Glottal Waveforms for Speaker Inference
&
A Regression Score Post-Processing
Method Applicable to General Classification Problems

David James Vandyke
27 June 2014

A Thesis Submitted in Requirement for the Degree of
Doctor of Philosophy (Information Sciences & Engineering)

Human-Centred Computing Lab
Faculty of Education, Science, Technology & Mathematics
University of Canberra
Glottal Waveforms for Speaker Inference
&
A Regression Score Post-Processing
Method Applicable to General
Classification Problems

SUPERVISORY PANEL:

Professor Michael Wagner
Faculty of ESTeM
University of Canberra

Associate Professor Roland Göecke
Faculty of ESTeM
University of Canberra

Associate Professor Girija Chetty
Faculty of ESTeM
University of Canberra

EXAMINERS:

DATE APPROVED:
Certificate of Authorship of Thesis

Form B
Certificate of Authorship of Thesis

Except where clearly acknowledged in footnotes, quotations and the bibliography, I certify that I am the sole author of the thesis submitted today entitled –

Glottal Waveforms for Speaker Inference & A Regression Score
Post-Processing Method Applicable to General Classification
Problems

[Thesis title]

I further certify that to the best of my knowledge the thesis contains no material previously published or written by another person except where due reference is made in the text of the thesis.

The material in the thesis has not been the basis of an award of any other degree or diploma except where due reference is made in the text of the thesis.

The thesis complies with University requirements for a thesis as set out in Gold Book Part 7: Examination of Higher Degree by Research Theses Policy, Schedule Two (S2). Refer to http://www.canberra.edu.au/research-students/goldbook

--------------------------------------------------------
Signature of Candidate

--------------------------------------------------------
Signature of chair of the supervisory panel

Date: .............................

NOTE: The wording contained in Form B must be bound into the thesis, preferably as the third page, and signed by the author of the thesis and the supervisor. The exact layout does not need to be duplicated.
Acknowledgements

I wish to thank my primary supervisor Professor Michael Wagner for his introducing me to speech as a biometric, and for his support, suggestions and guidance throughout the learning process that has been my doctoral studies. Thank you also to Associate Professor Roland Goecke for his active supervision, knowledge and advice over the last three years. My appreciation goes to Girija Chetty, Alice Richardson, Robert Cox, Serena Chong, Jason Webber and several other members of the University of Canberra’s Faculty of Education, Science, Technology & Mathematics (ESTeM) for their approachability, insights and assistance. Thanks also to Phil Rose for his expertise regarding the forensic content of this thesis and to Professor Elliot Moore II for graciously hosting my visit to ‘Georgia Tech’.

To my fellow research students I wish you all the best with your future endeavours. A special thanks to my fiancée Hannah and my family for their love and support (and proof reading talents!).

Finally I acknowledge and appreciate the receipt of the Australian Postgraduate Award from the Australian Government and the WJ Weeden Scholarship from the University of Canberra.

Sincerely, David Vandyke
Abstract

Contributions are made along two main lines. Firstly a method is proposed for using a regression model to learn relationships within the scores of a machine learning classifier, which can then be applied to future classifier output for the purpose of improving recognition accuracy. The method is termed \( r \)-norm and strong empirical results are obtained from its application to several text-independent automatic speaker recognition tasks. Secondly the glottal waveform describing the flow of air through the glottis during voiced phonation is modelled for the task of inferring speaker identity. A prosody normalised glottal flow derivative feature termed a source-frame is proposed with empirical evidence presented for its utility in differentiating speakers. Inferences are also made from the glottal flow signal regarding detection of the affective disorder depression. Comprehensive literature reviews of the fields of automatic speaker recognition, forensic voice comparison and the estimation of the glottal waveform are also presented.

**Keywords:** Glottal Waveform, Text-Independent Speaker Recognition, Forensic Voice Comparison, Score Post-Processing, Depression Detection

This thesis is available as a .pdf (~5 MB) with Chapter, Section and Reference PDF links from:
http://staff.estem-uc.edu.au/davidv/
Quotes and Allegories

“...every voice is making its own particular contribution to the whole...”

Johann Georg Sulzer, 1774, on the musical theory of polyphony in:
Allgemeine Theorie der Schnen Knste (General theory of the fine arts)

“Superb fairy-wren (Malurus cyaneus) females call to their eggs and
upon hatching nestlings produce begging calls with key elements from
their mothers incubation call. ... We conclude that wrens use a parent-
specific password learned embryonically to shape call similarity with their
own young and thereby detect foreign cuckoo nestlings.”

Diane Colombelli-Négre et. al., 2012, a beautiful example of the voice
as biometric in nature from: Embryonic Learning of Vocal Passwords in
Superb Fairy-Wrens Reveals Intruder Cuckoo Nestlings [84]

“The Death’s-head Hawkmoth’s squeak
is sufficiently similar to the sound
made by the hive’s queen bee to fool
the workers into believing that their
queen is instructing them to remain
passive.”

A theory for the use of the high pitched squeak produced by the Death’s-
head Hawkmoth (Acherontia atropos) in stealing honey. Notably this
squeak is generated through a source-filter process.
karlshuker.blogspot.com.au/2012/07/at-sign-of-deathshead

v
Publications

Peer-Reviewed Conferences

- **R-NORM: Improving Inter-Speaker Variability Modelling at the Score Level via Regression Score Normalisation.**
  *David Vandyke, Michael Wagner and Roland Goecke.*

- **Voice Source Waveforms for Utterance Level Speaker Identification using Support Vector Machines.**
  *David Vandyke, Michael Wagner and Roland Goecke.*
  **Awarded Best Conference Paper**

- **Speaker identification using glottal-source waveforms and support-vector-machine modelling.**
  *David Vandyke, Michael Wagner, Girija Chetty and Roland Goecke.*
  Proceedings of the Australasian International Conference on Speech Science and Technology (SST), pages 49-52, Sydney, Australia, 3-6 December 2012.
  **Awarded Best Student Paper**

Peer Reviewed Abstracts

- **The Voice Source in Forensic-Voice-Comparison: a Likelihood-Ratio based Investigation with the Challenging YAFM Database.**
  *David Vandyke, Phil Rose and Michael Wagner.*
  Proceedings International Association of Forensic Phonetics and Acoustics (IAFPA) 2013, Tampa, Florida.

