour (as described in Section 7.4.2) of typical noise and clean speech phase spectra under the experimental configuration used in this dissertation. Clean speech phase samples were generated from 100 randomly

chosen TIMIT test sentences [42], whilst samples of car noise from the NOISEX database [140] and randomly generated Additive White Gaussian Noise (AWGN) were used as examples of noise phase spectra.

To best visualise the stationarity behaviour, histograms describing the relationship of *observed* phase spectra in adjacent frames were generated for each DFT centre frequency. To do this, true differences in phase between adjacent frames were calculated and then normalised for the expected delay due to the observation sampling. This calculation equates to a modified version of Eq. (7.4):

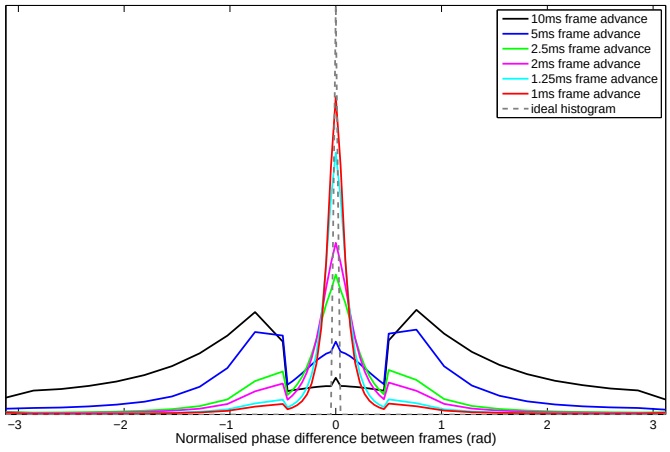
$$\Delta\phi = \phi^i - \phi^{i-1} - 2\pi\tau f \tag{7.5}$$

where each value of $\Delta\phi$ was wrapped to the range $-\pi < \Delta\phi \leq \pi$. By removing the effect of the observational delay, the histograms identify the accuracy of the stationarity assumption which is pivotal to this phase estimation technique.

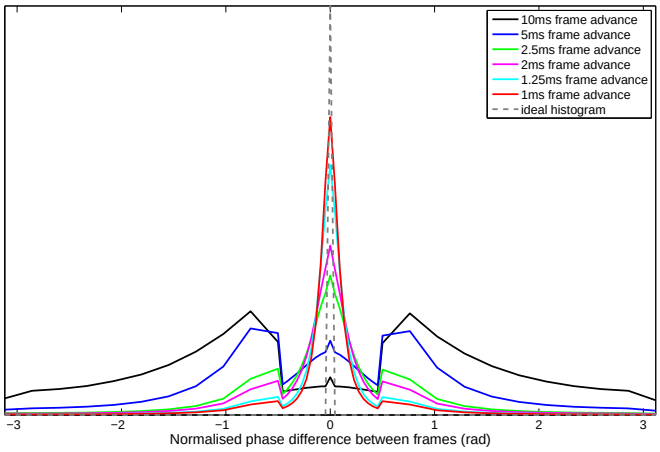
Figure 7.8 shows the *average* distributions for all frequencies of actual phase differences between adjacent frames for AWGN (Fig. 7.8(a)), car noise (Fig. 7.8(b)) and clean speech (Fig. 7.8(c)). Each histogram is compared against the ideal distribution which reflects the assumption of perfect signal stationarity between frames (i.e. $\Delta\phi = 0$). It can be observed that the distributions of the three different signals are very similar, as is their behaviour as the frame advance is decreased. This similarity in behaviour suggests that the PEDEP approach will be suitable for a wide range of noise and speech signals.

In general, the histogram for the standard speech processing frame rate of 10 ms exhibits main peaks at approximately $\pm\frac{\pi}{4}$, and only a minor peak at 0 radians which makes it significantly different in shape to the desired distribution. This observation led to decreasing the frame advance in order to improve the distribution. The motivation for increasing the frame rate was that as the observation points become closer, the more likely the signal remains stationary; in other words, there is less time for sinusoidal components change.

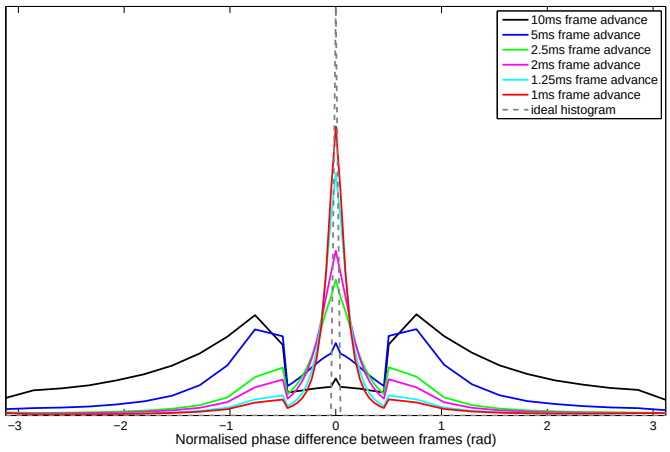
Examining Fig. 7.8, it can be noted that as the frame advance is reduced, the distribution slowly changes to become more and more like the ideal histogram, with dominant peaks appearing at 0 radians for frame advances less than – and including – 2.5 ms. Just as importantly, the secondary peaks around $\pm\frac{\pi}{4}$ are



(a)



(b)



(c)

Figure 7.8: Histograms showing normalised phase differences between adjacent frames for (a) AWGN, (b) car noise, and (c) clean speech.

attenuated, but they also shift towards $\pm\frac{\pi}{6}$. As the frame advance is continually decreased to 1 ms, only minor improvements are observed in the histograms which suggests only minor improvements in the accuracy of the phase estimation procedure would be observed.

It should also be noted that whilst only the average distributions across all DFT frequencies have been shown here, the behaviour of individual frequencies was found to be very similar to that observed in Fig. 7.8. Thus, it will be possible to use the outcomes of this investigation to perform PEDEP on a frequency-by-frequency basis as required by complex spectrum subtraction.

In summary, these histograms verify the assumption of stationarity necessary for the proposed PEDEP approach to be effective for integration with complex spectrum subtraction. In particular, this assumption is more closely approximated as the frame advances become smaller (i.e. the frame rate is increased).

7.5.2 "Oracle-style" ASR Experiments

Having determined that the underlying assumption of the PEDEP approach can be better approximated when using larger frame rates, it was also necessary to evaluate the performance of complex spectrum subtraction using PEDEP on an ASR task. A number of oracle-based recognition tests were designed to:

1. Predict an upper bound on ASR performance if perfect phase estimation was possible;

2. Determine the effects on ASR performance of different frame advances required by the PEDEP approach;

3. Compare the ASR performance using directly estimated and interpolated phase spectra; and

4. Compare the proposed phase estimation technique and CSS with conventional magnitude spectral subtraction.

This experiment used the same baseline speech recogniser detailed in Section 4.3. For all experiments, noise magnitude estimates were derived using

the time-recursive averaging method with soft-decision SAD as described in Section 3.4. Test data consisted of the same 100 TIMIT test sentences used for generating the histograms in the previous section. Each utterance was corrupted with varying levels of car noise from the NOISEX database as well as randomly-generated AWGN to satisfy a range of target SNRs from 20 dB to -5 dB. Using synthesised test data ensured clean speech, noisy speech and noise signals were all available to the oracle framework.

An explanation of the baseline and enhancement approaches compared in this oracle-based evaluation is as follows:

- *Clean Speech*: baseline ASR using the original TIMIT sentences.
- *Noisy Speech*: baseline ASR using the noise-corrupted TIMIT data over a range of SNR.
- *Magnitude Spectral Subtraction*: traditional magnitude spectral subtraction (i.e. $\gamma = 1$) incorporated into the front-end. For this experiment, the noise magnitude estimate was considered to be accurate, therefore $\alpha = 1$.
- *True Phase CSS*: complex spectrum subtraction using the true noise phase spectrum but estimated magnitude spectrum. This scenario simulates the upper bound on ASR performance when using CSS with perfect phase estimation in the ASR front-end.
- *Colinear CSS*: complex spectrum subtraction assuming the noise and speech signals are in-phase. This is the equivalent of magnitude spectral subtraction without the spectral flooring operation.
- *Direct CSS with PEDEP*: complex spectrum subtraction using PEDEP to estimate the noise phase spectrum.
- *Interpolated CSS with PEDEP*: complex spectrum subtraction using PEDEP to estimate the clean speech phase spectrum and interpolating using the tangent method described in Section 7.4.1. In frames deemed to be noise, the colinearity assumption was used to generate the subtraction result. The intersection method was not evaluated in this thesis as there is concern

over error propagation since the method relies on the previous clean speech magnitude estimate which will have some level of error associated with it.

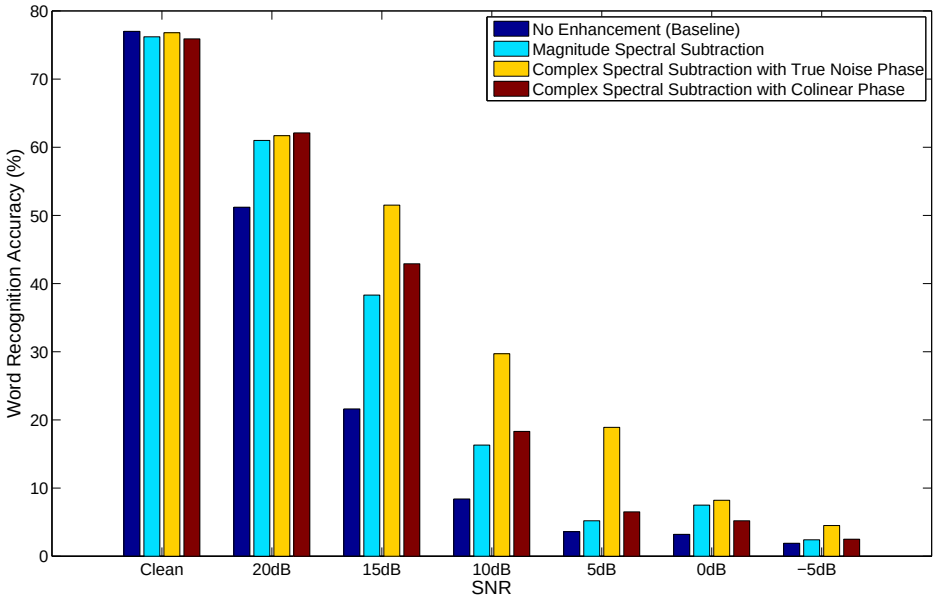
For both CSS configurations incorporating PEDEP, estimates were derived for a range of frame advances from 1 ms through to the standard processing rate of 10 ms. In order to maintain the 10 ms frame rate for ASR feature generation, only frames which correspond to those of the standard 10 ms advances are used – these frames are referred to as the analysis frames. For example, in the 1 ms frame advance case, every 10^{th} frame is considered an analysis frame. For each analysis frame, complex spectrum subtraction was performed using the phase estimate derived by projecting forward the *true* phase from the previous frame (regardless of the frame advance). This oracle experiment was designed to explore whether the PEDEP approach provides useful information for CSS when operating in its most accurate form (i.e. propagation errors are eliminated by projecting the phase forward one frame as opposed to n frames).

Whilst the effect of projecting the phase over multiple frames was not explicitly considered in the experiment design, some indications of the performance in this scenario exist within the results presented in this section. For example, using a 1 ms frame advance and projecting forward 2 frames is equivalent to using a 2 ms frame advance and projecting forward 1 frame.

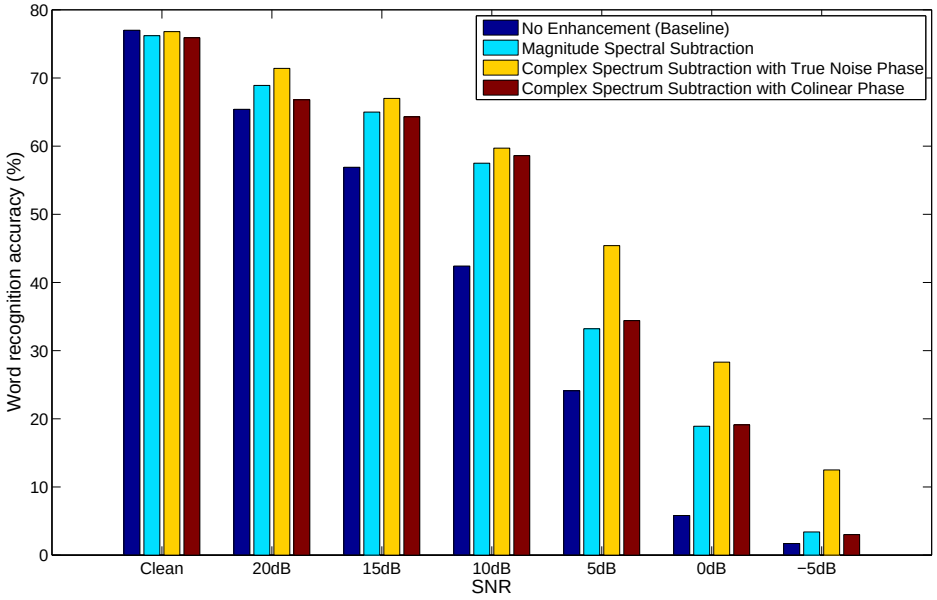
Complex Spectrum Subtraction Proof of Concept

The results of an initial proof-of-concept recognition experiment are provided in Fig. 7.9 for a range of SNR for both additive white Gaussian noise and car noise. The results for both types of noise indicate that if the phase spectrum of the noise signal is known, large improvements in speech recognition accuracy can be achieved compared to using just the magnitude information as in magnitude spectral subtraction. This result is true for all SNR, but is particularly noticeable for moderate SNR (5-15 dB) for AWGN and for low SNR (less than 5 dB) for car noise. At 5 dB, the relative improvements in word accuracy are 14.5% and 18.3% for AWGN and car noise respectively. The average relative improvement for both noise types are around 9.5% for all SNR.

It should also be noted that complex spectrum subtraction is able to recover



(a)



(b)

Figure 7.9: Proof-of-concept speech recognition results for complex spectrum subtraction using known (i.e. true) phase and assuming speech and noise colinearity for (a) AWGN and (b) car noise.

some of the performance loss which is brought about by distortion introduced by magnitude spectral subtraction in clean environments. Distortion arises due to the use of a noise magnitude estimate instead of the true instantaneous noise magnitude, as well as the spectral flooring process. Both of these factors are contributors to musical noise – a well-known artefact of spectral subtraction. By using CSS, the removal of the flooring process as well as the introduction of phase information is able to reduce these levels of distortion.

