Automatic Spoken Language Identification Utilizing Acoustic and Phonetic Speech Information

by

Kim-Yung Eddie Wong, BEng(Hons), BIT

PhD Thesis
Submitted in Fulfilment
of the Requirements
for the Degree of

Doctor of Philosophy

at the
Queensland University of Technology

Speech and Audio Research Laboratory
School of Electrical & Electronic Systems Engineering

June 2004
Substitute the “official” thesis signature page here.
Keywords

Language identification, universal background model, vocal tract length normalization, phone recognition, output score fusion, multi-lingual phone recognition, cross-lingual speech recognition
Abstract

Automatic spoken Language Identification (LID) is the process of identifying the language spoken within an utterance. The challenge that this task presents is that no prior information is available indicating the content of the utterance or the identity of the speaker. The trend of globalization and the pervasive popularity of the Internet will amplify the need for the capabilities spoken language identification systems provide. A prominent application arises in call centers dealing with speakers speaking different languages. Another important application is to index or search huge speech data archives and corpora that contain multiple languages.

The aim of this research is to develop techniques targeted at producing a fast and more accurate automatic spoken LID system compared to the previous National Institute of Standards and Technology (NIST) Language Recognition Evaluation. Acoustic and phonetic speech information are targeted as the most suitable features for representing the characteristics of a language. To model the acoustic speech features a Gaussian Mixture Model based approach is employed. Phonetic speech information is extracted using existing speech recognition technology. Various techniques to improve LID accuracy are also studied. One approach examined is the employment of Vocal Tract Length Normalization to reduce the speech variation caused by different speakers. A linear data fusion technique is adopted to combine the various aspects of information extracted from speech. As a result of this research, a LID system was implemented and presented for evaluation in the 2003 Language Recognition Evaluation conducted by the NIST.
Contents

Abstract i

List of Tables x

List of Figures xiii

Acronyms & Abbreviations xvi

Certification of Thesis xviii

Acknowledgments xix

Chapter 1 Introduction 1

1.1 Motivation ................................. 1

1.2 Scope and Aim ............................ 2

1.3 Outline of Thesis ......................... 4

1.4 Major Contributions of the Research 6
## CONTENTS

1.5 Publications Resulting from this Research .......................... 7

### Chapter 2 Background of Automatic Spoken Language Identification

2.1 Introduction .................................................. 9

2.2 Automatic Language Identification in General ...................... 11

2.3 Speech Information Utilized for Language Identification .......... 12

2.3.1 Spoken Level Speech Information ............................ 13

2.3.2 Word Level Speech Information .............................. 15

2.4 Automatic Language Identification Systems ....................... 16

2.4.1 Utilizing Acoustic Information ............................... 16

2.4.2 Utilizing Phonetic Information ................................ 18

2.4.3 Utilizing Prosodic Information ................................ 19

2.4.4 Utilizing Word Level Information ............................. 19

2.5 Difficulties of LID Research .................................. 20

2.6 Resources for LID Research .................................... 21

2.6.1 Oregon Graduate Institute Telephone Speech Corpus .......... 21

2.6.2 OGI 22 Language Telephone Speech Corpus .................. 23

2.6.3 LDC CALLFRIEND Telephone Speech Database ............... 23

2.6.4 International Phonetic Alphabet ............................... 24
2.7 NIST Language Identification/Recognition Evaluation 25
   2.7.1 1993/1994 Language Identification Evaluation 25
   2.7.2 1996 Language Recognition Evaluation 26
   2.7.3 2003 Language Recognition Evaluation 28

2.8 Summary and Discussion 28

Chapter 3 Language Identification Utilizing Efficient Gaussian Mixture Analysis

3.1 Introduction 31

3.2 Gaussian Mixture Model Based LID System 33
   3.2.1 Gaussian Mixture Model 34
   3.2.2 GMM based Language Identification 37

3.3 Universal Background Model Technique 38
   3.3.1 Background Speaker Modelling 38
   3.3.2 Universal Background Model 39
   3.3.3 Application of UBM Technique in LID 42

3.4 Acoustic Information based LID Experiments 45
   3.4.1 System Setting 45
   3.4.2 Experimental Results 46
3.5 Experiments Using Different Parameterization Techniques and Features ................................................. 49

3.5.1 Mel-Frequency Cepstral Coefficient ............................................. 50
3.5.2 Linear Prediction Cepstral Coefficient ............................................. 50
3.5.3 Mel-Cepstral Coefficient .............................................................. 51
3.5.4 Perceptual Linear Predictive Coefficients ........................................ 52
3.5.5 Band-Limiting ................................................................. 53
3.5.6 Experiment Setup ................................................................. 54
3.5.7 Experimental Results ................................................................. 54
3.5.8 NIST 1996 Language Identification Experiment ................................ 57

3.6 Summary .................................................................................. 58

Chapter 4 Study of Phonetic Details for a Phonetic Information based Language Identification System 61

4.1 Introduction ............................................................................. 61

4.2 Phonetic Information Based Test System ........................................ 63

4.2.1 PRLM System Structure ......................................................... 64

4.2.2 Parallel-PRLM .................................................................. 66

4.3 Incorporate Phone Duration ........................................................ 66

4.4 Fine Phonetic Details LID Experiments ......................................... 68
4.4.1 Test System Implementation .......................... 68
4.4.2 Experimental Results ................................. 70

4.5 Broad Phonetic Detail LID Experiments .......................... 71
4.5.1 Define Broad Phonetic Events – Phonetic Knowledge Approach .......................... 72
4.5.2 Define Broad Phonetic Events – Data Driven Approach .......................... 73
4.5.3 Test System Setting ................................. 75
4.5.4 Experimental Results ................................. 75

4.6 Summary .................................................. 78

Chapter 5 Techniques to Improve LID Accuracy 81

5.1 Introduction .................................................. 81

5.2 Vocal Tract Length Normalization ................................. 84
5.2.1 Speaker Normalization ...................................... 84
5.2.2 Normalization Factor Selection ................................. 86
5.2.3 LID Experiments .................................................. 89

5.3 Output Score Fusion .............................................. 94
5.3.1 Data Fusion .............................................. 95
5.3.2 Output Score Fusion Strategies Experiment ................. 97
5.3.3 Experiment on Combining Acoustic and Phonetic Speech Information ........................................ 101

5.4 Summary .......................................................................................................................... 102

Chapter 6 NIST 2003 Language Recognition Evaluation System 105

6.1 Introduction ...................................................................................................................... 105

6.2 NIST 2003 Evaluation Task ............................................................................................ 106

6.2.1 The Evaluation ............................................................................................................ 106

6.2.2 Test segment duration ............................................................................................... 107

6.2.3 Evaluation Rules ......................................................................................................... 108

6.3 QUT 2003 NIST Language Recognition Evaluation System .................................. 108

6.3.1 System Structure ........................................................................................................ 108

6.3.2 Generate Evaluation Output ....................................................................................... 110

6.4 Evaluation Results .......................................................................................................... 111

6.5 Summary .......................................................................................................................... 112

Chapter 7 Conclusions and Further Work 117

7.1 Conclusions .................................................................................................................... 117

7.1.1 Acoustic Speech Information Investigation .............................................................. 117

7.1.2 Phonetic Speech Information Investigation ............................................................... 118
7.1.3 Accuracy Improvement Techniques ................. 118
7.1.4 QUT 2003 NIST Evaluation System ................. 119
7.2 Further Work ......................................... 119

Bibliography .............................................. 121

Appendix A Application of LID - Multilingual Phone Recognition133
A.1 Abstract .............................................. 133
A.2 Introduction .......................................... 134
A.3 Multilingual Phone Recognition Approaches .......... 135
A.3.1 Explicit-LID - Approach 1 ......................... 135
A.3.2 Explicit-LID - Approach 2 ......................... 136
A.3.3 Explicit-LID - Approach 3 ......................... 136
A.4 Test Systems ............................................. 137
A.4.1 Baseline Monolingual Phone Recognizers .......... 137
A.4.2 Multilingual Phone Recognition Test Systems .... 138
A.5 Experiments and Results ................................ 139
A.5.1 Isolated Phone Recognition Results ............... 140
A.5.2 Continuous Phone Recognition Results ............ 141
A.6 Conclusions ............................................. 142
Appendix B  Cross-lingual Speech Recognition (Indonesian)  143

B.1 Abstract ................................................................. 143

B.2 Introduction ........................................................... 144

B.3 Multilingual Phone Model Clustering ............................... 146

B.3.1 Clustering Criteria ................................................ 146

B.3.2 Clustering Process ................................................. 147

B.4 Pronunciation Modelling ............................................. 148

B.5 Experiments and Results ............................................ 150

B.6 Discussion ............................................................ 152

B.7 Conclusions ........................................................... 153
List of Tables

2.1 The amount of speech data for each language in the OGI-TS database. .................................................. 22

3.1 Experimental results (% correct) of applying the UBM technique on GMM based LID system. ..................... 47

3.2 Experimental results (% correct) of creating Language-GMM with EM but initialized from UBM. ...................... 49

3.3 LID Results (% correct) comparing basic feature, band-limiting (BL, 300-3400Hz) and cepstral mean subtraction (CMS). .......................... 55

3.4 LID Results (% correct) comparing the use of delta (D), acceleration (A) and delta energy (DE) coefficients. The window size for both delta and acceleration is 9 frame. .................. 56

3.5 LID Results (% correct) including band-limiting, cepstral mean subtraction, delta, acceleration and delta energy coefficients to the feature vector. .................................................. 56

3.6 GMM-UBM LID results (% correct) on NIST 1996 Language Identification experiment. ............................. 58
4.1 The number of phones and the amount of training data used to train the phone recognizer front-ends.  

4.2 LID results (% correct) of the NIST 1994 LID experiment using phonetic information based LID system.  

4.3 LID results (% correct) of the NIST 1996 LID experiment using phonetic information based LID system.  

4.4 Chart of English vowel (not including retroflexes).  

4.5 LID results (% correct) of the NIST 1996 LID experiment using the BPRLM LID system.  

4.6 Comparison of LID performance utilizing fine and broad phonetic information in NIST 1996 LID experiment (% correct).  

5.1 Test results (% correct) for applying VTLN to the GMM-UBM LID system. The “Diff. $\alpha$” column is the difference in $\alpha$ of all speakers compared to the previous iteration. $i$ represents the number of iteration. There are 678 male and 268 female speakers in this task.  

5.2 Test results (% correct) and time requirement (parameterization + selection) of employing different a selection approaches during testing. Total length of test data (45s+10s test) is 16385s. (All tests were performed on a Pentium III 933MHz 512MB RAM Linux system.)  

5.3 Test results (% correct) of applying VTLN to the GMM-UBM and PPRLM LID system on the NIST 1996 LID experiment.  

5.4 LID accuracy (% correct) of the baseline GMM-UBM systems on the NIST 1994 LID experiment.
5.5 LID Results (% correct) comparing linear fusion methods. \( \omega \) is the score weighting applied to the first GMM-UBM system under Fusion column. ......................................................... 99

5.6 LID Results (% correct) comparing fusion of varies number of GMM-UBM systems using Linear Score Weighting. ......................... 100

5.7 Output Score Fusion results on NIST 1996 LID experiment. VTLN are applied to all LID systems (% correct). ................................. 102

A.1 Details and experimental results of the baseline monolingual phone recognizers. Continuous phone recognition results under “\% Accuracy” included insertion error while “\% Correct” does not. .... 139

A.2 Multilingual (isolated and continuous) phone recognition experimental results. “\% LID” is in percentage of correct. ...................... 140

B.1 The first twenty questions asked during the decision tree clustering (en - English, hi - Hindi, sp - Spanish). ................................. 148

B.2 A set of 47 multilingual phones obtained from the decision tree clustering process. Each string represents a cluster with the ‘-’ character used to connect each monolingual phone. The ‘-’ character is the language tag, where en - English, hi - Hindi, sp - Spanish. 149

B.3 Indonesian spontaneous speech recognition results. .................... 151
List of Figures

2.1 A general structure of a LID system. .................................................. 12
2.2 Table of consonants represented in IPA. .......................... 24
2.3 A sample Detection Error Tradeoff (DET) curve. .................... 27
3.1 Depicting the approximation of a “off axis” Gaussian with four
“axis aligned” Gaussians. ................................................................. 36
3.2 Block diagram of a GMM based LID system. ............................ 38
3.3 Steps to utilize the UBM technique to calculate the likelihood score
of a given feature vector against a speaker model during testing .
Step 1: Select the Gaussians from the UBM obtaining the highest
likelihood as the significant mixture components. Step 2: Sum the log likelihood scores of the corresponding significant mixture
components (those selected from the UBM) of this speaker model.
This summed score is then utilized as the approximated likelihood
of this speaker instead of calculating from all mixture components. 43
3.4 Block diagram of the GMM-UBM LID system. The top half of the
diagram shows the training process and the testing is shown at the
bottom half. ................................................................. 44
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5</td>
<td>LID performance variation with respect to the number of significant mixture components selected.</td>
<td>48</td>
</tr>
<tr>
<td>3.6</td>
<td>Confusion Matrix of the 10s test case of both NIST 1994 and NIST 1996 LID experiment using the GMM-UBM LID system.</td>
<td>59</td>
</tr>
<tr>
<td>4.1</td>
<td>Block diagram of a PRLM LID system.</td>
<td>65</td>
</tr>
<tr>
<td>4.2</td>
<td>Block diagram of a PPRLM LID system. The languages utilized to train the phone recognizer front-ends can be independent of the targeted languages.</td>
<td>67</td>
</tr>
<tr>
<td>4.3</td>
<td>The phonetic clustering tree of English phones produced by the phonetic questions clustering process.</td>
<td>74</td>
</tr>
<tr>
<td>4.4</td>
<td>Confusion Matrix of the 10s test case of the NIST 1996 LID experiment of both the PPRLM (2-gram LM) and BPRLM (data driven mapping, 3-gram LM) systems. Both systems have phone duration information appended.</td>
<td>77</td>
</tr>
<tr>
<td>5.1</td>
<td>Piecewise linear re-scaling the frequency axis to achieve speaker normalization. $\alpha$ is the speaker-specific normalization factor and $f_0$ is set to fix the bandwidth mismatch.</td>
<td>85</td>
</tr>
<tr>
<td>5.2</td>
<td>The normalization factor selection process where $X_{ni}$ are the set of feature vectors normalized with normalization factor $\alpha_i$. $N$ is the number of predefined $\alpha$ available for selection and $\lambda$ is the given model. Here, $p(X_{ni} \mid \lambda)$ represents the likelihood of the feature vectors $X_{ni}$, given the model $\lambda$.</td>
<td>87</td>
</tr>
<tr>
<td>5.3</td>
<td>Histogram of all speakers (in the training data set) sorted according to $\alpha$.</td>
<td>91</td>
</tr>
</tbody>
</table>
5.4 LID test results of using different length of test data to select $\alpha$. The dashed straight line represents the performance without VTLN. 92

5.5 Percentage of accuracy improvement after VTLN on each system of the NIST 1996 LID experiment. 94

5.6 Input Fusion System. $f_i, i = 1 \ldots N$, are the feature vectors that capturing different aspect of speech information. 96

5.7 Output Fusion System. $f_i, i = 1 \ldots N$, are the feature vectors that capturing different aspect of speech information. 96

6.1 Block diagram of the QUT 2003 NIST language recognition evaluation system. 109

6.2 The DET curve for the 2003 NIST language recognition evaluation 30s test case. 112

6.3 The DET curve for the 2003 NIST language recognition evaluation 10s test case. 113

6.4 The DET curve for the 2003 NIST language recognition evaluation 3s test case. 114

6.5 The weighting applied on the output scores generated by different sub-systems during fusion for different test segment duration. These weightings were calculated from the development test data set. 115

A.1 Block diagram of the three different approaches to achieve multi-lingual phone recognition. 137
Acronyms & Abbreviations

ANN  Artificial Neural Network

ASCII  American Standard Code for Information Interchange

BPRLM  Broad Phonetic Recognition followed by Language Modelling

CMS  Cepstral Mean Subtraction

DCT  Discrete Cosine Transform

DET  Detection Error Tradeoff

DFT  Discrete Fourier Transform

EER  Equal Error Rate

EM  Expectation-Maximization

GMM  Gaussian Mixture Model

HMM  Hidden Markov Model

IPA  International Phonetic Alphabet

LDC  Linguistic Data Consortium

LID  Language Identification

LM  Language Model

LPC  Linear Prediction Coefficient
LPCC  Linear Prediction Cepstral Coefficient
LSW   Linear Score Weighting
MCC   Mel-Cepstral Coefficient
MFCC  Mel-Frequency Cepstral Coefficient
NIST  National Institute of Standards and Technology
PLP   Perceptual Linear Predictive
PPRLM Parallel Phone Recognition followed by Language Modelling
QUT   Queensland University of Technology
SLP   Selective Linear Prediction
UBM   Universal Background Model
VTLN  Vocal Tract Length Normalization
Certification of Thesis

The work contained in this thesis has not been previously submitted for a degree or diploma at any other higher educational 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: _________________________
Acknowledgments

Firstly, I would like to thank my principal supervisor Professor Sridha Sridharan for his guidance and support. I also wish to acknowledge the support and encouragement my associate supervisor Professor Miles Moody has provided.

I would also like to thank the QUT super-computer services for their assistance in providing their computer facilities. Without their support, I would not have been able to complete the NIST language recognition evaluation.

All members of the Speech Laboratory must also be acknowledged. They have created a very enjoyable atmosphere for doing research and many other discussions which totally unrelated to work. Special thanks to John Dines, Terry Martin and Jason Pelecanos who have provided much invaluable advice and assistance, especially for all the kind help provided by Terry which made the completion of this thesis possible. I would also like to thank Simon Lucey for his help and the accommodation provided during my visit to Carnegie Mellon University. That was an unforgettable trip. Tim Wark, Iain McCowan, Michael Mason, David Cole, Darren Butler, Anthony Nguyen, Brendan Baker, Robbie Vogt, and Kitty Thambiratnam all deserve special mention for their support and friendships.

A final acknowledgment must be dedicated to my family (Mum, Dad, Fred and Roger) for their unconditional love and support not only during the course of my study but during my entire life. I must also thank my friends Leon Dang, Johnny Law, Chris Ng, “Home Ma”, and Sandy Chen who have provided much support.
to allow me to finish the Ph.D. thesis that I never thought I would be able to achieve.

Kim-Yung Eddie Wong

Queensland University of Technology

June 2004
Chapter 1

Introduction

1.1 Motivation

The phenomenon of globalization has brought together people from around the world. However, one barrier to this increase in global communication is that many people still speak different languages and effectively lack a common communication medium. That is, in order to communicate effectively, a language which is mutually understandable by both parties is required. Language Identification (LID) offers a means for providing this medium.

Generally, this identification task is performed by humans and is achieved by the utilization of different cues such as the location and environment, text, physiological characteristics, and various other types of information contained within the speech uttered by the speaker. Unfortunately, the LID task quite often requires familiarity with the subset of languages for the task, a requirement that is impractical for a large portion of the world’s population. Thus, the availability of an automated means for achieving this task presents unlimited advantages.

One application with great potential is the human to machine interface. At
present the most common form of input a computer can handle effectively is the mouse and keyboard. More recently speech enabled technology has emerged as a promising replacement. Unfortunately, existing speech processing technologies are dominantly monolingual or single language applications and the list of speech enabled technology is limited to only a few languages. Effectively this excludes any speaker who is not familiar with one of these languages from exploiting speech as a medium for human to machine communication. A Language Identification system can be used to redirect the speech to a speech recognition application that has been tailored for the detected language. This provides scope for a truly worldwide multilingual communication interface.

1.2 Scope and Aim

There are different aspects incorporated in speech which can be employed to represent the characteristics of a language. Some of these aspects are language specific such as the phoneme inventory, or rules for syllable structure. For instance, Japanese has quite a rigid Consonant-Vowel structure whereas German allows clustering of consonants within syllables.

However, exploiting this type of linguistically motivated information relies heavily on detailed linguistic knowledge for a number of languages in order to identify and extract it from the speech data. Whilst this language specific information is intuitively a good source of information for improving LID performance, the techniques studied in this thesis attempt to remain independent of the languages targeted for identification.

This research direction was chosen because there are still many languages which have not been studied systematically and therefore capturing their language specific characteristics is not feasible. Using this paradigm, speech features (which are intended to represent different aspects of speech information) are extracted
1.2 Scope and Aim

automatically from speech data using an identical method across all languages, neglecting (for the most part) language specific information. One advantage of this approach is that the resultant LID system is capable of identifying any language as long as sufficient training data is available. Nevertheless, depending on the aspect of speech information to be incorporated, a certain amount of transcribed speech data is required in order to develop a front-end system to extract the corresponding speech features. However, the language of these transcriptions does not have to be the same as the target languages.

The target scope for this research is limited to telephone quality speech data spoken by native speakers due to the limited resources that are available for LID research.

Based on the defined target scope, the aim of this research is to develop techniques focused on producing an efficient and more accurate automatic spoken LID system for telephone speech. The system accuracy will be compared to the 1996 National Institute of Standards and Technology (NIST) Language Recognition Evaluation. Both acoustic and phonetic speech information are selected as the source of information to represent the characteristics of each language. A Gaussian Mixture Model (GMM) based LID approach, which uses acoustic features to characterize a language, has been pinpointed for directed study in an attempt to achieve efficiency for the stated aim. The simple structure that the GMM based LID system provides is the underlying motivation for its use and the goal of this work is to develop techniques to further improve the identification speed whilst still producing comparable accuracy to other LID techniques.

In order to improve the accuracy of the LID system, the incorporation of phonetic information will be examined. The LID approach, “Parallel Phone Recognition followed by Language Modelling”, proposes a framework to extract and model the phonetic features. More importantly, this approach has proved to provide good LID performance. Hence, with the structure defined, the goal of this research is to study the usefulness of various levels of phonetic information which may
yield better LID accuracy. Other techniques, which are capable of improving system performance, will also be developed, especially in the area of combining the extracted acoustic and phonetic speech features. The accuracy of the final LID system, which combines all the developed techniques, will be compared to the 1996 NIST language recognition evaluation which focuses on achieving a more accurate identification rate.

1.3 Outline of Thesis

The remainder of the thesis is organized as follows:

Chapter 2 provides the background materials related to the research of automatic spoken language identification. A general LID approach is first discussed and followed by various aspects of speech information that are utilized for LID. This is followed by a brief description of the LID systems utilizing these speech features. The resources suitable for LID research are then outlined. Finally, an international LID technique benchmark event - “the NIST Language Recognition Evaluation” is described.

Chapter 3 presents an investigation based on utilizing an acoustic speech feature with a Gaussian Mixture Model (GMM) based LID system. This chapter proposes the application of a GMM based analysis technique called Universal Background Model (UBM) to LID (which has been successfully applied to the speaker recognition task) in an attempt to improve the efficiency and robustness of the GMM based LID system. Two sets of LID experiments are performed. The first studies the GMM-UBM LID system to show that the UBM technique is well suited for GMM based LID. Second, acoustic features extracted using a range of speech parameterization techniques for LID are examined.

Chapter 4 studies the use of phonetic speech information for LID. A phonetic based LID system, “Parallel Phone Recognition followed by Language Modelling”
1.3 Outline of Thesis

(PPRLM), is adopted as the basic system structure for extraction and utilization of phonetic features. The aim of this chapter is to study the effect of modelling various levels of phonetic detail on LID performance. The phonetic features examined includes both broad and fine phonetic events, and the incorporation of additional phonetic detail for each phonetic event. Finally, the length of phonetic context is also examined in comparison to the amount of phonetic detail given the condition of limited training data.

Chapter 5 investigates techniques focused on improving the accuracy of a LID system. The application of Vocal Tract Length Normalization (VTLN) is first proposed. The rationale behind this technique is that a large number of speakers are presented in the multilingual speech database, hence, speaker variation will be one of the major issues affecting the representative capability of the feature sets extracted from speech data. VTLN, therefore, can be applied to remove unwanted variation, leaving a feature set more suited to discriminating between languages. Another improvement technique studied is the combination of different aspects of speech information. A number of techniques to fuse acoustic and phonetic information at the output score level are examined.

Chapter 6 presents a LID system, which utilizes the acoustic and phonetic information in combination with a number of improvement techniques outlined in previous chapters. The design and implementation of this system was targeted towards the 2003 NIST language recognition evaluation. Full details of the system and its achievement in the evaluation are covered. Additionally, some analysis and improvements to the system after the evaluation event are discussed.

Chapter 7 concludes with a discussion of the achievements made in this thesis. These includes the use of UBM technique for a GMM based LID system, the usefulness of various acoustic and phonetic features, the application of VTLN, and the fusion technique to combine different aspects of speech information. Additionally, the utilization of the proposed LID techniques for the NIST language recognition evaluation, which achieved comparable identification accuracy, are
This research provides contributions which address a number of aspects within the field of LID. They are listed as follows:

- Proposing the use of UBM technique in a GMM based LID system. This has since been adopted by a number of leading researchers in the field as a standard technique for state-of-the-art LID systems [72]. The GMM-UBM system has significantly reduced the amount of computation during both the training and testing stages compared to the standard GMM approach with only a slight sacrifice in LID accuracy. Further, a more robust model can be obtained to represent the characteristics of each language. The proposed system is structured so adding new languages would be straightforward and robust even when the availability of the training data of the new languages is limited.

- The usefulness of the acoustic features that captured information on the physiological aspects of the speaker, such as vocal tract characteristics, in discriminating between languages was identified. This would be beneficial to LID systems exploiting acoustic features. The inclusion of energy and dynamic coefficients (delta and acceleration coefficients) was also shown to be capable of improving the LID accuracy.

- It was shown that in LID systems incorporating phonetic information, the use of more detailed phonetic events is more desirable than the incorporation of longer phonetic context. Thus, under the situation where the amount of training data available does not allow both to be employed, enhancing the phonetic details would produce a more robust LID system.
• Proposing the use of VTLN for LID, which provides higher accuracy especially on systems utilizing acoustic information. Additionally, a novel approach that performs VTLN rapidly during the testing stage is proposed. This approach makes use of the UBM technique to reduce significantly the time required in comparison to other VTLN approaches. This technique is also applicable to other speech processing tasks that require fast speaker normalization.

• It was shown that the incorporation of multiple acoustic and phonetic speech information provides significant improvement in LID accuracy. A linear score weighting fusion technique was shown to be efficient in combining both aspects of speech information.