Doctoral Consortium

- **Depression Detection & Emotion Classification via Data-Driven Glottal Waveforms.**
  *David Vandyke.*
# Table of Contents

Certificate of Authorship of Thesis .................................................. ii  
Acknowledgements ........................................................................... iii  
Abstract ............................................................................................. iv  
Quotes and Allegories ......................................................................... v  
Publications ......................................................................................... vi  
List of Figures ...................................................................................... xi  
List of Tables ....................................................................................... xiv  
Nomenclature ....................................................................................... xvi  

## 1 Introduction  
1.1 Motivation and Aims ................................................................. 1  
1.2 Chapter Outlines ........................................................................ 2  

## 2 Literature Review: Speaker Recognition ...................................... 4  
2.1 Introduction ................................................................................. 4  
2.2 What is Automatic Speaker Recognition? ................................. 5  
2.3 Human Processes and Ability in Recognising Speakers ............. 8  
2.4 Precursors: Signal Processing, Speech Recognition and Early Attempts ....................................................... 9  
2.5 Representing the Human Voice: Features for Speaker Recognition ................................................................. 11  
2.6 Variability Compensation and Robustness .............................. 15  
2.7 Statistically Modelling the Human Voice ................................. 18  
2.8 State of the Art Speaker Recognition Systems ....................... 19  
2.9 Cautions and the Future of Speaker Recognition .................... 23  

## 3 The Glottal Waveform ................................................................. 24  
3.1 Introduction ................................................................................. 24  
3.2 Speech Production and the Glottal Waveform .......................... 24  
3.2.1 Physiology of the Human Larynx and the Vocal Folds ....... 25  

vii
5.4 Experiment 3: Distance Metric on Mean Source-Frames ............... 102
5.4.1 Introduction ................................................. 102
5.4.2 Experimental Design ........................................... 103
5.4.3 Results and Discussion ........................................ 106
5.5 Experiment 4: Frame Level Identification with Support Vector Machines 109
5.5.1 Introduction ................................................ 109
5.5.2 Experimental Design ......................................... 109
5.5.3 Results and Discussion ........................................ 111
5.6 Experiment 5: Support Vector Machine Approach at the Utterance Level 113
5.6.1 Introduction ................................................ 113
5.6.2 Experimental Design ......................................... 113
5.6.3 Results ........................................................ 116
5.7 Experiment 6: Glottal Information with $\ell$-Norm ...................... 123
5.7.1 Introduction ................................................ 123
5.7.2 Experimental Design ......................................... 123
5.7.3 Results ........................................................ 130
5.8 Chapter Summary .................................................. 137

6 Glottal Waveforms: Forensic Voice Comparison 138
6.1 Introduction ..................................................... 138
6.2 Literature Review: Forensic Voice Comparison and the Glottal Waveform 139
6.2.1 The Evolving Paradigm of Forensic Voice Comparison ............ 139
6.2.2 The Glottal Waveform in Forensics .......................... 144
6.3 Experiment 1: YAFM Database Naive Listener Task .................... 145
6.3.1 Introduction ................................................ 145
6.3.2 The YAFM Database .......................................... 145
6.3.3 Experimental Design ......................................... 146
6.3.4 Results and Discussion ...................................... 147
6.4 Experiment 2: YAFM Database Statistical Forensic Voice Comparison 151
6.4.1 Introduction ................................................ 151
6.4.2 Experimental Design ......................................... 151
6.4.3 Results and Discussion ...................................... 154

7 Glottal Waveforms: Depression Detection and Severity Grading 158
7.1 Introduction: Depression and the Need for Quantitative Assessment Tools 158
7.2 Literature Review: Glottal Flow for Automatic Detection of Depression 160
7.3 Experiment 1: Investigation on the Black Dog Institute Dataset ........ 163
7.3.1 Introduction ................................................ 163
7.3.2 Experimental Design ......................................... 164
7.3.3 Results and Discussion ...................................... 167