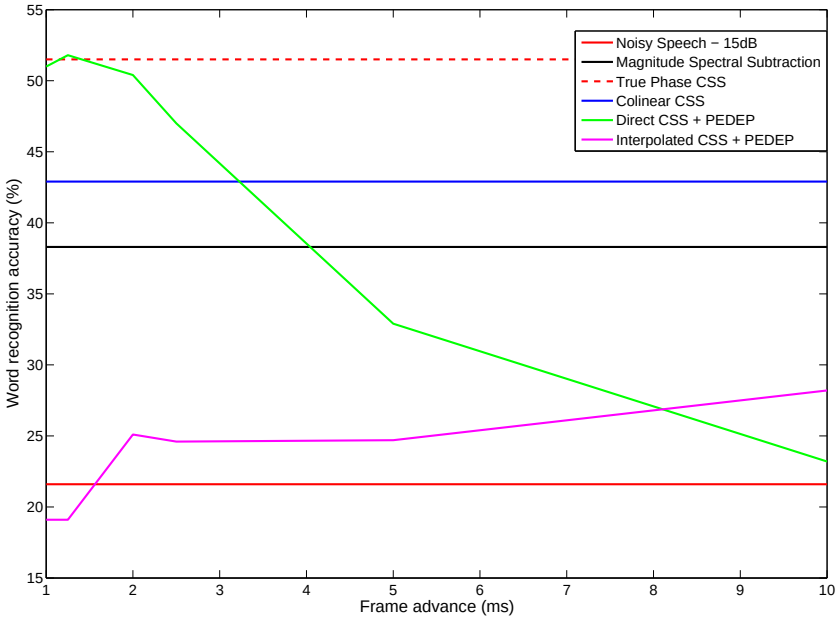
Utilising the colinearity assumption to generate the phase information (and thereby removing the flooring process) provides similar performance – generally within 1% – to magnitude spectral subtraction. In some instances, such as for AWGN, colinear CSS outperforms magnitude subtraction for SNR between 20 dB and 5 dB, but is inferior for additive car noise at SNR greater than 10 dB. Despite the different characteristics for the two different noise types, this result tends to suggest that the computationally expensive noise flooring operation can be removed by utilising colinear spectrum subtraction. This alteration may be highly beneficial for hardware implementations of spectral subtraction.

This proof-of-concept experiment justifies the proposal to perform subtraction in the complex frequency domain by demonstrating the ability to *at least* match the performance of traditional magnitude spectral subtraction. The noise phase information used in these experiments represent two extremes of the phase estimation problem – true phase information represents the objective of perfect estimation, however colinear phase provides a reasonable starting (as well as fall-back) position from which phase estimates can be improved.

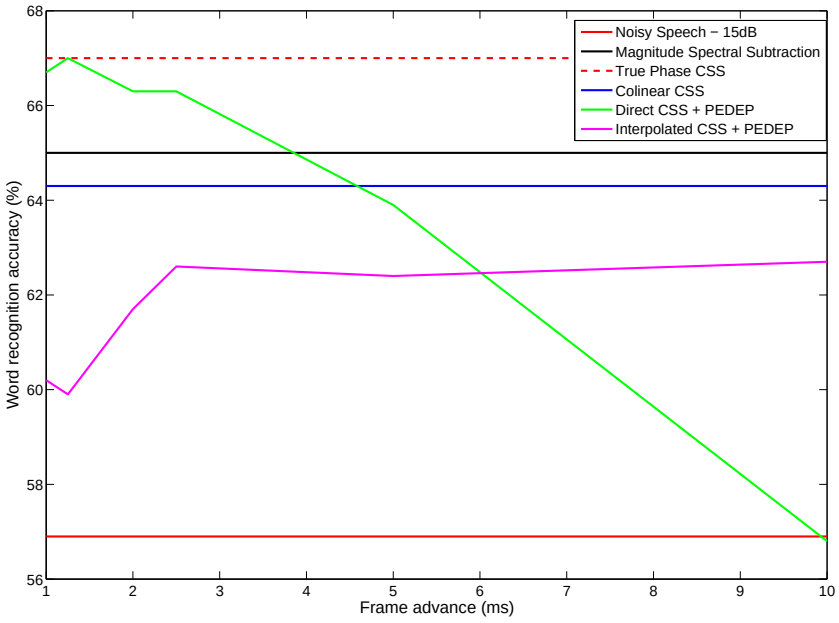
Phase Estimation

Having justified the complex spectrum subtraction approach using known phase information, the proposed PEDEP estimation technique was evaluated using the same noisy test data from the previous experiment. The typical recognition characteristics as the frame advance is increased from 1 ms to 10 ms are shown in Fig. 7.10 for 15 dB SNR. Similar characteristics for other SNR are provided in Appendix B.

For both types of additive background noise, the recognition characteristics



(a)



(b)

Figure 7.10: ASR performance of the proposed PEDEP phase estimation technique for increasing frame advances using (a) AWGN and (b) car noise at 15 dB SNR.

across the range of frame advances are very similar. In general, as the frame advance is increased (i.e. the frame rate is decreased), word accuracy using direct estimation of the noise phase decreases (green line), whilst the accuracy using estimation of the speech phase and interpolating the subtraction result using the tangent method increases (magenta line). This commonality between different noise types suggests that proposed phase estimation technique is applicable to a wide range of background noise and therefore suitable for a wide range of speech recognition applications.

The results of the investigation in Section 7.5.1 showed that the stationarity assumption is better approximated as the frame rate is increased for both speech and noise signals. Therefore, it is anticipated that as the phase estimate using PEDEP becomes more accurate, so too will the resulting clean speech magnitude (N.B. the noise magnitude estimate is the same for all frame rates as it only applies to the 10 ms analysis frames). From the results presented in Fig. 7.10, it is clear that this is the case when using PEDEP to directly estimate the noise phase. In this case, the ASR performance is inferior to both traditional magnitude spectral subtraction and colinear complex spectrum subtraction for the longer frame advances (more than 5 ms) where the stationarity assumption was seen to exhibit considerable sidelobes at phase differences of $\frac{\pi}{4}$. When the frame advances are decreased to 2.5 ms and below, direct CSS with PEDEP begins to outperform both reference methods in all SNR which suggests the phase estimates are sufficiently accurate to justify the use of CSS. Applying a 2.5 ms frame advance (i.e. 400 Hz frame rate), the proposed estimation technique produces average relative word accuracy improvements for all SNR of 6.7% and 4.8% compared to magnitude spectral subtraction in white noise and car noise respectively. If the frame rate is further increased to 1000 Hz (1 ms frame advance), these relative improvements increase to 9.4% and 7.4% respectively, however for this small dataset this increase required approximately six times the processing time compared to a 400 Hz frame rate. Therefore, a 2.5 ms frame advance is seen as an appropriate trade-off between processing time and the resulting improvements in ASR word accuracy.

Using PEDEP to estimate the clean speech phase and interpolating to obtain

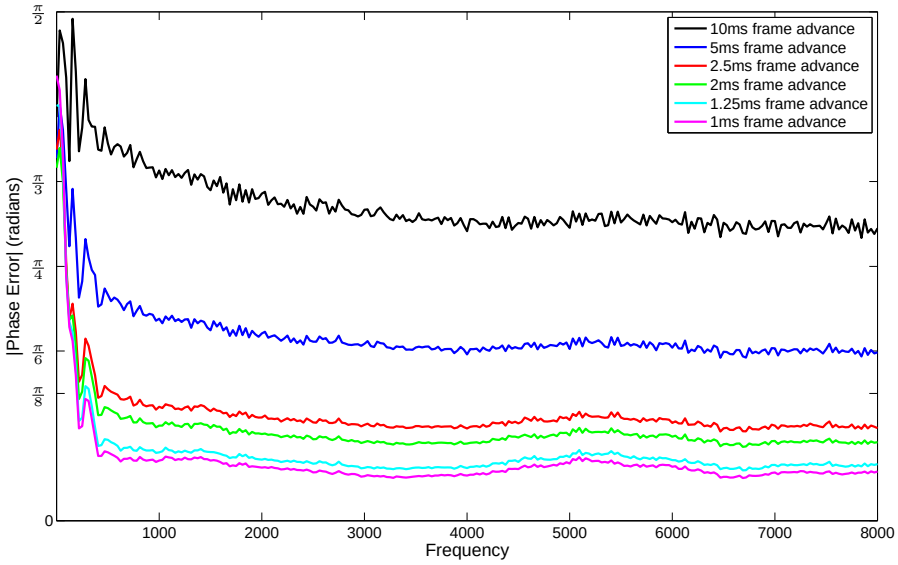


Figure 7.11: Frequency analysis of the average phase error between the true and estimated clean speech phase.

the subtraction result behaves opposite to what was expected given the speech phase characteristic between adjacent frames observed in Fig. 7.8(c). In order to determine the cause of this behaviour, first the accuracy of the estimated clean speech phase was determined. Ten of the noise-corrupted TIMIT test sentences at 10 dB were used to determine the average error between the estimated and true clean speech phase for each frequency. Figure 7.11 demonstrates that the accuracy of the clean speech phase estimate is increasing as the advance between frames is decreased; this is the behaviour which was expected given the results of the previous investigation. Therefore, the cause of the recognition performance seen in Fig. 7.10 is due to the method of interpolation and not the accuracy of the proposed delay-based phase estimation.

Further investigation of the tangent interpolation method reveals a significant flaw in the underlying assumptions. As described in Section 7.4.1, when the speech phase estimate falls between the two circle tangents, the subtraction result is taken as the point which corresponds to the estimated clean speech phase but has the *smallest magnitude*. This operation was based on the assumption that the magnitude of the resulting enhanced signal should be less than the noisy signal magnitude – this assumption is the basis of traditional magnitude spectral

subtraction. The shortfall of this operation occurs when the noise and clean speech signals are significantly out of phase. In this specific case, the original mixing procedure can cause the noisy speech vector to have a smaller magnitude than the clean speech; these cases require "spectral addition" in order to obtain a truer estimate of the clean speech magnitude. This effect was previously explained in Section 7.3.1 and was demonstrated graphically in Fig. 7.4.

Given the shortfalls of this interpolation method when estimating the clean speech phase, direct noise phase estimation appears more reliable for improving speech recognition accuracy at this stage. Improvements to the current tangent interpolation method to cater for cases where the speech and noise phase are considerably out of phase may enable further improvements in word recognition accuracy than what has been shown in these experiments. This is described further in Section 7.6.

7.5.3 "Real-World" ASR Experiments

The oracle experiments in the previous section demonstrated that directly estimating the noise phase spectrum using PEDEP improved ASR performance compared to traditional magnitude spectral subtraction for all frame rates greater than 400 Hz. In those experiments, the true noise phase was always projected forward to the next frame, however this procedure is not possible in "real-world" scenarios where the nature of the additive background noise (and therefore the true noise signal) is unknown.

To implement PEDEP on realistic data, the following algorithm is required.

Algorithm 1 Practical Implementation of PEDEP for Complex Spectrum Subtraction

```
Initialise noise phase reference for each frequency using phase of first frame which should contain only noise.
Initialise reference frame numbers to 1 for each frequency.
for all subsequent frames do
    Perform soft-decision SAD on each frequency.
    if frequency contains only noise then
        Update reference phase to phase of current frame.
        Update reference frame number to current frame.
    else
        Project phase from most recent reference frame forward to current frame.
        if current frame required for analysis then
            Perform complex spectrum subtraction.
        end if
    end if
end for
```

Table 7.1: Speech recognition performance of the practical implementation of the PEDEP phase estimation for complex spectrum subtraction.

		ASR Word Accuracy (%)					
		20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
AWGN	No enhancement	51.2	21.6	8.4	3.6	3.2	1.9
	Magnitude Spec Sub	61.0	38.3	16.3	5.2	7.5	2.4
	Direct CSS + PEDEP	**61.7**	**42.3**	**18.6**	**6.5**	4.7	**2.5**
Car	No enhancement	65.4	56.9	42.4	24.1	5.8	1.7
	Magnitude Spec Sub	68.9	65.0	57.5	33.2	18.9	3.4
	Direct CSS + PEDEP	67.1	64.5	56.8	**35.1**	**19.0**	**3.6**

This algorithm was used with a frame rate of 400 Hz, with analysis frames taken every 10 ms. This frame rate was chosen as the previous discussion deemed it a suitable trade-off between processing time and improvements in ASR word accuracy. The soft-decision SAD is the same as that which has been used to update the noise magnitude estimation throughout this dissertation.

To determine the realistic performance of the PEDEP phase estimator using this intermediate frame rate, the same 100 noise-corrupted TIMIT sentences were used for the speech recognition task. The results for AWGN and additive car noise are shown in Table 7.1. From these results it can be seen that the proposed complex spectrum subtraction with delay-based phase estimation improves word recognition accuracy in comparison with a baseline system without enhancement. Further, in approximately 66% of test SNR across the two different noise types (highlighted bold) the proposed method is able to outperform traditional magnitude spectral subtraction. For white Gaussian noise in particular, the proposed method is superior in most SNR, with the only exception being 0 dB. At 15 dB, the relative improvement in word accuracy for AWGN is 6.5%; this is the maximum improvement of all SNR cases.

Low signal-to-noise ratios are extremely common in vehicular environments due to the use of distant microphones [15]. The proposed CSS technique with phase estimation also shows promise in these environments as it can be seen to outperform magnitude spectral subtraction for SNR $\leq$ 5 dB. At 5 dB, the relative improvement in accuracy is 2.8%.

Whilst the proposed technique fails to outperform magnitude spectral subtraction in all environments, these results confirm that the approach described in Algorithm 1 is suitable for improving ASR performance by using complex spectrum subtraction in the ASR front-end. This result combined with those reported previously demonstrates the use of phase information in spectral subtractive-type speech enhancement can improve clean speech magnitude estimates which ultimately improves ASR performance.

The results presented throughout this section verify the benefits of the proposed PEDEP algorithm and its integration with complex spectrum subtraction in order to improve the performance of traditional spectral subtraction techniques.

7.6 Research Directions

The novel contributions contained in this chapter provide the first step to using phase spectrum information in spectral subtraction, and there are a number of research directions which can be taken to make this approach more effective.

As far as the work contained in this chapter is concerned, there were a couple of questions raised within the discussion that necessitate further investigation. The first of those is the secondary peaks at $\frac{\pi}{4}$ observed for the longer frame advances in the phase difference histograms in Fig. 7.8. Of equal interest is why these peaks tend to decrease in amplitude compared to the main peak at 0 radians as the frame rate is increased, and also why they move towards $\frac{\pi}{6}$. This phenomenon was not regarded as important for demonstrating the concept of the proposed delay-based phase estimation, however with extended knowledge of this occurrence, estimation accuracy may be improved at frame rates closer to those generally used for speech processing. From the perspective of reducing computational complexity, this would be an appealing outcome.

Another issue with the assumption that complex spectrum subtraction with PEDEP can be applied to individual frequencies is the effect of smearing which occurs when signal frequencies do not exactly match DFT centre frequencies. This effect was seen in Fig. 7.7. The assumption of independence between frequencies could be relaxed if the nature of this smearing effect can be determined using

information obtained by techniques such as pitch and true frequency calculation. This information could allow for the smearing effect to be reduced, and could result in more accurate estimates of phase.

In the experiments, it was seen that the speech recognition accuracy when estimating the clean speech phase and interpolating the subtraction output behaved in the opposite manner to that expected given the improved phase estimates as the frame rate was increased. Further analysis discovered that the method of interpolation was not suitable when the clean speech phase fell between the two circle tangent points and the noise and speech signals were significantly out of phase. Future research goals are to determine an approach whereby this interpolation can be improved; this is particularly important during long periods of speech where the speech phase estimate should be more accurate than a projected noise phase. The existing interpolation method could be extended to incorporate either the previous clean speech magnitude estimate or the noise phase estimate. In both cases, this extra information could be used to determine whether spectral subtraction or spectral addition (i.e. when the signals are significantly out of phase) is required on a frame-by-frame and frequency-by-frequency basis. Combining this information could potentially increase ASR performance at the standard speech processing frame rate of 10 ms, a result which is computationally attractive.