• Developed a LID system utilizing both the acoustic and phonetic information which has achieved very competitive results with the leading researchers in the 2003 NIST language recognition evaluation. By placing more emphasis on acoustic speech information and broad phonetic information, the developed system has obtained the best results in the 3s test case of the evaluation.

• Techniques developed in this research have also been utilized to perform multilingual speech processing tasks. These include “Multi-lingual Phone Recognition” (see Appendix A) and “Cross-lingual Speech Recognition” (Appendix B).

1.5 Publications Resulting from this Research


(ii) E. Wong, T. Martin, T. Svendsen and S. Sridharan. “Multilingual Phone


(v) E. Wong and S. Sridharan. “Methods to Improve Gaussian Mixture Model Based Language Identification System,” *International Conference on Spoken Language Processing*, vol. 1, pp. 93-96, 2002. (This paper obtained the third place of ISCA prize for the best student paper at ICSLP 2002.)


Chapter 2

Background of Automatic Spoken Language Identification

2.1 Introduction

Automatic spoken Language Identification (LID) is the process of identifying the language spoken within an utterance. The challenge that this task presents is that no prior information is available indicating the content of the utterance or the identity of the speaker. An ideal LID system should accurately exploit different aspects of speech information that are capable of discriminating between different languages and also flexible enough to withstand various speaking styles introduced by different speakers in order to utilize these information in a consistent manner. The main applications for LID systems include call centers (e.g., emergency and customer services) or information directories (e.g., airport, hotel, and tourist attractions) dealing with speakers speaking different languages. In this scenario, based on the automatic identification of the caller’s language, the call can be redirected to a human operator or alternatively, an automatic speech recognition system that is capable of recognizing speech in the caller’s language, thus establishing the intention of the user via a fully automated system. Apart
from applications performed under real time environments, the use of a LID sys-
tem is also applicable in an offline fashion. One important example is to index
or search huge speech data archives and corpora that contain multiple languages,
such as the Internet, thus exploiting the enormous amount of multilingual speech
resources that is available in this medium.

Background materials related to the research of automatic language identification
are provided in this chapter. A general LID approach is first discussed where the
task can be achieved with a simple maximum likelihood classifier. The approach
depicted can serve as a framework for any LID system. Given the definition of
the framework, it is clear that the aspect of speech information to be utilized to
characterize a language is the most important aspect of a LID system. Hence, var-
ious different types of speech information that are utilized for LID are discussed.
These speech information, which are intended to characterize a language, can be
divided into two levels: Spoken and Word Level. Facets of speech information at
the spoken level of a language contain acoustic/phonetic and prosody informa-
tion, which can be acquired directly from the raw speech data. Morphology and
syntax/grammar information belong to the word level of a language. Utilization
of the word level aspect of information would require a large vocabulary speech
recognizer to convert the speech data into word level before further analysis could
be performed.

Excellent reviews of automatic LID systems can be found at [49] and [95]. Hence,
only a brief description of the LID systems utilizing the aforementioned informa-
tion is discussed. Most of these systems employed a combination of techniques
developed for the task of speech recognition and speaker recognition, aiming to
extract and model speech related information. It is well known that the more
aspects of speech information to be utilized by the LID system, the better the
performance improvement. However, the approaches employed to exploit differ-
ent types of speech information, typically based on statistical techniques, require
a large amount of resources in order to estimate the statistics robustly. Hence,
it is obvious that the availability of speech resources for a LID system development presents a major problem and this issue is elucidated further in this thesis. The resources suitable for LID research are then outlined. Finally, an international LID technique benchmark event – “the NIST Language Identification/Recognition Evaluation” is described. The aim of this event is to establish a baseline of performance capability for language recognition of conversational telephone speech. Experiments performed in this thesis usually mirrored the guidelines of the evaluation so that the results can be meaningfully compared to other LID techniques.

2.2 Automatic Language Identification in General

As with many identification tasks, a maximum likelihood classifier can be employed to perform this task. Figure 2.1 depicts a classifier suitable for achieving automatic language identification. Before the classifier can be functional, a training stage is required (Top part of Figure 2.1). The purpose of this stage is to extract speech features, $X$, which represent a particular type of speech information, from the speech training data. A model, $\lambda$, is then created using a modelling technique (usually modelled statistically) to capture the characteristics of each language spoken within the speech data. During testing (bottom part of Figure 2.1), the same kind of speech features are first extracted from the unknown utterance. The feature set is then compared to the model set, $\lambda_l \quad l = 1 \ldots L$, where $L$ is the number of possible languages catered for by the system. The final step is to select the most likely model according to:

$$\hat{L} = \arg \max_{1 \leq l \leq L} P(\lambda_l \mid X).$$  (2.1)
The language, $\hat{L}$, represented by the selected model is identified as the language spoken in the unknown utterance. Therefore, according to this general structure, how and what feature to be extracted from the speech data represents an extremely important aspect of LID research. A correctly chosen modelling technique can properly represent the characteristics of a language described by the speech feature set. Nevertheless, the type of speech feature to be extracted and modelled by the system will contribute largely to the final performance of the system as it defines the speech feature space to discriminate between languages.

2.3 Speech Information Utilized for Language Identification

As aforementioned, the suitable selection of speech features to be utilized in the LID system heavily influences the final performance of the system. The purpose of this section is to describe various aspects of speech information that have been employed by a range of LID systems. The description of these types of information is presented according to a hierarchical level of speech representation, where the higher level speech information can be represented by a sequence of the lower
2.3 Speech Information Utilized for Language Identification

level one. Generally, various types of speech information can be divided into two aspects: Spoken and Word Level. At the spoken level, there is acoustic, phonetic and prosodic information. At the word level, difference between languages can be exploited based on morphology and sentence syntax.

Acoustic, phonetic and prosodic information has been widely used in LID research as they are integral components of the spoken aspect of a language. Importantly, statistical modelling techniques such as Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) can be employed to model their characteristics. The most common method used to extract phonetic features from the input speech is by utilizing a phoneme/phone recognizer. Prosodic features such as the pitch or duration can also be extracted from the speech data and used to perform LID. Different word forms (such as inflection and derivation) and syntax rules (e.g., grammar) belong to the written aspect of a language. These features require information at the word level of the speech data. A large vocabulary continuous speech recognizer therefore is the most effective tool to achieve this task as language specific vocabulary and syntax rules can be incorporated into the recognition network. More details of these LID systems are covered in Section 2.4.

2.3.1 Spoken Level Speech Information

Acoustic Information

An acoustic speech feature by itself is simply a compact representation of the raw speech sound (i.e., acoustic event) at a particular instant in time, characterized by a set of parameters which have unrelated information removed from the speech after a process called *Speech Parameterization*, leaving the more salient elements. Acoustic is one of the most primitive information which can be acquired directly from speech. It also serves as the basic building block for the extraction of other
higher level aspects of speech features such as phonetic and word information. Some of the more popular parameterization techniques employed for speech processing are Mel-Frequency Cepstral Coefficient (MFCC) [7] and Linear Prediction Cepstral Coefficient (LPCC) [41].

**Phonetic Information**

There is a finite set of meaningful sounds which appear in human languages that can be produced physically by humans. Not all of these sounds appear in any given language, hence each language has its own finite subset of meaningful sounds. A phoneme is the abstract sound unit of the phonetic system of a language capable of conveying a distinction in meaning. In addition, sounds that are different but accepted as the same phoneme in a language are called allophones (or a *phone* in terms of the physically produced sound). Phonetic features refer to a sequence of sound units that can be extracted from speech.

Different languages can have different sets of phones representing their phonemes. Phoneme/phone frequencies of occurrence may also differ, where a phoneme/phone may occur in two languages but it may be more frequent in one language than the other. Phonotactics, the rules governing the sequences of allowable phonemes/phones, can be different as well. Hence, phonetic information appears to be extremely a suitable information for exploiting the characteristics of a language.

**Prosody Information**

Prosodic information refers to the duration characteristics of phones, intonation (pitch variation) and stress patterns. Some phones are shared across multiple languages, however their duration characteristics will depend on the phonetic system of the language. Intonation is the variation of tone or pitch used when
speaking. Such a variation can convey different interpretations of a sentence in some languages such as English, where a sentence can change from a statement to a question by a rise in the tone at the end of the sentence. Variation of pitch can convey different meanings in some languages such as Mandarin where the syllables are contrasted by pitch. In some languages, the pattern of stress can determine the meaning of a word, as in English where a noun can become a verb by placing stress on a particular syllable. For example, *permit* means allow to but *PERmit* means permission.

### 2.3.2 Word Level Speech Information

**Morphology**

Morphology is the study of word structure. Hence, LID can be performed at the word level by examining the characteristics of word forms such as inflection, derivation and compounding. The word roots and lexicons are usually different from language to language. In addition, each language has its own vocabulary and its own manner of forming words.

**Syntax**

Syntax is the study of the rules governing the way words in a sentence are connected. The sentence patterns are different among languages. Even when two languages share a word, the word sequence that precedes or follows the word will be different.
2.4 Automatic Language Identification Systems

The majority of LID research has been conducted over the last 20 years and many of the LID systems incorporate the technology and techniques developed from the research on speech recognition and speaker identification/verification. As is the case when humans perform LID, utilization of more speech related information can result in better accuracy. Systems utilizing multiple aspects of speech information can expect to give a higher performance based on the assumption that the information extracted from speech can characterize the language and is not corrupted due to noise or channel effects. In practice, this performance depends on the speech features selected to model the characteristics of a language and the amount of resources available for development.

Here the term features, from a pattern recognition point of view, means the representation of the information extracted from the speech data utilized to discriminate between different classes, where in this case the classes are different languages.

2.4.1 Utilizing Acoustic Information

The acoustic speech features referred to in this thesis are a compact representation of the raw speech signal. The process of extracting these features from speech is usually referred to as Speech Parameterization (various approaches are described in Section 3.5). The basic goal of speech parameterization is to extract salient information from the speech data and disregard any unrelated information embedded in speech. Thus, not only the data rate is significantly reduced, most of the unwanted information can be removed from the speech as well. Instead of studying the speech in time domain, the Short Time Fourier Transform is utilized to analyze the speech in the frequency domain during parameterization. “Short time” in speech analysis means a small segment of speech, typically 20 to 30ms
in duration. The reason for using “short time” speech segment is because the speech signal is slowly varying in time. A short segment of speech therefore can be assumed to be stationary and relatively constant spectral characteristics of the speech can be obtained. Hence, the digitized speech data is first divided into short segments, where a segment is also referred to as a frame of speech, before performing parameterization. As a result of this process, the input speech is transformed into a set of feature vectors. Each feature vector has an array of coefficients representing the more significant information within a frame of speech.

Almost every task of speech processing uses acoustic speech features as the system input. Other speech features, such as phonetic, word level or speaker information, are extracted based on them. Therefore, acoustic features are the most basic building block for speech processing and have been applied to LID in several different ways.

Early LID systems used distance measure criteria between feature vectors to measure their similarity, such as Euclidean and Mahalanobis distance, to match the unknown utterance to the model representing a language. A model in these systems is typically represented by a set of templates which can be obtained through clustering algorithms such as K-mean clustering and Vector Quantization [13]. Systems employing distance measure have been proposed by Leonard and Doddington [35, 36, 37, 38], Chimarusti and Ives [4], Foil [9], Goodman et al [12], and Sugiyama [75].

More recently, more sophisticated modelling techniques were employed. Riek et al [64], Nakagawa et al [53] and Zissman [91] utilized Gaussian Mixture Models (GMMs) to model the feature vectors using a probability density. Techniques using Hidden Markov Models (HMMs) for LID have been outlined by House and Neuburg [18], Savic et al [65], Riek et al [64], Nakagawa et al [53] and Zissman [91]. Artificial Neural Networks (ANNs) have also been employed to model languages [30] using acoustic features.
2.4.2 Utilizing Phonetic Information

The most common method used to extract phonetic features from speech data is by utilizing a phoneme/phone recognizer. This phoneme/phone recognition process pertains to the “Extract Speech Information” block of Figure 2.1 in a LID system and the output phone sequence is the speech features to be modelled. Varies modelling techniques can be applied to the output phone sequence to capture the characteristics of a language. One of the limitations of such an approach is that the error rate of the recognition is relatively high. Therefore, LID systems which capture phonetic information by this means are hindered by the accuracy level of the transcription obtained from the input speech data.

Studies performed by Lamel and Gauvain [31] and Muthusamy et al [50] have shown that the likelihood scores generated from language dependent phone recognizers can discriminate between languages. These systems required phonetically transcribed speech data in order to develop the front-end phone recognizers. Hence, adding new languages into the system can become a problem due to lack of transcribed speech data. One solution is to utilize existing phone recognizer(s) to decode the speech data of those “un-transcribed” languages. A disadvantage of this solution is that the phone recognizer is required to decode speech data from multiple languages, and the use of language specific N-gram language model recognition networks would bias any resulting phone stream. Thus, an N-gram weighted phone recognizer cannot be used. Instead, a “phone-loop recognition network” is employed to obtain the phone sequence, with resultant higher phone recognition error rates. The majority of phonetic information based systems employed this approach to handle the lack of transcription problem and examples of such a system have been proposed by Hazen and Zue [14], Yan and Barnard [88], and Zissman and Singer [96].

Once the speech data of the languages are converted into phone sequences, the most popular modelling technique to capture the phonotactic information is by
2.4 Automatic Language Identification Systems

N-gram Language Modelling (LM) [26]. N-gram LM describes the probability of the occurrence of a phone given that the previous $N - 1$ phones were observed. Due to the limited training data, bi-gram and tri-gram language models are employed by most systems. An alternative method for modelling the phone sequence has been performed by Navratil [54], where a binary tree structure and acoustic pronunciation are employed.

2.4.3 Utilizing Prosodic Information

The most popular prosodic information utilized for LID is the pitch contour. The use of this feature has been studied by Foil [9], Savic et al [65], Hazen and Zue [14], Itahashi and Du [23], and Itahashi et al [24]. In addition, some systems extract pitch characteristics in a segmental approach such as syllable segments as proposed by Hutchins and Thyme-Gobbel [20] and Thyme-Gobbel and Hutchins [78] or broad phonetic event segments studied by Muthusamy [48]. Another prosodic information utilized for LID is the duration of a phonetic event, which has been employed as an enhancement to the phonetic feature set [14, 51].

2.4.4 Utilizing Word Level Information

An intuitive approach to extract word level information from speech is to utilize a large vocabulary continuous speech recognizer. This speech recognizer has implicitly incorporated both the acoustic and phonetic features to aid the speech recognition process. Language specific vocabulary and grammar rules can also be incorporated in the recognition network to determine the correct word sequence. Thus, this type of LID system can be expected to produce very high accuracies as it makes uses of many level of speech information. Mendoza et al [47], Schultz and Waibel [66, 68], and Hieronymus and Kadambe [17] have studied this type of LID system.
Apart from employing a large vocabulary continuous speech recognizer, which requires large amounts of resources to develop, word level information can alternatively be extracted by mapping a phone sequence against a word to phonetic lexicon. In this case, the system only requires a phone recognizer to obtain the phone sequence and a corresponding language dependent lexicon. Ramesh and Roe [60], Kadambe and Hieronymous [27], Lund and Gish [39], Lund et al [40], Thomas et al [77], Matrouf et al [45], and Braun and Levkovwitz [2] have all proposed similar systems. However, these systems also suffer from the same phone recognition error as the phonetic information based LID systems.

The main problem with this type of LID system is it cannot identify languages that do not have enough transcribed speech data to develop the phone recognizer or an even more resource-expensive large vocabulary continuous speech recognizer. In addition, a large amount of computation is required in order to obtain the word level information and therefore these LID systems have very limited applications.

2.5 Difficulties of LID Research

The major difficulty in LID research is the lack of suitable resources. An initial problem in the formative years of LID research was the lack of speech data across multiple languages (most of the earlier LID studies are based on acoustic speech features). Later, more speech data became available including multilingual speech databases [32, 52] suitable for LID research. However, recording across multiple languages is a start, and obtaining accurate phonetic transcriptions becomes the next pressing issue. Especially as most of the LID studies have focused on the utilization of phonetic information, for such a system phonetic transcriptions of the speech data are mandatory. The utilization of word-level information therefore, becomes a more serious problem due to the requirement of a large amount of word level transcriptions and the corresponding phonetic dictionary. For this
reason, the majority of LID techniques have been developed at the phonetic level and did not incorporate word level information such as vocabulary and grammar rules which requires a large vocabulary continuous speech recognition system in order to extract this information from the speech data. Available resources that are suitable for LID research are covered at Section 2.6.

2.6 Resources for LID Research

This section describes the resources suitable for LID research. While there are other speech databases containing multiple languages, such as the SpeechDat series databases [83], which covers most of the European languages, and the CALLHOME telephone speech database [33] released by LDC. Both databases are designed for speech recognition application and not aimed for LID. Thus, they were not employed for the experiments performed in this thesis.

2.6.1 Oregon Graduate Institute Telephone Speech Corpus

The first publicly available database designed for LID experiments was the Oregon Graduate Institute Telephone Speech (OGI-TS) Corpus [52]. There were ten languages (English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese) included in the initial stage of this corpus, where each language contained telephone calls from 90 native speakers. These 90 calls from each language were divided into 50 for training, 20 for development testing and the other 20 for final testing. The male to female speaker ratio was approximately 3:1, which varied by language. Each caller was asked a series of questions designed to elicit:

- fixed, useful vocabulary speech
Table 2.1: The amount of speech data for each language in the OGI-TS database.

<table>
<thead>
<tr>
<th>Language</th>
<th>Amount (Hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>3.47</td>
</tr>
<tr>
<td>Farsi</td>
<td>1.21</td>
</tr>
<tr>
<td>French</td>
<td>1.51</td>
</tr>
<tr>
<td>German</td>
<td>1.47</td>
</tr>
<tr>
<td>Hindi</td>
<td>2.32</td>
</tr>
<tr>
<td>Japanese</td>
<td>1.09</td>
</tr>
<tr>
<td>Korean</td>
<td>1.02</td>
</tr>
<tr>
<td>Mandarin</td>
<td>1.29</td>
</tr>
<tr>
<td>Spanish</td>
<td>1.72</td>
</tr>
<tr>
<td>Tamil</td>
<td>1.47</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>1.20</td>
</tr>
</tbody>
</table>

- domain-specific vocabulary speech
- unrestricted vocabulary speech.

The “story before tone” (max. duration 50 seconds) and “story after tone” (max. 10 seconds) utterances together form the 1-minute unrestricted vocabulary speech portion of each call.

The corpus was later extended with additional recordings for each of the ten languages above, plus 200 Hindi calls were added, making a total of 11 languages collected. The amount of speech data for each language are listed at Table 2.1.

The initial stage of the corpus had certain speech data transcribed in broad phonetic labels. 500 utterances labelled for each of the ten languages (two utterances per call for 25 calls per language). The seven broad phonetic classes were vowel, fricative, silence or closure, stops, pre-vocalic sonorant, inter-vocalic sonorant, post-vocalic sonorant. Fine phonetic transcriptions of six languages (English, German, Hindi, Japanese, Mandarin and Spanish), which transcribed only the “story before tone” utterances, were later released for the extended version of the corpus.
2.6.2 OGI 22 Language Telephone Speech Corpus

The 22 Language Telephone Speech Corpus [32] is another multilingual speech database that was designed for LID research. It consists of telephone speech from 22 languages: Eastern Arabic, Cantonese, Czech, Farsi, French, German, Hindi, Hungarian, Japanese, Korean, Malay, Mandarin, Italian, Polish, Portuguese, Russian, Spanish, Swedish, Swahili, Tamil, Vietnamese, and English. The corpus contained fixed vocabulary utterances (e.g. days of the week) as well as fluent continuous speech. Each of the 50,192 utterances was verified by a native speaker to determine if the caller followed instructions when answering the prompts. Approximately 13,000 utterances have corresponding orthographic transcriptions.

2.6.3 LDC CALLFRIEND Telephone Speech Database


Each language/dialect consisted of 20 telephone conversations (two sided) for training, 20 conversations for development and 20 conversations for testing. The duration of each conversation was up to 30 minutes. Unfortunately, none of this speech data is transcribed in any form.
2.6.4 International Phonetic Alphabet

Inevitably, researchers have to deal with multiple languages during the study of LID techniques. Thus, it would be convenient to have a set of unified phone symbols across all languages. The International Phonetic Alphabet (IPA) [22] and Worldbet [16] (the ASCII encoding of IPA) are designed for this purpose. Figure 2.2 shows a table of consonants represented in IPA.

The International Phonetic Association was established in 1886 with the aim of developing a set of symbols which would be convenient to use, but comprehensive enough to cope with the wide variety of sounds found in the languages of the world. This set of symbols named International Phonetic Alphabet (IPA) is based on the Roman alphabet and has been revised in 1993 and updated in 1996.

Worldbet is an attempt to have a phonetic alphabet which covers all of the world’s languages in a systematic fashion. It is an ASCII version of the IPA plus a number of symbols which were found useful in database labelling, which are not in the IPA set. Additionally, the symbols includes diacritics which can be easily stripped to become the base phone symbol.
2.7 NIST Language Identification/Recognition Evaluation

As stated at the 2003 evaluation, the aim of the National Institute of Standards and Technology (NIST) [55] Language Recognition evaluation is intended to establish a new baseline of current performance capability for language recognition of conversational telephone speech and to lay the groundwork for further research efforts in the field. A workshop was also hosted after the event which allows participants from around the world to discuss and exchange ideas. Therefore, LID experiments based on the evaluation specification form a good benchmark to evaluate the LID techniques studied and allowed the corresponding experimental results to be compared with other LID systems.

2.7.1 1993/1994 Language Identification Evaluation

The specification of the NIST 1993 Language Identification evaluation is defined according to the initial stage (10 language version) of OGI-TS corpus (Section 2.6.1). The development test data defined by the corpus are employed as the evaluation test data for the 93 LID evaluation and no development test data are defined. For the purpose of LID testing, only the “story before tone” utterance, which is the unrestricted vocabulary recording that is approximately 45 seconds long on average, is employed. Two different test cases are defined for this evaluation: whole-story (which is also designated as 45s in the experimental results presented in this thesis) and 10s, where a “story before the tone” utterance will be used for 45s testing if its duration is longer than 25 seconds and the 10s test segments are cuts from the same utterances used in the 45s test. This resulted in 178 utterances for 45s test and 575 utterances for the 10s test.

The extended version (11 languages) of the OGI-TS corpus was adopted by NIST for the 1994 LID evaluation including the definition of partitioning the develop-
ment and evaluation test data set defined by the corpus. As a result, 194 utter-
ances for 45s and 635 utterances for 10s are defined for development test, and
187 utterances for 45s and 625 utterances for 10s are defined for evaluation test.

The performance of the LID systems of these two evaluation are evaluated accord-
ing to the LID accuracy of “Language Pair” (a task that identify two languages
only) and all 10/11 languages.

2.7.2 1996 Language Recognition Evaluation

Language recognition research has focused on language identification tasks in the
past. The evaluation in 1996 directed emphasis to research aimed at developing
a general base of technology that could be ported to various language recogni-
tion tasks with minimum effort. The task to be evaluated is the detection of a
given target language. Given a test segment of speech, a target language will be
designated as a “test hypothesis”, and the task is to determine whether this test
hypothesis is true or false. The performance of a detection system is character-
ized by its miss and false alarm probabilities, and these probabilities will therefore
serve as the basis for evaluating system performance on the language detection
task using the Detection Error Tradeoff (DET) Curve. The DET curve [42] plots
the miss probabilities on the ordinate axis and the false alarm probabilities on the
abscissa axis. Both axes are in normal deviate scale, which resulted in a more lin-
ear curve than the Relative Operating Characteristic (ROC) curve that was used
to evaluate system performance prior to the DET curve being proposed. Figure
2.3 shows a sample DET curve of language detection results, where the curve
closer to the origin shows better performance. The Equal Error Rate (EER), as
indicated in Figure 2.3, can be used as a simpler indication of the performance
of the system.

The main training data set defined for this evaluation is based on the LDC CALL-
FRIEND (Section 2.6.3) database containing 12 languages, where three of these
have two different dialects. Participants are allowed to employ data from other sources.

The test segments have three different durations: 3s, 10s, and 30s. Development test data are taken from each of 20 conversations of the development data portion of CALLFRIEND database for each of the 15 target languages and dialects. There are over 3000 development test segments (1147 for 30s, 1172 for 10s and 1174 for 3s). Evaluation data comprised 80 test segments of each of the three test durations for each of the 15 target languages and dialects. These primary test data are supplemented with up to 320 segments of American English from other sources for each of the three test durations. Thus there are over 4000 evaluation
test segments (1492 for 30s, 1502 for 10s and 1503 for 3s).

The scale of this evaluation is larger and therefore capable of reflecting more realistic system performance under telephone speech.

2.7.3 2003 Language Recognition Evaluation

The format of the 2003 evaluation is very similar to 1996, where the primary condition is focused on the proposed 12 languages. In addition, an unknown language (Russian) is presented in the evaluation test data to test the ability of the language recognition system to reject language not known to the system.

The CALLFRIEND database is again defined as the primary training data set. The development test data set is expanded to include the development and evaluation test data from the 1996 evaluation. Hence, the amount of development test data is doubled compared to last evaluation. The evaluation test data is extracted from the unreleased version of CALLFRIEND database with addition English and Japanese data extracted from other sources. However, the primary condition evaluation included test data from the exposed CALLFRIEND database only. There are up to 3840 evaluation test segments (1280 for each test segment duration and reduced to 960 test segments for the primary condition evaluation). The performance of a language recognition system is again evaluated according to the DET curve.

2.8 Summary and Discussion

Background materials relevant to the field of LID research are covered in this chapter. A general approach for a LID system is first outlined. It highlights that how and what speech features to be extracted and utilized by the system represent the most important aspect of LID research. Various aspects of speech information
and the LID systems utilizing them are then discussed. Amongst these LID systems, the majority utilized the phonetic speech information to describe the characteristics of a language. This observation implies that the most effective LID approach for discriminating between different languages is based on the phonetic information. A concise overview outlining the difficulty of LID research follows. It concludes that the major problem faced in LID research is the lack of speech resources. Hence, a description of the resources available is given. Finally, an international LID system benchmark event that was participated in by most of the active LID researchers around the world is described. This evaluation has a set of guidelines for evaluating the performance of a LID system for telephone conversational speech data, and these guidelines have been adapted by this thesis for LID experiments.