8 Conclusion 173
8.1 Thesis Contributions ............................................. 173
8.2 Future Research Directions ....................................... 175
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Average number of annual IEEE speaker recognition publications.</td>
<td>4</td>
</tr>
<tr>
<td>3.1</td>
<td>Vocal tract diagram including location of vocal folds</td>
<td>25</td>
</tr>
<tr>
<td>3.2</td>
<td>Vocal fold cycle during voiced phonation</td>
<td>28</td>
</tr>
<tr>
<td>3.3</td>
<td>Waveform of airflow through the glottis during voicing</td>
<td>30</td>
</tr>
<tr>
<td>3.4</td>
<td>Diagram of the source-filter theory of speech production</td>
<td>33</td>
</tr>
<tr>
<td>4.1</td>
<td>Simple visual example of score trends $r$-norm adjusts for</td>
<td>53</td>
</tr>
<tr>
<td>4.2</td>
<td>Schematic outlining the steps of the $r$-norm method</td>
<td>55</td>
</tr>
<tr>
<td>4.3</td>
<td>Decision landscape: target and non-target score distributions</td>
<td>59</td>
</tr>
<tr>
<td>4.4</td>
<td>Schematic of $z$-norm and $t$-norm processes</td>
<td>60</td>
</tr>
<tr>
<td>4.5</td>
<td>NIST-2003 DET curves for raw and normalised scores</td>
<td>66</td>
</tr>
<tr>
<td>4.6</td>
<td>NIST-2003 relative frequency histograms of scores pre and post $r$-norm</td>
<td>67</td>
</tr>
<tr>
<td>4.7</td>
<td>Post $r$-norm EER surface over $i_T$ &amp; $i_{NT}$ values</td>
<td>72</td>
</tr>
<tr>
<td>4.8</td>
<td>NIST-2006 relative frequency histograms for target and non-target i-vector scores, pre &amp; post $r$-norm</td>
<td>76</td>
</tr>
<tr>
<td>4.9</td>
<td>NIST-2006 DET curves pre and post $r$-norm</td>
<td>76</td>
</tr>
<tr>
<td>4.10</td>
<td>Wideband relative frequency histograms for target and non-target scores, pre and post $r$-norm</td>
<td>81</td>
</tr>
<tr>
<td>4.11</td>
<td>G.711 relative frequency histograms for target and non-target scores, pre and post $r$-norm</td>
<td>82</td>
</tr>
<tr>
<td>5.1</td>
<td>EER &amp; misID rate for combinations of MFCC &amp; VSCC</td>
<td>91</td>
</tr>
<tr>
<td>5.2</td>
<td>DET plots for combinations of MFCC &amp; VSCC</td>
<td>92</td>
</tr>
<tr>
<td>5.3</td>
<td>Inter-speaker relative-frequency histogram for phonetic groups</td>
<td>97</td>
</tr>
<tr>
<td>5.4</td>
<td>Distribution of intra-speaker phonetic group scores</td>
<td>100</td>
</tr>
<tr>
<td>5.5</td>
<td>Distribution of intra-speaker phonetic group scores</td>
<td>101</td>
</tr>
<tr>
<td>5.6</td>
<td>Example mean source-frame</td>
<td>105</td>
</tr>
<tr>
<td>5.7</td>
<td>DET curve for mean source-frame/distance metric on YOHO</td>
<td>108</td>
</tr>
<tr>
<td>5.8</td>
<td>Source-frame variation covered by increasing PCA dimensions</td>
<td>110</td>
</tr>
<tr>
<td>5.9</td>
<td>Source-frame identification rates against PCA dimension size</td>
<td>112</td>
</tr>
<tr>
<td>5.10</td>
<td>Variance of Source-Frame data covered by principal component basis</td>
<td>115</td>
</tr>
</tbody>
</table>
A.7 Mean source-frame from a single speaker ............... 182
A.8 Mean source-frames from two distinct speakers ....... 182
A.9 Mean source-frames from three distinct speakers ....... 183
A.10 Mean source-frames from four distinct speakers ....... 183
A.11 Mean source-frames from five distinct speakers ....... 184
A.12 LDA Projection of 2 speakers ......................... 185
A.13 Distribution of LDA projections ...................... 185
List of Tables

2.1 Relaxation of restrictions that have occurred over time in speaker recognition systems ........................................... 10
2.2 Levels of speaker recognition features ................................................................. 11
2.3 Sources of nuisance variation occurring in speech waveforms ............................... 15

4.1 NIST-2003 EER and minDCF values for each normalisation method ............... 65
4.2 Breakdown of AusTalk participants used in experiment ................................. 68
4.3 AusTalk EERs pre and post r-norm ................................................................. 71
4.4 AusTalk pre r-norm score distribution summary statistics ................................ 71
4.5 NIST-2006 r-norm performance on i-vector baseline ....................................... 75
4.6 NIST 2006 score distribution summary statistics ................................................ 75
4.7 Wall Street Journal - Phase II paired conditions of enrol and test data ........... 78
4.8 WSJ1: EER and minDCF pre and post r-norm for the 4 experiments performed on the WSJ1 data. ................................................................. 80
4.9 Score distribution summary statistics for WSJ1 experiments. .......................... 80
4.10 Summary of r-norm performance measured by EER ....................................... 83

5.1 Parameter-Value pairs as used in VSCC replication. ........................................ 90
5.2 VSCC replication misidentification rates & EER on YOHO .............................. 90
5.3 Grouping of voiced English letters used for phonetic dependence testing on TI-46 database. ................................................................. 94
5.4 Inter-speaker Kolmogorov-Smirnov test p-values for phonetic groups ............. 96
5.5 Intra-speaker Kolmogorov-Smirnov test p-values for TI-46 speakers .................. 98
5.6 Summary results of source-frame/distance measure ANDOSL experiment .... 106
5.7 Number of source-frames for SVM training per cohort size .............................. 111
5.8 Best source-frame identification rates for each speaker cohort size .................. 111
5.9 Summary results for multiclass SVM modelling ............................................... 116
5.10 Summary results: mean identification rates using SVM regression ............... 119
5.11 AusTalk female individual systems EER and minDCF ..................................... 130
5.12 AusTalk male individual systems EER and minDCF ...................................... 131
5.13 Weighted fusion results for AusTalk females .................................................. 132
5.14 Logistic regression fusion results for AusTalk females .................................... 132
5.15 Weighted fusion results for AusTalk males .................................................... 133
5.16 Logistic regression fusion results for AusTalk males ................. 133
5.17 AusTalk male individual systems ID rates ......................... 134
5.18 AusTalk male individual systems ID rates ......................... 134