In this dissertation, speech enhancement performance has been considered only when it used as part of ASR front-end processing. For human intelligibility applications, the proposed technique presented in this chapter could also improve signal quality in two ways. The first is the improved magnitude estimates which result from using this approach – such improvements were seen throughout the discussion and confirmed by the widespread increase in ASR accuracy when speech features which rely on the magnitude spectrum are used. Another potential improvement for intelligibility applications is the use of an enhanced phase spectrum for reconstruction to the time domain. Traditionally, the noisy speech phase spectrum is left unaltered for reconstruction purposes; the improved phase spectrum information which results from complex spectrum subtraction could be used to improve the reconstruction operation. To test CSS in this particular

application, the effect of both enhanced magnitude and phase spectra can be assessed using Perceptual Speech Quality Measures (PSQM). Clean speech and noise-corrupted data from either the TIMIT test sentences used in this thesis, or from the more comprehensive Aurora experimental framework [57] would be appropriate for such an evaluation.

7.7 Summary

Traditional frequency-domain spectral subtraction fails to utilise phase spectrum information in deriving clean speech magnitude estimates. In this chapter it was shown that without phase spectrum information, the clean speech magnitude cannot be perfectly reconstructed. The errors in the resulting clean speech magnitude when using traditional magnitude spectral subtraction were found to be dependent on the phase of both the noise and speech signals, as well as the instantaneous signal-to-noise ratio.

Given this finding, it was proposed to include phase information into spectral subtractive-type algorithms, and perform the subtraction in the complex frequency spectrum. The estimated complex clean speech signal can be determined by directly estimating the noise phase or interpolating the clean speech phase estimate and combining with the noise magnitude estimate. A novel method termed Phase Estimation via DElay Projection (PEDEP) for estimating either phase spectrum was proposed; this approach is based on the stationarity of sinusoidal waveforms and the delay between observations. A preliminary investigation showed the underlying assumptions were better approximated as the time advance between adjacent frames was decreased.

The results of a proof-of-concept experiment demonstrated that the use of true phase information is beneficial to speech recognition performance when using spectral subtraction in the ASR front-end. This experiment provided an upper bound on the possible recognition performance (with this data) if both phase and magnitude estimation procedures were employed for complex spectrum subtraction. As the time advance between frames was decreased, ASR word accuracy was increased when using PEDEP to directly estimate the noise phase, however

the same was not true for interpolating the clean speech phase estimate. The latter behaviour was attributed to scenarios where the noise and speech signals were significantly out of phase; a solution to this problem is sought in future research.

A final experiment using speech activity detection to assist the noise phase spectrum estimation demonstrated – in a number of noisy conditions – that complex spectrum subtraction with proposed PEDEP algorithm was able to outperform traditional magnitude spectral subtraction. This experiment confirmed the suitability of the proposed approach for coupling spectral subtractive speech enhancement with ASR.

The research undertaken in this chapter should challenge the research community to reconsider the use of phase spectrum information to further improve both automatic speech recognition and human intelligibility in adverse environments.

Throughout this chapter and the two chapters prior, constant reference has been made to the practical implementation of spectral subtractive speech enhancement, particularly in terms of computational requirements. In Chapter 8, these considerations are used to direct a simplification of the traditional frequency-domain spectral subtraction algorithm to enable a resource-efficient implementation on FPGA hardware for in-car ASR applications.

Chapter 8

FPGA Hardware Implementation of Spectral Subtraction

8.1 Introduction

In previous chapters, review of the literature showed significant research effort devoted to creating novel ways of solving the problem of ambient noise on the speech signal. For a number of reasons however, much of this research is not implemented in commercial products. For instance, in the automotive industry, speech solutions need to be effective in a wide range of noise conditions, realisable in low-cost hardware and maintain real-time operation. The need for real-time processing makes many speech enhancement techniques unsuitable (including some proposed in this dissertation), and multi-microphone systems are still too expensive for wide-spread industry adoption at this point in time.

In Chapter 3, limited examples of spectral subtraction specifically applied to noisy signals recorded in an automotive environment were provided however none of these studies proposed any hardware implementations. Traditional magnitude spectral subtraction is an appropriate enhancement method for automotive applications as it requires the installation of only a single microphone, and the processing can be simplified considerably (Section 8.2) to satisfy low-cost and real-time requirements.

The majority of existing automotive electronics are powered by low-cost embedded processors that provide services such as car area networking and human-machine interfaces. To date, Field Programmable Gate Arrays (FPGA) have been used for only a small amount of these electronics primarily due to their higher single-unit cost compared to embedded processors. This difference in cost is becoming insignificant as multiple instantiations of embedded processors and other specialised hardware are possible in a single, modest-sized FPGA, making them suitable for automotive applications. A low-cost implementation of spectral subtraction on FPGA based on the simplified algorithm was performed by researchers at LaTrobe University, Melbourne, Australia (Section 8.3). Verification of this design and evaluation of its ASR performance are performed with reference to an equivalent floating-point model in Sections 8.4 and 8.5 respectively.

8.2 Spectral Subtraction for In-Car Applications

The FPGA implementation of spectral subtraction is based on a modified version of the frequency-domain formulation described in Eq. (3.3):

$$|\hat{S}^i(k)|^\gamma = \begin{cases} |Y^i(k)|^\gamma - a^i(k)|\hat{D}^i(k)|^\gamma & |Y^i(k)|^\gamma - a^i(k)|\hat{D}^i(k)|^\gamma > \beta|Z^i(k)|^\gamma \\ \beta|Z^i(k)|^\gamma & \text{otherwise} \end{cases} \tag{8.1}$$

where $|Z^i(k)|$ is either the instantaneous noisy speech signal magnitude $|Y^i(k)|$ or the noise magnitude estimate $|\hat{D}^i(k)|$. In this instance, the resulting clean speech magnitude estimate $|\hat{S}^i(k)|$ is recombined with the noisy signal phase $e^{j\angle Y^i(k)}$ for synthesis to the time-domain, which enables the enhanced signal to be used for playback or as input to further speech processing such as ASR using a commercial speech recognition engine.

The subtraction process described by Eq. (8.1) requires a lot of real-time multiplications since the frequency-dependent subtraction factors – and potentially the noise floor – are calculated on a frame-by-frame basis (i.e. every 10 ms). For a real-time, low-cost hardware implementation on FPGA the following simplifications are proposed:

1. Assuming the noise estimate $|\hat{D}^i(k)|^\gamma$ is sufficiently accurate, the frequency-dependent subtraction factors $\alpha^i(k)$ – which are introduced to reduce the effects of inaccurate noise estimates – are not required. Therefore, $\alpha^i(k)$ is set to 1 for all frames i and DFT frequencies k.

2. Assuming the initial N frames of each recording contain only noise (i.e. no speech components), the average of these initial silence frames produces a noise estimate which remains stationary for the remainder of the recording. Therefore, the noise estimate is only calculated prior to any signal enhancement and can be represented as $|\hat{D}(k)|^\gamma$ (N.B. the relaxation of the dependency on the frame number i). For simplicity in the hardware design, 8 frames are chosen as sufficient to calculate the noise estimate, allowing a simple 3-bit shift for determining the average of these frames. In order to maintain real-time processing, these 8 frames are discarded from the output waveform to avoid buffering in order to apply spectral subtraction to these frames after the estimate has been calculated.

3. Having generated a constant noise estimate $|\hat{D}(k)|$, this estimate can be used to represent $|Z^i(k)|$ in Eq. (8.1) for calculating the noise floor. Therefore, the noise floor also remains constant for the entire utterance instead of constantly changing through scaling of the noisy signal magnitude $|Y^i(k)|$.

Following these simplifications to Eq. (8.1), the spectral subtraction equation used in the FPGA implementation outlined in Section 8.3 is:

$$|\hat{S}^i(k)|^\gamma = \begin{cases} |Y^i(k)|^\gamma - |\hat{D}(k)|^\gamma & |Y^i(k)|^\gamma - |\hat{D}(k)|^\gamma > \beta|\hat{D}(k)|^\gamma \\ \beta|\hat{D}(k)|^\gamma & \text{otherwise} \end{cases} \tag{8.2}$$

Equation (8.2) leaves only two parameters (γ and β) to be further optimised for the FPGA implementation. Common values for these parameters were noted in Section 3.3.1. The values of γ are typically used for their conceptual meaning as opposed to ASR performance whilst β is often chosen to optimise SNR given a particular value of γ. In [68] it was established that in-car ASR performance differs greatly with various combinations of γ and β; therefore these values must be chosen carefully.

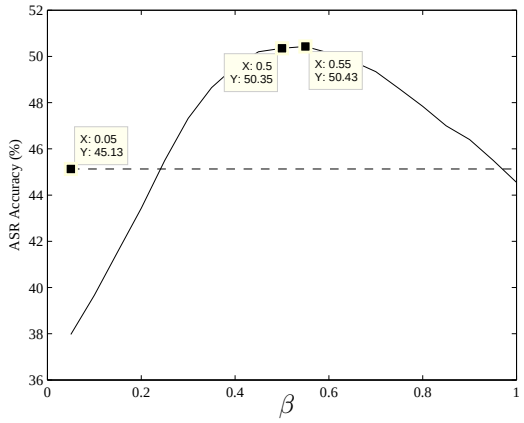


Figure 8.1: The effect of the noise floor scaling factor, β, on ASR accuracy averaged over a range of automotive noise conditions.

In order to reduce the processing requirements of the FPGA implementation in Section 8.3, magnitude spectral subtraction ($\gamma = 1$) was chosen. Using this parameter value avoids the need for the FPGA to perform resource-intensive square and square-root operations. Further, comparable ASR accuracy can be obtained for magnitude and power spectral subtraction if the values of β are optimised for each value of γ.

Using a suitable floating-point model, preliminary experiments similar to those reported in [68] were performed to determine the optimal value of β to use in the FPGA implementation. Using the first 5 experimental folds from the AVICAR evaluation protocol [66], values of β were varied in linear increments through the range [0, 1] with $\gamma = 1$. The average combined results for all noise conditions in the AVICAR database are shown in Fig. 8.1. It should be noted that the individual noise conditions exhibited similar characteristics to Fig. 8.1; this permits a constant β value to be applied to all in-car noise scenarios.

From this figure it can be seen that a wide range of (but not all) β values lead to improvements in word accuracy over a system with no enhancement (shown by the dashed line). Maximum recognition accuracy can be obtained by setting $\beta = 0.55$, however this performance is only marginally better than at $\beta = 0.5$ (less than 0.1%). As a result, a value of $\beta = 0.5$ was chosen for the FPGA implementation as this value is easily and accurately represented in fixed-point notation.

8.3 Hardware-Based Speech Enhancement

Researchers at LaTrobe University, Melbourne, implemented the simplified spectral subtraction algorithm presented in Section 8.2 on both a Xilinx Virtex-4 SX FPGA [146] and a Xilinx XA Spartan-3A DSP 1800A FPGA [147]. The latter device is the general production equivalent of its Xilinx Automotive cousin; therefore successful implementation on this device demonstrates the capability for implementation in an automotive-grade FPGA.

The FPGA design process consisted of the following steps:

1. Development of a MATLAB version of the spectral subtraction algorithm (Section 8.2) using high-precision, complex floating-point arithmetic.

2. Conversion to a fixed-point (data and operations) implementation in MATLAB, mirroring the major blocks expected in the FPGA implementation.

3. Comprehensive testing of the fixed-point MATLAB design against the floating-point version, both block-by-block and at the complete system level.

4. Implementation of the fixed-point design as Xilinx System Generator^{TM} (XSG) models.

5. Comprehensive testing of each major block of the XSG design against its fixed-point MATLAB equivalent, and testing of the complete XSG model against both the fixed-point and floating-point MATLAB versions.

6. From the completed XSG model a hardware description language representation was generated, synthesised using Xilinx ISE 9.2 tools, and implemented on the higher-end Xilinx Virtex-4 SX FPGA.

7. Following a check of the FPGA resource usage of the design, the XSG model was analysed block-by-block to identify resource inefficiencies and refined to use more appropriate resources.

8. Performance of the Virtex-4 realisation was checked against the XSG and floating-point models by comparing output waveforms for common input.

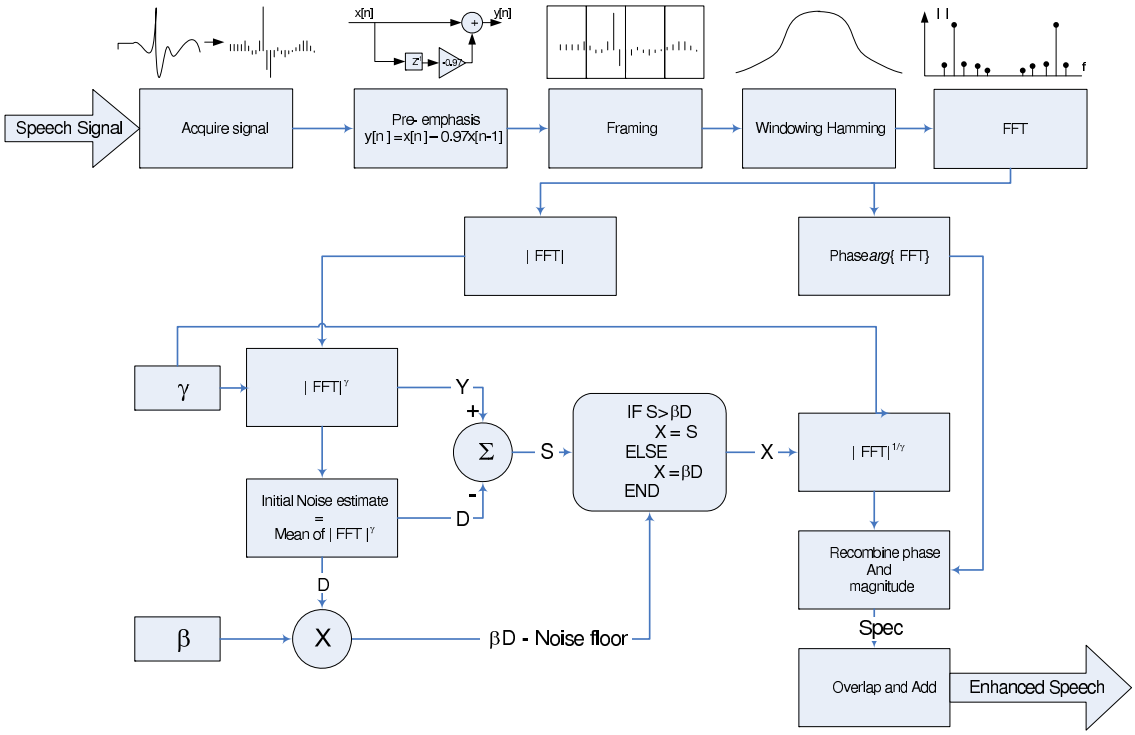


Figure 8.2: Block diagram of hardware implementation of spectral subtraction algorithm.

9. The validated design was synthesised into the low-end Xilinx Spartan-3A DSP FPGA, and tested against the Virtex-4 implementation on a sample-by-sample basis for a range of signal inputs including basic ramps, modulated chirps and speech samples.

Figure 8.2 shows the block diagram of the FPGA implementation of the spectral subtraction algorithm. Specific details of the implementation and optimisation are not provided here – the reader is directed to the publications resulting from this work for such details [146, 147].