The trend of the LID research has shown that in order to develop a state-of-the-art LID system, both the acoustic and phonetic speech information should be utilized. This stems from the fact that transcribed training data, either phonetically or orthographically, are usually very limited (as discussed at Section 2.5). Hence, the use of higher level speech information becomes infeasible. Further, in the field of linguistics, languages can be defined via a number of phonetic events, providing a succinct representation for the sounds of that language. Given that speech recognition technology provides a mechanism for extracting phonetic information, the use of phonetic information seems to be better suited in comparison to higher level features such as word. Study of phonetic speech information is presented in Chapter 4.

Nevertheless, there is a major disadvantage in using phonetic speech information. It is a difficult task to identify phonetic events correctly within a given utterance due to insufficient transcribed training data necessary to estimate robustly the parameters for the phonetic event models. Additionally, the realization of a phonetic event can vary according to the context of the neighboring events. These context differences are hard to encapsulate within the model parameters. In
addition, these allophonic variations produced in conversational or rapid spoken speech are exacerbated due to additional co-articulatory effects such as deletion, substitution and insertions. LID system performance, therefore, would suffer due to the error induced in the recognized phonetic output. Hence, one of the experiments performed in Chapter 4 addresses this issue and studies the effect on LID performance by increasing the recognition accuracy of phonetic events with a smaller set but broad sense of events.

Another disadvantage imposed in phonetic based LID systems is the extra processing time required to extract the phonetic sequences from speech. When efficiency is required by the LID system, the use of acoustic information for LID become apparent as the process of extracting acoustic features is relatively faster. Further, the underlying concept of recognizing speech and speaker are more or less based on characterizing the variation of acoustic features and it appears that the same theory can be extended to the case of a language. Moreover, systems based on acoustic features achieved the best results in speaker recognition applications, while reaffirming the fact that acoustic features are capable of discriminating between different classes. Nevertheless, the utilization of acoustic features does not take advantage of the phonetic information presented in speech. A study performed on utilizing acoustic speech information is given in Chapter 3.

Based on the aforementioned motivations, research in this thesis is focused on acoustic and phonetic speech information. Other types of speech information, such as prosody, can then be formed as complementary information to improve the accuracy of the system. However, this also incurred extra computation time in order to extract these speech features.
Chapter 3

Language Identification Utilizing Efficient Gaussian Mixture Analysis

3.1 Introduction

As outlined in Chapter 2, Acoustic speech features are one of the major sources of information utilized for the LID task. Techniques which have been employed to model the characteristics of these features include Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM). One key difference between these two modelling techniques is that HMM can incorporate time sequence information about the feature vectors. When considering LID, however, it is difficult to select an appropriate HMM topology because the acoustic events of a language do not occur in a particular time order. Additionally, it is hard to determine which characteristics of a language a HMM state should be representing and the corresponding transition between these states. Given this, previous studies into modelling acoustic features with HMM employed an ergodic (fully connected) HMM topology.
The use of GMMs, however, does not require such a selection of topology. Although a GMM totally discards the temporal information about the acoustic features, this loss of information will result in less identification error compared to the errors introduced by an inappropriate selection of HMM topology. Similar outcomes were obtained by Riek et al [64] and Zissman [91], where LID systems employed GMMs achieved better performance than HMM based systems.

In this chapter a LID system which uses GMMs to model acoustic speech features is described. GMM based LID systems have previously been studied by Riek et al [64], Nakagawa et al [53] and Zissman [91], however following these studies into GMM based LID, LID research refocused on the study of phonetic information. One reason is that the utilization of acoustic information did not perform as well as phonetic information. This performance difference was imposed by the hardware limitation at the time of study, as the amount of computation required to train a GMM is dependent on the number of mixture components employed and the dimension of the feature vector. Due to the speed limitation imposed by the hardware only around 100 mixture components and 24 dimension feature vectors were employed in these studies. The characteristics of a language are far more complex than this “size” of GMM could model, hence, with the help of today’s faster computing, the feature space of a language can be modelled using a “larger” GMM.

Another reason to extend the work on GMM based LID approach is the Universal Background Model (UBM) technique. This GMM technique has been successfully applied to the task of speaker recognition. In this technique, a GMM (or the UBM) comprising a high number of mixture components is created (which will be trained with data from all the different classes to be identified) in an attempt to provide prior information of the entire acoustic feature space. With a slight loss of accuracy, due to certain approximations made, this technique can improve the efficiency of GMMs significantly at both the training and testing stages. The reduction in computational requirements allows more mixture components to be
employed to fulfil the need of a “larger” GMM to model the characteristics of a language as aforementioned. In addition, most LID applications require faster than real-time operation in order to perform the task.

In previous studies of GMM based LID, the only speech features considered were Mel-Frequency Cepstral Coefficients (MFCCs). Even though MFCCs have been successfully used for the task of speech and speaker recognition, the acoustic information extracted from this technique does not necessarily retain the information required to best discriminate between languages. To date there have been no studies conducted which examine what representation of acoustic information is most useful for language discrimination. In an effort to address this question, the LID performance achieved using four different feature extraction techniques is presented in this chapter.

Section 3.2 describes the details of a GMM based LID system. This is followed by the description of the UBM technique and its application to the GMM LID system. Two LID experiments are performed in this chapter. The first outlines the GMM-UBM LID system and shows that the UBM technique is well suited for GMM based LID. Compared to the standard GMM approach, comparable accuracy can be achieved with a significant reduction in time for system training and testing. In the second experiment, acoustic features calculated using different speech parameterization techniques are examined. The purpose of this experiment is to establish which parameterization technique provides better language discrimination. The experimental results have show that techniques which exploit vocal tract characteristics provide better discrimination between languages.

### 3.2 Gaussian Mixture Model Based LID System

Basically, a GMM based LID system is a classifier with each class (a language in this task) modelled by a GMM. Language classification is performed according
to the likelihood score calculated by the *Language-GMMs* against a given feature vector. To determine the language in LID testing, multiple feature vectors are used. That is, likelihood scores are accumulated for each language and the decision making is delayed until all the feature vectors are processed.

The system structure of a GMM based LID system is very simple. Accordingly, the computational requirements for processing are low. This simplicity advantage also extends to the development phase for such a system. The parameters which are required for estimation of the classifier are those used to describe the *Language-GMMs*. An important advantage of utilizing GMMs is that no phonetic transcriptions of the training data are required. This is a very important feature since transcribing speech data phonetically is an expensive and time consuming task.

### 3.2.1 Gaussian Mixture Model

A Gaussian Mixture Model (GMM) is a weighted sum of multi-variate Gaussian mixture components which model the probability density function for a set of feature vectors given by:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} w_i b_i(\vec{x})$$

where $\vec{x}$ is a $D$-dimensional feature vector, $i$ is the mixture component index $(1 \leq i \leq M)$, $w_i$ is the weighting applied to the Gaussian mixture component $i$, such that $\sum_{i=1}^{M} w_i = 1$ and $b_i(\vec{x})$ is a $D$-variate Gaussian density function:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2}(\vec{x} - \mu_i)' \Sigma_i^{-1}(\vec{x} - \mu_i) \right\}$$

with mean vector $\mu_i$ and covariance matrix $\Sigma_i$. The complete GMM is described
by

$$\lambda = \{w_i, \bar{\mu}_i, \Sigma_i\} \quad i = 1, \ldots, M. \quad (3.3)$$

The estimation of the GMM parameters, \( \lambda \), is accomplished by an iterative process, termed the Expectation-Maximization (EM) algorithm [8]. The EM algorithm, given the initial parameters \( \lambda \), will estimate the new parameters such that

$$p(X \mid \hat{\lambda}) \geq p(X \mid \lambda),$$

where \( X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\} \) is a set of training feature vectors and \( \hat{\lambda} \) are the new parameters. On each EM iteration, the parameters are updated using the following equations:

**Mixture Component Weights:**

$$\hat{w}_i = \frac{1}{T} \sum_{t=1}^{T} P(i \mid \vec{x}_t, \lambda) \quad (3.4)$$

**Mixture Component Means:**

$$\hat{\mu}_i = \frac{\sum_{t=1}^{T} P(i \mid \vec{x}_t, \lambda) \vec{x}_t}{\sum_{t=1}^{T} P(i \mid \vec{x}_t, \lambda)} \quad (3.5)$$

**Mixture Component Covariances:**

$$\hat{\Sigma}_i = \frac{\sum_{t=1}^{T} P(i \mid \vec{x}_t, \lambda) \vec{x}_t \vec{x}_t'}{\sum_{t=1}^{T} P(i \mid \vec{x}_t, \lambda)} - \hat{\mu}_i \hat{\mu}_i' \quad (3.6)$$

where \( \hat{w}_i, \hat{\mu}_i \) and \( \hat{\Sigma}_i \) are the updated GMM parameters.

The *a posteriori* probability for the Gaussian mixture component \( i \) is given by

$$P(i \mid \vec{x}_t, \lambda) = \frac{w_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} w_k b_k(\vec{x}_t)} \quad (3.7)$$
An important property of GMMs is that the linear combination of Gaussians is capable of providing a smooth approximation for any arbitrarily shaped distributions [62]. As “real world” data have multi-modal distributions, GMM provide a great tool to model the characteristics of the data. Another extremely useful property of GMMs is the possibility of employing diagonal covariance matrix instead of a full covariance matrix. Modelling a distribution with a diagonal covariance matrix only provides scope for the use of mixture components whose eigenvectors are aligned with the feature space axes. Thus, if a distribution has “off axis components”, a full covariance matrix is required to represent the correlation between different elements of the feature vectors. Nevertheless, the same “off axis components” distribution can also be fitted approximately using a linear combination of “axis aligned components”. That is, a set of $M$ off axis Gaussians can be represented alternatively with a larger set of axis aligned Gaussians. Figure 3.1 depicts a simple example where an “off axis” Gaussian is fitted with four “axis aligned” Gaussians. This means that a diagonal covariance matrix GMM is capable of approximately modelling the correlations between the elements of the feature vector, which originally required a full covariance matrix, and therefore, reduced the amount of computation significantly.
3.2 Gaussian Mixture Model Based LID System

3.2.2 GMM based Language Identification

Given an unknown speech utterance, \( X = \{ \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_T \} \), the language spoken within that utterance is classified as

\[
\hat{L} = \arg \max_{1 \leq l \leq L} P(\lambda_l | X)
\]  

(3.8)

where \( \lambda_l \) is the GMM of language \( l \) and \( L \) is the number of languages the system can identify. Apply Bayes’ rule, Equation 3.8 becomes

\[
\hat{L} = \arg \max_{1 \leq l \leq L} \frac{p(X | \lambda_l)P(\lambda_l)}{P(X)}
\]  

(3.9)

For the task of LID, the chance for a language to be spoken in a test utterance is usually assumed to be equally probable. In addition, \( P(X) \) is the same across all languages, hence the final LID decision rule can be reduced to

\[
\hat{L} = \arg \max_{1 \leq l \leq L} p(X | \lambda_l)
\]  

(3.10)

or alternatively based on the average log likelihood score

\[
\hat{L} = \arg \max_{1 \leq l \leq L} \frac{1}{T} \sum_{t=1}^{T} \log p(\bar{x}_t | \lambda_l)
\]  

(3.11)

where \( p(\bar{x}_t | \lambda_l) \) is given in Equation 3.1. Figure 3.2 shows a block diagram of both the training and testing stages of a GMM based LID system.
3.3 Universal Background Model Technique

3.3.1 Background Speaker Modelling

For the task of speaker verification, a likelihood score of the input speech data against the model of the claimed speaker is calculated to decide whether to accept or reject the speaker. In order to obtain a more robust likelihood score with the purpose of protecting the system from different types of impostor, such as very similar or dissimilar voices to the real speaker, a set of background speaker models are prepared. These background speaker models are employed to normalize the likelihood score in such a way that the real speaker will obtain a higher score and impostors a lower score. One disadvantage of this approach is that the variation of impostor characteristics is limited by the number of background speaker models available during score normalization. To resolve this, Reynolds [61] proposed the use of a single GMM to act as background speaker model, namely *Universal Background Model* (UBM) to replace the background speaker set. As the name implies, a UBM represents the characteristics of all the background speakers or impostors. It is a single GMM that is comprised of a large number of mixture components (typically >128) in order to model the characteristics of a large number of speakers.

![Block diagram of a GMM based LID system.](image-url)
3.3 Universal Background Model Technique

3.3.2 Universal Background Model

The UBM technique has been applied successfully to both speaker verification and identification tasks and has become the dominant technique used in the NIST speaker recognition evaluation [55]. The main benefit of this technique is it provides computation reduction in both training and testing stages. In addition, this technique allows a higher number of mixture components for each GMM (for both the speaker model and background speaker model), compared to the standard GMM approach, so that more details of the acoustic features can be captured.

To implement the UBM technique, a large amount of background speaker data is required in order to estimate robustly all the parameters of the very high order GMM. To be able to train a UBM there must be sufficient variation in the speech data in order to cover the general acoustic features of all speakers and not overly tune the model to a particular class of speakers.

Speaker Model Training

As aforementioned, a UBM represents the characteristics for a lot of speakers. Hence, the UBM can be employed as a base model to create the GMM for each speaker through Bayesian Adaptation [10]. The advantage of creating speaker model by adaptation is the amount of computation is a lot less than the standard EM algorithm. The number of mixture components employed in a standard GMM based speaker verification system is usually limited (typically around 64 Gaussians only, compared to >128 in UBM technique) due to the computational requirement of the EM algorithm. Hence, the adaptation approach solved this problem and allows the speaker model to capture finer details of the acoustic features. In addition, the use of adaptation instead of EM training has the advantage that the UBM has “seen” many different types of data. Therefore, a larger acoustic space is created and any unseen data (acoustic events) will not
result in bias to any particular models, as each model is adapted from the UBM. This produces a more robust speaker model. The standard GMM approach cannot handle this case, as the data used to train the GMM is limited to that particular speaker.

Bayesian adaptation is performed using a fixed relevance factor $r$ to adapt the mixture component weights, means and variances of the GMM in the following way [61]:

Given a UBM (which acts as the prior distribution) and training observations from the claimant speaker, $X = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_T\}$, this process determines the probabilistic alignment of the training data with the prior mixture components in the UBM. For mixture component $i$, the probabilistic count of training observations, $n$, mapping into the prior parameters $(\omega_p, \mu_p, \sigma_p^2)$ is defined as

$$n = \sum_{i=1}^{T} P(i \mid x_i). \quad (3.12)$$

(Since all adaptation equations refer to a single mixture component, the $i$ notation on parameters is omitted from the following equations for clarity.) The adaptation coefficient for this mixture component is

$$\alpha = \frac{n}{n + r} \quad (3.13)$$

where $r$ is a fixed relevance factor. The adapted mixture weight is defined as

$$\omega_a = [\alpha \omega_s + (1 - \alpha) \omega_p] \gamma \quad (3.14)$$

where
is the mixture weight for the new observations and $T$ is the total number of training observations. The scale $\gamma$ is computed over all adapted mixture weights to ensure they sum to unity.

The adapted mixture mean is defined as

$$\mu_a = \alpha \mu_s + (1 - \alpha) \mu_p$$  \hspace{1cm} (3.16)$$

where

$$\mu_s = \frac{1}{n} \sum_{t=1}^{T} P(i \mid x_t)x_t$$  \hspace{1cm} (3.17)$$

is the mixture mean of the new observations.

The adapted mixture variance is defined as

$$\sigma_a^2 = \alpha E\{x_i^2\} + (1 - \alpha)(\sigma_p^2 + \mu_p^2) - \mu_a^2$$  \hspace{1cm} (3.18)$$

where

$$E\{x_i^2\} = \frac{1}{n} \sum_{t=1}^{T} P(i \mid x_t)x_t^2$$  \hspace{1cm} (3.19)$$

is the expected squared value of the new observations in the mixture component.
Testing

The computation benefits of the UBM technique are not just useful for the training of speaker model, they are also applicable during the testing stage by employing the follow two properties:

- Only a few of the mixture components within a GMM contribute significantly to the likelihood score for a feature vector.
- The mixture components of the adapted speaker model share a certain correspondence with the UBM, since the speaker model is adapted from it.

Therefore, the likelihood score of the speaker model against the input feature vector can be calculated by scoring only the more significant mixture components (the top five mixture components in [61]). According to the correspondence of mixture components between the UBM and speaker model, those significant mixture components can be obtained by selecting the mixture components from the UBM having the highest score. Figure 3.3 shows a diagram depicting the process of using the significant mixture components testing technique. By employing this mixture testing strategy, the computation required for testing is reduced significantly. A pitfall of this significant mixture components testing technique is the possible degradation of accuracy. However, McLaughlin et al [46] has shown that the sacrifice in accuracy is small.

3.3.3 Application of UBM Technique in LID

The discussion of the UBM technique provided above is in the context of the speaker verification task and some modifications are required for the task of speaker or language identification. In an identification task, no background model is required due to the close-set selection nature. Hence, the meaning of the
Figure 3.3: Steps to utilize the UBM technique to calculate the likelihood score of a given feature vector against a speaker model during testing. Step 1: Select the Gaussians from the UBM obtaining the highest likelihood as the significant mixture components. Step 2: Sum the log likelihood scores of the corresponding significant mixture components (those selected from the UBM) of this speaker model. This summed score is then utilized as the approximated likelihood of this speaker instead of calculating from all mixture components.

The aforementioned Universal Background Model is slightly changed. In this task, the UBM should be renamed Universal Speaker Model for speaker identification or Universal Language Model for the LID task, in which the model is essentially representing the characteristics of all different classes of the identification task. But to avoid confusion, the same term UBM will be employed thereafter because in spite of slightly different interpretations, the Universal Background Model performs the same function in both cases.

To implement the UBM technique for LID, first the UBM is trained by pooling data from all languages. Note that the LID task does not require the UBM to perform background normalization, hence the UBM can be trained with all available data to obtain a more complete description of the acoustic feature space. In contrast to speaker verification/identification, a set of background speaker data, which is separated from the claimant speaker’s data, is required to be prepared to train the UBM. Thus, the implementation of the UBM technique for
LID is a lot simpler in terms of data preparation. Once a UBM is obtained, the Language-GMM is then created by performing adaptation from the UBM using language specific data. The process of LID system training is depicted in the top section of Figure 3.4.

The same significant mixture components test strategy from the UBM technique is directly applicable to LID as both of the GMM-UBM properties mentioned in Section 3.3.2 are fulfilled. The process of LID testing utilizing the UBM technique is depicted at the bottom part of Figure 3.4. The reduction in computational requirements benefit is the most important motivation to apply the UBM technique, because most of the LID applications require faster than real-time operation in order to perform the task.

By employing this approach in an LID system, the computation speed was rapidly improved. This can be shown as follows. Given that both the GMM and UBM have \( N \) mixture components, we choose to test the top \( C \) mixture components for \( L \) languages using the UBM technique. The number of mixture components required to test \( R \) is:

\[
R = N_{UBM} + C_{GMM} \times L
\]
Alternatively, for the standard GMM system utilizing $N$ mixture components, the number of mixture components required to test will be

$$R = N_{GMM} \ast L$$

(3.21)

For a case of 10 languages using a 512 mixture component GMM and determined the top five mixture components from the adapted models. This gives only $R = (512 + 5 \ast 10) = 562$ mixture component tests compared to $R = (512 \ast 10) = 5120$ mixture component tests for standard GMM system. This gives us a 900% computation improvement.

### 3.4 Acoustic Information based LID Experiments

The aim of these experiments is to show the GMM-UBM LID approach, which utilized acoustic speech information, is viable to replace the standard GMM approach without significant sacrifice in LID accuracy, so that all the advantages provided by the UBM technique can be utilized. Additionally, experiments are performed to show the simplicity of adding new languages to the GMM-UBM LID system.

#### 3.4.1 System Setting

The feature vectors used comprised 12 order Mel-Frequency Cepstral Coefficients (MFCCs) (more details in Section 3.5.1) derived from 20 filterbanks. Each feature vector is extracted at 10ms intervals using a 32ms window of bandlimited (300-3400 Hz) speech. Each frame of speech is pre-emphasized by the first order difference equation $s'_n = s_n - 0.97s_{n-1}$. A Hamming window was then applied
to the speech frame. Since the experiment involved telephone speech, Cepstral Mean Subtraction (CMS) was applied to the MFCCs to reduce the linear channel effects. The corresponding delta coefficients were computed over a window length of 15 frames. Initially a shorter delta coefficient window length was utilized, however, preliminary experiments indicated an improvement by extending this window length. A longer delta window length may be able to encapsulate more of the temporal information that is specific to language discrimination, particularly when the GMM does not use information across frames. Finally, the delta coefficient of the frame energies (over the same window size) was appended to the vector.

All GMMs (including UBM and Language-GMM) comprised 512 mixture components with diagonal covariance matrix. The estimation of the UBM parameters is accomplished by the EM algorithm. For more rapid GMM convergence, the mixture means, weights and variances are seeded by statistics determined by a K-means vector quantization estimate of the feature vectors [59]. Language-GMMs are created with one iteration of adaptation from the UBM with all the available speech data from the OGI-TS database. A frame energy based silence detection approach is employed to remove silence from all speech data [63]. Experiments are based on the NIST 1994 LID evaluation guidelines, which included 11 languages and 187 utterance for 45s test case and 625 utterance for 10s.

3.4.2 Experimental Results

Several GMM based LID approaches were tested: the standard GMM system (both 64 and 512 mixture components), the GMM-UBM system, the GMM-UBM system with the top five mixture components being tested and the use of a standard English GMM in place of the UBM. The variation in performance due to the number of significant mixture components selected for UBM testing was also analyzed.
3.4 Acoustic Information based LID Experiments

Table 3.1: Experimental results (% correct) of applying the UBM technique on GMM based LID system.

<table>
<thead>
<tr>
<th>GMM based LID approach</th>
<th>45s</th>
<th>10s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard GMM (64 mixture components)</td>
<td>53.5</td>
<td>50.1</td>
</tr>
<tr>
<td>Standard GMM (512 mixture components)</td>
<td>57.2</td>
<td>55.4</td>
</tr>
<tr>
<td>GMM-UBM</td>
<td>53.5</td>
<td>51.8</td>
</tr>
<tr>
<td>GMM-UBM with top 5 mixture components test</td>
<td>53.5</td>
<td>51.5</td>
</tr>
<tr>
<td>Replace UBM by standard English-GMM</td>
<td>53.5</td>
<td>50.4</td>
</tr>
<tr>
<td>Replace UBM by standard English-GMM with top 5 mixture components test</td>
<td>53.5</td>
<td>50.7</td>
</tr>
</tbody>
</table>

Table 3.1 shows all the experimental results given as the percentage of utterances correctly identified. First, by comparing the results between the two different numbers of mixture components employed by the standard GMM approach, the increase in mixture components from 64 to 512 has resulted in an improvement of LID accuracy. Thus, the results suggest that a higher number of mixture components are required in order to model properly the characteristics of a language with a GMM.

The results of the UBM-GMM system and the equivalent system employing the speed up testing approach have shown that these systems can obtain a comparable performance to the standard GMM systems (both the traditional setting with 64 mixture components and the large computational requirement setting with 512 mixture components). Results also show that using a few of the more significant mixture components for likelihood score calculation resulted in little to no measurable degradation in performance.

The purpose of the experiment replacing the UBM with a standard Language-GMM (an English model) is to show whether the system could be made to accommodate further Language-GMMs without retraining the UBM. In this configuration, the use of the English model to act as the UBM only reduced the LID performance slightly compared to the use of real UBM. Hence, as supported by the experimental result, this is a viable approach to building a GMM based LID system under the situation of lack of training data or when flexibility of adding new languages to the system is required.
Another experiment examined the varying performance of the GMM-UBM LID system when the number of significant UBM mixture components tested was altered. The results are shown in Figure 3.5 and it indicates that the system obtained almost constant accuracy with a varying number of significant mixture components tested.

Finally, an alternative way to create Language-GMM is studied where models are initialized using the parameters of the UBM instead of the K-means vector quantization approach. This is a hybrid approach between the standard-GMM and the GMM-UBM technique where EM algorithm is required when training the Language-GMMs, however the top five mixture components test technique is applicable during testing. The results are shown in Table 3.2. It shows that when all the Language-GMMs are initialized with the UBM parameters, system accuracy is similar to the VQ initialization approach. In addition, utilizing the top five mixture components test induces only a slightly drop in LID accuracy. Nevertheless, by comparing this hybrid approach to the GMM-UBM technique, the extra computation required at training time does not improve the system’s performance.
3.5 Experiments Using Different Parameterization Techniques and Features

Table 3.2: Experimental results (% correct) of creating Language-GMM with EM but initialized from UBM.

<table>
<thead>
<tr>
<th>GMM based LID approach</th>
<th>45s</th>
<th>10s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard GMM (512 mixture components)</td>
<td>57.2</td>
<td>55.4</td>
</tr>
<tr>
<td>Language-GMM initialized from UBM</td>
<td>57.8</td>
<td>55.8</td>
</tr>
<tr>
<td>Language-GMM initialized from UBM with top 5 mixture components test</td>
<td>54.0</td>
<td>52.5</td>
</tr>
<tr>
<td>GMM-UBM with top 5 mixture components test</td>
<td>53.5</td>
<td>51.5</td>
</tr>
</tbody>
</table>

3.5 Experiments Using Different Parameterization Techniques and Features

The GMM based LID system solely depends on acoustic speech features. Thus the acoustic feature vectors employed directly influence the accuracy of the system. So far there are no studies conducted which examine what type of acoustic features are useful for discriminating between languages. Therefore, these experiments are focused on examining which acoustic features are more suitable for the GMM based LID systems. Mel-Frequency Cepstral Coefficient (MFCC) and Linear Prediction Cepstral Coefficient (LPCC) have been previously studied for GMM based LID approaches and many other speech processing tasks, and hence are included in this comparison. Other speech parameterization techniques selected are Mel-Cepstral Coefficient (MCC) and Perceptual Linear Predictive coefficient (PLP). MCC is a representation of the LPCC in Mel-scale which, in other words, represents LPCC perceptually. Hence, it is selected for comparison to find out whether representing speech perceptually will provide better language discrimination compared to the linear-frequency approach. PLP attempts to mimic the properties of the human auditory spectrum by transforming the spectrum. Additionally, it provides a more compact representation than the other parameterization techniques mentioned above. The selection of PLP is motivated by its perceptual representation and the successful results achieved in many speech recognition tasks. More detailed descriptions of each of these parameterization techniques are given in the following sections.
3.5 Experiments Using Different Parameterization Techniques and Features

3.5.1 Mel-Frequency Cepstral Coefficient

Mel-Frequency Cepstral Coefficients (MFCC) [7] are one of the more popular filter-bank based parameterization methods used by researchers in the speech technology field. The advantage of applying the Mel-scale is that it approximates the nonlinear frequency resolution of the human ear. As with any filter-bank based speech analysis technique, an array of bandpass filters are utilized to analyze the speech in different frequency bandwidths. In MFCC parameterization, the positions of the bandpass filters along the linear frequency scale are mapped to a Mel-scale according to:

$$\text{Mel}(f) = 2595 \log_{10} \left(1 + \frac{f}{700}\right).$$ (3.22)

Triangular filters are employed in MFCC to weight the Discrete Fourier Transform (DFT) of the speech such that the outputs are approximately the energies of the filter-bank signals. The outputs of the bandpass filters are then used to calculate the MFCC using the Discrete Cosine Transform (DCT), which is defined as:

$$\text{MFCC}_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos \left(\frac{i \pi}{N} (j - 0.5)\right)$$ (3.23)

where $N$ is the number of bandpass filter, $m_j$ is the log of the $j^{th}$ bandpass filter’s output amplitude.