6.1 Summary statistics for all YAFM naive listening task responses ...... 148
6.2 Summary statistics of responses to YAFM naive listening task - English
   L1 speakers only .................................................. 148
6.3 \( C_{llr} \) for the MFFC, glottal and fused systems ...................... 155

7.1 Summary of depression classification results .......................... 167
7.2 Original paper accuracies .......................................... 168
7.3 Accuracies of the logistic regression classifier ................. 169
Nomenclature

CC        Complex-Cepstrum
CP        Closed-Phase
DET       Detection Error Tradeoff (Curve)
EER       Equal-Error Rate
EM        Expectation-Maximisation (Algorithm)
FAR       False Acceptance Rate
FRR       False Rejection Rate
F0        Fundamental Frequency
GMM       Gaussian Mixture Model
GCI       Glottal Closure Instant
GOI       Glottal Opening Instant
IF        Inverse-Filtering/Filtered
JFA       Joint Factor Analysis
LF        Liljencrants-Fant
LL        Log-Likelihood
LLR       Log-Likelihood Ratio
MAP       Maximum A Posteriori (Adaptation)
MFCC      Mel Frequency Cepstral Coefficients
SF        Source Frame
SVM       Support Vector Machine
UBM       Universal Background Model
V-V       Volume-Velocity (Airflow in cm$^3$/s)

Corpora Abbreviations

ANDOSL    Australian National Database of Spoken Language [260]
AusTalk   An audio-visual corpus of Australian English [85]
BDI       Black Dog Institute [43]
NIST SRE  National Institute of Standards and Technology - Speaker Recognition Evaluations [281]
TIMIT     TIMIT Acoustic-Phonetic Continuous Speech Corpus [157]
TI-46     TI 46-Word [239]
WSJ       Continuous Speech Recognition - Wall Street Journal [296]
YAFM      Young Australian Female Map-task [332]
YOHO      YOHO Speaker Verification [71]
Chapter 1

Introduction

1.1 Motivation and Aims

The focus of this thesis is on the information contained within a speaker’s glottal waveform, primarily for aiding in the recognition of a speaker’s identity. This is a considerably underutilised source of information for the problem of speaker recognition and only a limited number of studies have been published exploring the extent to which this signal is beneficial for the task. It is understood that information pertaining to the speaker’s vocal-tract (MFCC) is the most revealing of speaker identity but, particularly for high security systems, the inclusion of additional complementary descriptors of identity have the potential to increase recognition accuracies and also improve robustness to environmental factors and spoofing.

To this end the glottal waveform is described, further empirical evidence of its potential for speaker recognition presented and its ability to complement MFCC quantified. In almost all investigations a normalised time domain representation of the derivative of the glottal flow is used. This feature is referred to as a source-frame and is described in Section 3.5. It is a data-driven representation that captures speaker idiosyncrasies that functional forms (theoretic models) for the glottal flow likely spurn.

The use of glottal flow signal is also investigated for the automatic detection of depressive illnesses. A promising method whose results were reported on a small dataset is replicated on a larger clinical dataset. This is particularly important in developing tools to aid patients and clinicians given the difficulty in recording and creating valid datasets for research purposes and the parallel problem given increased ethical sensitivities of sharing such corpora. These are both important problems and their efficient solutions will result in several considerable benefits to society, as later outlined.
A fundamentally related general aim of this thesis is to improve the recognition accuracy of statistical classification methods, to which the standard treatment of the specific problems of text-independent speaker recognition and depression detection belong. To this end a score post processing method based on regression and termed r-norm is proposed. It is described and then its ability to improve classification results explored empirically in several text-independent speaker recognition experiments. Although we focus on the primary problem of interest, namely speaker recognition, the r-norm method is applicable to any general classification problem where scores are output in quantifying unknowns. As such it has the potential to be beneficial in a considerable range of pattern recognition problems from diagnostic medical imaging to spam email detection.

1.2 Chapter Outlines

An overview is now provided outlining the key points and purpose of each chapter.

1. Chapter 2 → The focus problem of automatic speaker recognition is introduced before providing a comprehensive review of the research literature covering the development up to modern state-of-the-art. Emphasis is placed on features for representing the human voice and the limited use of the glottal waveform, despite its potential, is highlighted.

2. Chapter 3 → An overview of the speech production process with a focus on the larynx is given before describing the quasi-periodic airflow through the glottis during voiced phonation, the so called glottal waveform. A literature review of methods for estimating this signal from digitised speech, parameterising it and the results of its use for speaker recognition is then provided. After the parameterisation section a novel feature is introduced that is based on a normalised representation of obtained glottal flow estimates and that enables the application of various modelling techniques. These features are termed source-frames.

3. Chapter 4 → A novel score post-processing algorithm is introduced that employs a regression function for learning systematic errors in the scores output by general machine learning classifiers. The method is termed r-norm. The results of the application of the r-norm method to the scores of four different recognition experiments are then reported, providing strong empirical evidence for the potential of the method to increase a systems recognition accuracy.
4. Chapter 5 — Several experiments are reported all relating to the use of glottal information for speaker recognition. First a replication of a previously published promising feature based on a cepstral representation obtained without inverse-filtering is presented followed by five experiments employing the proposed source-frame features. In the final experiment the proposed score post-processing method $r$-norm is also applied to the classifier’s scores. The results reported in this chapter provide evidence for the speaker dependent information contained within the glottal flow waveform.

5. Chapter 6 — To complement the Chapter 2 overview of speaker recognition a review of the evolving field of forensic voice comparison (FVC) is presented, concluding with a description of the very limited research into the use of glottal flow features for FVC. Following this the results of a naive listening task are reported in order to introduce the new YAFM database [332] and to obtain an understanding of the data as well as a small insight into its’ suitability for testing of FVC methods. A statistical FVC experiment is then reported demonstrating the ability of source-frame features to complement a MFCC baseline.