8.4 Design Verification & Resource Usage

8.4.1 Verification

The accuracy of the FPGA implementation was verified using the USB test harness developed in [146]. Various signals were passed through the FPGA and

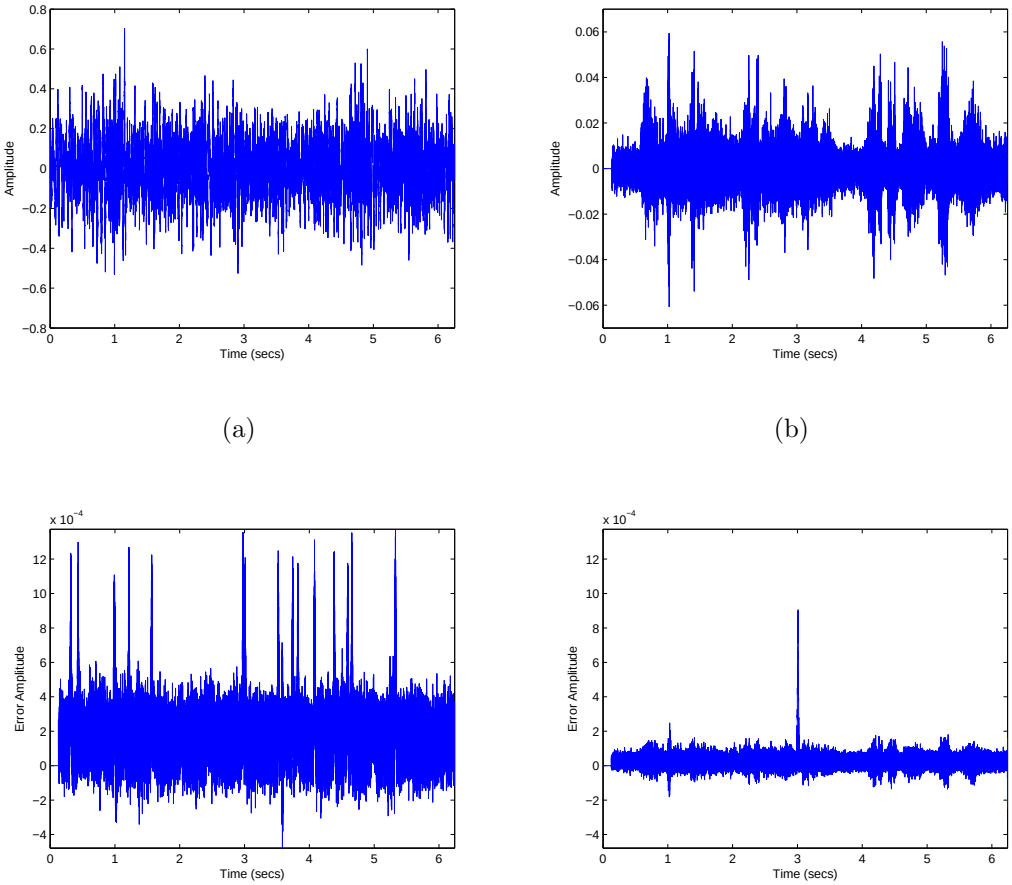


Figure 8.3: (a) Noisy speech signal from AVICAR database, (b) output of spectral subtraction algorithm, (c) difference between floating-point and initial FPGA design, and (d) difference between floating-point and optimised FPGA design.

floating-point models, and were compared on a sample-by-sample basis. An example of the output waveforms of one such test on a speech sample from the 35 mph with window down noise condition of the AVICAR database is shown in Fig. 8.3. The original waveform is shown in Fig. 8.3(a), and the corresponding spectral subtraction output in Fig. 8.3(b). The sample-by-sample differences of the original FPGA design in [146] and the optimised design in [147] with respect to the floating-point output are shown in Fig. 8.3(c) and Fig. 8.3(d) respectively.

Observing the waveforms in Figs. 8.3(a)-8.3(b), it can be seen spectral subtraction provides noticeable signal enhancement of the noisy in-car speech. Prior to enhancement, the time-domain structure of the speech signal is not visible –

Table 8.1: Spartan-3A DSP 1800A FPGA resource usage summary.

Resource Type	Available	Usage (%)	
		Initial	Optimised
Slices	16640	1622 (9%)	2196 (13%)
Flip Flops	33280	2581 (7%)	3093 (9%)
4-input Look-Up Tables	33280	2419 (7%)	3010 (9%)
Block-RAM	84	10 (11%)	10 (11%)
Digital Clock Manager	8	1 (12.5%)	1 (12.5%)
DSP48	84	21 (25%)	25 (29%)

after enhancement the regions of speech are more pronounced. This observation was further validated through analysis of the output spectrograms in [146].

Analysing the difference signals created by the two FPGA designs (Figs. 8.3(c)-8.3(d)), the average gain of 13.5 dB represents an approximate improvement of 2-bits between the initial and optimised designs. The continued presence of spikes in the difference signal (e.g. around the 3 second mark in Fig. 8.3(d)) are attributed to the Xilinx FFT block outputting a quantised version of its internal scaling factor which occasionally leads to larger sample errors. Through optimisation of the FPGA design (i.e. from Fig. 8.3(c) to Fig. 8.3(d)), the frequency of these spikes was reduced to less than 1 in every 200 samples.

8.4.2 Resource Usage

Table 8.1 shows the total resources required to implement the spectral subtraction algorithm design in a Spartan-3A DSP FPGA. The "initial" and "optimised" designs are identical in terms of the design architecture – the only difference is the bit resolution used within the designs. Overall, the initial design used 9% of the total (i.e. general FPGA logic fabric) slices available, and 25% of the DSP48 XtremeDSP^{TM} blocks. The larger percentage use of the DSP48 blocks is expected due to the intensive DSP requirements of the algorithm (namely the FFT/IFFT block). The percentage use of other key resources, block-RAM and digital clock manager blocks is of a similar level to the slice usage.

Each sub-block of the original design was optimised so that the number of bits

used was sufficient for processing a range of types of speech input with minimum quantization error and no arithmetic overflow. This resulted in a slight increase in resource usage to 13% of slices and 29% of DSP48 blocks due to the use of larger bit widths within some sub-blocks.

The relatively low resource usage shown here enables other in-car services (such as car area networking, human-machine interfacing or other speech processing) to be incorporated into the same FPGA which will assist in minimising overall manufacturing costs for these services.

8.5 Experimental Results & Discussion

To test the true effectiveness of the FPGA implementation for use in in-car speech recognition, the FPGA processed waveforms were evaluated under the speech recognition protocols outlined in Chapter 4 and compared to a floating-point equivalent of the spectral subtraction algorithm. It should be noted that the results shown here are different from those presented in Chapter 4 since the first 8 frames are removed from all signals in order to reflect the hardware processing during the noise estimation period. Speech recognition results are shown in Tables 8.2 and 8.3 for the AVICAR and Australian English In-Car Speech databases respectively. The results for the AEICS corpus are for command and navigation tasks combined – it was not deemed necessary to separate them for this validation.

Analysing the results in these tables it can be seen that all versions of the spectral subtractor provide improvements in recognition performance across the

Table 8.2: ASR results (% word accuracy) for FPGA validation on the AVICAR database.

	ASR Word Accuracy (%)				
	IDL	35U	35D	55U	55D
No Enhancement	71.5	49.6	37.2	42.8	24.6
Floating-Point	74.8	54.7	40.9	50.7	30.7
Initial FPGA	70.5	54.9	41.6	50.7	30.8
Optimised FPGA	74.7	54.8	40.9	50.6	30.7

Table 8.3: ASR results (% word accuracy) for FPGA validation on the AEICS database.

	ASR Word Accuracy (%)						
	C0	C6	C1	C2	C3	C4	C5
No Enhancement	84.9	41.2	69.7	34.1	53.0	53.9	30.5
Floating-Point	86.9	52.8	76.2	48.3	60.6	61.5	45.2
Initial FPGA	86.2	52.8	70.6	47.6	34.2	60.5	45.7
Optimised FPGA	86.9	52.9	76.2	48.4	59.8	61.6	45.4

full range of in-car noise scenarios. Most importantly, the optimised FPGA design performs closely to the floating-point algorithm, proving this hardware design is more than suitable for in-car speech recognition systems.

The exception to this observation is the 50-60 km/h with window down condition in the AEICS database. It can be seen that the initial FPGA design failed to deal with the noise and word accuracy performance reduced by almost 20% from the no enhancement case. Further analysis showed this noise condition is highly susceptible to microphone vibration due to wind (from the open window) which causes very high amplitude values in the low-frequency range (compared to higher frequencies). These high amplitudes were unable to be handled by the lower-precision FPGA design due to overflow in some of the hardware blocks, particularly the FFT/IFFT block. This shortfall of the original design was corrected in the optimised FPGA design where the speech recognition performance is only 0.8% inferior to the floating-point version but still improves on the case without enhancement by almost 7%.

8.6 Research Directions

Despite the considerable improvements in speech recognition accuracy provided by the optimised FPGA design, the greater deviation from the floating-point model in the 50-60 km/h with driver's window down (C3) warrants further investigation. Despite the optimised design considerably reducing the effect of high amplitudes in the low-frequency components of the signal, it appears that the output from the FFT block – which has been set to the maximum bit-width

available from the Xilinx IP core used in this design – is still experiencing some arithmetic overflow, causing the noise estimation and subtraction processes to become less accurate.

The spikes observed in the difference between floating-point and FPGA designs (Fig. 8.3(c)-8.3(d)) appear to be another artefact resulting from limitations of the FFT block. Using the maximum bit-width available has reduced the occurrence of these significant deviations, but has not eliminated them. To improve the design further and correct these two problems, a new, higher resolution FFT/IFFT block would be needed – the implementation of which would require significantly more FPGA resources. Alternatively, a redesign of the pre-emphasis filter for greater attenuation at low frequencies could lead to some improvement at a more modest increase in resources.

Despite the already low resource usage demonstrated in Section 8.4.2, the spectral subtraction implementation presented in this chapter could be made even more resource efficient through the use of the colinear complex spectral subtraction method presented in Chapter 7. This method removes the need for the flooring operation by assuming the speech and noise signals are in-phase (i.e. colinear). This method was shown to provide similar speech recognition performance to traditional magnitude-based spectral subtraction, and could be used in this application as a way of decreasing FPGA resource usage.

Preliminary experiments with floating-point models have shown that combining a dual-channel delay-and-sum beamformer with spectral subtraction acting as a beamformer post-filter can yield even better ASR performance. Work undertaken in parallel with this research successfully implemented a dual-channel delay-and-sum beamformer in a Xilinx Virtex-4 FPGA [152]. Having shown low resource usage in each of these two designs, and acknowledging the common processing blocks between the two enhancement techniques (i.e. pre-emphasis filtering, framing and FFT/IFFT), a low-cost FPGA implementation of this combination will further improve in-car speech recognition.

8.7 Summary

In this chapter a simplified spectral subtraction algorithm has been presented which has been designed specifically for low-cost FPGA hardware implementation in automotive environments. This algorithm uses an initial silence period to calculate the noise estimate which is assumed stationary and accurate throughout the speech recording and is also used in the noise flooring operation. Brief details of the resulting implementation in a Xilinx XA Spartan-3A 1800A DSP FPGA have been provided, with verification through waveform analysis on in-car speech samples validating the effectiveness of the design. Speech recognition experiments further validate the optimised FPGA design; results show it is able to perform within 0.1% of a floating-point equivalent in almost all in-car noise conditions.

A number of future research directions based on the current implementation have also been discussed including further analysis of artefacts observed in the output waveforms, and analysis of the windows down noise condition which causes the microphones to vibrate. Future implementations could include simplification of the algorithm through the removal of the noise flooring operation, or incorporation of this implementation as a post-filter for a dual-channel delay-and-sum beamformer.

In the previous four chapters, a number of novel contributions in the field of robust speech recognition using speech enhancement have been presented. A number of potential future research directions have also been proposed to extend the work contained in this dissertation. The next chapter summarises the overall contribution of this dissertation, and highlights the most significant of the proposed research directions.

Chapter 9

Conclusions and Future Research

9.1 Introduction

This chapter summarises the work presented in this dissertation. Particular reference is made to the primary aims and the novel contributions which were introduced in Chapter 1 are highlighted in detail. A summary of the major avenues for future research is also provided.

9.2 Conclusions

9.2.1 General Findings

The scope of this research was to investigate the performance of single-channel speech enhancement for robust speech recognition in real automotive environments. The spectral subtraction algorithm was chosen as the enhancement technique of interest due to its common usage throughout the speech research community, and also its computational simplicity which was seen to be important for in-car hardware implementations.

Within this scope, the primary aims of this dissertation as defined in the introductory chapter were:

1. To demonstrate that speech enhancement techniques optimised for human intelligibility are sub-optimal for integration with state of the art speech recognition systems.

2. To propose novel techniques which improve current speech enhancement algorithms when used as part of the front-end processing for in-car ASR.

3. To consider the implementation of speech enhancement algorithms within the constraints of the automotive environment.

Specific research objectives were also documented in order to achieve each of the primary aims. Each of these objectives are listed below, along with the major findings of this thesis:

1. *To quantify the word accuracy performance of ASR systems when speech is collected in real car environments and is therefore corrupted by a wide range of in-vehicle noise conditions.*

   The performance of automatic speech recognition in automotive environments degrades rapidly as the level of noise is increased. In very noisy conditions (such as driving at 55 mph with the windows down) the word accuracy can drop to as low as 25% which is well below consumer expectation. Increases in vehicle speed, changes to window positions and air-conditioning systems were all seen to provide major challenges for in-car environments. Appropriate microphone placement was judged as an important consideration when attempting to deal with noise from the open window and/or the air-conditioning vents.

2. *To analyse the effectiveness of traditional speech enhancement and model adaptation techniques for increasing noise-robustness of ASR systems.*

   Various implementations of spectral subtraction were able to improve the speech recognition performance of a wide range of in-car noise conditions, increasing the worst-case scenario to approximately 30% word accuracy. Model adaptation was found to be a more effective method for robust in-car ASR, with worst case recognition increasing to approximately 60%. Combining the two approaches resulted in varying recognition responses; the conflict between enhancement and adaptation approaches was found to be particularly problematic for conditions with lower levels of background noise.

3. *To analyse how subtractive-type enhancement algorithms have been previously used for speech enhancement and ASR and identify the shortfalls of these approaches in terms of the resulting ASR performance.*

   Through an expansive literature review, it was discovered that the majority of spectral subtractive-type algorithms were designed initially to improve human intelligibility rather than ASR. This is the direct result of considering the speech enhancement and recognition systems as separate entities. In recent studies, and the work contained in this dissertation, it was observed that optimisation for signal-level criteria was sub-optimal for speech recognition applications. As a result, it was enforced that the nature of both the enhancement and recognition systems need to be considered when designing a robust ASR front-end.

4. *To propose novel speech enhancement algorithms which directly improve the performance of the underlying speech recognition engine.*

   Two separate novel contributions were made within the dissertation to address this objective – the application of likelihood-maximising (LIMA) speech enhancement to the more computationally efficient Mel-filterbank noise subtraction (Chapter 5), and the introduction of phase spectrum information which incorporates a pioneer phase estimation procedure termed Phase Estimation via Delay Projection (Chapter 7).