3.5.2 Linear Prediction Cepstral Coefficient

Linear Prediction Cepstral Coefficients (LPCC) [41] are Linear Prediction Coefficients (LPC) represented in the cepstral domain. LPCC have been widely used for speech research for a few decades and have been proven to be more robust and reliable than LPC.
The idea of LPC is based on a speech production model in which the characteristic of the vocal tract can be modelled by an all-pole filter $H(z) = 1/A(z)$. $A(z)$ is called the inverse filter and is defined as:

$$A_m(z) = 1 + a_1 z^{-1} + \ldots + a_m z^{-m}$$

(3.24)

where $m$ is the order of $A_m(z)$ and $a_i$ are the LPC. The autocorrelation method is usually employed as it allows the LPC to be computed efficiently using the Levinson-Durbin recursion.

The LPCC, $c_i$, are obtained from the LPC as:

$$c_i = a_i + \sum_{k=1}^{i-1} \frac{k-i}{i} c_{i-k} a_k.$$  

(3.25)

### 3.5.3 Mel-Cepstral Coefficient

Mel-Cepstral coefficients (MCC) have been applied successfully to both speech coding and speech recognition. They do not have the disadvantage of LPCC, which approximate speech linearly at all frequencies. Instead, the cepstrum is mapped to the Mel-scale to model the auditory nonlinear frequency response.

There are several different methods of obtaining MCC, and the frequency transformation approach converting the linear cepstrum to the Mel-scale was adopted in these experiments. The frequency transformation approach calculates the MCC by applying a recursion formula to the LPCC, and they are defined as [80]:

$$g_m^{(i)} = \begin{cases} 
  c_{(-i)} + \alpha g_0^{(i-1)}, & m = 0 \\
  (1 - \alpha^2) g_0^{(i-1)} + \alpha g_1^{(i-1)}, & m = 1 \\
  g_{m-1}^{(i-1)} + \alpha(g_m^{(i-1)} - g_{m-1}^{(i)}), & m = 2, \ldots, M_2.
\end{cases}$$

(3.26)

This calculation is repeated from $i = -M_1$ to $i = 0$, where $M_1$ and $M_2$ are the
order of the LPCC and MCC respectively, $\alpha$ is the transformation scale ($\alpha = 0.31$ for 8kHz sampling frequency, $\alpha = 0.42$ for 16kHz sampling frequency), and the MCC are obtained as $g_{m}^{(0)}$. One drawback of the frequency transformation is that the MCC obtained using this approach have errors caused by truncation in the quefrency domain. This is because the original cepstrum is an infinite sequence and only the first 256 coefficients are used in this experiment. Therefore, the LPCC order must be large enough to compensate for the truncation error.

### 3.5.4 Perceptual Linear Predictive Coefficients

Perceptual Linear Predictive (PLP) [15] coefficients have been widely used in speech recognition and have been shown to give good accuracy in different applications. Instead of modelling the spectrum of the speech as with LPC, the same inverse filter is utilized to model the auditory (critical-band) spectrum in PLP. This critical-band spectrum is obtained by first mapping the linear frequency spectrum to the Bark-scale:

$$\Omega(\omega) = 6 \ln\left\{\omega/1200\pi + [(\omega/1200\pi)^2 + 1]^{0.5}\right\}, \quad (3.27)$$

where $\omega$ is the angular frequency in rad/s. Alternatively, a simpler Bark-scale frequency mapping equation [82] can be used in place of Equation 3.27, This alternative is given by:

$$\Omega(\omega) = \frac{26.81\omega/2\pi}{(1960 + \omega/2\pi)} - 0.53 \quad (3.28)$$

and is employed in this experiment. The critical-band spectrum is then derived by applying the critical-band masking curve, $\Psi(\Omega)$,
3.5 Experiments Using Different Parameterization Techniques and Features

\[ \Psi(\Omega) = \begin{cases} 
0 & \text{for } \Omega < -1.3, \\
10^{2.5(\Omega+0.5)} & \text{for } -1.3 \leq \Omega \leq -0.5, \\
1 & \text{for } -0.5 < \Omega < 0.5, \\
10^{-1.0(\Omega-0.5)} & \text{for } 0.5 \leq \Omega \leq 2.5, \\
0 & \text{for } \Omega > 2.5.
\]  
(3.29)

the equal-loudness curve, \( E(\omega) \),

\[ E(\omega) = \frac{[\omega^2 + 56.8 \times 10^6 \omega^4]/[(\omega^2 + 6.3 \times 10^6)^2 \times (\omega^2 + 0.38 \times 10^9)]}{(3.30)} \]

and the intensity-loudness power law, \( \Phi(\Omega) \),

\[ \Phi(\Omega) = \Xi(\Omega)^{0.33} \]  
(3.31)

where all are based on the psychophysics of human hearing.

The PLP coefficients are then obtained by modelling this critical-band spectrum with the inverse filter, \( A(z) \), defined in Equation 3.24, and converted to the cepstrum domain using Equation 3.25. Thus, PLP is a technique of parameterization that combines filter-bank and linear predictive analysis.

### 3.5.5 Band-Limiting

As the database used in the experiments was collected over a telephone line, band-limiting is required in order to remove unnecessary information from the recorded data. For filter-bank based speech analysis techniques such as MFCC and PLP, this can be easily achieved by limiting the filter-banks within the required bandwidth. For the LPC based methods such as LPCC and MCC, band-limiting is
difficult to perform and requires extra computation. The approach for performing band-limiting on LPC is termed Selective Linear Prediction (SLP) [41]. The essential idea of this approach is to translate the selected bandwidth of the spectrum to the angular range from zero to $\pi$ on the unit circle in the $z$-plane. The LPC are then obtained by computing the autocorrelation coefficients from this new spectrum.

### 3.5.6 Experiment Setup

The main objective of this study is to establish which acoustic features are useful for discriminating between different languages. Band-limiting, Cepstral Mean Subtraction (CMS) as well as appending delta and acceleration coefficients to the feature vectors were examined. The setup of these experiments are identical to the GMM based LID experiments described at Section 3.4.

### 3.5.7 Experimental Results

Table 3.3 gives the accuracy of the basic setup for the different parameterization techniques. The specifications for each method are: MFCC (20 filters, 12 cepstral coefficients), LPCC (14th order LPC, 12 cepstral coefficients), MCC (14th order LPC, 256th order LPCC, 12 cepstral coefficients) and PLP (17 filters, 5 cepstral coefficients). The results after applying the band-limiting (BL, 300-3400Hz) and CMS are also included.

Without any form of band-limiting and channel compensation, the LPC based techniques (LPCC and MCC) performed better than the others, and PLP has the poorest performance. By examining the band-limiting and CMS results, it can be seen that all techniques obtained better performance. These results also show that CMS is essential when performing experiments on telephone speech database as the average improvement in accuracy is 12% compared to 3% using only band-
3.5 Experiments Using Different Parameterization Techniques and Features

Table 3.3: LID Results (% correct) comparing basic feature, band-limiting (BL, 300-3400Hz) and cepstral mean subtraction (CMS).

<table>
<thead>
<tr>
<th>Technique</th>
<th>Basic 45s</th>
<th>Basic 10s</th>
<th>BL 45s</th>
<th>BL 10s</th>
<th>CMS 45s</th>
<th>CMS 10s</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>26.7</td>
<td>26.9</td>
<td>29.9</td>
<td>31.7</td>
<td>39.0</td>
<td>39.4</td>
</tr>
<tr>
<td>LPCC</td>
<td>27.8</td>
<td>29.1</td>
<td>31.6</td>
<td>30.2</td>
<td>42.8</td>
<td>41.3</td>
</tr>
<tr>
<td>MCC</td>
<td>32.1</td>
<td>30.7</td>
<td>31.6</td>
<td>31.5</td>
<td>44.4</td>
<td>41.9</td>
</tr>
<tr>
<td>PLP</td>
<td>25.7</td>
<td>24.2</td>
<td>31.0</td>
<td>29.1</td>
<td>38.0</td>
<td>38.4</td>
</tr>
</tbody>
</table>

limiting. Again the LPC based techniques (LPCC and MCC) exhibited in higher accuracy.

The RASTA filter for PLP, which is capable of removing channel affects, has also been tested. However, this filter needs a certain length of silence to exist in front of each speech file for filter initialization. The results for this test were not included in the comparison, as the testing database does not fulfil this condition.

Table 3.4 shows the results of adding delta coefficients, acceleration coefficients and delta energy to the basic feature vector. The window size for both delta and acceleration calculation is nine frames of speech. The accuracy of the GMM-UBM LID system has been increased with the addition of these coefficients for all tested parameterization techniques. The accuracy was almost double that for the basic features, which means these coefficients provide useful information for discriminating languages. By adding the delta energy to the feature vector, an average of 2.5% improvement in accuracy was achieved for all methods. PLP again performs the worst in all the tests. However, the accuracy of adding delta energy to the PLP had a 5.4% improvement in the 45s test. This occurs because the critical-band spectrum which modelled by the PLP has been normalized by the intensity-loudness power law (Equation 3.31). Hence, the energy changes are uncorrelated to the PLP itself and therefore the utilization of delta energy introduces more information.

Table 3.5 shows the results of combining all improvement techniques from the above tests. They include band-limiting, CMS, delta, acceleration and delta
3.5 Experiments Using Different Parameterization Techniques and Features

Table 3.4: LID Results (% correct) comparing the use of delta (D), acceleration (A) and delta energy (DE) coefficients. The window size for both delta and acceleration is 9 frame.

<table>
<thead>
<tr>
<th>Technique</th>
<th>D</th>
<th>45s</th>
<th>10s</th>
<th>D_A</th>
<th>45s</th>
<th>10s</th>
<th>D_A_DE</th>
<th>45s</th>
<th>10s</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>41.2</td>
<td>41.6</td>
<td></td>
<td>48.1</td>
<td>45.9</td>
<td></td>
<td>49.7</td>
<td>47.8</td>
<td></td>
</tr>
<tr>
<td>LPCC</td>
<td>48.1</td>
<td>45.1</td>
<td></td>
<td>52.4</td>
<td>51.8</td>
<td></td>
<td>55.1</td>
<td>54.2</td>
<td></td>
</tr>
<tr>
<td>MCC</td>
<td>44.4</td>
<td>44.3</td>
<td></td>
<td>50.8</td>
<td>49.1</td>
<td></td>
<td>53.5</td>
<td>51.5</td>
<td></td>
</tr>
<tr>
<td>PLP</td>
<td>41.7</td>
<td>37.1</td>
<td></td>
<td>46.5</td>
<td>43.2</td>
<td></td>
<td>51.9</td>
<td>46.1</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.5: LID Results (% correct) including band-limiting, cepstral mean subtraction, delta, acceleration and delta energy coefficients to the feature vector.

<table>
<thead>
<tr>
<th>Technique</th>
<th>45s</th>
<th>10s</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>59.9</td>
<td>56.6</td>
</tr>
<tr>
<td>LPCC</td>
<td>63.1</td>
<td>61.8</td>
</tr>
<tr>
<td>MCC</td>
<td>62.6</td>
<td>59.0</td>
</tr>
<tr>
<td>PLP</td>
<td>69.5</td>
<td>61.8</td>
</tr>
</tbody>
</table>

energy coefficients. The performances of all parameterization techniques have greatly increased from their basic features. This has proved that modelling the characteristics of a language with a GMM requires a very complicated acoustic feature space in order to cover all the differences between languages. In the 45s test, PLP outperformed other methods with 69.5% accuracy. For the 10s test, PLP and LPCC performed the best with 61.8% accuracy. Note that PLP has obtained the highest accuracy in this test where in previous tests its performance was poorer. This shows that PLP might contain more discriminative information in the delta and acceleration coefficients after the CMS was applied. In addition to the advantage obtained from delta energy, it is seen that adding delta and acceleration coefficients as well as performing band-limiting and CMS increases the accuracy of PLP based LID from 25.7% in the basic feature to 69.5% in the final system in the 45s test. The performance of MCC is slightly worse than LPCC, hence the perceptual representation does not help to improve LID accuracy. However, the lower performance of MCC could be due to the truncation in the quefrency domain as mentioned at Section 3.5.3, which resulted in a loss of information.
3.5 Experiments Using Different Parameterization Techniques and Features

3.5.8 NIST 1996 Language Identification Experiment

The purpose of this experiment was to investigate the performance of the techniques studied above on a larger scale LID experiment. An additional intention was to compare the performance of the GMM-UBM LID system to the best system in the NIST 1996 evaluation.

The setup of the GMM-UBM LID system is identical to the above experiments except that the training data is extracted from both the OGI-TS and CALLFRIEND database. In addition, 10 iterations of adaptation were performed during Language-GMM creation because multiple iteration of MAP adaptation has been shown to be able to create better models in a GMM based speaker recognition system [58]. Testing is based on the NIST 1996 language recognition evaluation guidelines, which included 12 languages and 1492 utterances for 30s test case, 1502 utterances for 10s and 1503 utterances for 3s. Table 3.6 shows the results of this experiment. LPCC performed the best out of all the parameterization techniques in all test cases, which is consistent with the previous experiment. The performance of MCC was again shown to lag the LPCC performance, which could suggest that perceptual representation of speech might not be important. Interestingly, PLP obtained the worst accuracy in contrast to the previous experiment where the best performance was achieved. A possible reason for such a contract is that the NIST 1994 test case is small and the test data set happened to favor the information provided by PLP.

Finally, comparing the LID accuracy of the GMM-UBM LID system to the reference phonetic based system, both the 10s and 3s test cases obtained better results but not in the 30s case. Hence, this result suggests that the GMM-UBM system can not take advantage of the longer test segment duration as well as the phonetic based system, where the GMM based LID system refers to performing the identification solely based on the log likelihood score averaged from all the feature vectors.
Table 3.6: GMM-UBM LID results (% correct) on NIST 1996 Language Identification experiment.

<table>
<thead>
<tr>
<th>Technique</th>
<th>30s</th>
<th>10s</th>
<th>3s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonetic Reference System [94]</td>
<td>74.3</td>
<td>53.4</td>
<td>34.8</td>
</tr>
<tr>
<td>GMM-UBM MFCC</td>
<td>62.7</td>
<td>56.3</td>
<td>43.6</td>
</tr>
<tr>
<td>GMM-UBM LPCC</td>
<td>66.9</td>
<td>60.0</td>
<td>46.1</td>
</tr>
<tr>
<td>GMM-UBM MCC</td>
<td>63.9</td>
<td>57.5</td>
<td>43.9</td>
</tr>
<tr>
<td>GMM-UBM PLP</td>
<td>62.6</td>
<td>54.1</td>
<td>40.4</td>
</tr>
</tbody>
</table>

As shown by the experimental results, parameterization techniques that derive the vocal tract and formant characteristics (LPCC) provided better discrimination between languages. A similar finding was obtained by Foil [9] where the formant information was found to be important for LID. In addition, the utilization of channel compensation and dynamic acoustic features (delta energy, delta and acceleration coefficients) are equally important to the selection of parameterization technique. Figure 3.6 shows the confusion matrices of the 10s test case of both the NIST 1994 (Section 3.4) and NIST 1996 LID experiment. Around 60% correct is achieved in both experiments and there is no trend of confusion between any two or more languages. Thus, the GMM-UBM LID system seems to work independently of the languages targeted for identification.

3.6 Summary

This chapter discusses the use of GMMs to model the acoustic speech characteristics of a language. The GMM-UBM technique is employed as an efficient way to develop a GMM based LID system. This technique has the benefit that the model training time is decreased due to the use of the adaptation procedure. In testing, by employing the computation speed-up method of the GMM-UBM system, the efficiency of testing was improved dramatically while accuracy remained comparable to the standard GMM LID system. Additionally, with the use of the UBM, adding new languages to the system is flexible even under the condition...
3.6 Summary

### Confusion Matrix for 10 Second Test of NIST 1994

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>fa</th>
<th>fr</th>
<th>ge</th>
<th>hi</th>
<th>ja</th>
<th>ko</th>
<th>ma</th>
<th>sp</th>
<th>ta</th>
<th>vi</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>en(69)</td>
<td>59</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>85.5</td>
</tr>
<tr>
<td>fa(58)</td>
<td>2</td>
<td>42</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>72.4</td>
</tr>
<tr>
<td>fr(62)</td>
<td>0</td>
<td>1</td>
<td>39</td>
<td>6</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>62.9</td>
</tr>
<tr>
<td>ge(65)</td>
<td>8</td>
<td>4</td>
<td>10</td>
<td>34</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>52.3</td>
</tr>
<tr>
<td>hi(65)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>37</td>
<td>8</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>56.9</td>
</tr>
<tr>
<td>ja(61)</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>7</td>
<td>33</td>
<td>5</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>54.1</td>
</tr>
<tr>
<td>ko(45)</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>33.3</td>
<td></td>
</tr>
<tr>
<td>ma(52)</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>33</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>63.5</td>
</tr>
<tr>
<td>sp(58)</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>33</td>
<td>1</td>
<td>3</td>
<td>56.9</td>
</tr>
<tr>
<td>ta(43)</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>36</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>83.7</td>
</tr>
<tr>
<td>vi(47)</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>25</td>
<td>53.2</td>
</tr>
</tbody>
</table>

(625)

Overall Average ID Rate: 61.8 %
Equally Weighted by Language: 61.3 %

### Confusion Matrix for 10 Second Test of NIST 1996

<table>
<thead>
<tr>
<th></th>
<th>ANA</th>
<th>ENG</th>
<th>FAR</th>
<th>FRE</th>
<th>GER</th>
<th>HIN</th>
<th>JAP</th>
<th>KOR</th>
<th>MAN</th>
<th>SPA</th>
<th>TAM</th>
<th>VIE</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANA(80)</td>
<td>37</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>46.3</td>
</tr>
<tr>
<td>ENG(479)</td>
<td>33</td>
<td>372</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>11</td>
<td>3</td>
<td>5</td>
<td>17</td>
<td>18</td>
<td>4</td>
<td>5</td>
<td>77.7</td>
</tr>
<tr>
<td>FAR(80)</td>
<td>7</td>
<td>5</td>
<td>39</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>9</td>
<td>7</td>
<td>0</td>
<td>4</td>
<td>48.8</td>
</tr>
<tr>
<td>FRE(80)</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>57</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>71.3</td>
</tr>
<tr>
<td>GER(80)</td>
<td>17</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>46</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>57.5</td>
</tr>
<tr>
<td>HIN(78)</td>
<td>4</td>
<td>7</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>32</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>41.0</td>
</tr>
<tr>
<td>JAP(80)</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>39</td>
<td>13</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>48.8</td>
</tr>
<tr>
<td>KOR(79)</td>
<td>9</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>11</td>
<td>27</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>34.2</td>
</tr>
<tr>
<td>MAN(156)</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>2</td>
<td>4102</td>
<td>4</td>
<td>2</td>
<td>13</td>
<td>65.4</td>
<td></td>
</tr>
<tr>
<td>SPA(155)</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>26</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>78</td>
<td>3</td>
<td>12</td>
<td>50.3</td>
</tr>
<tr>
<td>TAM(75)</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td>21</td>
<td>11</td>
<td>28.0</td>
</tr>
<tr>
<td>VIE(80)</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>3</td>
<td>51</td>
<td>63.8</td>
</tr>
</tbody>
</table>

(1502)

Overall Average ID Rate: 60.0 %
Equally Weighted by Language: 52.8 %

Figure 3.6: Confusion Matrix of the 10s test case of both NIST 1994 and NIST 1996 LID experiment using the GMM-UBM LID system.
of a lack of speech data. The proposed GMM-UBM LID system achieved better accuracy, except for 30s test case, than the phonetic based LID system which achieved the best results in the NIST 1996 language recognition evaluation.

Experiments were also performed to compare different techniques of speech parameterization for GMM-UBM LID system. Results show that LPCC obtained the best performance in two different LID experiments. This indicates that the vocal tract and formant information derived from speech have better ability to discriminate between different languages and the use of perceptual representation of speech does not result in improved LID accuracy.
Chapter 4

Study of Phonetic Details for a Phonetic Information based Language Identification System

4.1 Introduction

In spoken human languages there is a finite set of meaningful sounds that can be produced physically. Not all of these sounds appear in any given language and hence, each language has its own finite subset of meaningful sound units. Phonetic information refers to the characteristics of a sequence of sound units extracted from speech. As outlined in Chapter 2, Phonetic speech information is the most common source of speech information utilized for the LID task. It is the most studied area in LID and systems utilizing phonetic information have achieved the best results in the past NIST language recognition evaluations [94, 96]. The most common technique to extract phonetic information from speech data (which is borrowed from speech recognition technology) is to use a Hidden Markov Model (HMM) based phonetic event recognizer. In such a recognizer, the HMM is employed to model the characteristics of each phonetic event. To perform the
recognition, all possible combinations of HMMs are compared against the input speech, with the most probable combination deemed as the underlying sequence of phonetic events. The sequences of output phonetic labels are referred to as the phonetic speech features (the *Speech Feature Set* in Figure 2.1).

A phonetic based LID system, namely "Parallel Phone Recognition followed by Language Modelling" (PPRLM) [96], is adopted as the test system to study the influence of phonetic details on LID performance. This system is selected due to the superior performance it achieved in the NIST 1996 language recognition evaluation. In a PPRLM system, a front-end phone recognizer (or a broad phonetic event recognizer, depending on the amount of phonetic detail required from the speech) is used to extract the phonetic information. The phonetic characteristics of a language are then modelled using an $N$-gram Language Model (LM), such a model allows for the estimation of the probability of occurrence of a given phonetic sequence. Thus, the discriminative power of the system relies on how much phonetic information can be captured and modelled by the LM. Two main approaches are employed to enhance the ability of $N$-gram LMs to accurately describe the characteristics of a language. The first is to increase the order of the $N$-gram LM, where a larger number of phonetic events are examined at once, this is done in order to more precisely capture the variation in the occurrence of phonetic events.

The other approach is to enrich the information carried by each phonetic event (one example is to incorporate the duration information of the event) in an attempt to represent the speech data in more detail. This results in a larger set of phonetic events that can be used to describe the phonetic information. Nevertheless, due to the limited training data that is available to estimate robustly the statistics of the LMs, a trade off needs to be made between these two approaches. The aim of this chapter is to examine the amount of phonetic detail that should be carried by every phonetic event in order to extract better LID performance. Hence, experiments performed in this chapter will be focused on comparing the
utilization of different level of phonetic information.

Full details of the phonetic information based test system adopted for experimentation are given in the next section. This is followed by a discussion of the approach used to enrich the phonetic detail of phonetic events by incorporating duration information. Experiments are first performed at the fine phonetic level where a phone recognizer is employed as the front end of the test LID system. LID accuracy is compared between different order $N$-gram LMs and the addition of duration information to the phone event. The next set of experiments examined utilizes broad phonetic events. Due to the limited availability of transcribed data, broad phonetic transcriptions of the speech data are obtained by mapping from the existing fine phonetic transcriptions. Two broad phonetic event mapping approaches are studied. The first of which utilizes phonetic knowledge and the second is data driven based. The phonetic knowledge approach simply maps the fine phonetic events to their corresponding broad phonetic category. The data driven approach makes use of a clustering algorithm based on similarity between the fine phonetic events to define the mapping. Experimental results have shown that, when the same amount of training data is considered, it is more desirable to employ a set of finer detailed phonetic events than to use a longer phonetic context.

### 4.2 Phonetic Information Based Test System

“Parallel Phone Recognition followed by Language Modelling” (PPRLM) [96] has proved to be one of the most successful techniques used in LID systems. It models a language via its phonotactic characteristics, where phonotactic refers to the frequency and the possible order of occurrence of a sequence of phonetic events. The PPRLM system comprises of a bank of identical “Phone Recognition followed by Language Modelling” (PRLM) sub-systems running in parallel. Each of the sub-systems performs the same LID function, however they are trained
individually with speech data from different languages. The following section
describes the PPRLM system in more detail, followed by a discussion outlining
the rationale for the utilization of the multiple front-end phone recognizers.