6. Chapter 7 — A description of the prevalent problem of depressive disorders is given to motivate the large potential benefits for practitioners and patients alike of developing objective tools for the automatic detection of depression. A brief literature review of the use of speech, specifically focusing on the glottal flow, for the detection of such illnesses is given. The results of the replication of a promising small study [267] are then reported as well as an investigation of the features proposed therein for not just the classification of depressed/non-depressed but for prediction of the severity of the illness.

7. Chapter 8 — The primary contributions of the thesis are listed. Future research directions related to unexplored ideas and possibilities opened by the presented results are also discussed.

8. Appendix A — Multiple plots of source-frames from same and different speaker’s are presented in order to provide a qualitative insight into their typical shape and also their intra and inter speaker variation which makes them suitable as features for recognition. Also presented are the results of preliminary investigations into otherwise unreported modelling approaches.
Chapter 2

Literature Review: Speaker Recognition

2.1 Introduction

This chapter examines the historical and influential research published over the last fifty years that has formed the understanding we have today of the text-independent automatic speaker recognition task.

The volume of research into speaker recognition continues to grow each year, and this growth will presumably maintain its momentum at least until techniques are established with sufficiently low error rates to allow the safe adoption of the technology by society. The rate of this growth in research is suggested by the number of publications annually from the Institute of Electrical & Electronics Engineers (IEEE) with “speaker recognition” as a keyword. This is shown visually in Figure 2.1.

Figure 2.1: The average number of annual IEEE publications with “speaker recognition” as a keyword has increased exponentially since the 1950s. The decade from 2010 is based only on data from 2010,11 & 12.
The need society has for increased security and access control in the modern data driven, digital world has stimulated interest in the biometric field and clearly the utility of speech as a biometric can be seen by an ever increasing group of modern day academics.

2.2 What is Automatic Speaker Recognition?

Primarily speech is used to convey linguistic meaning. However it is an information rich signal which typically also carries indicators of emotional state, age, health, language, culture, gender, personality, dialect, education, intelligence and identity. Speaker recognition is a generic label for any task of differentiating people by their voice, as such it is concerned with this last indicator, identity.\(^1\)

Speaker recognition, first explored in the 1960s, is one of a growing number of biometrics, where the term biometric is a portmanteau of 'biology' and 'metric' and describes any quantifiable system whereby personal identity is determined by measurements related to the persons biologic make-up \(^2\). Biometrics can also include a behavioural aspect which may reflect flexible, learned behaviour expressed within the constraints specified by one’s biology. Examples include gait, keystroke and handwriting. Speech is interesting in that it combines factors that are both physiological and behavioural. What may be termed one’s ‘natural sounding voice’ is predominately shaped by the physiology of the vocal tract, mouth and nasal cavities and the tongue and lip articulators.\(^3\) However there is a strong behavioural aspect to the quality of one’s voice. A person can learn one or more languages, with geography and social cohort shaping a certain dialect. Further more, in the process of uttering a phrase a person has a certain linguistic intent and emotional basis for shaping sounds in a specific manner. Along with the physiological changes induced by age and health, all of these factors combine to render speech a highly dynamic process that poses many challenges to reliably and consistently solving the speaker recognition problem whilst similarly possessing the variability to allow discrimination of persons even in large cohorts.

Automatic speaker recognition describes the application of engineering and statistical

---

\(^1\) Speaker diarisation is the related but distinct task of separating out the speech of each individual speaker present within a given recording. In the majority of speaker recognition research it is assumed that only a single speaker is present.

\(^2\) The earliest biometrics were suggested in the 19th century, the best know being suggested by a French policeman named Alphonse Bertillon whose anthropometry system consisted of taking measurements of the size and proportion of human features. This was surpassed by fingerprinting, which has a history of use dating back to antiquity for biometric related applications such as contract seals. Today research continues to improve the use of fingerprints along with a host of other biometrics of justifiable and speculative quality. These include deoxyribonucleic acid (DNA), iris, retina, face, vein, ear and teeth.

\(^3\) The speech production process and its’ scientific modelling are described in detail in Section 3.2.
methods implemented in software for the purpose of determining the identity of the person who uttered a given recording of speech, using only the information contained within that recording. This determination falls into the two categories of speaker identification and speaker verification. The speaker identification (SI) process involves determining the speaker from a group of people (possibly with a null option that the speaker is not present within the group), while speaker verification (SV), also termed authentication, consists of making an accept or reject decision against a specifically claimed identity. Almost all real world applications of speaker recognition are authentication tasks [223].

Almost all automatic speaker recognition systems involve a training phase (also called enrolment) where models are constructed to represent each speakers voice and a testing phase where unknown speech is presented and an identity decision is made. The first step in training is to find a concise, descriptive and stable representation of the speech samples that captures the identity information and disregards to the highest extent all other non identity informative characteristics. This process of feature extraction is reviewed in detail in Section 2.5. An accurate quantitative description of these extracted features for each speaker is then constructed and represented by statistical model or template, a process termed speaker modelling, approaches to which are reviewed in Sections 2.7 and 2.8. The decision logic step at testing then involves quantifying the degree of similarity between the newly presented, previously unseen speech sample for which the identity of the generating speaker is unknown.

A fundamental difference between the task of SI and SV is that whilst a closest match is sufficient in the SI task the SV task requires a much greater understanding of the variability of the measured features [36]. SV tasks quantify the similarity between the test speech and the model for the claimed identity, producing a score which is then compared against a pre determined threshold, making accept decisions if the score falls above this threshold, mutatis mutandis.