5. *To design frameworks which are suitable for integration with existing in-car speech systems, and where possible, are designed with computational and hardware requirements in mind.*

   Two novel contributions were made within the dissertation to address this objective – LIMA-based speech enhancement which fully utilises the characteristics of speech dialogue system interaction (Chapter 6), and the simplification of frequency-domain spectral subtraction for low-resource implementation on FPGA hardware (Chapter 8).

6. *To assess each of the proposed techniques and report their performance based on speech recognition accuracy and computational requirements.*

   Each of the novel contributions were assessed in terms of their ASR word accuracy performance; each approach demonstrated improvements over baseline enhancement techniques in a wide range of noise conditions. For LIMA-based speech enhancement in Chapter 5, a comparative evaluation of the required processing time for the observed levels of ASR performance was also performed. Details of both performance aspects can be found in the relevant contributions in Section 9.2.2.

9.2.2 Summary of Original Contributions

Some of the major findings detailed in the previous section directed the novel contributions which have arisen from this research. These contributions are spread between Chapters 4-8. The following sections detail each of novel contributions and provides references back to the relevant chapters/sections.

Major Contributions

1. **The collection and validation of the first in-car speech database recorded with native Australian speakers in Australian driving conditions.**

   Prior to this dissertation, no in-car speech data existed within Australia. Consequently, this dissertation collected the Australian English In-Car Speech corpus (documented in Section 4.2.2 and [67]) which incorporates multi-channel recordings of 50 native Australian English speakers (30 male, 20 female) in seven different noise environments common to current Australian driving conditions. An evaluation protocol enabling speech recognition experiments on command-and-control and navigation tasks was also developed. This protocol, which enables model adaptation, development and evaluation testing was validated experimentally, and it was seen that acoustic model adaptation from a well-trained American English acoustic model produced a minimum 8% absolute improvement in ASR performance.

This corpus will have a major impact on both the Australian spoken language community and the automotive industry.

2. **Application of Mel-filterbank noise subtraction to a likelihood-maximising speech enhancement framework specifically for in-car speech recognition.**

   Previously, enhancement techniques have been optimised based on signal-level criteria such as signal-to-noise ratio (SNR) or minimal speech distortion which is sub-optimal for speech recognition applications. The LIMA approach optimises the parameters of an enhancement technique specifically for improved speech recognition accuracy. In this dissertation, it was shown theoretically that the application of the LIMA framework to Mel-filterbank noise subtraction produces a computationally more efficient solution – O(log(K)) versus O(K)+O(K) – when compared to an equivalent frequency-domain Multi-band spectral subtraction approach (Chapter 5).[1]

   It was previously reported that varying dynamic ranges of the cepstral coefficients is detrimental to gradient descent optimisation. As a solution to this problem, cepstral liftering was incorporated into the LIMA framework and was shown to be computationally inexpensive, however failed to show any improvements in ASR performance. It was hypothesised that the cepstral lifter used in this dissertation is not suited to in-car noise environments; the search for an appropriate lifter is sought in future research.

   Unique to previous spectral subtraction implementations within the LIMA framework, the use of *both* oversubtraction and spectral flooring factors (α and β respectively) was proposed and mathematically derived (see Appendix A) and was shown to provide superior speech recognition accuracy compared to other parameter combinations. Absolute improvements in word accuracy performance ranging from 0.5% to 3.4% over an optimised implementation of MFNS were observed.

   Previous application of the LIMA framework has almost solely relied on

[1] These two LIMA-based systems were unable to be compared experimentally due to implementation issues.

a calibration session which – for in-car environments – would be required for each speaker in each noise condition in order to provide the best ASR accuracy. For practical implementation in car environments, this dissertation identified that a grounding process is necessary in speech dialogue systems to confirm the desired human responses prior to performing actions such as route navigation. A new framework was subsequently proposed in Chapter 6 which utilises this grounding procedure to direct the parameter optimisation process which avoids the need for an infinitely large number of enhancement parameter sets. Rather than using an explicit adaptation session, the dialogue-based framework waits for affirmative feedback from the driver before optimisation can occur. This mode of operation was seen to be robust to continually changing noise conditions between utterances; a scenario which was developed in order to simulate normal operation of the vehicle. Improvements in ASR word accuracy of between 1.2% and 2.8% relative to the traditional calibration-style framework were observed despite optimisation occurring only 3% of the time (due to the 100% accuracy requirement). This framework has the added advantage over calibration-style frameworks in that optimisation (a lengthly computation) occurs without the user's awareness, and optimisation only occurs on state sequences that are *known* to be correct rather than hypothesised.

3. **The use of the short-time phase spectrum to improve the ASR performance of frequency-domain spectral subtraction.**

   For the past 30 years, the phase spectrum in spectral subtractive-type algorithms has been disregarded as it was originally shown to provide no useful information for human intelligibility applications. Given the focus of this research being on speech enhancement specifically for ASR, Chapter 7 demonstrates that ignoring the phase spectrum information leads to errors in the cleaned magnitude spectrum, information which is commonly used for speech feature extraction in ASR applications (e.g. in MFCCs). It was shown that clean speech magnitude estimates are particularly sensitive to decreases in SNR as well as the degree of phase difference between the noise

and clean speech signals.

Magnitude spectral subtraction was therefore reformulated to be performed in the complex frequency domain which incorporates both magnitude and phase information. Oracle-type experiments demonstrated that with perfect phase spectrum estimation, average relative improvements in word accuracy of approximately 9.5% could be obtained in both additive white Gaussian and car noise for a range of SNR on a subset of the TIMIT database. These experiments justified the need for phase information within the algorithm.

In order to perform Complex Spectrum Subtraction, it is necessary to estimate either the noise or clean speech phase spectrum; no research to date has proposed a method for achieving this. Consequently, a novel estimation method termed Phase Estimation via DElay Projection (PEDEP) was proposed in Section 7.4. This approach uses the assumption of stationarity of sinusoidal signals (which are themselves assumed to be accurately represented by the DFT centre frequencies) and the known time delay between adjacent frames to project forward a known reference phase to a new time instance. This technique was shown to improve monotonically as the frame advance was reduced from the standard processing rate of 10 ms to 1 ms, demonstrating better approximation of the stationarity assumption. Speech recognition performance on an oracle task showed that estimating the noise phase produced average global improvements over optimised magnitude spectral subtraction of 4.8% for car noise and 6.7% for additive white Gaussian noise when the frame advance was reduced to 2.5 ms. Further improvements in ASR performance were observed when further reducing the frame advance, however the computational overhead incurred was deemed greater than the improvements in word recognition accuracy. The performance of the proposed method when estimating the clean speech phase was inconclusive, and corrections to the algorithm are sought in future research.

The proposed Complex Spectrum Subtraction incorporating PEDEP algorithm was incorporated with soft-decision speech activity detection in order

to realise the proposed approach in real-world conditions. The ASR performance of this practical implementation was seen to improve on magnitude spectral subtraction in 66% of the tested noise conditions, with a maximum relative word accuracy improvement of 6.5% for AWGN at 15 dB.

The proposed technique, whilst incorporating moderate amounts of extra computation also has the added advantage of removing some of the enhancement parameters of traditional spectral subtraction. This makes it an attractive alternative to algorithms which require careful tuning for every encountered operating environment.

Other Contributions

1. **A speaker-independent, continuous ASR evaluation protocol for the AVICAR database.**

   The freely-available AVICAR database is distributed without an accompanying evaluation protocol for the continuous speech phone numbers and TIMIT sentences tasks. This research developed a new protocol for these two tasks which allows for model adaptation, development and evaluation testing. The effectiveness of this protocol for speech recognition was confirmed experimentally (both Chapter 4 and [66]). Importantly, this protocol allows for reliable comparisons between both single- and multi-channel enhancement techniques which is often difficult with existing easily-available, large-scale corpora. The evaluation protocol has been made publicly available to the wider research community.

2. **Evaluation of LIMA-based enhancement on test data incorporating multiple layers of acoustic mismatch.**

   LIMA-based speech enhancement was previously demonstrated to be suitable for use with noise-adapted acoustic models. No studies however, had assessed the effect of a second level of acoustic mismatch between training and testing data. In Chapter 5, a second level of mismatch due to unseen speaker dialects was simulated using the AEICS corpus. It was observed that for dialect-mismatched clean speech models, the ASR performance of

the LIMA framework was inferior to an equivalent static enhancement parameter system in all in-car noise conditions. Even though noise was being removed through speech enhancement (as this system outperformed the unprocessed noisy speech), the optimisation process was performed using unreliable state sequences due to the remaining mismatch between training and testing dialects. Performing model adaptation to account for this extra mismatch was unable to correct this performance characteristic due to insufficient coverage of the triphone model space. The experiments in this dissertation have demonstrated the sensitivity of the LIMA framework to a second layer of acoustic mismatch. Correcting this shortfall of likelihood-maximisation is seen as an important future research direction.

3. **Simplification of the frequency-domain spectral subtraction algorithm for cost-effective, real-time implementation in FPGA hardware.**

   Prior to this work, few reports of implementation of speech enhancement algorithms in FPGA hardware existed in the research community. This project – in conjunction with researchers at LaTrobe University, Melbourne, Australia – developed an appropriate implementation of spectral subtraction for use in automotive-grade FPGAs.

   The work performed in this dissertation focused on reducing the computational complexity of traditional frequency-domain spectral subtraction and its associated noise estimation procedure in order to enable a low-resource FPGA solution. This resulted in a solution which:

   - relies solely on the initial silence period for noise estimation;
   - assumes the noise estimate to be sufficiently accurate; and
   - utilises the constant noise estimate to generate the spectral floor.

   The resulting FPGA solution (implemented by LaTrobe University – see Chapter 8 and [146, 147]) matches the ASR performance of an equivalent floating-point model to within 0.2% in over 90% of the evaluated noise conditions. Further, this implementation produced global improvements in

ASR performance over a system without speech enhancement. The overall resource consumption on the chosen automotive-grade Spartan-3A DSP FPGA was less than 30% which enables this solution to be integrated with other in-car electronics in a single FPGA.

9.3 Future Work

A number of future research directions were proposed at the end of each chapter, particularly those chapters containing original contributions. In this section it has been chosen to highlight two major research focuses which extend the work in this dissertation and which will contribute most to the body of scientific knowledge. These two areas are derived one each from the work on likelihood-maximising speech enhancement, and the use of the phase spectrum for spectral subtraction.

1. The sensitivity of the likelihood-maximising approach to the nature of both the training and testing data was uncovered in this research when a second type of acoustic mismatch was present. This observation highlighted a concern about the overall sensitivity of this technique to other aspects of the acoustic model for the LIMA-based enhancement approach. For instance, how does the performance differ when the basic model units (i.e. triphones, phones, or broad phonetic classes) are changed? How much test data is required to provide sufficient coverage of the acoustic model space? How are the effects of secondary acoustic mismatches counteracted? Such questions are yet to be answered, with researchers choosing a certain acoustic model and testing data without a broader perspective of real-world operating conditions. The work contained in this dissertation is no exception. Full understanding of these sensitivities is essential to devising a solution which is applicable to *any* speech recognition system in *any* adverse operating environment.

2. This dissertation highlighted the importance of phase spectrum information for estimating the true clean speech magnitude in spectral subtraction.

An original attempt at estimating phase was investigated and showed significant promise in solving this problem. The simplicity of this approach however, means there is considerable room for improvement. Of particular importance are methods which can improve the subtraction process during speech periods, namely by improving the method of interpolating the clean speech phase to an appropriate spectral subtraction result. Further, the current state of implementation requires that the frame rate be increased to 400 Hz (i.e. four times the standard speech processing frame rate) – methods which can improve the performance of this approach at standard frame rates are therefore of particular practical and commercial significance.

In general, this dissertation has shown that by closely coupling speech enhancement processing with ASR systems, speech recognition performance can be improved. This method of thinking applies to any speech enhancement technique, and some approaches to this problem such as the likelihood-maximisation framework discussed in this dissertation are applicable to almost any enhancement scheme. Adopting a unified approach for improving robust speech recognition is an essential step forward in reaching the science-fiction dream of speech recognition systems which meet the expectations of customers in *any* conceivable environment.

Appendix A

Derivation of the Jacobian Matrix for LIMA-Based Mel-Filterbank Noise Subtraction

A.1 Introduction

The Jacobian matrix is required for any speech enhancement algorithm applied under the likelihood-maximisation framework. It consists of the partial derivatives of the observed feature vector $\boldsymbol{o}$ with respect to the set of speech enhancement parameters $\boldsymbol{\xi}$. In this appendix, the Jacobian elements for Mel-filterbank noise subtraction are derived for an HMM-based recognition system using MFCC feature extraction.

The Jacobian matrix $\boldsymbol{J}_i$ in frame i is a $P \times C$ matrix where P denotes the length of the set of enhancement parameters $\boldsymbol{\xi} = \{\xi_0, \xi_1, ..., \xi_P\}$ and C denotes the length of the feature vector $\boldsymbol{o}$. Therefore:

$$\mathbf{J}_i = \frac{\partial \mathbf{o}_i}{\partial \boldsymbol{\xi}} = \begin{bmatrix} \frac{\partial o_0^i}{\partial \xi_0} & \frac{\partial o_1^i}{\partial \xi_0} & \cdots & \frac{\partial o_{C-1}^i}{\partial \xi_0} \\ \frac{\partial o_0^i}{\partial \xi_1} & \frac{\partial o_1^i}{\partial \xi_1} & \cdots & \frac{\partial o_{C-1}^i}{\partial \xi_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial o_0^i}{\partial \xi_{P-1}} & \frac{\partial o_1^i}{\partial \xi_{P-1}} & \cdots & \frac{\partial o_{C-1}^i}{\partial \xi_{P-1}} \end{bmatrix}. \tag{A.1}$$

To derive the expressions for each Jacobian element, it is necessary to consider both the feature extraction process and the speech enhancement algorithm. In this appendix the feature extraction process is MFCCs as described in Chapter 2 and the MFNS speech enhancement algorithm described in Chapter 3.

A.2 Computing the Elements of the Jacobian Matrix

For any of the enhancement parameters in the MFNS speech enhancement algorithm, the Jacobian elements are calculated as:

$$\frac{\partial o_c^i}{\partial \boldsymbol{\xi}} = \sum_{l=0}^{L-1} \frac{\Phi_{cl}}{E^i_{\hat{S}}(l)} \frac{\partial E^i_{\hat{S}}(l)}{\partial \boldsymbol{\xi}}. \tag{A.2}$$

In this section, the partial derivative terms $\frac{\partial E^i_{\hat{S}}(l)}{\partial \boldsymbol{\xi}}$ are derived for both the subtraction factors α, and the flooring factor β.

A.2.1 Subtraction Factors, α_l

In this instance, the set of enhancement parameters $\boldsymbol{\xi}$ is defined as:

$$\boldsymbol{\xi} = \{\alpha_1, \alpha_2, \ldots, \alpha_L\} \tag{A.3}$$

where a subtraction factor α is applied to each Mel-filterbank (N.B. in this instance $L = P$). Throughout the derivation, all references to the l^{th} filterbank are dropped for simplicity since the derivation applies to each filterbank independently of all other filterbanks.