4.2.1 PRLM System Structure

Figure 4.1 shows a block diagram which depicts the structure of the “Phone
Recognition followed by Language Modelling” (PRLM) system. Phonetic level in-
f ormation is first extracted from the speech data in preparation for the character-
ization of the phonotactic information of the language at a later stage. To achieve
this, a HMM based fine phonetic event (phone) recognizer is used to decode the
unknown utterance into a sequence of phone labels, \( \Psi = \{ \psi_1, \psi_2, \ldots, \psi_P \} \), where
\( P \) denotes the number of phone labels produced to represent the utterance. With
the speech data represented at the phonetic level, the next step is to capture the
phonotactic characteristics from the decoded phone sequences for each language.
The \( N \)-gram Language Model (LM) [26] modelling technique is employed here to
estimate the probability of the occurrence for a particular phone sequence within
a language. Given the observation of a sequence of \( N - 1 \) consecutive phonetic
events, an \( N \)-gram LM estimates the probability that a particular phonetic event
occurs next, defined as

\[
P(\psi_p | \psi_{p-1}, \psi_{p-2}, \ldots, \psi_{p-(N-1)}).
\]

(4.1)

Hence, the probabilities for all possible phone sequences that occur in a language
can be estimated. The resulting language model, \( \lambda \), captures the phonotactic
information for each language. To perform LID, the likelihood score of a phone
sequence, \( \Psi \), can be calculated as
Figure 4.1: Block diagram of a PRLM LID system.

\[ \mathcal{L}(\Psi \mid \lambda_l) = \sum_{p=1}^{P} \log P(\psi_p \mid \psi_{p-1}, \psi_{p-2}, \ldots, \psi_{p-(N-1)}, \lambda_l) \]  

(4.2)

where \( \lambda_l \) is the LM of language \( l \), and \( P(\psi_p \mid \psi_{p-1}, \psi_{p-2}, \ldots, \psi_{p-(N-1)}, \lambda_l) \) is the probability of the \( N \)-gram event \((\psi_{p-(N-1)} \text{ to } \psi_p)\) estimated from \( \lambda_l \). Identification is then performed as

\[ \hat{L} = \arg \max_{1 \leq l \leq L} \mathcal{L}(\Psi \mid \lambda_l). \]  

(4.3)

In order to exploit the phonotactic characteristics to discriminate between different languages, it is necessary to examine the phonetic events up to a certain time period. The larger the number of consecutive phonetic events that can be examined at one time, the more distinct the characteristics of a language that can be observed and captured. However, the number of consecutive phonetic events that can be examined is limited by the amount of training data that is available to robustly estimate the necessary statistics. In practice, the amount of training data available usually only allows for bi-gram \((N = 2)\) or tri-gram \((N = 3)\) LMs to be estimated robustly. Equation 4.2 is then simplified to

\[ \mathcal{L}(\Psi \mid \lambda_l) = \sum_{p=3}^{P} \log P(\psi_p \mid \psi_{p-1}, \psi_{p-2}, \lambda_l) \]  

(4.4)

for the case of tri-gram LM, where the probability of a \( N \)-gram event is defined
4.2.2 Parallel-PRLM

The PRLM LID system utilizes a single language-dependent phone recognizer to decode speech data from multiple languages. This approach is limited because the phonetic events that can be recognized are limited to those from a single language. Ideally a phone recognizer should be created for each language, however, as stated in Section 2.5, the lack of phonetically transcribed data for most languages is a significant hurdle to achieving this. Subsequently, it is very common to have no transcribed data available for some of the languages that are identification targets. One method used to compensate for this problem is to run multiple PRLM systems in parallel with the phone recognizers, each trained with a different language. The languages employed to create these PRLM systems are independent of the target languages. The aim of utilizing multiple language-dependent phone recognizers is to characterize each target language in terms of the independent languages. This solution is analogous to the approach that a human might use to identify unknown languages given that their own version of phonology information is developed based on the language(s) they know. Figure 4.2 depicts this parallel-PRLM (PPRLM) LID system.

4.3 Incorporate Phone Duration

One limitation of the PRLM approach is the amount of phonetic detail that can be extracted from speech data. This is due to the limited phone set utilized within the front-end phone recognizer. In turn, this gives rise to the situation where out of language phone sequence from different languages can be scored similarly by the LM. Hence, this limited phonetic description prevents the system from distinguishing between the candidate languages. Which phone set can be
4.3 Incorporate Phone Duration

utilized is confined by the language used to train the front-end phone recognizer, however one way to compensate for this limitation is to expand the phone set, by extracting more specific speech information from the phone events, and to incorporate it by modifying the corresponding phone label. With these annotated phone sequence, each language can be described with more detail and therefore reduce the chance of obtaining similar probabilities across different languages of a given phone sequence.

One of the byproducts of the front-end phone recognizer is the length of each phone event. This duration information can be incorporated easily at the phonetic level [92]. From the phonetically decoded training data, the average duration of each phone utilized within the phone recognizer is first computed. A duration tag “–L” is then appended to the phone label if it is longer than the average duration. e.g. /a/ → /a–L/. Similarly, a “–S” tag is added to the label if its duration is shorter than the average. In this way, two phonetically identical phone sequences can be discriminated between by their duration information.

Though it is desirable to utilize a phone set that is more descriptive to represent the speech data, it introduces a problem when employing N-gram LMs because the possible number of N-gram events is increased proportionally to the size of the phone set. Hence, in a real world situation where training data is limited,
a trade off will be made between the order of the $N$-gram LM (the number of previous phonetic events to examine) and the amount of details incorporated within the phone set.

### 4.4 Fine Phonetic Details LID Experiments

Experiments are performed to compare the LID performance when different orders of $N$-gram LM are applied to the PPRLM system. In addition, the use of duration information to enhance the phonetic detail of every phone events is studied. The purpose of these experiments is to examine, with the same amount of training data, the trade off required between the length of phonetic context and the amount of phonetic detail carried by phonetic events. Two different experimental setups performed according to the guideline of NIST 1994 and 1996 evaluation are included in an attempt to show the performance of the phonetic information based system when different amount of training data are available.

#### 4.4.1 Test System Implementation

**Front-end Phone Recognizers**

Six different phone recognizers were trained using the OGI-TS database. The languages included are English, German, Hindi, Japanese, Mandarin and Spanish. All diacritic labels were removed from the original transcriptions, leaving only the phonemic base labels. After the diacritic labels were removed, any consecutive labels having the same symbols were then merged. The number of occurrences of each phone were then tallied. Phones that occurred less than 100 times were mapped to a similar counterpart phone according to the phonetic system of the language. The resultant transcriptions were then used to train the phone recognizer front-ends. Table 4.1 shows the number of phones that resulted in each
Table 4.1: The number of phones and the amount of training data used to train the phone recognizer front-ends.

<table>
<thead>
<tr>
<th>Language</th>
<th># Phones</th>
<th>Training Data (Hour)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>42</td>
<td>1.46</td>
</tr>
<tr>
<td>German</td>
<td>42</td>
<td>0.80</td>
</tr>
<tr>
<td>Hindi</td>
<td>45</td>
<td>0.73</td>
</tr>
<tr>
<td>Japanese</td>
<td>32</td>
<td>0.59</td>
</tr>
<tr>
<td>Mandarin</td>
<td>41</td>
<td>0.49</td>
</tr>
<tr>
<td>Spanish</td>
<td>39</td>
<td>0.93</td>
</tr>
</tbody>
</table>

language and the approximate amount of training data available.

The Hidden Markov Model Toolkit (HTK) [19] was used to develop the HMM based phone recognizers. Three state left-to-right HMMs were employed, where each state distribution was modelled by eight Gaussian mixture components. The phone recognizer was required to decode data from multiple languages, hence a language-dependent “N-gram LM recognition network” (which is usually employed in a recognition system to improve accuracy) is not applicable in this task and an “open-phone-loop network” was incorporated inside the recognizer. Each feature vector contained 12\(^{th}\) order PLPs with energy included. Additionally, the delta and acceleration coefficients were appended which resulted in a 39-dimension vector.

\textit{N-gram Language Models}

The CMU-Cambridge Statistical Language Modelling toolkit [5] was used to estimate the statistics of each LM. The resultant model is a Back-off N-gram LM using Good Turing discounting method. The phonetically decoded training data used to create the LMs for the NIST 1994 LID experiments included all the speech data from the OGI-TS database. The NIST 1996 experiments used the speech from both sides of all conversations of the CALLFRIEND database.
Table 4.2: LID results (% correct) of the NIST 1994 LID experiment using phonetic information based LID system.

<table>
<thead>
<tr>
<th>Technique</th>
<th>45s</th>
<th>10s</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-gram</td>
<td>89.8</td>
<td>83.4</td>
</tr>
<tr>
<td>2-gram + Phone Duration</td>
<td>80.2</td>
<td>74.6</td>
</tr>
<tr>
<td>3-gram</td>
<td>72.7</td>
<td>65.3</td>
</tr>
</tbody>
</table>

### 4.4.2 Experimental Results

For the NIST 1994 LID experiment, 2-gram, 2-gram with phone duration information and 3-gram LM were tested. Due to the larger size of training data set, the NIST 1996 experiment also included the 3-gram, 4-gram and the addition of phone duration information.

Table 4.2 shows the LID results of the NIST 1994 experiment. Out of all the LMs tested, the use of 2-gram models achieved the best performance. The inclusion of extra phonetic detail or phonetic context does not help to provide better accuracy. However, the accuracy was consistently reduced in both 45s and 10s test cases when the degree of phonetic detail or context is increased. This suggested that the amount of training data available from the OGI-TS is not large enough to properly estimate the statistics for the LMs to incorporate these amount of phonetic information.

Table 4.3 shows the results of the NIST 1996 LID experiment. In contrast to the NIST 1994 experiment, the use of extra phonetic information (2-gram + phone duration) and longer phonetic context (3-gram) have provided better accuracy in all test segment duration tests (except for the 3s test on 3-gram). By looking at the 3-gram and 3-gram + phone duration tests (where we expected that the increase in phonetic detail would provide better accuracy in all test cases) even though the 30s test achieved an improvement, however, this is not the case for 10s and 3s tests. This might suggest that the shorter test segment does not provide enough phonetic information to accurately discriminate between different languages. A similar trend can be observed by comparing the results between
Table 4.3: LID results (% correct) of the NIST 1996 LID experiment using phonetic information based LID system.

<table>
<thead>
<tr>
<th>Technique</th>
<th>30s</th>
<th>10s</th>
<th>3s</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-gram</td>
<td>82.3</td>
<td>69.7</td>
<td>49.6</td>
</tr>
<tr>
<td>2-gram + Phone Duration</td>
<td>86.3</td>
<td>75.1</td>
<td>53.7</td>
</tr>
<tr>
<td>3-gram</td>
<td>84.6</td>
<td>71.6</td>
<td>49.2</td>
</tr>
<tr>
<td>3-gram + Phone Duration</td>
<td>87.1</td>
<td>71.2</td>
<td>47.1</td>
</tr>
<tr>
<td>4-gram</td>
<td>81.3</td>
<td>62.9</td>
<td>41.0</td>
</tr>
<tr>
<td>4-gram + Phone Duration</td>
<td>81.1</td>
<td>58.9</td>
<td>37.9</td>
</tr>
</tbody>
</table>

2-gram and 3-gram LMs where the improvement is gradually reduced when the test segment duration is decreased. Further increases in phonetic context (4-gram) and detail (4-gram + phone duration) have resulted in performance drop, again, due to the problem of data sparsity similar to the NIST 1994 LID experiment. Additionally, performance drop is higher when the test segment duration is shorter. Thus, the use of higher phonetic detail is more desirable than longer phonetic context when the trade off between these two is needed to be made.

4.5 Broad Phonetic Detail LID Experiments

The aim of these experiments is to study the use of broad phonetic events to perform LID instead of the fine phonetic counterpart. Since the same test system that is utilized for fine phonetic studies can also be adopted for this purpose, the resultant system is named as “Broad Phonetic Recognition followed by Language Modelling” (BPRLM) LID system. The advantage of utilizing broad phonetic events is the recognizer can obtain higher accuracy where there is less confusion between the models, resulting in the production of more accurate phonetic sequences. In addition, higher order $N$-gram modelling can be applied with the same amount of training data given the reduced number of possible phonetic labels. However, this also results in the reduction of phonetic detail available to model the characteristics of a language. The existing phonetically transcribed data are described in fine phonetic details, hence techniques to define the map-
ping rules (language-dependent) required to map the labels from fine phonetics into broad phonetics are also studied. In this experiment, two broad phonetic event mapping approaches are examined. The first utilizes phonetic knowledge and the second is data driven based.

### 4.5.1 Define Broad Phonetic Events – Phonetic Knowledge Approach

The simplest way to perform the label mapping is based on the phonetic system of the language. Each phone can be categorized into its corresponding broad phonetic group. As an example, Table 4.4 shows a chart of the English vowel where the mid front vowel /ɪ/ can be assigned to the vowel group. Therefore, by preselecting the broad phonetic groups intended to be modelled, the fine phonetic transcriptions can easily be mapped in terms of broad phonetic events. In this experiment, 10 broad phonetic groups are defined:

- Vowel (Front, Middle and Back)
- Diphthong
- Semi-vowel
- Stop (voiced and unvoiced)
- Fricative (voiced and unvoiced)
- Nasal
4.5 Broad Phonetic Detail LID Experiments

Table 4.4: Chart of English vowel (not including retroflexes).

<table>
<thead>
<tr>
<th></th>
<th>Front</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>i:</td>
<td>ux</td>
<td>u</td>
</tr>
<tr>
<td>Mid</td>
<td>I</td>
<td>I₂</td>
<td>U</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>&amp; &amp;0</td>
<td>&gt;</td>
</tr>
<tr>
<td>Low</td>
<td>@</td>
<td>∧</td>
<td>A</td>
</tr>
</tbody>
</table>

4.5.2 Define Broad Phonetic Events – Data Driven Approach

The purpose of the broad phonetic event defining process in this study is to categorize similar fine phonetic events. The approach described in the previous section performed this process based on the phonetic knowledge which is defined by linguists as summarised in IPA [22]. While the basic idea is to find the similarity between fine phonetic events, an alternative approach to achieve the same task is to compare them acoustically. In another word, their similarity can be measured by examining the speech data.

A top-down binary phonetic clustering algorithm is used in this study. A set of phonetic questions are first prepared for the clustering process, where each question defines how a pool of HMM phone models should be split according to their phonetic natures. The questions range from broad phonetic definition such as “Is this phone a vowel?” to fine phonetic such as “Is this phone /I/?”. Figure 4.3 depicts the clustering tree produced for the English phones by this phonetic questions clustering technique. At the beginning of the clustering process, all the phones are first pooled together. Each phonetic question is then applied to the pool to calculate the likelihood of representing the training data with the split two groups of model. According to the highest likelihood question, the models are then split into two clusters. The same “question asking” process is repeated at each new cluster and the splitting is stopped when the highest likelihood score calculated is below a threshold. The process will be terminated when no
more clusters can be split. Each of the resultant clusters is then defined as a broad phonetic event. Although the term "broad phonetic event" is employed here, however it is possible that a single phone resides in a cluster by itself, if it is distinct to the others. The broad phonetic mapping rules are then defined according to the clustering results. The phonetic questions related to broadest phonetic events are usually asked first, which in turn splits the phones into finer phonetic groups. For example, the very first question asked for English is to split the pool of phones into vowel and consonant groups.

In order to compare the performance of this data driven mapping approach to the phonetic knowledge counterpart, the same number of broad phonetic events are defined for the experiment.
4.5.3 Test System Setting

The test system is identical to the one described at Section 4.4.1 with the exception that the transcriptions used to train the front-end phone recognizer are mapped to broad phonetic details. Hence, the resultant recognizers are generating broad phonetic labels as output.

4.5.4 Experimental Results

The experimental results of the NIST 1996 LID experiment using the broad phonetic information based system are shown in Table 4.5. Both the phonetic knowledge and data driven approaches have been tested on 2-gram, 3-gram and 4-gram LMs with phone duration information appended. The results show that the BPRLM system employing the data driven mapping outperformed the phonetic knowledge approach in all tests. Two explanations, which relate to the mapping definition, can be given.

First, by comparing the mapping defined by both approaches, it can be observed that the consonant phone models obtained very similar mapping across both approaches. However, the mapping defined by the data driven approach has shown that the vowel and diphthong mapping are language-dependent. The phonetic knowledge approach, which applies the same mapping definition to all languages, would have lost this information and thus, induced more LID errors. Second, the resultant HMM broad phonetic models trained by the data driven approach would obtain better recognition accuracy (especially the vowel models) because both the recognition and the process of phonetic clustering are based on measuring phonetic similarity acoustically. Hence, the front-end phone recognizer trained with the data driven mapping can produce more accurate and consistent broad phonetic event sequences.

Across all the LMs tested, the 3-gram LM has obtained the best result. Compared
Table 4.5: LID results (% correct) of the NIST 1996 LID experiment using the BPRLM LID system.

<table>
<thead>
<tr>
<th>Technique</th>
<th>30s</th>
<th>10s</th>
<th>3s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonetic Knowledge Mapping</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-gram + Phone Duration</td>
<td>64.7</td>
<td>52.3</td>
<td>33.5</td>
</tr>
<tr>
<td>3-gram + Phone Duration</td>
<td>75.2</td>
<td>61.1</td>
<td>39.3</td>
</tr>
<tr>
<td>4-gram + Phone Duration</td>
<td>54.6</td>
<td>52.7</td>
<td>38.4</td>
</tr>
<tr>
<td>Data Driven Mapping</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-gram + Phone Duration</td>
<td>72.1</td>
<td>56.7</td>
<td>39.1</td>
</tr>
<tr>
<td>3-gram + Phone Duration</td>
<td>79.8</td>
<td>65.7</td>
<td>45.0</td>
</tr>
<tr>
<td>4-gram + Phone Duration</td>
<td>61.3</td>
<td>53.7</td>
<td>38.2</td>
</tr>
</tbody>
</table>

to the PPRLM system examined at Section 4.4, where both experiments used the same amount of speech data to extract phonetic information, the use of broad phonetic information allowed the system to incorporate an extra length of phonetic context details. Further increases to the order of $N$-gram LM to 4-gram do not achieve better accuracy. One strange behavior observed in the 4-gram LID experiment was the identification confusion was biased to English, Mandarin and Spanish. This could be due to the amount of data available for training where English, Mandarin and Spanish have double the amount of training data (two different dialects in these languages) compared to the other languages. Hence, the LMs created for these languages were a lot more robust than the others and therefore introduced the bias.

Table 4.6 shows the LID results of comparing fine and broad phonetic information. The PPRLM LID system outperformed all BPRLM systems, hence, the extra phonetic context information provided by BPRLM was not as important as the capture of more phonetic details. This is supported by the results obtained in fine phonetic experiment section (Section 4.4.2) where the use of higher phonetic detail is more desirable than longer phonetic context when the trade off between these two needs to be made. Figure 4.4 shows the confusion matrices of the 10s test case of NIST 1996 LID experiment. Looking at the accuracy of individual languages of both matrices, the phonetic information based LID systems does not favor any particular language.
4.5 Broad Phonetic Detail LID Experiments

CONFUSION MATRIX OF PPRLM LID SYSTEM FOR 10 SECOND TEST OF NIST 1996

<table>
<thead>
<tr>
<th>Language</th>
<th>ARA</th>
<th>ENG</th>
<th>FAR</th>
<th>FRE</th>
<th>GER</th>
<th>HIN</th>
<th>JAP</th>
<th>KOR</th>
<th>MAN</th>
<th>SPA</th>
<th>TAM</th>
<th>VIE</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANA (80)</td>
<td>46</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>57.5</td>
</tr>
<tr>
<td>ENG (479)</td>
<td>5</td>
<td>434</td>
<td>1</td>
<td>8</td>
<td>11</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>90.6</td>
</tr>
<tr>
<td>FAR (80)</td>
<td>4</td>
<td>6</td>
<td>42</td>
<td>3</td>
<td>42</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>52.5</td>
</tr>
<tr>
<td>FRE (80)</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>59</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>73.8</td>
</tr>
<tr>
<td>GER (80)</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>60</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>75.0</td>
<td></td>
</tr>
<tr>
<td>HIN (78)</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>48.7</td>
</tr>
<tr>
<td>JAP (80)</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>52</td>
<td>8</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>65.0</td>
</tr>
<tr>
<td>KOR (79)</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>51</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>64.6</td>
</tr>
<tr>
<td>MAN (156)</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>121</td>
<td>5</td>
<td>1</td>
<td>8</td>
<td>77.6</td>
<td></td>
</tr>
<tr>
<td>SPA (155)</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>77.4</td>
<td></td>
</tr>
<tr>
<td>TAM (75)</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>46</td>
<td>8</td>
<td>61.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIE (80)</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>59</td>
<td>73.8</td>
<td></td>
</tr>
</tbody>
</table>

Total number of correct: 1128
Overall Average ID Rate: 75.1%
Equally Weighted by Language: 68.2%

CONFUSION MATRIX OF BPRLM LID SYSTEM FOR 10 SECOND TEST OF NIST 1996

<table>
<thead>
<tr>
<th>Language</th>
<th>ARA</th>
<th>ENG</th>
<th>FAR</th>
<th>FRE</th>
<th>GER</th>
<th>HIN</th>
<th>JAP</th>
<th>KOR</th>
<th>MAN</th>
<th>SPA</th>
<th>TAM</th>
<th>VIE</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARA (80)</td>
<td>32</td>
<td>5</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>4</td>
<td>40.0</td>
</tr>
<tr>
<td>ENG (479)</td>
<td>2</td>
<td>398</td>
<td>5</td>
<td>28</td>
<td>15</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>83.1</td>
</tr>
<tr>
<td>FAR (80)</td>
<td>4</td>
<td>9</td>
<td>34</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>1</td>
<td>18</td>
<td>42.5</td>
<td></td>
</tr>
<tr>
<td>FRE (80)</td>
<td>3</td>
<td>14</td>
<td>4</td>
<td>43</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>53.8</td>
</tr>
<tr>
<td>GER (80)</td>
<td>1</td>
<td>9</td>
<td>8</td>
<td>2</td>
<td>54</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>67.5</td>
</tr>
<tr>
<td>HIN (78)</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>5</td>
<td>24</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>11</td>
<td>9</td>
<td>30.8</td>
</tr>
<tr>
<td>JAP (80)</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>11</td>
<td>10</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>44.4</td>
</tr>
<tr>
<td>KOR (79)</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>39</td>
<td>11</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>49.4</td>
</tr>
<tr>
<td>MAN (156)</td>
<td>0</td>
<td>11</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>112</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>71.8</td>
</tr>
<tr>
<td>SPA (155)</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>114</td>
<td>6</td>
<td>13</td>
<td>73.6</td>
</tr>
<tr>
<td>TAM (75)</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>39</td>
<td>11</td>
<td>52.0</td>
</tr>
<tr>
<td>VIE (80)</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>56</td>
<td>70.0</td>
</tr>
</tbody>
</table>

Total number of correct: 986
Overall Average ID Rate: 65.7%
Equally Weighted by Language: 57.2%

Figure 4.4: Confusion Matrix of the 10s test case of the NIST 1996 LID experiment of both the PPRLM (2-gram LM) and BPRLM (data driven mapping, 3-gram LM) systems. Both systems have phone duration information appended.
Table 4.6: Comparison of LID performance utilizing fine and broad phonetic information in NIST 1996 LID experiment (% correct).

<table>
<thead>
<tr>
<th>Technique</th>
<th>30s</th>
<th>10s</th>
<th>3s</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPRLM 2-gram + Phone Duration</td>
<td>86.3</td>
<td>75.1</td>
<td>53.7</td>
</tr>
<tr>
<td>BPRLM (Knowledge) 3-gram + Phone Duration</td>
<td>75.2</td>
<td>61.1</td>
<td>39.3</td>
</tr>
<tr>
<td>BPRLM (Data Driven) 3-gram + Phone Duration</td>
<td>79.8</td>
<td>65.7</td>
<td>45.0</td>
</tr>
<tr>
<td>Phonetic Reference System [94]</td>
<td>74.3</td>
<td>53.4</td>
<td>34.8</td>
</tr>
</tbody>
</table>

4.6 Summary

This chapter studied the utilization of various levels of phonetic details for a phonetic information based LID system. The system developed for the experiment achieved lower identification error rates than the best phonetic LID system in the NIST 1996 language recognition evaluation. Experiments have been performed using different amount of phonetic details (broad and fine phonetic details, and the inclusion of phone duration information) and length of phonetic context (the order of \(N\)-gram LM) to describe the characteristics of a language. Since there are no phonetic transcriptions available in broad phonetic details, both a phonetic knowledge based and a data driven based approach was used in order to map the existing fine phonetic transcriptions into broad phonetic details. In the data driven mapping approach, a phonetic questions based clustering algorithm was used to group similar fine phonetic events into broad phonetic events. Results have shown that the utilization of data driven mapping not only produced higher LID accuracy, additionally language-dependent phonetic information was also encapsulated within the front-end phonetic event recognizers. This language-dependent information was incorporated via the different resultant clusters of the vowel and diphthong phones. Hence, this might suggest that phonetic differences between languages is mainly determined by the difference and variations of the vowel phones.

In reality, the amount of training data is usually limited. Hence, a trade off needs to be made between the amount of phonetic detail and the length of phonetic
context in order to robustly estimate the statistics to model the characteristics of a language. While both approaches improve LID accuracy, experimental results suggested that the increase of phonetic details is preferable. This increase in phonetic details can be achieved by utilizing fine phonetic information and the addition of extra phonetic information such as the duration of the phonetic event. Further, the employment of longer phonetic context could result in a performance drop when the test segment duration is short (shorter than 30s in experiment performed in this chapter). This fact is important, as in real life LID applications the amount of speech that can be collected from the speaker could vary in length and therefore, better utilization of phonetic details would produce a more robust LID system.
Chapter 5

Techniques to Improve LID Accuracy

5.1 Introduction

In previous chapters, studies were conducted examining extracting and utilizing acoustic (Chapter 3) and phonetic (Chapter 4) speech information. The research emphasis for these chapters was focused on determining the most discriminative speech characteristics and then extracting them from speech. Although the aspects of speech information to be extracted and utilized to represent a language is the most important aspect of a LID system as mentioned at Chapter 2, the direction of research in this chapter, however, is directed to techniques capable of enhancing the performance of the identification task. The goal of this chapter, therefore, is to develop techniques to improve the final accuracy of the LID systems developed in previous chapters which are independent of the speech information and modelling techniques employed within those systems. Two approaches are identified for this purpose; one is to pre-process the raw speech data so that speech features can be extracted more accurately. The other approach is to post-process the output generated by the LID systems in an attempt to obtain
a more confident identification decision.

Pre-processing of raw speech data is a common step in all speech processing tasks. One of the most important procedures is to compensate for the variation of speech features introduced by different recording and transmitting channels. This is especially evident in telephone speech data, where variation can be imposed by a number of sources such as different handsets, encoding schemes, and telephone lines.

Another source of variation which can affect the speech features is the influence of gender. One approach to compensate for this problem is to introduce gender dependent models, where two separate models are created for each language using speech data from each gender. This approach is capable of improving LID accuracy [92], however, the required computation is doubled. This is not feasible if the amount of training data is limited. Hence, the application of Vocal Tract Length Normalization (VTLN) to the task of LID is proposed in this work.