The authentication system is thus capable of making two distinct types of error: False Rejections (Type I) and False Acceptances (Type II). These errors are in conflict with each other; for example one can always achieve a false rejection rate (FRR) of zero for any system by simply raising the threshold to authenticate every test claim, in turn guaranteeing a 100% false acceptance rate (FAR). Thus in describing the performance

---

4Performance will always be better for verification tasks than identification (with N classes) tasks as the information gained from the recognition task is $\ln_2 2 = 1$ versus $\ln_2 N$ respectively, given the respective uniform priors in both cases. Phrased differently, the chance of a making a correct SV decision at random is $\frac{1}{N}$ against $\frac{1}{2}$ in the SI case.

5This test sample can be text-dependent where a specific word or phrase is required or text-independent, which is the focus of this thesis and the majority of research to date.

6Assuming higher scores indicate greater similarity, as occurs for any probabilistic classifiers.
of a SV system one must make reference to both error rates at the specified threshold. More often the equal-error rate (EER) of the system is quoted, being the rate at which FAR is equal to FRR. How these errors relate to client (‘target’) and impostor (‘non-target’) scores is described elegantly by the pioneer of iris recognition J. Daugman [88].

General reasons for the adoption of the ‘what you are’ biometric systems are their convenience over the ‘what you have’ paradigm of keys and tokens (can lose, theft target) or the ‘what you know’ method of passwords (can forget). Evidence of growing adoption rates and increased research funding is given by the USA Defence Department which has budgeted $3.5 billion for biometric spending for the 2007 through 2015 fiscal years.

Speech has several advantages over other biometrics due to the facts that it can be collected via non-invasive methods and easily transmitted rapidly over long distances via already prevalent technologies. This is in contrast with many other biometrics which are visual based and require larger bandwidths for data transmission and greater computer processing power. Providing the accuracy and reliability of automatic speaker recognition continues to increase, it is feasible to see the technology integrated into more and more facets of daily life. Current and foreseeable applications of speaker recognition include:

- **Access control**: telephone banking, daily transactions, building admittance, customs and immigration checks.
- **Forensics**: identifying a speaker at a recorded crime or clearing a suspect, surveillance for investigation.
- **Automatic transcription**: meetings, interviews, and court proceedings can be associated with a specific person.

Methods for displaying system accuracy over a range of operating points include ROC and DET curves. Receiver Operating Characteristic (ROC) graphs show on linear scales the trade off of error rates in classification problems, plotting the true acceptance rate against the false alarm rate [136]. A scalar measure of a classifiers average performance can be taken as the area under the ROC curve. As an average measurement of classifier performance it is possible for classifier X with higher AUC than classifier Y to perform worse than Y in regions of ROC space. More commonly DET curves are used, which show the error trade offs on a non linear scale [255]. Assuming scores are normally distributed, plots are given for normal deviates corresponding to the probability of false/true acceptance. As a result straight line DET curves indicate that the target/non-target scores are normally distributed.

Where the cost of misclassification is prohibitively high these can be adopted in parallel such that you are required to present ‘something you have, something you know and something you are’ to meet authentication requirements, creating systems that are potentially extremely hard to spoof.

2.3 Human Processes and Ability in Recognising Speakers

The human brain is the most complex machine known to man. In particular there remains much to learn with respect to the processes it employs to perform identity recognition and language understanding tasks. The ear coupled with the brain’s auditory processing of sound enables humans to form an awareness of their environment, perform spatial location, communicate, enjoy music and rapidly recognise linguistic meaning and speaker identity. Sound enters through the ear canal and is processed in narrow bandwidths with various neurons responding to certain complex sounds as this representation is passed through the various brain regions [321]. There is no complete model for any of these speech or speaker recognition tasks and thus no clear mapping from the existing biological speaker recognition solution into the machine domain.

“A much greater understanding of the human speech process is required before automatic ... speaker recognition systems can approach human performance” [152]. This viewpoint was suggested in 2005 however since then, as we shall see in Section 2.8, the development of state of the art automatic systems has followed a path with little analogue to the human process. It is well acknowledged that automatic systems are much better able to recognise speakers [100, 355], even from small familial groups where the target voices have been known to human listeners for years [206] and human assistance to automatic systems has even been found to lower their performance [192]. Humans however are better able to recognise speech in general [27] and both speech and speakers in the presence of increasing distortions to the speech signal [27].

Interesting results of potential relevance to the SV task and that demonstrate the very limited understanding science has of these processes include that speaker recogni-

---

There exists some evidence that both temporal poles of the brain are involved in speaker recognition [193] and others in signal processing, speech processing, language production and comprehension such as the connected Broca’s and Wernicke’s areas.

Even fundamental pre-processing methods such as analysis of fixed length short segments of speech is in contrast to the results of studies regarding humans perception of speech [427].

In a closed identification task using samples of 2.5 minutes [184] found that a group of ten listeners identified familiar voices in 98% of cases. However using only the short stimuli ‘hello’ a success rate of 31% was achieved with a similar sized group [233]. Evidence suggests that lay people tend to significantly overestimate the ability of listeners to identify familiar voices [424]. This is also of relevance to the evolving domain of forensic applications of speaker recognition, where for many years forensic speaker recognition experts have performed the analysis of speech evidence by listening to crime scene and suspect samples [273]. This is changing as discussed in Section 6.2.