To calculate the gradient term $\frac{\partial E^i_{\hat{S}}}{\partial \boldsymbol{\xi}}$, first recall the equation for Mel-filterbank noise subtraction:

$$\hat{E}^i_S = \begin{cases} E^i_Y - \alpha E^i_{\hat{D}} & E^i_Y - \alpha E^i_{\hat{D}} > \beta E^i_Y \\ \beta E^i_Y & \text{otherwise} \end{cases} \tag{A.4}$$

To calculate the required partial derivatives, the second case in Eq. (A.4) is

made to be 0 by subtracting the term βE_Y^i from all sides:

$$\hat{E}_S^i - \beta E_Y^i = \begin{cases} E_Y^i(1-\beta) - \alpha E_{\hat{D}}^i & E_Y^i(1-\beta) - \alpha E_{\hat{D}}^i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{A.5}$$

which has the form of half-wave rectification originally proposed by Boll [14]. Following the lead of [14], the half-wave rectified input-output relationship of the spectral subtractor has the form:

$$H_R = \frac{H + |H|}{2} \tag{A.6}$$

where H defines the general input-output relationship and its absolute value is required in order to enforce the rectification. Dividing Eq. (A.5) by E_Y^i and rearranging, the term for H is found to be:

$$H = \frac{\hat{E}_O^i}{E_Y^i} = 1 - \beta - \frac{\alpha E_{\hat{D}}^i}{E_Y^i} \tag{A.7}$$

where $\hat{E}_O^i = \hat{E}_S^i - \beta E_Y^i$. The spectral subtraction output equivalent to Eq. (A.4) but in filter form is:

$$\hat{E}_S^i = H_R E_Y^i = \frac{E_Y^i(1-\beta) - \alpha E_{\hat{D}}^i}{2} + \frac{|E_Y^i(1-\beta) - \alpha E_{\hat{D}}^i|}{2} + \beta E_Y^i \tag{A.8}$$

The expression for $\frac{\partial \hat{E}_S^i}{\partial \boldsymbol{\xi}}$ can be determined more easily from this equation as opposed to the representation in Eq. (A.4). The derivatives for the first and third terms are trivial, however the second term requires the substitution $|X| = \sqrt{X^2}$. The full partial derivative of $\hat{E}_S^i$ w.r.t. α is therefore determined to be:

$$\begin{aligned} \frac{\partial \hat{E}_S^i}{\partial \alpha} &= -\frac{E_{\hat{D}}^i}{2} \times \left(1 + \frac{E_Y^i(1-\beta) - \alpha E_{\hat{D}}^i}{|E_Y^i(1-\beta) - \alpha E_{\hat{D}}^i|}\right) \\ &= -\frac{E_{\hat{D}}^i}{2} \times \left(1 + \text{sign}\{E_Y^i(1-\beta) - \alpha E_{\hat{D}}^i\}\right) \end{aligned} \tag{A.9}$$

since $\text{sign}\{X\} = X/|X|$. The use of the sign function in Eq. (A.9) instead of the expanded form is very important as it provides support for the special case when $X = 0$. If using the expanded form $0/|0|$, a discontinuity would be observed in the gradient function due to the division by zero; this would be problematic for gradient-descent optimisation.

The gradient function in Eq. (A.9) can be substituted directly into Eq. (A.2) to form the full expression for each Jacobian element. This gradient derivation was verified using appropriate tools in MATLAB and was found to provide convergence in the gradient-descent optimisation.

From examination of Eq. (A.9) it can be seen that the gradient function is dependent on the flooring factor β. This dependence and the successful verification of the gradient leads us to believe the gradient derivation by BabaAli *et al.* [10] is inaccurate as it fails to include the flooring factor. In Section A.3, it is shown that the derivation of the gradient functions for both approaches are equivalent in the case where $\beta \neq 0$.

A.2.2 Flooring Factor, β

In this instance, the set of enhancement parameters $\boldsymbol{\xi}$ is defined as:

$$\boldsymbol{\xi} = \{\beta_1, \beta_2, \ldots, \beta_L\} \tag{A.10}$$

where a flooring factor β is applied to each Mel-filterbank (N.B. in this instance $L = P$). Again, reference to the l^{th} filterbank is removed for clarity.

A similar procedure to that of Section A.2.1 can be followed to determine the expression for the partial derivative of $\hat{E}_S^i$ w.r.t. the energy flooring factor β from Eq. (A.8). The gradient expression is found to be:

$$\frac{\partial \hat{E}_S^i}{\partial \beta} = \frac{E_Y^i}{2} \times \left(1 - \text{sign}\{E_Y^i(1-\beta) - \alpha E_{\hat{D}}^i\}\right) \tag{A.11}$$

which was also validated using the appropriate MATLAB tools.

A.3 Comparison with Frequency-Domain MBSS Derivation

The subtraction algorithms for the Mel-spaced Multi-Band Spectral Subtraction (MBSS) proposed in [10] and the MFNS-based method proposed in this dissertation are identical except for the domains in which the subtraction takes place. In the MBSS method, subtraction takes place in the frequency-domain,

whereas this research performs the subtraction on the Mel-filterbank energies. Applying the following substitutions to the notation in Eq. (A.4) results in the same basic algorithm:

$$\begin{aligned} E_Y^i &\sim |Y^i|^2 \\ E_{\hat{D}}^i &\sim |\hat{D}^i|^2 \\ \hat{E}_S^i &\sim |\hat{S}^i|^2 \end{aligned} \tag{A.12}$$

except that the subtraction parameters α apply to overlapped filterbanks in the case of MBSS. This overlapping of subtraction parameters in the frequency-domain leads to a non-diagonal gradient matrix as opposed to the purely diagonal matrix resulting from the use of MFNS.

Appendix B

Supporting Results

B.1 Introduction

This appendix provides tables which support explanations of results which are provided in the main text. Data contained here was either deemed too large to be included in the main body of the dissertation, or has only warranted passing reference within the discussion.

B.2 Tables of Supporting Results

B.2.1 Chapter 5

Table B.1: Performance evaluation of LIMA framework on the AEICS commands task.

Experiment	GD. Iter.	ASR Word Accuracy (%)						
		C0	C6	C1	C2	C3	C4	C5
Baseline	NA	93.1	54.1	85.4	50.5	79.3	81.4	43.8
Static MFNS	NA	93.8	73.2	86.4	68.6	88.8	85.1	65.0
Constrained	4	93.8	70.2	87.6	64.0	87.8	85.5	60.7
Unconstrained	4	94.2	69.7	86.6	63.3	87.6	85.5	59.6

Table B.2: Performance evaluation of LIMA framework on the AEICS commands task with MAP adaptation.

Experiment	GD. Iter	ASR Word Accuracy (%)						
		C0	C6	C1	C2	C3	C4	C5
Baseline	NA	98.2	96.7	98.5	95.0	99.0	99.0	95.2
Static MFNS	NA	97.8	98.3	98.7	97.6	99.2	98.8	97.3
$26 \times \alpha$	1	97.3	98.3	98.5	96.9	99.2	98.8	96.5
$26 \times \beta$	1	97.6	98.7	98.9	98.0	99.2	99.2	97.6
$[26 \times \alpha] + [26 \times \beta]$	1	97.6	97.8	98.5	97.0	99.2	98.6	96.5

B.2.2 Chapter 6

Table B.3: Performance evaluation of various noise estimation techniques in a LIMA framework on the AVICAR phone numbers task.

Noise Estimator	Experiment	ASR Word Accuracy (%)				
		IDL	35U	35D	55U	55D
NA	Baseline	70.4	48.8	36.2	41.8	23.5
Initial Silence	Static MFNS	73.3	47.8	36.8	44.5	26.1
Initial Silence	LIMA	73.8	50.3	38.5	45.6	27.2
TRA	Static MFNS	72.5	49.7	37.9	44.1	26.1
TRA	LIMA	72.4	50.0	37.6	43.6	26.0
TRA+SAD	Static MFNS	73.8	48.3	37.9	44.8	27.2
TRA+SAD	LIMA	74.1	50.2	38.5	45.3	27.2

B.3 Supporting Figures

B.3.1 Chapter 7

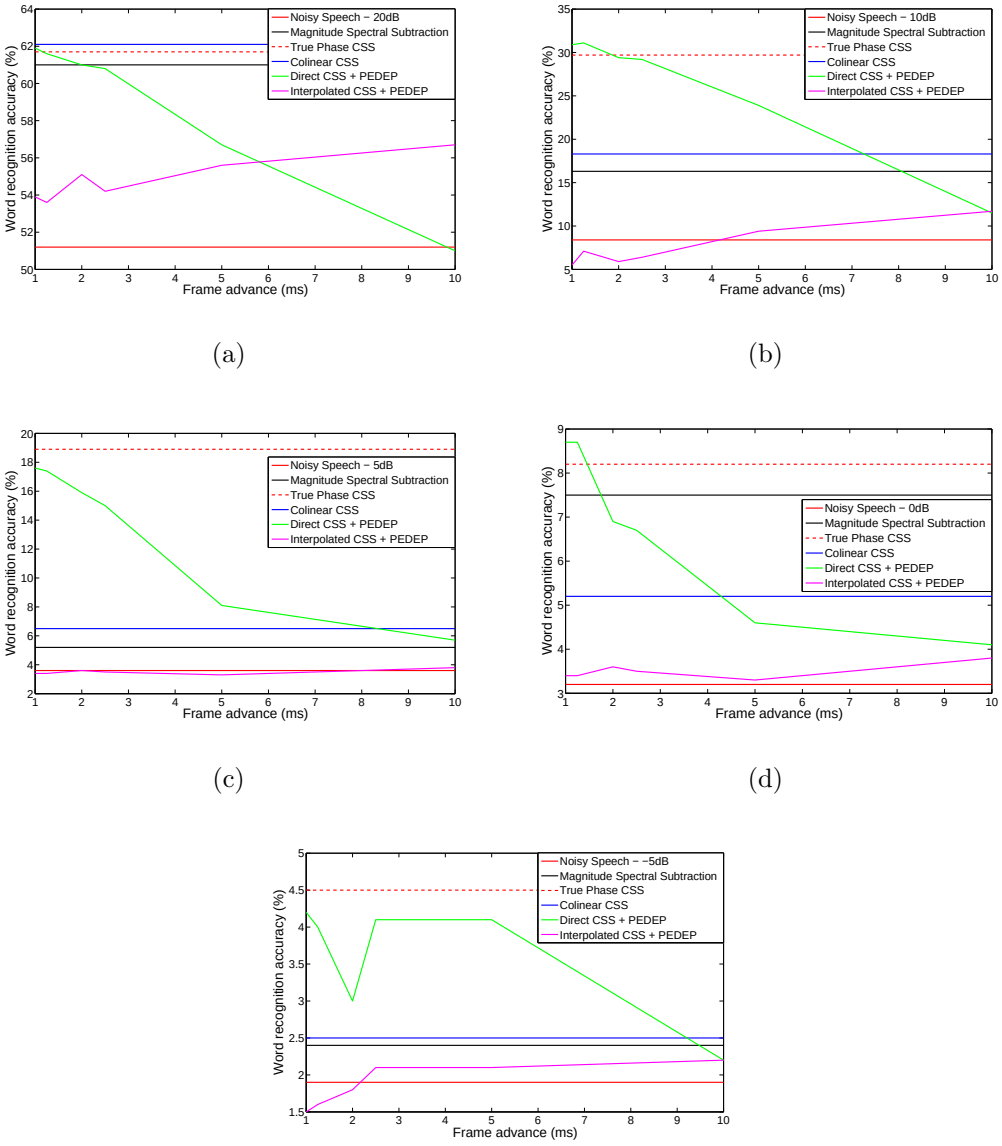


Figure B.1: ASR performance of the proposed PEDEP phase estimation technique for increasing frame advances using AWGN at SNR of (a) 20 dB, (b) 10 dB, (c) 5 dB, (d) 0 dB, and (e) -5 dB.

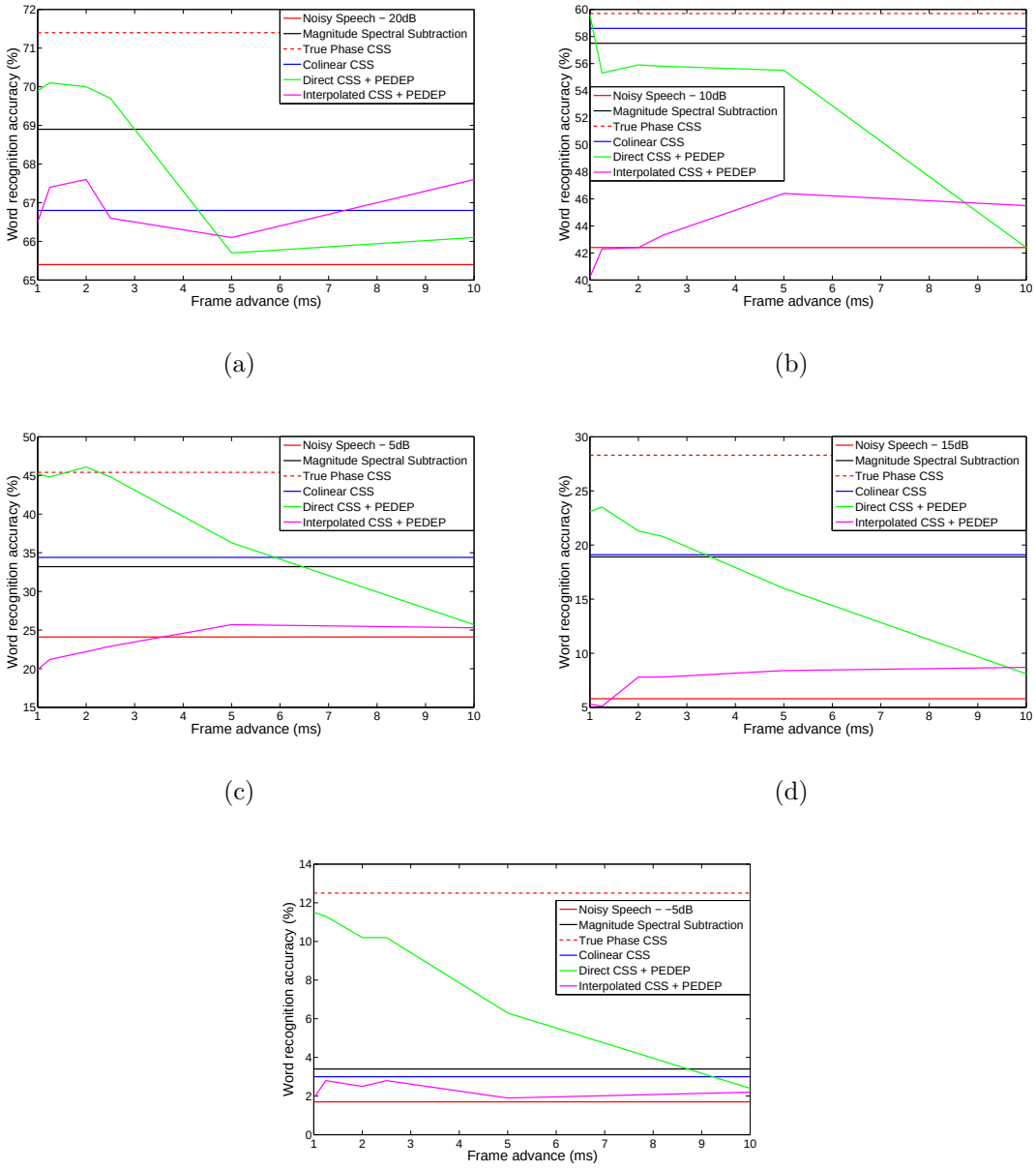


Figure B.2: ASR performance of the proposed PEDEP phase estimation technique for increasing frame advances using car noise at SNR of (a) 20 dB, (b) 10 dB, (c) 5 dB, (d) 0 dB, and (e) -5 dB.