The idea of VTLN is to normalize the speech data such that the resultant speech can be treated as if uttered by speakers with a unified vocal tract length. This technique not only compensates for the variation due to different genders, but more generally, alleviates the effect imposed by different speakers. This is especially useful for the use of acoustic information in LID as identification is performed solely based on the acoustic feature space. Hence, it is important to avoid this situation as a larger acoustic variation of a language would introduce more overlap to other languages in the acoustic feature space and thus increase the confusion of the identification task. This reduction of acoustic feature variation would also benefit the extraction of phonetic information where in this case the confusion between the phonetic events is reduced. VTLN has been successfully applied to speech recognition [34, 84, 90] to improve recognition accuracy. Section 5.2 will present the study of employing VTLN for LID. Experimental results have shown that the application of VTLN has greatly improved the accuracy of the GMM-UBM LID system with the PPRLM system also achieving positive results.
A major contribution which has resulted from this study is a VTLN technique (which applicable to any speech processing tasks) that can rapidly normalize speech data during the testing stage.

After the investigation into pre-processing speech data, Section 5.3 then presents a study on post-processing of the output generated by LID systems, which could more succinctly be phrased “fusion of information provided by different systems”. As the thesis title indicates, the research conducted is focused on the utilization of “both” acoustic and phonetic speech information for performing LID. This implies that at some stage within the LID process, both sources of information will be combined in an attempt to improve the identification. The technique of combining different information can be referred to as Data Fusion.

Fusing data from different data sources is capable of increasing system performance in many different tasks including speech processing related recognition and identification tasks. The basic idea is to combine different complementary sources of information to improve the accuracy of the final decision. In this study, fusion of different aspects of speech information is performed at the output level. The reason is that the complexities of the LID system can be significantly reduced. If the information is to be combined at the input level, each speech feature will require expansion into larger dimensions. Additionally, acoustic features (a set of coefficients) are represented in a different form to the phonetic features (a sequence of labels) and combining these features at the model level will require some form of modelling technique capable of handling both continuous and discrete data. The aim of this study, therefore, is to develop techniques to enable fusion of the output likelihood scores generated by different systems to subsequently improve LID accuracy.
5.2 Vocal Tract Length Normalization

In many speech applications, speaker variation is one of the significant problems that degrade the performance of the system. One common method to alleviate this effect is to employ gender-dependent models; however, while this method considers gender differences it ignores variations within each gender class. Vocal Tract Length Normalization (VTLN) can accommodate for some of the intra-class variations and experimental results in speech recognition have shown that applying VTLN performs better than the gender-dependent models [34, 84, 90]. For this reason, VTLN has been widely used in speech recognition, especially for large vocabulary speaker independent continuous speech recognition. This is important when the computational requirement for employing gender-dependent models is very large for a high complexity system. For LID, where the system complexity is typically lower, the motivation of applying VTLN to LID is the reduction of speaker variations. This is essential, as the number of speakers within a multilingual speech database is generally very high.

The idea of VTLN is to normalize the speech data such that the resultant speech can be treated as uttered by speakers with a unified vocal tract length. Achieving this result immediately creates two implementation problems:

1. How to process the speech data from each speaker to achieve the normalized vocal tract length? (Section 5.2.1)

2. How to establish the length of vocal tract for each speaker? (Section 5.2.2)

5.2.1 Speaker Normalization

It is known that the length of the human vocal tract has an inverse relationship to each formant frequency [84]. Therefore, speaker normalization can be performed by re-scaling the frequency axis according to the length of the vocal tract or
Figure 5.1: Piecewise linear re-scaling the frequency axis to achieve speaker normalization. $\alpha$ is the speaker-specific normalization factor and $f_0$ is set to fix the bandwidth mismatch.

A corresponding normalization factor. This frequency re-scaling process can be achieved using a piecewise linear re-scaling of the frequency axis [90] as depicted in Figure 5.1 and is as follows:

$$f' = \begin{cases} 
\alpha^{-1}f & \text{if } f < f_0 \\
bf + c & \text{if } f \geq f_0 
\end{cases} \quad (5.1)$$

where $f'$ is the normalized frequency, $\alpha$ is the speaker-specific normalization factor, $f_0$ is a fixed frequency which is set to handle the bandwidth mismatching problem during re-scaling ($f_0 = 3500$ for this study) and $b, c$ can be calculated with a known $f_0$. Given the normalization factor of a speaker, the aim of the normalization process therefore is to re-scale the spectrum of the speech according to Equation (5.1) such that the resultant spectrum would have similar scale as a speaker with $\alpha = 1.0$. For instance, longer a vocal tract would result from a speaker with a low tone voice and hence requires a stretch of the spectrum (which implies shortening the vocal tract length) to be performed as depicted at Figure 5.1 where $\alpha = 0.88$. The opposite would be applied to speaker with a high tone voice.
An alternative method to re-scale the frequency axis can be achieved by varying the spacing and width of the filter-banks [90]. However, this approach is restricted to the filter-bank based speech parameterization techniques such as MFCC and PLP. Thus, this method is not examined in this study because in Chapter 3 LPCC has been found to provide more discriminating information in the GMM-UBM LID system and is employed to obtain acoustic features.

### 5.2.2 Normalization Factor Selection

An additional problem which confronts effective normalization of speaker variation is to obtain the vocal tract length or equivalently the normalization factor, $\alpha$, for each speaker during both the training and testing phases. The most direct method is to obtain the vocal tract length according to the formant frequencies; however, formant frequencies can vary under different speech contexts [90] and thus, is not robust. The most common approach is to select $\alpha$ by performing a sequential grid search of a set of predefined normalization factors (e.g., $0.88 \geq \alpha \geq 1.12$ and stepping by 0.02). The value of $\alpha$ that maximizes the likelihood score against a given model is selected as the normalization factor for the speaker as follows:

$$\hat{\alpha} = \arg \max_i p(X_{\alpha_i} | \lambda)$$

(5.2)

Figure 5.2 shows a block diagram of this selection process. Depending on application, the given model, $\lambda$, can be a set of Hidden Markov Models (HMMs) from a speech recogniser or phoneme recogniser, or a single Gaussian Mixture Model (GMM) to represent a generic speech model as described in [84]. The selection process is then repeated until there are no more significant changes for all speakers. This maximum likelihood approach does not select an $\alpha$ that is directly related to the physical vocal tract length of each speaker. However, it does guarantee that the normalized version of the speech will provide equal or
Figure 5.2: The normalization factor selection process where $X_{\alpha i}$ are the set of feature vectors normalized with normalization factor $\alpha_i$. $N$ is the number of predefined $\alpha$ available for selection and $\lambda$ is the given model. Here, $p(X_{\alpha i} | \lambda)$ represents the likelihood of the feature vectors $X_{\alpha i}$, given the model $\lambda$.

better likelihood scoring against the given model.

In this study, the normalization factor, $\alpha$, for each speaker is selected against a single GMM. The motivation is that prior information of the speech data, such as the phonetic transcriptions, is not required in order to perform the $\alpha$ selection process. In addition, the resultant normalized GMM can be employed as the UBM for the GMM-UBM LID system discussed at Chapter 3. Hence, $\alpha$ selection for each speaker involves choosing the $\alpha$ value that maximizes the likelihood score against the UBM, which is a GMM representing the characteristic of all languages. In order to obtain a more accurate normalization factor for each speaker and a UBM that represents the normalized speech from the training data, the selection process is performed iteratively and the entire process is as follows:

1. Train UBM with data from all languages and select $\alpha$ for all speakers using Equation (5.2).

2. Normalize speaker data with the selected $\alpha$ and re-train the UBM with the normalized data.

3. Select $\alpha$ using Equation (5.2) for all speakers against the normalized UBM from step 2.
4. Repeat steps 2 and 3 until there are no further significant changes in $\alpha$ selected for the speakers.

During system testing, the input speech data is again normalized using Equation (5.2) with the normalized UBM obtained from the training stage. An alternative approach of $\alpha$ selection during testing has been proposed by Lee and Rose [34]. Instead of parameterizing the input test speech with each allowable normalization factor and then calculating the corresponding likelihood scores against the normalized UBM as described in Equation (5.2), the un-normalized training data are first grouped according to the selected $\alpha$ value. A GMM is then trained for each of these “$\alpha$ groups”. No normalization is performed at this stage because the idea of this approach is to model the characteristics of a particular type of $\alpha$ value through a GMM. If the data used to train these GMMs are normalized, the resultant GMMs would have modelled the normalized speech instead of the characteristics of speakers with a particular vocal tract length or in another term, $\alpha$ value. Hence, the $\alpha$ selection process during testing becomes

$$\hat{\alpha} = \arg \max_i p(X | \lambda_{\alpha i})$$  \hspace{1cm} (5.3)

where $X$ is the un-normalized feature vectors and $\lambda_{\alpha i}$ is a GMM representing the speech possessing the characteristics of normalization factor $\alpha i$. An extra step is then added to the $\alpha$ selection process described previously:

5. Train a set of GMMs for each normalization factor with the corresponding un-normalized version of the data that obtained the same normalization factor.

This “$\alpha$-model” selection approach requires less computation compared to Equation (5.2) because normalization is not performed until the selection process has selected an $\alpha$ value for the test data.
5.2 Vocal Tract Length Normalization

Rapid Test Data Normalization Factor Selection Approach

Based on the “α-model” normalization factor selection approach discussed in the previous section, a significantly faster version can be implemented by employing the GMM-UBM technique covered in Chapter 3. Firstly, instead of performing the EM training to obtain the GMMs, these α-models can be created by performing an adaptation from the un-normalized UBM (which is analogous to the creation of Language-GMM of the GMM-UBM LID system). According to the relationship between the UBM and α-models, the top-5 mixture component likelihood approximation technique can be applied to Equation (5.3) and thus significantly reduce the computation required by the α selection process. In addition, any “α groups” which do not have enough data to robustly estimate the statistics of a GMM can now have an appropriate α-model.

5.2.3 LID Experiments

The purpose of this set of experiments was to investigate the improvements of LID accuracy achievable by employing VTLN to the speech data. Since VTLN is believed to improve the performance of LID systems utilizing acoustic information as mentioned in Section 5.1, hence, experiments are first performed on the GMM-UBM system using the NIST 1994 LID task. Investigations are conducted on the effect of the number of α selection iterations on training data and the amount of test data that is required in order to obtain an accurate α value estimation, and also study the performance of different α selection approaches during the testing stage. The NIST 1996 LID task is then performed to compare the improvements that can be made on both the GMM-UBM and PPRLM LID systems.
NIST 1994 LID Experiment

PLP cepstral coefficients (5th order) were used in this experiment, where each feature vector was extracted at 10ms intervals using a 32ms window. Delta energy, delta and acceleration coefficients were appended to the feature vectors with mean and variance normalization applied to compensate for the channel effect. Speaker normalization during VTLN was performed using the piecewise linear re-scaling approach described by Equation (5.1) and there were 13 different normalization factors \(0.88 \geq \alpha \geq 1.12\), and stepping by 0.02) selected for each speaker. All models consisted of GMMs using 512 mixture components.

Table 5.1 shows the results of applying VTLN to the GMM-UBM system. The \(\alpha\) selection process at both training and testing are based on Equation (5.2). The results showed that most of the improvement made to the LID accuracy is achieved after just one iteration of \(\alpha\) selection on training data. Further iterations yield only a slight increase in accuracy. The \(\alpha\) selection process converged after five to six iterations and the employment of VTLN after five iterations improved the accuracy from 77.0% to 84.0% for 45s test segments and 65.6% to 72.8% for 10s test segments. This is the improvement achievable by just limiting some of the inter-speaker variations presented in the speech data. Figure 5.3 shows the histogram of \(\alpha\) selected for all speakers after five iterations. It is clear that apart from the gender difference, the VTLN process identified the speaker difference within each gender class as well.

Figure 5.4 shows the results for using different lengths of data, which were taken from a given test segment, to determine the best \(\alpha\) value. The maximum length of test data used in this experiment was 45 seconds. Figure 5.4 indicates that acceptable performance is achieved by using 10 seconds or more of the test data to select \(\alpha\). Note that using only two seconds of test data for \(\alpha\) selection already achieved an accuracy that outperformed the system without VTLN. This scenario is well suited for applications required to operate in real-time.
Table 5.1: Test results (% correct) for applying VTLN to the GMM-UBM LID system. The “Diff. $\alpha$” column is the difference in $\alpha$ of all speakers compared to the previous iteration. $i$ represents the number of iteration. There are 678 male and 268 female speakers in this task.

<table>
<thead>
<tr>
<th>Number of Iteration</th>
<th>45s</th>
<th>10s</th>
<th>Diff. $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No VTLN</td>
<td>77.0</td>
<td>65.6</td>
<td>-</td>
</tr>
<tr>
<td>VTLN $i = 1$</td>
<td>83.4</td>
<td>70.9</td>
<td>586</td>
</tr>
<tr>
<td>VTLN $i = 2$</td>
<td>85.0</td>
<td>72.2</td>
<td>345</td>
</tr>
<tr>
<td>VTLN $i = 3$</td>
<td>85.6</td>
<td>72.5</td>
<td>164</td>
</tr>
<tr>
<td>VTLN $i = 4$</td>
<td>84.5</td>
<td>72.5</td>
<td>84</td>
</tr>
<tr>
<td>VTLN $i = 5$</td>
<td>84.0</td>
<td>72.8</td>
<td>75</td>
</tr>
<tr>
<td>VTLN $i = 6$</td>
<td>84.0</td>
<td>72.2</td>
<td>61</td>
</tr>
<tr>
<td>VTLN $i = 7$</td>
<td>85.6</td>
<td>73.3</td>
<td>59</td>
</tr>
<tr>
<td>VTLN $i = 8$</td>
<td>85.0</td>
<td>73.1</td>
<td>57</td>
</tr>
<tr>
<td>VTLN $i = 9$</td>
<td>84.0</td>
<td>72.8</td>
<td>41</td>
</tr>
<tr>
<td>VTLN $i = 10$</td>
<td>84.0</td>
<td>73.0</td>
<td>49</td>
</tr>
</tbody>
</table>

Figure 5.3: Histogram of all speakers (in the training data set) sorted according to $\alpha$. 
5.2 Vocal Tract Length Normalization

Figure 5.4: LID test results of using different length of test data to select $\alpha$. The dashed straight line represents the performance without VTLN.

Table 5.2 presents the results of three different $\alpha$ selection approaches which utilized the entire test segment. The “UBM” approach requires that for each test segment, 13 parameterization sets (13 normalization factors) be generated first and then the likelihood scores against the UBM using Equation (5.2) are calculated.

In contrast, the “$\alpha$-models” method requires that for each test segment one parameterization is first performed to calculate the likelihood scores against all 13 $\alpha$-models using Equation (5.3), then one more parameterization to normalize the test segment using the selected $\alpha$ value. The “$\alpha$-models with top-5 test” method is the same as “$\alpha$-models” except that the top-5 mixture component testing technique is used for likelihood scoring. All $\alpha$-models are created by adapting from the un-normalized UBM since the amount of data assigned to some of the normalization factors was small as depicted at Figure 5.3.

Experimental results shows that the “UBM” approach obtained the best performance as expected although the performances of the other two approaches are similar. This is attributed to the consistency of the $\alpha$ selection process employed across both training and testing. The main difference between these approaches
Table 5.2: Test results (% correct) and time requirement (parameterization + selection) of employing different α selection approaches during testing. Total length of test data (45s+10s test) is 16385s. (All tests were performed on a Pentium III 933MHz 512MB RAM Linux system.)

<table>
<thead>
<tr>
<th>Selection Approach</th>
<th>45s</th>
<th>10s</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBM</td>
<td>84.0</td>
<td>72.8</td>
<td>12,976</td>
</tr>
<tr>
<td>α-models</td>
<td>83.4</td>
<td>72.0</td>
<td>8593</td>
</tr>
<tr>
<td>α-models with top-5 test</td>
<td>82.9</td>
<td>72.0</td>
<td>1502</td>
</tr>
</tbody>
</table>

is their computation requirement. In this experiment, the “UBM” approach required 79% of real-time for completing the parameterization (including normalization) and α selection, whereas the “α-models” required 52% and the “α-models with top-5 test” approach required only 9% of real-time.

**NIST 1996 LID Experiment**

In this experiment, two versions of GMM-UBM system were implemented. One employed PLP (5th order) and the other was LPCC (12th order). Both features have delta energy, delta and acceleration coefficients appended with mean and variance normalization applied. The system setup for PPRLM was the same as used in the experiment described in Section 4.4. All models consisted of GMMs using 512 mixture components. 7 different normalization factors (0.88 ≥ α ≥ 1.12, and stepping by 0.04) were selected for each speaker. A smaller set of α values was selected for this experiment because the amount of training data required to normalize was too large in this task. In addition, preliminary experimentation showed that the LID accuracy remained similar with this smaller set of α values. 5 iteration of α selection were performed on each feature.

Table 5.3 shows the results of the improvement made to both the GMM-UBM and PPRLM systems on the NIST 1996 LID task and the percentage of improvement for each system is shown in Figure 5.5. Results have shown that all LID systems improved their accuracy as result of VTLN. From Figure 5.5, the experimental results also confirmed that the use of VTLN would benefit more to acoustic
Table 5.3: Test results (% correct) of applying VTLN to the GMM-UBM and PPRLM LID system on the NIST 1996 LID experiment.

<table>
<thead>
<tr>
<th>LID System</th>
<th>30s</th>
<th>10s</th>
<th>3s</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM-UBM PLP</td>
<td>62.9</td>
<td>53.8</td>
<td>40.9</td>
</tr>
<tr>
<td>GMM-UBM PLP with VTLN</td>
<td>68.4</td>
<td>59.0</td>
<td>45.6</td>
</tr>
<tr>
<td>GMM-UBM LPCC</td>
<td>66.9</td>
<td>60.0</td>
<td>46.1</td>
</tr>
<tr>
<td>GMM-UBM LPCC with VTLN</td>
<td>74.7</td>
<td>68.0</td>
<td>53.4</td>
</tr>
<tr>
<td>PPRLM 2-gram + Phone Duration</td>
<td>84.7</td>
<td>72.8</td>
<td>51.0</td>
</tr>
<tr>
<td>PPRLM 2-gram + Phone Duration with VTLN</td>
<td>86.3</td>
<td>75.5</td>
<td>54.3</td>
</tr>
</tbody>
</table>

Figure 5.5: Percentage of accuracy improvement after VTLN on each system of the NIST 1996 LID experiment.

information based system where both GMM-UBM systems have obtained better improvement than the PPRLM system.

5.3 Output Score Fusion

Fusing data from different data sources has been shown to be capable of increasing a system’s performance in many different tasks. Within speech technology, it has been successfully applied to gender identification [73], speaker identification, speaker verification system [11] and speech recognition systems [79]. Data fusion can be generally divided into input fusion and output fusion methods. Input fusion is simply the concatenation of different feature vectors for use by a classifier,
while output fusion is the utilization of the output from several classifiers to form the final scores. LID typically processes a large amount of data in order to model a language’s characteristics. Due to this, input fusion makes the input vector size larger and thus reduces the efficiency of the system. Moreover, preliminary LID experiments show that output fusion performs better than the input counterpart.

The main purpose of this study is to examine the effectiveness of combining acoustic and phonetic speech information using output score fusion. To achieve this aim, investigation is first performed to determine the linear fusion strategy that is suitable for fusing output score generated by different LID systems. The GMM-UBM systems studied in Section 3.5 are employed here to serve as the baseline LID system for fusion. Hence, this experiment is based on the NIST 1994 LID task. Once a suitable fusion strategy is established, the fusion of the GMM-UBM and PPRLM system is then examined on the NIST 1996 LID experiment to study the performance of the combination of the acoustic and phonetic speech information for LID.

### 5.3.1 Data Fusion

The basic idea of data fusion is to combine different views or decisions generated by different experts (a classifier in this case) in an attempt to improve the discriminability of the overall system. Data fusion can be generally divided into input fusion and output fusion methods.

#### Input Fusion

Input fusion is simply the concatenation of different feature vectors into a single vector prior to processing by a classifier. A block diagram of an input fusion system is shown in Figure 5.6. The appending of energy, delta energy, delta and acceleration coefficients to the feature vector can be treated as a special case of
5.3 Output Score Fusion

Figure 5.6: Input Fusion System. $f_i, i = 1 \ldots N$, are the feature vectors that capturing different aspect of speech information.

Figure 5.7: Output Fusion System. $f_i, i = 1 \ldots N$, are the feature vectors that capturing different aspect of speech information.

input fusion.

Output Fusion

Output fusion is the utilization of the outputs from several classifiers to give a final score. Figure 5.7 shows a block diagram of such a system. Depending on the type of output the classifier generates, different styles of fusion can be performed. When continuous outputs such as likelihood scores are obtained from the classifier, both linear or non-linear combination methods can be applied. Although non-linear combinations like neural networks were shown to perform better than the linear methods [28, 71], they have the disadvantage of a larger computational expense and the score modelling is more complex. Therefore only the linear combination and other simple related methods were investigated.
5.3 Output Score Fusion

5.3.2 Output Score Fusion Strategies Experiment

As shown in Figure 5.7, the input feature vectors of each classifier was generated by different means. This implies that different feature extraction methods were employed and therefore the output likelihood scores will have a different dynamic range across the classifiers. Thus as a means of normalization, we calculated the \textit{a posteriori} probabilities instead of the likelihood scores as the output of each classifier. The \textit{a posteriori} probability of model $\lambda_i$ given a sequence of feature vectors $X = x_1, x_2, \ldots, x_T$ is defined as

$$P(\lambda_i \mid X) = \frac{P(\lambda_i) \ p(X \mid \lambda_i)}{\sum_{l=1}^{L} P(\lambda_l) \ p(X \mid \lambda_l)}$$  \hspace{1cm} (5.4)$$

where $L$ is the number of models for each classifier or the number of languages in this study and $P(\lambda_i)$ is the \textit{a priori} probability of model $\lambda_i$. Also $p(X \mid \lambda_i)$ is the joint likelihood of observations $x_1, x_2, \ldots, x_T$ given the model $\lambda_i$.

Fusion Strategies

Four different linear output combination strategies are investigated. They are defined as follows:

- Simple Addition (SA)

$$Z_{SA} = \sum_{i=1}^{N} Y_i$$  \hspace{1cm} (5.5)$$

- Simple Multiplication (SM)

$$Z_{SM} = \prod_{i=1}^{N} Y_i$$  \hspace{1cm} (5.6)$$
Table 5.4: LID accuracy (% correct) of the baseline GMM-UBM systems on the NIST 1994 LID experiment.

<table>
<thead>
<tr>
<th>Technique</th>
<th>45s</th>
<th>10s</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>66.8</td>
<td>62.4</td>
</tr>
<tr>
<td>LPCC</td>
<td>69.5</td>
<td>63.0</td>
</tr>
<tr>
<td>MCC</td>
<td>74.3</td>
<td>65.8</td>
</tr>
<tr>
<td>PLP</td>
<td>74.9</td>
<td>66.2</td>
</tr>
</tbody>
</table>

- Max Rule (MAX)

\[ Z_{MAX} = \max_{i=1}^{N} Y_i \] (5.7)

- Linear Score Weighting (LSW)

\[ Z_{LSW} = \sum_{i=1}^{N} w_i Y_i \] (5.8)

where \( Z \) is the final output score, \( Y_i \) is the output score generated by classifier \( i \), \( N \) is the number of classifiers and \( w_i \) is the score weighting for classifier output \( i \) such that

\[ \sum_{i=1}^{N} w_i = 1. \] (5.9)

**Experiment Setup**

The LID accuracies of each of the GMM-UBM systems on the NIST 1994 LID experiment are shown in Table 5.4. Note that by employing “mean and variance normalization” for channel compensation and 10 iterations of adaptation during Language-GMM creation, the performance of these systems has been improved compared to the results in Section 3.5.

The *a priori* probability, \( P(\lambda) \), is set to equal probabilities \( P(\lambda) = (1/11) \) for score normalization.
Table 5.5: LID Results (% correct) comparing linear fusion methods. \( \omega \) is the score weighting applied to the first GMM-UBM system under Fusion column.

<table>
<thead>
<tr>
<th>Fusion</th>
<th>SA 45s</th>
<th>SA 10s</th>
<th>SM 45s</th>
<th>SM 10s</th>
<th>MAX 45s</th>
<th>MAX 10s</th>
<th>LSW 45s</th>
<th>LSW 10s</th>
<th>( \omega )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP-MCC</td>
<td>75.9</td>
<td>70.6</td>
<td>75.9</td>
<td>70.6</td>
<td>74.3</td>
<td>66.2</td>
<td>79.1</td>
<td>0.80</td>
<td>71.4</td>
</tr>
<tr>
<td>PLP-LPCC</td>
<td>74.9</td>
<td>69.0</td>
<td>74.9</td>
<td>68.6</td>
<td>69.5</td>
<td>63.4</td>
<td>78.6</td>
<td>0.75</td>
<td>70.1</td>
</tr>
<tr>
<td>PLP-MFCC</td>
<td>70.1</td>
<td>65.0</td>
<td>70.1</td>
<td>64.8</td>
<td>67.9</td>
<td>62.2</td>
<td>76.5</td>
<td>0.93</td>
<td>67.7</td>
</tr>
<tr>
<td>MCC-LPCC</td>
<td>72.2</td>
<td>67.4</td>
<td>72.2</td>
<td>67.2</td>
<td>74.3</td>
<td>67.4</td>
<td>74.9</td>
<td>0.84</td>
<td>67.8</td>
</tr>
<tr>
<td>MCC-MFCC</td>
<td>72.7</td>
<td>65.8</td>
<td>72.7</td>
<td>65.6</td>
<td>73.8</td>
<td>64.8</td>
<td>75.4</td>
<td>0.89</td>
<td>67.2</td>
</tr>
<tr>
<td>LPCC-MFCC</td>
<td>70.1</td>
<td>64.0</td>
<td>70.1</td>
<td>64.3</td>
<td>67.9</td>
<td>63.8</td>
<td>71.1</td>
<td>0.68</td>
<td>65.1</td>
</tr>
</tbody>
</table>

Experimental Results

Experimental results are shown in Table 5.5. The best result is the fusion of PLP (\( \omega = 0.8 \)) with MCC (\( \omega = 0.2 \)) using LSW with an accuracy of 79.1% for the 45 second test and 71.4% for the 10 second test. The LSW method obtained the highest improvement amongst the tested fusion methods. The performance of SA and SM are approximately the same. Both methods have a slight improvement in accuracy after fusion. The MAX method does not show any improvement in accuracy.