Other studies have shown that it may be that local feature processing which is uncoupled across frequencies may be responsible for the increased robustness of the human listener to noise and reverberation when performing speech recognition [21], that human voices converge perceptually when shouting [46], that the octave band from 1 kHz to 2 kHz is most useful for human listeners for recognising speakers [100] and that humans use different acoustic parameters to recognise different speakers [234].
tion has been found to be a function of both speaker and listener [396], the existence of otherwise perfectly cognitively functional people with the complete inability to recog-
nise people by their voice (phonagnosia) [394], that there is a difference in the brain activity between speaker recognition and speaker discrimination (as there is with faces) [395], non-agreement about whether distinct areas of the brain are responsible for the processing of speech and musical sounds [173], that language ability is related to speaker recognition ability with dyslexics finding the task more difficult [298] and that the human ear finds it even more difficult a task to tell brothers and sisters unknown to the listener apart [353].

These early results regarding developing an understanding of the human process in recognising identity suggest that it is a much more diffuse problem with various approaches used in different situations and for different speakers. It remains relevant to consider with specific regard to speaker identity many of the twenty fundamental speech concepts outlined in 1995 in [269]. Greater understanding of the human process will likely be of use in either better understanding current systems or in improving their performance particularly with respect to robustness.

2.4 Precursors: Signal Processing, Speech Recognition and Early Attempts

Automatic speaker recognition has been a focus of study since the 1960s and initially developed from the work of several fundamentally related areas. Strongly coupled with the incremental progress of computing power, automatic speech recognition (ASR) and digital signal processing (DSP) became established fields in the 1950s and have introduced several key concepts for processing digital representations of speech that automatic speaker recognition systems have built from.¹⁴

Today ASR systems are significantly more prevalent, being progressively incorporated into almost all modern day computers, smart phones and call centres over the last decade. This is interesting in light of the fact that speaker recognition systems perform

---

¹⁴The progression of speech recognition has been from isolated digit and phoneme recognition systems designed for a single user, such as Bell Laboratories 1952 vowel formant frequency approach [89], towards a relaxation of restrictions enabling speaker-independent, continuous-word and sentence level understanding in unconstrained environments as available in modern products. Techniques introduced allowing this progression included dynamic time warping for improved temporal matching of speech signals [406], the cepstrum [49], delta cepstrum [151] and the application of the hidden Markov model (HMM) for modelling the temporal evolution of known linguistic output, which was first widely recog-
nised in the 1980s [137, 314]. Many of these tools are also prevalent in the modern speaker recognition systems.
better than human counterparts [100, 346], they are significantly worse across all conditions at ASR [27, 243]. This may be seen as a paradox but is almost certainly explained by the relative costs of errors for each task. It may also reflect a greater economic valuation of the ASR technology for the consumer market and the fact that much of the research has been done by major corporations and organisations such as IBM, Bell Laboratories, Microsoft and DARPA. By the end of this chapter we will see that the progress made by the field leads one to be optimistic about this decade seeing a significant uptake of speaker recognition technologies.

As noted automatic speaker recognition owes much to the early investigation of speech carried out by ASR researchers, and its development as an active field of academia began with restricted efforts in the 1960s, around a decade after the earliest ASR work.\textsuperscript{15} Since these early efforts automatic speaker recognition has made a gradual progression towards relaxing restrictions [223], enabling the technology to move from the laboratory to where we are beginning to see real world usage. These changes are summarised in Table 2.1. We begin the review of these insights and adoption of methods that has allowed these relaxations with consideration of feature extraction.

| Small Vocabulary | → Text Independent |
| Small Cohort     | → Large Cohort     |
| Controlled Environment | → Natural Environment |
| Much Training & Testing Data | → Realistic Quantities of Data |
| Same Channel &/or Microphone | → Mixed Transmission Mediums |

Table 2.1: Restrictions imposed upon automatic speaker recognition systems have been relaxed over time as our understanding has improved.

\textsuperscript{15} The first biometric speech systems included Bell Laboratories 1963 method where voice comparisons were based on the correlation of various filter bank spectrograms [307], and Texas Instruments 1971 system based on formant analysis [101], as had previously been exploited by speech recognition. Quantifying and modelling the range of intra-speaker variability was also first considered during this preliminary period. This included how age affects spectrograms of the human voice [123] and other features extracted for speaker identity [150]. Other approaches included comparing spectral magnitude vectors via distance metrics [100] and attempting to determine the acoustic correlates of perceptually relevant features [409], a method quickly realised to be incredibly difficult [83]. A nice review of approaches published in 1976 may be found in [29].
2.5 Representing the Human Voice: Features for Speaker Recognition

All automatic speaker recognition systems must aim to extract the helpful sources of variance that capture identity whilst managing and compensating for the many sources of variance within the speech waveform that hinder the recognition task. The feature extraction process aims to capture in a concise way a representation of the speaker that aids in distinguishing them [29]. The ideal feature would be highly discriminative, independent of age, health or mood, easily and reliably extracted from the speech digitisation and robust to environmental noise, transmission-channel effects and mimicry. Due to the highly dynamic nature of speech, such a feature is utopian and features that display minimal intra-speaker variability in comparison to inter-speaker variation are searched for.

The sources of variation within the speech waveform that help shape a person’s identity originate from several levels due to the mixed anatomical and behavioural nature of the speech biometric. Features that relate to a person’s anatomy are acknowledged as low-level, while high-level refers to the more pliable behavioural and learnt aspects. Typical examples of features from each level are shown in Table 2.2.

<table>
<thead>
<tr>
<th>Levels of Features for Speaker Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>High</td>
</tr>
</tbody>
</table>

Table 2.2: Features for speaker recognition are typically grouped with reference to the primary explanatory factors: from anatomical (low-level) to behavioural (high-level).