Appendix C

In-Car Speech Data in Changing In-Car Noise Conditions

C.1 Motivation

The collection of in-car speech data under purposely annotated noise conditions (including both constant and varying noise environments) was motivated by:

1. A desire to discretely analyse and correlate the performance of speech enhancement techniques operating under likelihood-maximisation frameworks in specific noise conditions. Noise conditions of interest include the long-term effects produced by changes in speed, window position and the use of air-conditioning systems.

2. Given that dialect mismatch was seen to be a potential problem when assessing these techniques on the AEICS corpus, this data was required to validate algorithmic performance observed on the AVICAR database by recording native American English speakers to complement the data used for acoustic model training in this dissertation.

3. It is realistic for drivers to communicate with in-car speech systems in any possible driving scenario, including during heavy acceleration or whilst passengers make changes to cabin acoustics by opening windows or changing

the status of the air-conditioning system. Speech systems should be sufficiently intelligent to deal with these extreme changes in noise levels, however a previous lack of in-car speech data recorded under such conditions has restricted ASR evaluations to "constant" noise conditions. This dataset aims to provide data suitable for conducting pilot studies in these conditions.

4. The UTDrive vehicle is fitted to capture data from a microphone array and CAN-bus (among others – see [6]). Since CAN-bus and audio signals are becoming readily available in current generation vehicles, improvements in speech systems may be possible if CAN-bus information can be used to assist noise environment classification, alleviating the need for complex noise estimation procedures. This data collection will also be suitable for pilot studies into methods for combining these two data streams.

C.2 Collection Description

Driver speech was recorded under a range of different noise conditions using the instrumented UTDrive vehicle – a Toyota RAV4 Sports Utility Vehicle. This vehicle is equipped with a number of sensors – for this particular data collection signals were recorded from a close-talk microphone, a 4-channel microphone array and the CAN-bus.

In total, 10 native speakers of American English (6 females, 4 males) were chosen in order to avoid issues with modeling accents in such a small dataset. Each speaker was asked to complete four circuits of the route shown in Fig. C.1 which originates from the University of Texas at Dallas campus. During each lap, at specific points on the route drivers were asked to recite one of three phone numbers which they had previously committed to memory for ease of recall to minimise the induced cognitive load. The three phone numbers were randomly assigned to each segment (A1, A2, B or C) on each lap (see Table C.1).

Speech data was collected in nine discrete and constant noise conditions as shown in Table C.2. The passenger-side window position and air-conditioning (A/C) system were varied on each lap.

To analyse changes to noise conditions during recordings, a series of realistic

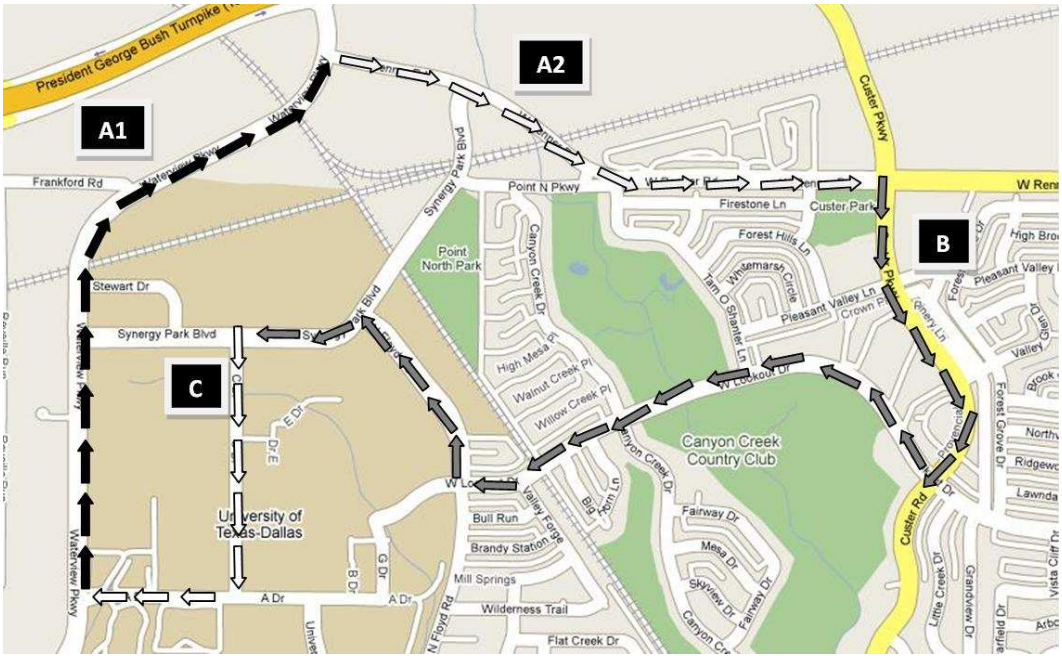


Figure C.1: Route used for collecting speech data in a range of constant and changing noise conditions.

Table C.1: ID of phone numbers recalled during each lap and route segment.

Lap No.	Route Segment			
	A1	A2	B	C
1	1	2	3	1
2	2	1	2	3
3	3	1	2	2
4	3	2	1	NA

scenarios were devised and collected across all four laps. These scenarios are summarised in Table C.3. On all laps, participants were asked to accelerate heavily from 0-40 mph after stopping at traffic signals or stop signs. On the final lap, focus was placed on the effects of passengers opening or closing their window or changing the air-conditioning system whilst the driver is speaking – this was controlled by the research assistant.

Table C.2: Constant noise conditions collected in this study.

Segment	Lap	Car Speed	Window Position	A/C Status
C	1	15 mph	Closed	Off
	2	15 mph	Open	Off
	3	15 mph	Closed	On
B	1	25 mph	Closed	Off
	2	25 mph	Open	Off
	3	25 mph	Closed	On
A1, A2	1	40 mph	Closed	Off
	2	40 mph	Open	Off
	3	40 mph	Closed	On

Table C.3: Changing noise conditions collected in this study.

Lap	Start Speed	End Speed	Window Position	A/C Status
1	0 mph	40 mph	Closed	Off
2	0 mph	40 mph	Open	Off
3	0 mph	40 mph	Closed	On
4	40 mph	40 mph	Closed to Open	Off
4	40 mph	40 mph	Open to Closed	Off
4	40 mph	40 mph	Closed	Off to On
4	40 mph	40 mph	Closed	On to Off
4	25 mph	25 mph	Closed to Open	Off
4	25 mph	25 mph	Open to Closed	Off
4	25 mph	25 mph	Closed	Off to On
4	25 mph	25 mph	Closed	On to Off

Bibliography

[1] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 4, pp. 1763–1773, 2004.

[2] A. Agarwal and Y. M. Cheng, "Two-stage Mel-warped Wiener filter for robust speech recognition," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, (Keystone, CO, USA), pp. 67–70, 1999.

[3] M. Akbacak and J. H. L. Hansen, "General issues in environmental noise tracking for robust in-vehicle speech applications: Supervised vs unsupervised acoustic noise analysis," in *Proceedings of the Biennial on Digital Signal Processing for In-Vehicle and Mobile Systems*, paper M2-2, 2005.

[4] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," in *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[5] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: a review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.

[6] P. Angkititrakul, M. Petracca, A. Sathyanarayana, and J. H. L. Hansen, "UTDrive: Driver behavior and speech interactive systems for in-vehicle environments," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, (Istanbul, Turkey), pp. 566–569, 2007.

[7] T. Arakawa, M. Tsujikawa, and R. Isotani, "Model-based Wiener filter for

noise robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, (Toulouse, France), pp. 537–540, 2006.

[8] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America*, vol. 50, no. 2, pp. 637–655, 1971.

[9] B. BabaAli, H. Sameti, and M. Safayani, "Spectral subtraction in likelihood-maximizing framework for robust speech recognition," in *Proceedings of INTERSPEECH*, (Brisbane, Australia), pp. 980–983, 2008.

[10] B. BabaAli, H. Sameti, and M. Safayani, "Likelihood-maximizing-based multiband spectral subtraction for robust speech recognition," *EURASIP Journal on Advances in Signal Processing*, no. 878105, pp. 1–15, 2009.

[11] L. Baum, "An inequality and associated maximization technique occurring in statistical estimation for probabilistic functions of a Markov process," *Inequalities*, vol. 3, pp. 1–8, 1972.

[12] S. K. Beack, B. Lee, M. Hahn, and S. H. Nam, "Blind source separation and Kalman filter-based speech enhancement in a car environment," in *Proceedings of the International Symposium on Signal Processing and Communication Systems*, (Seoul, Korea), pp. 520–523, 2004.

[13] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Washington, DC, USA), pp. 208–211, 1979.

[14] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[15] H. Bořil, P. Boyraz, and J. H. L. Hansen, "Multi-modal signal processing system for driver's stress detection," in *Proceedings of the Biennial Workshop*

on DSP for In-Vehicle Systems & Safety, DSP14, (Dallas, TX, USA), pp. 1–9, 2009.

[16] M. Brandstein and D. Ward, eds., *Microphone Arrays*. New York, NY, USA: Springer-Verlag, 2001.

[17] G. J. Brown and M. P. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, no. 44, pp. 297–336, 1994.

[18] B. Chen and P. Loizou, "A Laplacian-based MMSE estimator for speech enhancement," *Speech Communication*, vol. 49, pp. 134–143, 2007.

[19] J. Chen, K. K. Paliwal, and S. Nakamura, "Sub-band based additive noise removal for robust speech recognition," in *Proceedings of EUROSPEECH*, (Aalborg, Denmark), pp. 571–574, 2001.

[20] A. Chen, S. Vaseghi, and P. McCourt, "State based sub-band lp wiener filters for speech enhancement in car environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, (Istanbul, Turkey), pp. 213–216, 2000.

[21] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, 1953.

[22] J. Cho and A. Krishnamurthy, "Speech enhancement using microphone array in moving vehicle environment," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, (Columbus, OH, USA), pp. 366–371, 2003.

[23] I. Cohen, "On the decision-directed estimation approach of Ephraim and Malah," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Montreal, Canada), pp. 293–296, 2004.

[24] M. P. Cooke, P. G. Green, and M. D. Crawford, "Handling missing data in speech recognition," in *Proceedings of the International Conference on Spoken Language Processing*, (Yokohama, Japan), pp. 1555–1558, 1994.

[25] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[26] A. Das and J. H. L. Hansen, "Generalized parametric spectral subtraction using weighted Euclidean distortion," in *Proceedings of INTERSPEECH*, (Brisbane, Australia), pp. 399–402, 2008.

[27] G. M. Davis, ed., *Noise Reduction in Speech Applications*. Boca Raton, FL, USA: CRC Press, 2002.

[28] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for mono-syllabic word recognition in continuously spoken utterances," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[29] L. Deng and X. Huang, "Challenges in adopting speech recognition," *Communications of the ACM*, vol. 47, no. 1, pp. 69–75, 2004.

[30] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proceedings of EUROSPEECH*, vol. 2, (Madrid, Spain), pp. 1513–1516, 1995.

[31] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short time spectral amplitude estimator," *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[32] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[33] European Telecommunications Standards Institute, "ETSI ES 202 050 - Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," 2007.

[34] N. W. D. Evans, J. S. D. Mason, W. M. Liu, and B. Fauve, "An assessment on the fundamental limitations of spectral subtraction," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Toulouse, France), pp. 145–148, 2006.

[35] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, (Santa Barbara, CA, USA), pp. 347–354, 1997.

[36] D. B. Fry, *The Physics of Speech*. Cambridge, UK: Cambridge University Press, 1 ed., 1979.

[37] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[38] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[39] M. Gales and S. Young, "Robust continuous speech recognition using Parallel Model Combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 352–359, 1996.

[40] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 373–385, 1998.

[41] Y. Gao, J. Lu, K. Yu, and B.-L. Xu, "Codebook constrained iterative noise cancellation with applications to speech enhancement," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Salt Lake City, UT, USA), pp. 645–648, 2001.

[42] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM*. NIST, 1986.

[43] R. Gemello, F. Mana, and R. De Mori, "Automatic speech recognition with a modified Ephraim-Malah rule," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 56–59, 2006.

[44] Z. Goh and K. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 510–524, 1999.

[45] Z. Goh, K.-C. Tan, and T. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 287–292, 1998.

[46] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[47] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. AP-30, no. 1, pp. 27–34, 1982.

[48] J. Haigh and J. Mason, "Robust voice activity detection using ceptral features," in *Proceedings of IEEE TENCON*, (Beijing, China), pp. 321–324, 1993.

[49] J. H. L. Hansen, *Analysis and compensation of stressed and noisy speech with application to robust automatic recognition.* PhD thesis, Georgia Institute of Technology, 1988.

[50] J. Hansen, P. Angkititrakul, J. Plucienkowski, S. Gallant, U. Yapanel, B. Pellom, W. Ward, and R. Cole, "'CU-Move': Analysis & corpus development for interactive in-vehicle speech systems," in *Proceedings of EUROSPEECH*, vol. 3, (Aalborg, Denmark), pp. 2023–2026, 2001.

[51] J. H. L. Hansen and M. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 795–805, 1991.

[52] M. Hasan, S. Salahuddin, and M. Khan, "A modified a priori SNR for speech enhancement using spectral subtraction rules," *IEEE Signal Processing Letters*, vol. 11, no. 4, pp. 450–453, 2004.

[53] C. Hein and J. Rillings, "Spoken dialogue technologies for drivers," in *Convergence 2002 Proceedings: Transportation Electronics = Process + Business + Technology*, 2002.

[54] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[55] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[56] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Detroit, MI, USA), pp. 153–156, 1995.

[57] H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of Automatic Speech Recognition: Challenges for the new Millenium*, (Paris, France), 2000.

[58] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.

[59] F. Jelinek, "Continuous speech recognition by statistical methods," in *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556, April 1976.

[60] B. H. Juang, "Speech recognition in adverse environments," *Computer Speech and Language*, vol. 5, no. 3, pp. 275–294, 1991.

[61] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 7, pp. 947–954, 1987.

[62] J. C. Junqua, B. Reaves, and B. Mak, "A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognizers," in *Proceedings of EUROSPEECH*, (Genova, Italy), pp. 1371–1374, 1991.

[63] S. Kamath, *A multi-band spectral subtraction method for speech enhancement*. Masters thesis, University of Texas at Dallas, 2001.

[64] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Orlando, FL, USA), pp. 4160–4163, 2002.

[65] T. Kleinschmidt, P. Boyraz, H. Bořil, S. Sridharan, and J. H. L. Hansen, "Assessment of speech dialog systems using multi-modal cognitive load analysis and driving performance metrics," in *Proceeding of the IEEE International Conference on Vehicular Electronics & Safety*, (Pune, India), pp. 167–172, November 2009.

[66] T. Kleinschmidt, D. Dean, S. Sridharan, and M. Mason, "A continuous speech recognition protocol for the AVICAR database," in *Proceedings of the 1st International Conference on Signal Processing and Communication Systems*, (Gold Coast, Australia), pp. 339–344, 2007.

[67] T. Kleinschmidt, M. Mason, E. Wong, and S. Sridharan, "The Australian English speech corpus for in-car speech processing," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Taipei, Taiwan), pp. 4177–4180, 2009.