The results shown in the LSW column of Table 5.5 are the best accuracies obtained by exhaustively searching through all the score weighting combinations. The search is done by stepping through the \( \omega \) value of Equation (5.8) from 0.0 to 1.0 with each step incremented by 0.01. This searching approach is only possible when the correct results are available. Unfortunately, these results do not indicate the robustness of the fusion approach. However, it is employed in this experiment to show that a proper choice of score weighting is important to the LSW fusion method. For actual system operation, the optimized score weighting should be obtained by performing the aforementioned searching approach on the output scores generated from a set of development test data.

From Table 5.5, LSW yielded the best improvement when any of the two GMM-UBM systems were fused. In this experiment the fusion method is fixed at LSW.
Table 5.6: LID Results (% correct) comparing fusion of varies number of GMM-UBM systems using Linear Score Weighting.

<table>
<thead>
<tr>
<th>Fusion</th>
<th>LSW 45s</th>
<th>LSW 10s</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Systems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLP-MCC</td>
<td>79.1</td>
<td>71.4</td>
</tr>
<tr>
<td>PLP-LPCC</td>
<td>78.6</td>
<td>70.1</td>
</tr>
<tr>
<td>PLP-MFCC</td>
<td>76.5</td>
<td>67.7</td>
</tr>
<tr>
<td>MCC-LPCC</td>
<td>74.9</td>
<td>67.8</td>
</tr>
<tr>
<td>MCC-MFCC</td>
<td>75.4</td>
<td>67.2</td>
</tr>
<tr>
<td>LPCC-MFCC</td>
<td>71.1</td>
<td>65.1</td>
</tr>
<tr>
<td>3 Systems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLP-MCC-LPCC</td>
<td>79.7</td>
<td>71.8</td>
</tr>
<tr>
<td>PLP-MCC-MFCC</td>
<td>79.1</td>
<td>71.5</td>
</tr>
<tr>
<td>PLP-LPCC-MFCC</td>
<td>78.6</td>
<td>70.4</td>
</tr>
<tr>
<td>MCC-LPCC-MFCC</td>
<td>75.4</td>
<td>68.3</td>
</tr>
<tr>
<td>4 Systems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLP-MCC-LPCC-MFCC</td>
<td>79.7</td>
<td>71.8</td>
</tr>
</tbody>
</table>

and the number of systems vary. Table 5.6 shows the results of this experiment. Again the *a priori* probability, $P(\lambda)$, is set to equally probable and the optimized score weighting searching approach was applied to the LSW results.

The result shows that as the number of GMM-UBM systems increases, the accuracy remains the same or increases. This is consistent with information theory. Comparing the PLP-MCC row with PLP-MCC-MFCC row, there are no obvious improvements in accuracy. However, comparing the PLP-MFCC row with the PLP-MCC-MFCC row, more improvement was obtained. It is suggested that the complementary information provided by different systems is not entirely independent. As in this experiment, fusion of PLP with MCC covered most of the complementary information that other systems were capable of providing. Further increasing the number of systems in the fusion stage will only add a small amount of information to the final decision. Therefore, by choosing the systems that carry more distinct complementary information (which might be indicated by their LID accuracy) the number of systems required to achieve the optimal performance by fusion can be reduced.
5.3.3 Experiment on Combining Acoustic and Phonetic Speech Information

Section 5.3.2 has investigated a number of linear fusion strategies for the fusing of output scores generated by different LID systems. Experimental results have shown that the LSW fusion technique can achieve the most optimal accuracy improvement. Since the main purpose of this section is to examine the effectiveness of combining acoustic and phonetic speech information using output score fusion, this experiment focused on fusing output scores produced by the GMM-UBM and PPRLM LID systems. The setting of the systems is the same as described in Section 5.2.3 where VTLN are applied to those systems and the experiment is based on the NIST 1996 LID task.

Table 5.7 shows the LID accuracies of individual systems and after fusion. Experimental results have shown that output score fusion is capable of utilizing the information provided by both the acoustic and phonetic based system to improve LID accuracy. The amount of improvement provided by combining two GMM-UBM systems is not as significant as combining the GMM-UBM system with PPRLM. This is more obvious in the case of shorter test segments where the accuracies of both the 10s and 3s test cases have improved in a larger scale. This might suggest that more distinct complementary information is carried by systems utilizing different aspects of speech information.

Note that combining the GMM-UBM PLP system with PPRLM does not provide much improvement over fusion. One explanation could be due to the PPRLM system also employing PLP as its features and thus the complementary information provided by the GMM-UBM PLP system overlaps the PPRLM system, which utilized PLP as well. Although PPRLM modelled each language with phonetic information, this result might indicate the PPRLM system also captured certain amounts of acoustic information.
Table 5.7: Output Score Fusion results on NIST 1996 LID experiment. VTLN are applied to all LID systems (% correct).

<table>
<thead>
<tr>
<th>LID System</th>
<th>30s</th>
<th>10s</th>
<th>3s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonetic Reference System [94]</td>
<td>74.3</td>
<td>53.4</td>
<td>34.8</td>
</tr>
<tr>
<td>GMM-UBM PLP</td>
<td>68.4</td>
<td>59.0</td>
<td>45.6</td>
</tr>
<tr>
<td>GMM-UBM LPCC</td>
<td>74.7</td>
<td>68.0</td>
<td>53.4</td>
</tr>
<tr>
<td>PPRLM</td>
<td>86.3</td>
<td>75.5</td>
<td>54.3</td>
</tr>
<tr>
<td>GMM-UBM (PLP+LPCC)</td>
<td>76.1</td>
<td>70.1</td>
<td>55.6</td>
</tr>
<tr>
<td>GMM-UBM PLP + PPRLM</td>
<td>86.5</td>
<td>76.8</td>
<td>58.4</td>
</tr>
<tr>
<td>GMM-UBM LPCC + PPRLM</td>
<td>88.0</td>
<td>79.8</td>
<td>62.5</td>
</tr>
<tr>
<td>GMM-UBM (PLP+LPCC) + PPRLM</td>
<td>88.1</td>
<td>79.8</td>
<td>62.9</td>
</tr>
</tbody>
</table>

5.4 Summary

This chapter studied two techniques to improve the final accuracy of a LID system that is independent of the aspect of speech information and modelling techniques employed within the system. The first technique is the application of VTLN to compensate for the speaker variation presented in speech data. An iterative maximum likelihood approach was adapted to normalize the data against a UBM, which attempted to model the characteristics of speech data uttered by a unified vocal tract length. A number of normalization factor selection approaches during testing was investigated. One of these proposed approaches ($\alpha$-models) was capable of normalizing the testing data in a very rapid fashion that is applicable to any speech processing tasks. Experimental results have shown that VTLN is able to improve the accuracies of both acoustic and phonetic information based LID systems, with a more significant enhancement provided to systems that utilized acoustic information. In addition, using only two seconds of test data for normalization is capable of achieving an accuracy that outperformed the systems without VTLN.

Another LID accuracy improvement technique studied was to combine different complementary sources of speech information at the system output level. Investigations have been performed on selecting the linear fusion strategies suitable
for fusing output likelihood scores generated by different LID systems. With the selected fusion strategies, experiments are then performed on combining the acoustic and phonetic speech information. Experimental results have shown that the LSW fusion technique can achieve the most optimal accuracy improvement. Additionally, combining information from different levels of speech information (acoustic and phonetic in this case) can obtain improved results rather than solely utilizing the acoustic information.
Chapter 6

NIST 2003 Language Recognition Evaluation System

6.1 Introduction

As stated in Chapter 2, the aim of the NIST language recognition evaluation is to establish a new baseline of current performance capabilities for language recognition of conversational telephone speech. The LID technologies developed in previous chapters were employed to develop an evaluation system which was submitted by QUT to the 2003 NIST evaluation. This system included the utilization of acoustic speech information (Chapter 3), phonetic speech information (Chapter 4), vocal tract length normalization and output score fusion (Chapter 5). There were a total of six research groups from around the world participating in this evaluation and this was the first time that QUT joined the event. Twelve languages were targeted to be recognized and there were three different test segment durations (30s, 10s and 3s).

The purpose of this chapter is to describe the detail and the performance of the evaluation system developed by the author as the QUT submission for the 2003
NIST language recognition evaluation. A description of the evaluation task is given in next section. This section is extracted from the evaluation plan [56] to give an insight of the scale and the rules of the task. The full details of the QUT system then follow.

The QUT system comprised of three LID sub-systems and they are GMM-UBM (utilized acoustic information), BPRLM (utilized broad phonetic information) and PPRLM (utilized fine phonetic information). Each sub-system operates independently and the scores they output are linearly fused together to generate the final result. The acoustic based and phonetic based sub-systems have employed different parameterization techniques. Additionally, VTLN are applied to all sub-systems to alleviate the speaker variation. Section 6.4 then covers the results of the evaluation. By placing more emphasis on acoustic speech information and broad phonetic information, the evaluation system has obtained the best results in the 3s test case of the evaluation.

6.2 NIST 2003 Evaluation Task

6.2.1 The Evaluation

The target languages of this task were: Egyptian Arabic, American English, Farsi, Canadian French, German, Hindi, Japanese, Korean, Mandarin, Latin American Spanish, Tamil, and Vietnamese. The task evaluated the detection of a given target language. That is, given a test segment of speech and a target language, the task was to determine whether or not the speech was from the target language.

The performance of a detection system was characterized by its miss and false alarm probabilities; thus these probabilities served as the basis for evaluating system performance on the language detection task. Performance was measured using a detection cost function, $C_{Det}$, which represented the expected cost of
making a detection decision:

\[
C_{Det} = \left( C_{Miss} \times P_{Miss|Target} \times P_{Target} \right) + \\
\left( C_{FalseAlarm} \times P_{FalseAlarm|Non-Target} \times P_{Non-Target} \right)
\] \hspace{1cm} (6.1)

where \( C_{Miss} \) and \( C_{FalseAlarm} \) represent the relative costs of a miss and a false alarm, respectively. For this evaluation \( C_{Miss} \) and \( C_{FalseAlarm} \) were both be 1 and the \textit{a priori} probability of the target language, \( P_{Target} \), was 0.5.

The system under evaluation must be tested on all test segments. For each test segment, all of that system’s target language hypotheses were applied in turn. Thus, there were a total of 12 different trials for each test segment.

For each trial, two outputs had to be provided by the evaluation system. The first output was the actual decision (“true” or “false”) regarding whether or not the language spoken during the test segment was the target language. The second output was a score indicating how likely the language of the test segment was the same as the target language.

### 6.2.2 Test segment duration

The test segments were of three nominal durations, namely 3s, 10s, and 30s. Actual durations were varied but constrained to be within the ranges of 2-4 seconds, 7-13 seconds, and 25-35 seconds, respectively. Shorter test segments were subsets of longer test segments, i.e., each 10-second test segment was a subset of a corresponding 30-second test segment, and each 3-second test segment was a subset of a corresponding 10-second segment.
6.2.3 Evaluation Rules

The following is a list of evaluation rules and restrictions on system development and tests:

- Each test segment is to be processed separately, independently, and without use of any knowledge of other test segments. Especially, normalization over multiple test segments is not allowed.

- Use of the knowledge of the whole set of target languages is allowed. Thus normalization over multiple target languages is allowed, as is limiting (to say, one) the number of languages for which a “true” decision is made on any given test segment. Note, however that there may be test segments from nontarget languages, which are unknown to the system. Use of the knowledge of these languages is not allowed.

- Side knowledge of the sex or other characteristics of test speaker (except as obtained by automatic means) is not allowed.

- Listening to the evaluation data, or any other experimental interaction with the data, is not allowed before all test results have been submitted.

6.3 QUT 2003 NIST Language Recognition Evaluation System

6.3.1 System Structure

The evaluation system submitted for QUT by the author incorporated both the acoustic and phonetic speech information from the speech data. This system contains three sub-systems, where each of them is an independent LID system by
Figure 6.1: Block diagram of the QUT 2003 NIST language recognition evaluation system.

itself, which attempts to capture language characteristics from speech in different aspects. The VTLN technique “α-model” (described in Section 5.2.2), which is suitable for each individual sub-system, was utilized to reduce the inter- and intra-speakers speech variation. Figure 6.1 shows a block diagram of the overall system.

The PPRLM (described in Section 4.2.2) sub-system was utilized to exploit fine phonetic information. Implementation was the same as described in Section 4.4.1 with the LMs utilized bi-gram with phone duration information. The BPRLM sub-system was a variant of PPRLM. The difference between these two sub-systems was that BPRLM captures broad phonetic events instead of fine phonetic details. The “phonetic knowledge approach” (described in Section 4.5.1) was employed to obtain the broad phonetic transcriptions from the fine phonetic counterpart. The LMs of BPRLM utilized tri-gram with phone duration information.

The GMM-UBM (described in Section 3.3.3) sub-system was used to exploit acoustic speech information. This system employed two different parameterization methods to capture acoustic information in different aspects. These are 5\textsuperscript{th} order PLP and 12\textsuperscript{th} order LPCC with the delta, acceleration and delta energy coefficients appended to the final feature vector. Hence, two GMM-UBM sub-
systems were running in parallel, which differ only in speech features employed as shown at Figure 6.1; all GMMs (including VTLN $\alpha$-models) using 512 mixture components.

Finally, different aspects of speech information utilized by the aforementioned sub-systems were then combined at the output level using the “Linear Score Weight” fusion technique (described in Section 5.3) to produce the final likelihood score for each language. The score weighting applied to each sub-system were calculated from the development test data.

### 6.3.2 Generate Evaluation Output

After fusion, every language would have a likelihood score calculated for each test segment. To fulfill the evaluation tasks, which were required to perform a detection on all target languages given a test segment, the evaluation system also needed to produce a “true” or “false” decision. To achieve this task, a threshold was defined where any likelihood scores larger than this threshold were determined as “true” and vice versa. Since the performance of a system was evaluated according to the detection cost function as defined in Equation (6.1), this threshold was selected so that this function’s result would be minimized. In addition, before the likelihood score was compared to the threshold, the scores (across all languages of a given test segment) were mean and variance normalized so that all likelihood scores, which included the threshold, would be measured in the same scale. In this evaluation, the threshold was calculated from the development test data and a different threshold was computed for every test segment duration.

As mentioned in Section 6.2.3 where the use of the knowledge of the whole set of target languages was allowed for each test segment, it was possible to restrict the decision to allow only one “true” to be assigned. However, in order to handle the case of unknown language test segments, the decision was made not to apply this restriction to the evaluation system so that the unknown language test case
could be handled.

6.4 Evaluation Results

Figures 6.2, 6.3, and 6.4 shows the evaluation results of the test cases for 30s, 10s, and 3s respectively. The performances of all six sites which participated in the evaluation are included in these figures and the results are showed in terms of the Detection Error Tradeoff (DET) curves (described in Section 2.7.2). The circle on the DET curve shows the system operating point where the detection cost is minimum and the diamond indicates the actual decision selected for the system. Note that we are not permitted to name the individual participants. The participant list included sites from world leading organizations in the area of language recognition research. From the evaluation results, it is seen that our system has achieved very competitive performance in both the 30s and 10s test cases. In addition, the operating point selected was very close to the minimum detection cost compared to other systems. More importantly, the best result was obtained in the 3s test cases by the QUT system ahead of all other sites.

Figure 6.5 shows a plot of the weighting applied to the output scores during fusion in different test segment durations. It indicated that when the test segment duration is long (30s), the weighting has emphasized mostly on fine phonetic information. However, when the duration of the test segment is reduced, the weighting on the fine phonetic was gradually decreased with both the broad phonetic and acoustic information receiving more emphasis. By looking at the 3s test segment duration, the weighting of the fine phonetic information was even lower than the acoustic counterpart. Hence, this might be suggest that when the test segment is short, the information provided by fine phonetics become unreliable and therefore the use of lower level speech information is more desirable.
6.5 Summary

This chapter described the details and the performance of the evaluation system developed for the 2003 NIST language recognition evaluation. The evaluation system comprised of three LID sub-systems which utilized acoustic, broad phonetic and fine phonetic information. Output score fusion was employed to combine all these different aspects of information and VTLN was applied to all sub-systems to reduce speaker variation. A score threshold, which would minimize the detection cost of the system, was calculated to turn the LID system into a language detection system. The evaluation system achieved competitive performance in the 30s and 10s test cases. More importantly, the best results were achieved in
Figure 6.3: The DET curve for the 2003 NIST language recognition evaluation 10s test case.

the 3s test case.
Figure 6.4: The DET curve for the 2003 NIST language recognition evaluation 3s test case.
Figure 6.5: The weighting applied on the output scores generated by different sub-systems during fusion for different test segment duration. These weightings were calculated from the development test data set.
Chapter 7

Conclusions and Further Work

7.1 Conclusions

This research covers the utilization of two different aspects of speech information for the task of automatic spoken language identification. This involved the use of acoustic speech information to obtain an efficient LID system and the use of phonetic information to improve the LID accuracy. This work provides a number of original contributions in utilizing acoustic and phonetic speech information for LID.

7.1.1 Acoustic Speech Information Investigation

This study proposed the use of the UBM technique in GMM based LID system as an alternative approach to the standard GMM system to model the acoustic speech information. Study has been performed on an 11 language identification task. It was shown that the proposed system achieved similar accuracy as the standard GMM system with the computational requirements at both the training and testing stages significantly reduced. Further, a more robust model can be
obtained to represent the characteristics of each language.

Different speech parameterization techniques have also been studied for the GMM-UBM LID system as an attempt to examine acoustic features that emphasized different speech characteristics. It was found that acoustic features which captured vocal tract characteristics are more discriminative between different languages. In addition, the inclusion of energy and dynamic coefficients (delta and acceleration coefficients) was shown to be capable of improving LID accuracy.

### 7.1.2 Phonetic Speech Information Investigation

Different aspects of phonetic speech information were studied for the PPRLM LID system. Experiments were performed using different amounts of phonetic detail (broad and fine phonetic details, and the inclusion of phone duration information) and length of phonetic context (the order of $N$-gram LM) to describe the characteristics of a language. It was concluded that the use of more detailed phonetic events is more desirable than the incorporation of longer phonetic context. Further, the employment of longer phonetic context could result in a performance drop when the test segment duration is short.

### 7.1.3 Accuracy Improvement Techniques

Two improvement techniques applied to different stages of a LID system were studied. The first technique is targeted towards pre-processing the speech data to obtain more accurate features to represent their characteristics. For this purpose, VTLN was studied as an approach to alleviate the variation introduced by different speakers. It was found that the accuracy of the GMM-UBM was greatly improved and positive results were obtained by the PPRLM system. Hence, VTLN is mandatory for LID systems utilizing the acoustic speech information.
7.2 Further Work

Additionally, a novel approach that performs VTLN rapidly during the testing stage was proposed. This technique is applicable to other speech processing tasks that require fast speaker normalization.

The second studied technique was focused on combining information provided by different aspects of speech information at the output level. A linear score weighting fusion technique was applied to fuse the output scores generated by LID systems that utilized different speech information. Experimental results showed that incorporating multiple acoustic and phonetic speech information provides significant improvement in LID accuracy. It was concluded that utilizing multiple aspects of speech information would produce a more accurate LID system.

7.1.4 QUT 2003 NIST Evaluation System

An LID system utilizing both the acoustic and phonetic information and the improvement techniques was developed to participate in the 2003 NIST language recognition evaluation. Competitive performance was achieved with the $3s$ test case obtaining the best result. It was found that LID systems placing more emphasis on acoustic speech information and broad phonetic information could handle shorter duration test segments more robustly.

7.2 Further Work

As the GMM-UBM LID system provided an efficient means to identify spoken languages automatically, it is worth the effort to develop techniques to further improve the accuracy of this system since most LID applications require faster than real-time operation. The use of Shifted Delta Coefficients (SDC) [81] with the GMM-UBM approach was proposed recently and achieved a better performance than phonetic information based LID system. The basic idea behind SDC
is to capture the delta coefficient’s trajectory characteristics. Hence, study can be performed on incorporating temporal information of the acoustic features by the GMM.

The use of more detailed phonetic events was shown to be more important for discriminating between languages. As a result of this work, other aspects of phonetic information can be investigated to further enrich the details of the phonetic events; for example, the exploitation of formant characteristics, duration and pitch variation and context-dependent information. Another line of research following the study of broad phonetic event is to investigate the number of phonetic events sufficient to discriminate between different languages, which is similar to the study performed by Berkling [1]. Techniques to improve phone recognition accuracy would also improve performance of phonetic information based LID systems.

In addition to improving the speech features to be extracted from the speech data, it is important to employ a technique that can accurately capture and combine information from different speech features. In this study, a linear data fusion technique was investigated. As an extension to this work, the relationship between the output scores generated by different LID systems can be captured non-linearly. One such approach is to treat the likelihood scores calculated for all languages as a feature vector. Statistical modelling techniques, such as Neural Networking, are then employed to model the characteristics of these output scores. With a better understanding of the relationship between LID systems utilizing different speech information an improvement in LID accuracy can be expected.
Bibliography


Appendix A

Application of LID - Multilingual Phone Recognition

This work was presented at the 2003 International Conference on Acoustics, Speech and Signal Processing Conference under the title of “Three Approaches to Multilingual Phone Recognition” [87].

A.1 Abstract

This paper investigates and compares three different approaches of multilingual phone recognition (MPR). Two types of MPR approach are defined according to the Language Identification (LID) process of the system: Explicit-LID where language identification is mandatory, and Implicit-LID where LID is an integrated part of the MPR process. The OGI-TS database is employed to perform the isolated and continuous MPR experiments. Three of the world’s most spoken languages; English, Mandarin and Spanish are selected as the target languages for the system. Experimental results indicate that different MPR approaches should be employed for different applications according to the degree of LID ac-
curacy that can be achieved from the input test utterance. If high LID accuracy is achievable, the MPR approach that depends on LID can obtain better performance. Conversely, the Implicit-LID MPR approach is more appropriate.

A.2 Introduction

The ability to process speech in multiple languages by a single speech recognition system has become increasingly desirable due to the trend of globalization and the popularity pervasive of the Internet. This multilingual feature not only extended the usability of the system, but also allows it to process a larger range of speech data.

One of the more popular approaches to perform multilingual speech recognition is the utilization of a multilingual phone set. These multilingual phones are usually created by merging phones across the target languages that are acoustically similar in an attempt to obtain a minimal phone set that covers all the sounds that exist in all the target languages [21, 29, 67]. One important application of this approach is that the multilingual phone set can be adapted to recognize speech of an unknown language with no or limited adaptation speech data. Unfortunately, performance of the system employing this approach was not comparable to its monolingual counter part. Therefore, more investigation on this and other multilingual approaches are necessary.

In this paper, research on multilingual speech recognition is focused at the phone level using the OGI-TS corpus [52]. Three of the most spoken languages in the world; English, Mandarin and Spanish are selected as the target languages. In this study it is assumed that the unknown input speech data is monolingual. Three different approaches that perform multilingual phone recognition are investigated and descriptions of these approaches are given in Section A.3. Details of the multilingual test systems are given in Section A.4. Section A.5 presents the
A.3 Multilingual Phone Recognition Approaches

For the task of Multilingual Phone Recognition (MPR) the language of the input speech utterance is unknown. Thus, Language Identification (LID) will be performed at some stage during the recognition process in order to produce the final monolingual results. This LID procedure in the system can generally be divided into two different categories: Explicit-LID and Implicit-LID. Explicit-LID implies that the MPR system performs LID on the unknown input data explicitly. Implicit-LID means that LID is an integral part of the MPR process.

A.3.1 Explicit-LID - Approach 1

One approach to Explicit-LID employs an external LID system to first identify the language of the input utterance and the corresponding monolingual system is then selected to perform the phone recognition (as shown at Figure A.1, Approach 1). This is one of the most straightforward approaches to achieve MPR. When no LID errors occur, this approach achieves the same performance as the monolingual systems. Therefore the accuracy of the external LID system is the main concern to the overall system performance. The advantage of this approach is that it can employ language-dependent speech recognition techniques (e.g. different acoustic and language models) on each monolingual recognizer. However, it can not handle the case where the input utterance contains multiple languages as this system can only give monolingual results.
A.3.2 Explicit-LID - Approach 2

Another approach to Explicit-LID is to run all of the monolingual recognizers in parallel and select the output generated by the recognizer that obtained the highest likelihood score as the final result (as shown at Figure A.1, Approach 2). LID in this approach is performed at the end of the MPR process. The selection of final output not only results in the phone transcription, but also identifies the language of the input utterance. This approach shares similar characteristics with Approach 1, differing only in the strategy of the LID process.

A.3.3 Explicit-LID - Approach 3

A MPR approach that belongs to Implicit-LID utilizes a set of multilingual phones within a single recogniser (as shown at Figure A.1, Approach 3). These multilingual phones can be created by merging monolingual phones amongst the target languages that share the same phonetic symbol (e.g. IPA [22] and Worldbet [16]) or acoustically similar according to a certain distance measure or similarity criteria.

The aim of this phone merge procedure is to obtain a minimal phone set that covers all of the sounds that exist in the target languages, however, certain language-specific information are removed from the system after the phone merge procedure. The symbols of these merged monolingual phones will be mapped to a new symbol in the transcriptions. Therefore, before process the speech data for training and testing, all phonetic transcriptions will be mapped to the new multilingual phone set. Note that LID is not required to perform MPR with this approach, as the generated output can be directly compare to the correct (mapped) transcriptions. Nevertheless, language can be explicitly identified when this approach is applied to multilingual speech recognition. The recognized output in that case will indicate the language of the input utterance at the text level.
Figure A.1: Block diagram of the three different approaches to achieve multilingual phone recognition.

The advantage of this MPR approach is the recognition system can be configured to handle input utterances that contain multiple languages, however language-specific speech recognition techniques can not be easily applied as some of the language specific information is merged together (at the phone level) in the system.

A.4 Test Systems

A.4.1 Baseline Monolingual Phone Recognizers

The OGI-TS database [52] and hidden Markov model toolkit (HTK) [19] are used to train and test the phone recognizer for the languages of English, Mandarin and Spanish. Each feature vector is extracted at every 10ms and comprises Perceptual Linear Predictive (PLP) cepstral coefficients with energy plus it’s delta and acceleration coefficients. Phone model (monophone) topology comprises a 3 state Hidden Markov Model (HMM) with 8 Gaussian mixture components per state. Original transcriptions, which are transcribed using the Worldbet [16] phonetic symbol set, are modified such that all diacritics are removed. In addition, some phones are mapped to similar sounds to ensure sufficient data coverage. A language tag is appended to each phone symbol to preserve language information while performing multilingual experiments. The amount of training data and the
number of phones used in each monolingual recognizer are shown in Table A.1.

### A.4.2 Multilingual Phone Recognition Test Systems

The Approach 1 MPR system in this study employs an external LID system that utilizes efficient Gaussian Mixture Model (GMM) analysis, as described in [86]. The characteristics of each language is modelled by a single GMM. During LID, the language represented by the GMM that obtained the highest likelihood score against the unknown test utterance is deemed as the identified language. The Universal Background Model (UBM) technique [61] was employed to improve the efficiency of the system during both LID training and testing. In addition, Vocal Tract Length Normalization (VTLN) was applied to reduce speaker variation presented in the speech data. All GMMs, including the UBM for both VTLN and testing and GMM for each language, are trained using the OGI-TS database. The advantages of this GMM-LID system are that transcriptions are not required for model training and faster than real time performance can easily be obtained on regular computer hardware. The error rate of this GMM-LID system tested on an 11-language experiment with 45-second test segment is around 14% [86].

The Approach 2 MPR test system makes direct use of the baseline monolingual phone recognizers without any modification. It has the simplest structure among the three approaches but has the highest computational cost.

A multilingual phone set is created for the MPR test system that utilized Approach 3 by merging monolingual phones with the same Wordlbot symbol. For each merged phone, a new model is trained with the data pooled from the merged languages. After the phone merge procedure, the number of phones is reduced from a total of 118 to 73. As mentioned at Section A.3.3, all phonetic transcriptions are mapped to the multilingual phone set with the original language tag at each phone symbol removed.
A.5 Experiments and Results

Table A.1: Details and experimental results of the baseline monolingual phone recognizers. Continuous phone recognition results under “% Accuracy” included insertion error while “% Correct” does not.

<table>
<thead>
<tr>
<th>Language</th>
<th>Training Data (Hour)</th>
<th># Phone Model</th>
<th>% Correct</th>
<th>% Accuracy</th>
<th>Deletion</th>
<th>Substitution</th>
<th>Insertion</th>
<th>Total Test Phones</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Isolated Phone Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>1.5</td>
<td>41</td>
<td>49.2</td>
<td>-</td>
<td>-</td>
<td>4103</td>
<td>-</td>
<td>8079</td>
</tr>
<tr>
<td>Mandarin</td>
<td>0.5</td>
<td>40</td>
<td>50.5</td>
<td>-</td>
<td>-</td>
<td>2744</td>
<td>-</td>
<td>5539</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.9</td>
<td>37</td>
<td>58.0</td>
<td>-</td>
<td>-</td>
<td>3092</td>
<td>-</td>
<td>7357</td>
</tr>
<tr>
<td></td>
<td>Continuous Phone Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>1.5</td>
<td>41</td>
<td>47.9</td>
<td>32.8</td>
<td>1432</td>
<td>3487</td>
<td>1430</td>
<td>9448</td>
</tr>
<tr>
<td>Mandarin</td>
<td>0.5</td>
<td>40</td>
<td>49.6</td>
<td>34.3</td>
<td>1032</td>
<td>2368</td>
<td>1028</td>
<td>6740</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.9</td>
<td>37</td>
<td>56.3</td>
<td>45.9</td>
<td>904</td>
<td>2893</td>
<td>904</td>
<td>8689</td>
</tr>
</tbody>
</table>

A.5 Experiments and Results

The test data set contains 17 test utterances (continuous speech) for each language and each utterance has a duration of around 45 seconds. Both isolated and continuous phone recognition experiments were performed. For isolated phone recognition, a single phone token is prepared as input for each test system. Each of these tokens are extracted from the test data set according to the transcription, with those tokens with a duration of less than 30ms (or 3 feature vectors) removed. For continuous phone recognition experiments, the entire test utterance is utilized as input to the test systems. A phone-loop recognition network is used to perform the continuous recognition.

Table A.1 shows the experimental results of the baseline monolingual phone recognizers along with the number of deletion, substitution and insertion errors for the continuous phone recognition experiment. All results are obtained with approximately the same deletion and insertion rate using the insertion penalty option in HTK. The average number of correctly recognized phones is about 53% for isolated phone recognition and 51% for continuous phone recognition. No direct comparison of the results across the languages can be made as the amounts of train and test data varies. However, it does indicate that phone recognition in Spanish might be an easier task than the other two languages due to lower number of phones.
Table A.2: Multilingual (isolated and continuous) phone recognition experimental results. “% LID” is in percentage of correct.

<table>
<thead>
<tr>
<th>MPR Approach</th>
<th>English</th>
<th>Mandarin</th>
<th>Spanish</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% Correct</td>
<td>% LID</td>
<td>% Correct</td>
<td>% LID</td>
</tr>
<tr>
<td>Isolated Phone Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approach 1</td>
<td>27.6</td>
<td>(4339/8079)</td>
<td>53.7</td>
<td>27.0</td>
</tr>
<tr>
<td>Approach 2</td>
<td>32.8</td>
<td>(4959/8079)</td>
<td>61.4</td>
<td>26.7</td>
</tr>
<tr>
<td>Approach 3</td>
<td>38.7</td>
<td>-</td>
<td>32.2</td>
<td>-</td>
</tr>
<tr>
<td>Baseline</td>
<td>49.2</td>
<td>-</td>
<td>50.5</td>
<td>-</td>
</tr>
<tr>
<td>Continuous Phone Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approach 1</td>
<td>43.3</td>
<td>(16/18)</td>
<td>88.9</td>
<td>49.6</td>
</tr>
<tr>
<td>Approach 2</td>
<td>45.8</td>
<td>(18/18)</td>
<td>94.4</td>
<td>46.4</td>
</tr>
<tr>
<td>Approach 3</td>
<td>37.3</td>
<td>-</td>
<td>33.5</td>
<td>-</td>
</tr>
<tr>
<td>Baseline</td>
<td>47.9</td>
<td>-</td>
<td>49.6</td>
<td>-</td>
</tr>
</tbody>
</table>

The experimental results for both the multilingual isolated and continuous phone recognition tasks are shown at Table A.2. LID accuracy for Approach 1 and 2 were also included. No LID was required for Approach 3 as mentioned in Section A.3.3 and also for the baseline system where the language information is known a priori (or equivalently 100% correct LID).

A.5.1 Isolated Phone Recognition Results

Input speech data in this task typically contains fewer frames of speech and therefore provided very little information for the system to correctly identify its language. Therefore, multilingual isolated phone recognition is a difficult task and this is supported by the experimental results, where in average 31% of phones are correctly recognized across all the approaches compared to 53% for the baseline system. LID accuracy of Approach 1 and 2 are only 49% and 50% respectively. The phone recognition performance is greatly affected by the poor LID results as these approaches depend on LID to decide the final monolingual output. Approach 3 performs the best amongst all MPR approaches with 36% of phones correctly recognized. This approach does not depend on LID and its decline in performance is mainly contributed by the confusion between the 73
multilingual phone models.

A.5.2 Continuous Phone Recognition Results

Approximately 45 seconds of speech data is used as input in this experiment. Experimental results indicate that typically a LID accuracy of around 90% was achieved. The exception to this was Spanish, which resulted in 77% for Approach 1 and only 35% for Approach 2. Phone recognition rates for MPR Approach 1, 2 and 3 are 47%, 39% and 36% respectively compared to 51% for the baseline system. The success of Approach 1 hinges on the LID accuracy achieved by the system with an average of 89%. It can be seen that when applied to continuous phone recognition the increase in available input data duration results in an improvement in LID performance and subsequently an improvement in overall phone recognition for Approach 1 and 2. Conversely Approach 3 does not benefit from this increase in utterance duration.

Comparing the experimental results across the languages in Approach 3, the performance drop in Spanish compared to the baseline system was 18%. This compares to the drop of around 11% and 16% for English and Mandarin respectively. One explanation for these results is that the Spanish phone set (37 phones) has a higher degree of confusion when contained within the multilingual phone set (73 phones) compared to English and Mandarin (41 and 40 phones respectively). This is reinforced by the results achieved by Approach 1 and 2 where the LID accuracy for Spanish was similarly poor. Conversely for English and Mandarin, with higher LID accuracy of around 92% and 97% respectively averaged across Approach 1 and 2, phone recognition rate of Approach 3 did not suffer to as large degree. Therefore, although Approach 3 did not perform LID, the language-specific information presented in the data are contributed implicitly to the overall performance. In contrast to Approach 1 and 2 where the language-specific information are extracted explicitly from the input data, these
A.6 Conclusions

This paper compares three different approaches to achieve multilingual phone recognition and present results from experiments from isolated and continuous phone recognition. Three of the world’s most spoken languages; English, Mandarin and Spanish were selected as the target languages. For the isolated phone recognition experiment, MPR Approach 3 (Implicit-LID, which utilizes multilingual phone set in a single recognizer) has performed the best with 36% phone recognition rate. It obtains superior results because it does not suffer from the difficulty of identifying the language from a single phone token. In the continuous phone recognition experiment, MPR Approach 1 (Explicit-LID, which utilized an external LID system to select monolingual recognizer) obtains the best phone recognition rate of 47% compared to 51% for the baseline monolingual system. Its superior performance is mainly due to the higher LID accuracy obtained by the external GMM-LID system in the experiment.

Experimental results indicate that different MPR approaches should be employed for different applications according to the degree of LID accuracy that can be achieved from the input test utterance. If high LID accuracy is achievable, Explicit-LID MPR approach that depends on LID can obtain better performance, otherwise the Implicit-LID MPR approach is more appropriate.
Appendix B

Cross-lingual Speech Recognition (Indonesian)

This work was presented at the 2003 EuroSpeech Conference under the title of “Multilingual Phone Clustering for Recognition of Spontaneous Indonesian Speech Utilising Pronunciation Modelling Techniques” [85].

B.1 Abstract

In this paper, a multilingual acoustic model set derived from English, Hindi, and Spanish is utilized to recognize speech in Indonesian. In order to achieve this task we incorporate a two tiered approach to perform the cross-lingual porting of the multilingual models to a new language. In the first stage, we use an entropy based decision tree to merge similar phones from different languages into clusters to form a new multilingual model set. In the second stage, we propose the use of a cross-lingual pronunciation modelling technique to perform the mapping from the multilingual models to the Indonesian phone set. A set of mapping rules are derived from this process and are employed to convert the original Indone-
sian lexicon into a pronunciation lexicon in terms of the multilingual model set. Preliminary experimental results show that, compared to the common knowledge based approach, both of these techniques reduce the word error rate in a spontaneous speech recognition task.

B.2 Introduction

The resource necessary to produce Automatic Speech Recognition (ASR) systems for a new language are considerable. Of all the resources required, obtaining sufficient transcribed acoustic data and lexicons presents a major problem for many languages. There are still several languages with major population bases which have insufficient resources for the development of speech enabled applications. Our research is focused on producing generic techniques that exploit existing resources from Source languages for porting ASR technology to Target languages.

Indonesia has a population of 190 million and is the largest Moslem nation on earth. When ranking languages to include in the Global-Phone speech database, population; variability; distribution; religious circumstances and linguistic aspects were the factors considered [70]. Indonesian/Malay was ranked ninth suggesting it also ranks highly in the speech science community. Accordingly, there is particular interest in producing speech enabled applications for the Indonesian languages, and so a secondary focus of our research is to extend the generic methods developed for cross-lingual porting of resources to the Indonesian languages.

The process of utilizing resources from a data rich language (referred to as Source language in this paper) to a data poor language (referred to as Target language), can be broken into two basic steps. The first step involves the creation of the Source language/s model set. Monolingual models can be used for cross-lingual ASR, however in [29] and [69] it was reported that improved recognition results can be achieved using multilingual models. One explanation for this is that
multilingual models offer better coverage of the Target language because a model set covering multiple languages is more likely to provide a better match to the Target language than a monolingual Source set.

The second step in the process is to then express the lexicon of the Target language (Indonesian) in terms of the Source language models (a multilingual model set in this experiment). Past attempts to produce the lexicon have relied on either knowledge based or data driven methods. Knowledge based methods exploit linguistic knowledge to extract mappings between the representational symbols for the sounds of languages. The symbolic representation for the various sounds, and a description of articulatory features is provided by the International Phonetic Alphabet (IPA) [22]. Data driven techniques typically rely on either confusion matrix based approaches or alternatively on distance measures, usually based on relative entropy. Results reported in [3] indicated that data driven methods produce superior results in comparison to knowledge based methods.

In both cases, a single representative from the Source language/s is chosen to represent a phonetic event in the Target language. However, variation due to context can result in allophonic variants, which are often predictable but not captured. These variants will not be captured if we consider each phonetic event in isolation, such as the one to one mapping provided by the aforementioned methods. To address this issue we implement a technique introduced in a companion paper [44] which utilizes pronunciation modelling to capture the systematic variation that occurs in the Target language with the models from the Source language/s.

The rest of the paper is organized as follows: Section B.3 discusses the approach used to create a multilingual phone set from multiple languages. Section B.4 then presents the technique used to perform the lexicon mapping from the Target to the Source language. We then describe the Indonesian speech recognition experiments including results in Section B.5. Finally a discussion and conclusions are given in Section B.6 and B.7 respectively.
B.3 Multilingual Phone Model Clustering

The multilingual model set in this investigation is derived from the languages of English, Hindi, and Spanish. The process of merging similar phone models is achieved via a relative entropy based decision tree clustering technique, using a set of phonetic related questions. The purpose of employing the decision tree in our experiments differs from that reported in [57] for monolingual context dependent modelling, and that used in [69] for multilingual context dependent modelling. In [69], the underlying motivation was to use the decision tree process to cluster phones with similar contexts in order to overcome the problem of insufficient training data. In our case, the decision tree is used to merge phone models based on acoustic similarity estimated from the training data. The resultant clusters then maximally represent the Source languages acoustic evidence with a minimal set of models.

B.3.1 Clustering Criteria

The phonetic question set is arranged hierarchically from broad questions such as “Is the phone a Fricative” to more specific questions such as “Is the phone the Spanish /b/”. Effectively the question set places no restriction on the possible number of phone models within each cluster. Language tags are added to each phone so that both language dependent and language independent clusters are possible. No restrictions prevent phone models from the same language being clustered together. In other studies [89], emphasis was placed on ensuring that phones from the same language were not merged. The rationale for this restriction was that if a resultant cluster was too broad, and included phones from the same language, then the possibility for different words having the same phonetic transcription (homophones) increases. However, the intended aim of creating multilingual models for cross-lingual speech recognition is to increase the coverage of the Target language’s acoustic/phonetic space. To this end, letting the
clustering process be unconstrained, based on the acoustic evidence, is preferable to imposing constraints based on language specific phonemic requirements that may not be applicable to the Target language.

### B.3.2 Clustering Process

In the beginning of the clustering process, all monolingual phone models are pooled together at the first node of the decision tree and then split according to the phonetic questions. A threshold is used to stop the splitting process and in turn controls the number of clusters. At the end of the splitting process, phone models that belong to the same cluster are tied together to form a new multilingual model and their corresponding acoustic data are used to re-estimate it’s parameters. Table B.3.2 shows the first twenty questions that were asked during the decision tree clustering process. Questions that belong to broader phonetic events were asked first then progressively followed by more specific questions.

The aim of creating a multilingual phone set in this experiment is to perform Indonesian speech recognition, therefore we would like to choose the threshold to obtain a model set that maximizes the acoustic coverage for Indonesian. In addition, the threshold must be set so that the subsequent models are flexible enough to cope with allophonic variation. The last property is important as capturing variation is fundamental to the idea of pronunciation modelling discussed in Section B.4. The criteria we adopted for selecting the threshold was based on linguistic consideration, where the cluster size was empirically adjusted until sufficient coverage was obtained for the Target language. In our experimental setup, the baseline Indonesian system used 30 phone models. We selected a threshold that yielded 47 multilingual phone models. Table B.3.2 shows the details of the assignment of Source monolingual phone models to 47 clusters after the clustering process. The left column indicates the number of phones assigned to a single cluster and the right column illustrates the corresponding clusters which belong
Table B.1: The first twenty questions asked during the decision tree clustering (en - English, hi - Hindi, sp - Spanish).

<table>
<thead>
<tr>
<th>Phonetic Question</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Vowel_en_hi_sp</td>
<td>All vowel phones</td>
</tr>
<tr>
<td>2. Low_en_hi_sp</td>
<td>Low vowel phones</td>
</tr>
<tr>
<td>3. UnStrident_en_hi_sp</td>
<td>Stops, nasals, liquids</td>
</tr>
<tr>
<td>4. Stop_en_hi_sp</td>
<td>Stops</td>
</tr>
<tr>
<td>5. unvoiced_closures_en_hi_sp</td>
<td>Unvoiced closures</td>
</tr>
<tr>
<td>6. Fronting_en_hi_sp</td>
<td>Front Diphthongs</td>
</tr>
<tr>
<td>7. FrontVowel_en_hi_sp</td>
<td>Front vowel phones</td>
</tr>
<tr>
<td>8. unvoiced_closures_hi_sp</td>
<td>Unvoiced closures</td>
</tr>
<tr>
<td>9. Continuent_en_hi_sp</td>
<td>Fricative, liquids, glides</td>
</tr>
<tr>
<td>10. Rounded_en_sp</td>
<td>Rounded vowel phones</td>
</tr>
<tr>
<td>11. voiced_closures_en_hi_sp</td>
<td>Voiced closures</td>
</tr>
<tr>
<td>12. voiced_closures_hi_sp</td>
<td>Voiced closures</td>
</tr>
<tr>
<td>13. Medium_en_hi_sp</td>
<td>Medium vowel phones</td>
</tr>
<tr>
<td>14. UnvoicedStop_en</td>
<td>Unvoiced Stop</td>
</tr>
<tr>
<td>15. Liquids_en</td>
<td>Liquids</td>
</tr>
<tr>
<td>16. High_en_hi_sp</td>
<td>High vowel phones</td>
</tr>
<tr>
<td>17. Glottals_en_hi_sp</td>
<td>Glottals</td>
</tr>
<tr>
<td>18. Short_en_hi_sp</td>
<td>Short vowel phones</td>
</tr>
<tr>
<td>19. EVowel_en_hi_sp</td>
<td>Vowels involved /e/</td>
</tr>
<tr>
<td>20. UnvoicedFricative_en_hi_sp</td>
<td>Unvoiced Fricative</td>
</tr>
</tbody>
</table>

to each category. Each string in the right column represents a single cluster, where the ‘-’ character is used to group the monolingual phone models of that cluster and the ‘_’ character is the language tag.

B.4 Pronunciation Modelling

In a companion paper, [44], we investigated using pronunciation modelling to produce an Indonesian lexicon which captured the pronunciation of Indonesian words, expressed using symbolic representation for the Source language acoustic models. The basic procedure for achieving this is as follows:

1. Use data driven pronunciation modelling to obtain a pronunciation lexicon
Table B.2: A set of 47 multilingual phones obtained from the decision tree clustering process. Each string represents a cluster with the ‘-’ character used to connect each monolingual phone. The ‘-’ character is the language tag, where en - English, hi - Hindi, sp - Spanish.

<table>
<thead>
<tr>
<th># Phone merged in a cluster</th>
<th>Phone list</th>
</tr>
</thead>
</table>
| > 5                         | U.en_hi_sp-&.en_hi_sp-1x_hi-hs_sp  
Gottal.en_hi_sp-T_en-v_en-z_en-z_hi  
Dr_hi-tr_hi-r(.hi-rr_hi-L_sp-r_.sp-r_sp  
Ts_en_hi_sp-dZ_en-dZ_hi-tSh_hi  
g_en_hi_sp-kH_hi-k_hi-k_sp |
| 5                           | d.en-D_en-d{.hi-t{H_hi-d{.sp  
b.en_hi_sp-p_hi-p_sp  
l=en-n=en->i_en-o_u_en-x3_sp |
| 4                           | sil-nos1-nos2-bH_hi  
w_hi-V_sp-G_sp-D_sp  
n_en-n_hi-N_hi-n{.sp  
j_en_hi_sp-nj_sp |
| 3                           | h.en-h_hi-dZ_sp  
F_en_hi_sp  
vcl_en_hi_sp ph_en-kh_en-x_sp  
vuvcl_en_hi_sp m_en_hi_sp  
u_h-o_hi-o_sp  
S_en_hi_sp I_en_hi_sp |
| 2                           | s_hi-s_sp  
w_en-w_sp  
l_hi-l_sp  
N_en-N_sp  
T{.hi-t{.sp  
1_en-u_sp  
9r_en-kr_en  
i_hi-i_sp  
E_hi-E_sp  
e1_en-ai_hi  
`en-`hi  
A_en-a_hi  
>_{en->hi  
A_en-al_sp |
| 1                           | s_en  
th_en  
u_en  
3r_en  
i_en  
e_sp  
e_hi  
E_en  
aU_en  
a_sp  
@_hi  
@_en |

for the Target language training data in terms of the symbols used to represent the Source language acoustic models. This was only done for those lexicon units for which there was sufficient training examples.

2. Use this pronunciation lexicon to derive a set of rules that map the Target language phones to the Source language. Prune these rule according to predetermined criteria.

3. Apply the rule set derived to the entries which occur in the Target language vocabulary but were not optimized in Step 1. This results in full coverage of the desired Target language vocabulary in terms of the symbols used to represent the Source language acoustic models.
The methodology used in [44] employed a modification of monolingual techniques outlined in [6] and [76]. We use the syllable to constrain the boundary of context width which is then used to derive the rule set. The rationale for the selection of the syllable is outlined in [44]. We use 1.3 hours of Indonesian training data, which is segmented into isolated syllables. An optimization technique is then used to select the Source language phone-string (baseforms) for each syllable, which maximizes the likelihood of the training data.

Context dependent pronunciation rules are derived from the optimized baseforms. These rules are then pruned, and subsequently used to produce a lexicon based on the multilingual acoustic data.

### B.5 Experiments and Results

We conducted experiments using spontaneous telephone speech from both the 11 language (for English, Hindi and Spanish data), and the 22 language (Indonesian data) versions of the Oregon Graduate Institute Multi Language Telephone Speech Corpus [32]. Using data recorded in a similar environment provided the opportunity to standardize the training and test environment and hopefully reduce the impact of train/test mismatch and variations in channel effects. No transcriptions for the Indonesian acoustic data existed originally and so two native speakers were employed to transcribe three hours of speech data. This was then verified and corrected for errors. The speech data was split into a training set (1.3hrs), a development test set (54mins) and a test set (25mins). The Indonesian acoustic data transcribed included all utterance categories such as stories, age, routes, climates etc. We used a subset of a commercially produced 20 000 word Indonesian lexicon which included syllable demarcation. Further details of the transcription process and lexicon development are outlined in [43]. To avoid out-of-vocabulary errors the subset provided orthographic transcriptions for all the 2519 words that occurred in the train, development test and test data.
A speech recognition engine was developed using the multilingual monophone models derived from English, Hindi, and Spanish. Additionally a baseline Indonesian system was developed which also uses context independent acoustic models. For both recognition systems, the HMM topology was a 3 state left-to-right, with each state emission density comprising 8 Gaussian mixture components. A bi-gram Language Model, trained using the training and development test set, is incorporated. Speech was parameterized using a 12th order MFCC analysis plus normalized energy, 1st and 2nd order derivatives, and a frame size/shift of 25/10ms. Cepstral Mean Subtraction (CMS) was employed. The word recognition performance for this baseline Indonesian system is outlined in the first row of Table B.5.

Knowledge based mapping experiments were performed for the purpose of comparison with the pronunciation modelling technique. For the knowledge based approach, the Target languages lexicon is mapped in terms of the Source language phone set using linguistic knowledge. Experimental results are shown in the second row of Table B.5. Compared to the pronunciation modelling technique (the third row), the results suggest that system performance can be improved by incorporating pronunciation modelling. This suggests that it is important to handle the variation in pronunciation in the Target language’s lexicon.

The same pronunciation modelling experiment was performed using only English
as the *Source* language. The corresponding testing results are shown at the last row of Table B.5. This reinforces results previously reported in [69] where the multilingual models provided better recognition results and highlights that this phenomena is applicable for the Indonesian language. The more diverse coverage of acoustic space provided by the multilingual models increase the chances of better modelling the *Target* language and therefore would yield higher accuracy.

All experiments were repeated using the models after adaptation, where the adaptation used a development data set. The adaptation procedural sequence is as follows: conduct a global mean only MLLR; mean only MLLR for each model; mean and variance MLLR for each model; 5 iterations of mean an variance MAP adaptation. The experimental results are depicted in the second column of Table B.5. Consistent outcomes were obtained after adaptation supporting that both the multilingual phone set approach and the pronunciation modelling technique can improve the accuracy of a cross-lingual speech recognition system.

### B.6 Discussion

The limited amount of training data meant that the acoustic models could only reasonably be based on monophones. This subsequently constrained the possible recognition performance to a lower level. Using continuous telephone speech data also served to restrict the recognition accuracy. However, the preliminary indications are that the combined use of multilingual model set and pronunciation modelling are an effective means for obtaining improved cross lingual recognition performance.

In our experiment concerning the creation of the multilingual phone set, we wanted to examine the impact of not restricting the merge of phone models from the same language during the clustering process. As outlined in Section B.3.1 the disadvantage of this technique is that it could result in a cluster which is
phonetically too broad, with the potential loss in discrimination between words.

Interestingly, in an Indonesian lexicon, defined using English models and pronunciation modelling, two homophones occurred. In comparison, the combination of multilingual phone set with pronunciation modelling introduced only an additional two homophones. Although the lexicon is comparatively small, (2.3k words), this indicates that the impact of our clustering technique does not significantly increase the possibility for homophones, yet allows for a decrease in confusability by removing redundant clusters.

The pronunciation modelling technique we employed constrained the number of variants included in the Target language lexicon to one. If the clustering process (without language restriction) produces homophones, then the pronunciation modelling technique can produce an N-best list of pronunciations which can be traversed and substituted into the lexicon.

**B.7 Conclusions**

In this paper we presented a two tiered approach to cross-lingual porting of multilingual models. We investigated the use of an unconstrained decision tree based approach, which relies on a set of phonetic related questions, to merge acoustically similar phone models from multiple languages. The utilization of the multilingual phone set in this task reinforced previous reported results and showed that this technique is applicable for the Indonesian language.

The addition of pronunciation modelling to the multilingual clustering produced improved recognition results. Additionally it has the potential in future extensions to minimize the homophones which can occur because of the unconstrained clustering process by substituting alternate pronunciation variants.