[68] T. Kleinschmidt, S. Sridharan, and M. Mason, "A modified LIMA framework for spectral subtraction applied to in-car speech recognition," in *Proceedings of the 1st International Conference on Signal Processing and Communication Systems*, (Gold Coast, Australia), pp. 335–338, 2007.

[69] T. Kleinschmidt, S. Sridharan, and M. Mason, "Likelihood-maximising frameworks for enhanced in-car speech recognition," in *Proceedings of the*

4th Biennial Workshop on DSP for In-Vehicle Systems & Safety, DSP08, (Dallas, TX, USA), pp. 1–8, 2009.

[70] B. Kouhi-Jelehkaran, H. Bakhshi, and F. Razzazi, "Improvement in speech recognition using phone-based filter and sum parameter optimization," *IEICE Electronics Express*, vol. 6, no. 8, pp. 437–442, 2009.

[71] V. Krishnan, S. M. Siniscalchi, D. V. Anderson, and M. A. Clements, "Noise robust Aurora-2 speech recognition employing a codebook-constrained Kalman filter preprocessor," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. I, (Toulouse, France), pp. 781–784, 2006.

[72] C. Y.-K. Lai and P. Aarabi, "Multiple-microphone time-varying filters for robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. I, (Montreal, Canada), pp. 233–236, 2004.

[73] R. H. Lambert, *Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures*. PhD thesis, Dept. of Electrical Engineering, University of Southern California, 1996.

[74] F.-K. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 4, pp. 599–609, 1990.

[75] C.-H. Lee and J.-L. Gauvain, "Bayesian adaptive learning and MAP estimation of HMM," in *Automatic speech and speaker recognition : Advanced topics*, pp. 83–107, Boston, Massachusetts, USA: Kluwer Academic Publishers, 1996.

[76] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: Audio-visual speech corpus in a car environment," in *Proceedings of INTERSPEECH*, (Jeju Island, Korea), pp. 2489–2492, 2004.

[77] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[78] R. G. Leonard, "A database for speaker independent digit recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 9, (San Diego, CA, USA), pp. 328–331, 1984.

[79] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Noise reduction based on adaptive β-order generalized spectral subtraction for speech enhancement," in *Proceedings of INTERSPEECH*, (Antwerp, Belgium), pp. 802–805, 2007.

[80] W. Li, K. Takeda, and F. Itakura, "Robust in-car speech recognition based on nonlinear multiple regressions," *EURASIP Journal on Advances in Signal Processing*, no. 16921, pp. 1–10, 2007.

[81] H. Liao, *Uncertainty Decoding for Noise Robust Speech Recognition*. PhD thesis, University of Cambridge, September 2007.

[82] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[83] L. Lin, W. Holmes, and E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement," *Electronics Letters*, vol. 39, no. 9, pp. 754–755, 2003.

[84] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, (Dallas, TX, USA), pp. 705–708, 1987.

[85] P. Lockwood, C. Baillargeat, J. M. Gillot, J. Boudy, and G. Faucon, "Noise reduction for speech enhancement in cars: non-linear spectral subtraction / Kalman filtering," in *Proceedings of EUROSPEECH*, (Genova, Italy), pp. 83–86, 1991.

[86] P. Lockwood and J. Boudy, "Experiments with non-linear spectral subtractor (NSS), hidden Markov models, and the projection distance, for robust speech recognition in cars," *Speech Communication*, vol. 11, no. 2-3, pp. 215–228, 1992.

[87] P. C. Loizou, *Speech enhancement: theory and practice*. Boca Raton, FL, USA: CRC Press, 2007.

[88] E. Lombard, "Le signe de lelevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.

[89] B. Magladry and D. Bruce, *In-Vehicle Corpus and Signal Processing for Driver Behavior*, ch. Improved vehicle safety and how technology will get us there, hopefully, pp. 1–8. Springer Science+Business Media, LLC, 2009.

[90] D. Malah, R. Cox, and A. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Phoenix, AZ, USA), pp. 789–792, 1999.

[91] R. Martin, "Spectral subtraction based on minimum statistics," in *Proceedings of EUSIPCO*, (Edinburgh, UK), pp. 1182–1185, 1994.

[92] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[93] R. Martin, "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. I, (Orlando, FL, USA), pp. 253–256, 2002.

[94] M. Matassoni, G. Mian, M. Omologo, A. Santarelli, and P. Svaizer, "Some experiments on the use of one-channel noise reduction techniques with the Italian SpeechDat Car database," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, (Madonna di Campiglio, Italy), pp. 139–142, 2001.

[95] I. A. McCowan, *Robust speech recognition using microphone arrays*. Phd thesis, Queensland University of Technology, 2001.

[96] A. J. McKnight and B. B. Adams, "Driver education task analysis, volume 1: Task description report," tech. rep., DOT HS 800 367, Department of Transportation, 1970.

[97] M. F. McTear, *Spoken Dialogue Technology*. London: Springer-Verlag, 2004.

[98] J. Meyer and K. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, (Munich, Germany), pp. 1167–1170, 1997.

[99] J. Meyer, K. Simmer, and K. Kammeyer, "Comparison of one- and two-channel noise estimation techniques," in *5th International Workshop on Acoustic Echo and Noise Cancellation*, vol. 1, (London, UK), pp. 137–145, 1997.

[100] J. B. Millar, J. P. Vonwiller, J. M. Harrington, and P. J. Dermody, "The australian national database of spoken language," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, (Adelaide, Australia), pp. 97–100, 1994.

[101] B. Milner and S. Vaseghi, "Comparison of some noise-compensation methods for speech recognition in adverse environments," *IEE Proceedings on Vision, Image and Signal Processing*, vol. 141, no. 5, pp. 280–288, 1994.

[102] D. Moore, *Speech enhancement using microphone arrays*. Masters thesis, Queensland University of Technology, 2000.

[103] B. Nasersharif and A. Akbari, "A framework for robust MFCC feature extraction using SNR-dependent compression of enhanced Mel filter bank energies," in *Proceedings of INTERSPEECH*, (Pittsburgh, PA, USA), 2006, paper 1632-Mon1A2O.3.

[104] J. Nocedal and S. Wright, *Numerical Optimization*. New York: Springer, 1999.

[105] J. A. Nolazco Flores and S. J. Young, "Adapting a HMM-based recogniser for noisy speech enhanced by spectral subtraction," in *Proceedings of EUROSPEECH*, (Berlin, Germany), pp. 829–832, 1993.

[106] S. Nordholm, I. Claesson, and B. Bengtsson, "Adaptive array noise suppresion of hands-free speaker input in cars," *IEEE Transactions on Vehicular Technology*, vol. 42, no. 4, pp. 514–518, 1993.

[107] S. Oh, V. Viswanathan, and P. Panamichalis, "Hands-free voice communication in an automobile with a microphone array," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, (San Francisco, CA, USA), pp. 281–284, 1992.

[108] K. Onoe, H. Segi, T. Kobayakawa, S. Sato, T. Imai, and A. Ando, "Filter bank subtraction for robust speech recognition," in *Proceedings of the International Conference on Spoken Language Processing*, (Denver, CO, USA), pp. 1021–1024, 2002.

[109] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," in *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, 1981.

[110] K. K. Paliwal and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153–170, 2005.

[111] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6, (Dallas, TX, USA), pp. 177–180, 1987.

[112] B. Pellom and J. H. L. Hansen, "An improved constrained iterative speech enhancement for colored noise environments," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 6, pp. 573–579, 1998.

[113] D. Pisoni, R. Bernacki, H. Nusbaum, and M. Yuchtman, "Some acoustic-phonetic correlates of speech produced in noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Tampa, FL, USA), pp. 1581–1584, 1985.

[114] L. R. Rabiner, K. C. Pan, and F. K. Soong, "On the performance of isolated word speech recognizers using vector quantization and temporal energy contours," *AT&T Bell Labs Technical Journal*, vol. 63, no. 7, pp. 1245–1260, 1984.

[115] L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "High performance connected digit recognition using hidden Markov models," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, (New York, NY, USA), pp. 119–122, 1988.

[116] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 275–296, 2004.

[117] D. R. Reddy, "Speech recognition by machine: a review," in *Proceedings of the IEEE*, vol. 64, no. 4, pp. 501–531, 1976.

[118] C. Ris and S. Dupont, "Assessing local noise level estimation methods: application to noise robust ASR," *Speech Communication*, vol. 34, no. 1-2, pp. 141–158, 2001.

[119] M. Sala, F. Sanchez, H. Wengelnik, H. van den Heuvel, A. Moreno, E. Deregibus, G. Richard, and E. Le Chevalier, "SpeechDat-Car: speech databases for voice driven teleservices and control of in-car applications," in *Proceedings of the European Automotive Engineers Congress*, (Barcelona, Spain), pp. 90–98, 1999.

[120] H. Saruwatari, S. Kurita, K. Takeda, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1135–1146, 2003.

[121] V. Schless and F. Class, "Adaptive model combination for robust speech recognition in car environments," in *Proceedings of EUROSPEECH*, (Rhodes, Greece), pp. 1091–1094, 1997.

[122] V. Schless and F. Class, "SNR-dependent flooring and noise overestimation for joint application of spectral subtraction and model combination," in *Proceedings of the International Conference on Spoken Language Processing*, paper 0138, (Sydney, Australia), 1998.

[123] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, (Salt Lake City, UT, USA), pp. 133–136, May 2001.

[124] R. Schwartz, Y. Chow, S. Roucos, M. Krasner, and J. Makhoul, "Improved hidden Markov modeling of phonemes for continuous speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 9, no. 1, (San Diego, CA, USA), pp. 21–24, 1984.

[125] M. L. Seltzer, *Microphone Array Processing for Robust Speech Recognition*. PhD thesis, Dept. of Electrical and Computer Engineering, Carnegie Mellon University, 2003.

[126] M. L. Seltzer, B. Raj, and R. M. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

[127] M. Seltzer, B. Raj, and R. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 489–498, 2004.

[128] M. Seltzer and R. Stern, "Subband likelihood-maximizing beamforming for speech recognition in reverberant environments," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2109–2121, 2006.

[129] J.-L. Shen, W.-L. Hwang, and L.-S. Lee, "Robust speech recognition features based on temporal trajectory filtering of frequency band spectrum," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, (Philadelphia, PA, USA), pp. 881–884, 1996.

[130] G. Shi, P. Aarabi, and H. Jiang, "Phase-based dual-microphone speech enhancement using a prior speech model," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 109–118, 2007.

[131] M. Shozakai, S. Nakamura, and K. Shikano, "A speech enhancement approach E-CMN/CSS for speech recognition in car environments," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, (Santa Barbara, CA, USA), pp. 450–457, 1997.

[132] L. Singh and S. Sridharan, "Speech enhancement using critical band spectral subtraction," in *Proceedings of the International Conference on Spoken Language Processing*, (Sydney, Australia), pp. 2827–2830, 1998.

[133] O. Siohan, C. Chesta, and C. Lee, "Joint maximum a posteriori estimation of transformation and hidden Markov model parameters," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Istanbul, Turkey), pp. 965–968, 2000.

[134] J. Smolders, T. Claes, G. Sablon, and D. van Compernolle, "On the importance of the microphone position for speech recognition in the car," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, (Adelaide, Australia), pp. 429–432, 1994.

[135] S. So, K. K. Wójcicki, J. G. Lyons, A. G. Stark, and K. K. Paliwal, "Kalman filter with phase spectrum compensation algorithm for speech enhancement," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Taipei, Taiwan), pp. 4405–4408, 2009.

[136] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, (Seattle, WA, USA), pp. 365–368, 1998.

[137] A. P. Stark, K. K. Wójcicki, J. G. Lyons, and K. K. Paliwal, "Noise driven

short-time phase spectrum compensation procedure for speech enhancement," in *Proceedings of INTERSPEECH*, (Brisbane, Australia), pp. 549–552, 2008.

[138] S. S. Stevens, J. Volkman, and E. Newman, "A scale for the measurement of the psychological magnitude of pitch," *Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[139] R. Tucker, "Voice activity detection using a periodicity measure," in *Proceedings of the IEEE*, vol. 139, no. 4, pp. 377–380, 1992.

[140] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," tech. rep., Speech Research Unit, Defence Research Agency, Malvern, UK, 1992.

[141] S. Vaseghi, Q. Yan, and A. Ghorshi, "Speech accent profiles: Modeling and synthesis," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 69–74, 2009.

[142] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. IT-13, no. 2, pp. 260–269, 1967.

[143] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[144] L. Wang, S. Ohtsuka, and Nakagawa, "High improvement of speaker identification and verification by combining MFCC and phase information," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Taipei, Taiwan), pp. 4529–4532, 2009.

[145] J. C. Wells, *Accents of English*. Cambridge, UK: Cambridge University Press, 1982.

[146] J. Whittington, K. Deo, T. Kleinschmidt, and M. Mason, "FPGA implementation of spectral subtraction for in-car speech enhancement and recognition," in *Proceedings of the 2nd International Conference on Signal Processing and Communication Systems*, (Gold Coast, Australia), 2008.

[147] J. Whittington, K. Deo, T. Kleinschmidt, and M. Mason, "FPGA implementation of spectral subtraction for automotive speech recognition," in *Proceedings of the IEEE Workshop on Computational Intelligence in Vehicles and Vehicular Systems*, (Nashville, TN, USA), pp. 72–79, 2009.

[148] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*. Cambridge, MA, USA: MIT Press, 1949.

[149] K. Wójcicki, M. Milacic, A. Stark, J. Lyons, and K. Paliwal, "Exploiting conjugate symmetry of the short-time Fourier spectrum for speech enhancement," *IEEE Signal Processing Letters*, vol. 15, pp. 461–464, 2008.

[150] K.-G. Wu and P.-C. Chen, "Efficient speech enhancement using spectral subtraction for car hands-free applications," in *Proceedings of the International Conference on Consumer Electronics*, (Los Angeles, CA, USA), pp. 220–221, 2001.

[151] U. Yapanel, X. Zhang, and J. H. L. Hansen, "High performance digit recognition in real car environments," in *Proceedings of the International Conference on Spoken Language Processing*, (Denver, Colorado), pp. 793–796, 2002.

[152] H. Ye, J. Whittington, I. Himawan, T. Kleinschmidt, and M. Mason, "FPGA implementation of dual-microphone delay-and-sum beamforming for in-car speech enhancement and recognition," in *Proceedings of the AutoCRC Conference*, (Melbourne, Australia), 2009.

[153] D. Yellin and E. Weinstein, "Multichannel signal separation: methods and analysis," *IEEE Transactions on Signal Processing*, vol. 44, no. 1, pp. 106–118, 1996.

[154] C. You, S. Koh, and S. Rahardja, "Adaptive β-order MMSE estimation for speech enhancement," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. I, (Hong Kong, Hong Kong), pp. 900–903, 2003.

[155] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge University Engineering Department, 3.4 ed., December 2006.

[156] S. J. Young, N. H. Russell, and J. H. S. Thornton, "Token passing: a simple conceptual model for connected speech recognition systems," Tech. Rep. Technical Report CUED/F-INFENG/TR38, Cambridge University Engineering Dept., 1989.

[157] X. Zhang and J. H. L. Hansen, "CSA-BF: a constrained switched adaptive beamformer for speech enhancement and recognition in real car environments," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 733–745, 2003.

[158] D. Zhu and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, (Montreal, Canada), pp. 125–128, 2004.