2.5.1 Low-Level Anatomical Features

Low-level features are typically text-independent, easier to extract and require less data to model however they are also more easily corrupted by noise, transmission channels and other intersession variabilities [64]. They are typically extracted from analysis of short speech segments of the order of 10ms [418] termed frames\(^\text{16}\) where the speech signal can be approximated as a quasi stationary waveform (for voiced speech) [315]. The greatest

\(^{16}\)Frame size and its implicit resolution of frequency content is important for all recognition tasks [405]. Mid-level phonetic information [203] and syllable boundaries [48] have also been used to demarcate frames for improved low-level features.
levels of speaker discriminative information have been found in these low-level features which principally relate to the speakers anatomy, such as their vocal tract dimensions which principally shape the frequency content within a speech waveform, and hence are implicit within its spectrum. Deriving from the spectrum, and almost definitely the most important feature know to all speech processing fields are the mel-frequency\(^1\) cepstral coefficients (MFCC).\(^2\) Today the MFCC is acknowledged as the reference feature for speaker recognition, and is endemic in all speech processing fields, including mobile telephony, where there are published standards on the variables involved in the MFCC extraction process, such as from the European Telecommunications Standards Institute [126]. This is important as the choice of variables such as the number of mel-filters across the bandwidth and their spacings have been shown to strongly affect the EER in verification experiments [154].

The use of first and second order derivatives of cepstral coefficients, known respectively as deltas or double deltas, has been shown to improve speaker recognition performance [364] being only weakly correlated to the cepstral coefficient information alone and more robust to channel effects [364]. This requires sufficient training data to overcome the dimensionality increase in the feature vector. The effect is greatest in text-dependent conditions where the deltas are able to encompass some of the temporal patterns in reproducing fixed phrases or words. Indeed double deltas have been sometimes found to be ineffectual otherwise [104].

Several other low level features have also been investigated including Linear predic-

\(^{1}\) The non-linear perception of the human ear is commonly modelled by the mel frequency-scale, proposed in 1937 (mel from melody, as it relates to the comparison of pitch) [366]. Whilst the human ear can typically perceive sounds from 20 Hz to 20 kHz, it has a much greater sensitivity over the first few thousand Hz. Indeed this first few kHz of the spectrum (0-4 kHz) contains the majority of speech related information, and it may be that evolution has optimised the ear and brain in this respect. Other perceptual scales exist, but are less common, such as the related Bark scale [430] based rather on the arctan of frequencies over the twenty four critical bands of hearing (the mel scale uses base 10 logarithms).

\(^2\) The cepstrum (spectrum), was introduced in 1963 in a highly significant and linguistically creative paper [49]. The power cepstrum, the most common cepstrum in speech research, is obtained from an inverse Fourier transform of the log magnitude spectrum of the signal. Applied to speech signals, the cepstrum performs a homomorphic analysis, converting the convolved source and filter signals into sums of their cepstra. The cepstra are the magnitudes in the transformed frequency domain of the cepstrum, which is termed the quefrency domain. The distribution of individual cepstral coefficients is typically uni modal and approximately normal [387]. The cepstrum performs a de-correlation of the spectral coefficients, akin to a principal component analysis of the spectrum, which enables the frequency content of a signal to be concisely represented. The cepstrum, without any perceptual warping, was first used in the early 1980s for speech recognition [151], and speaker recognition [189, 263]. The first use of the combination of the mel-frequency scale and the cepstrum for speech processing was published in 1980 [90]. Linear cepstral features have been shown to outperform mel warped features in certain speaker recognition conditions [192]. The number of cepstra retained is important to avoid effectively low-pass filtering the signal or overfitting the spectrum [322].
tion (LP) parameters [130, 253] which also strongly relate to the vocal tract (also used for low bandwidth encoding [44] and speech recognition [196]). Early discriminative modelling of these parameters was investigated in a small speaker identification task [341]. LP parameters however have poor interpolation properties which renders them unsuitable for use in statistical modelling such that when used they are equivalently restated as line-spectral pairs, area ratios or reflection coefficients [295]. Typically it is the linear filter parameters only that are used and the voice source information in the LP residual is disregarded despite it relating to the action at the glottis. Speaker identifying information from glottal fold vibrations as one of the central research areas of this thesis is reviewed in detail in Chapter 3.

Other low level information present within the short term spectral analysis of the speech signal is the phase which is also typically disregarded, retaining only the magnitudes from any Fourier decomposition [421]. This is generally due to the fragility of the phase spectrum [294] however this signal has also been found to contain speaker dependent information in several studies [159, 293, 353, 415]. Formant locations and bandwidths had initially been examined in isolation [101, 422] although this information is contained within the spectral envelope.

2.5.2 Mid-Level Prosodic Features

Speakers have been found to differ sufficiently in the patterns of stress and intonation during communication, so called prosody, making these characteristics potentially suitable as speaker recognition features [299, 226]. Prosodic information falls within the middle of the low-high labelling, containing the greatest mixture of behavioural and anatomical sources. While our lung capacity and diaphragm size strongly determines the energy typical of ones speech, this can be adjusted by choice, shaped by personality, or manipulated for effect. Such control is also present over our pitch, despite being described on average by of vocal fold physiology. Prosodic features representative of pitch, energy (loudness) and speaking rate have all been shown to contain information relating to a speakers identity [312]. Using the NIST 2006 speaker recognition evaluation (SRE) data, containing nearly 600 speakers, polynomial approximations to F0 contours of syllables

\[^{19}\text{Cepstral parameters may also be obtained from the lp estimated spectral envelope [253] or from a perceptual linear prediction (PLP) [181] however results have been presented for the claim that only small differences in recognition accuracies result from these and other methods of obtaining cepstral coefficients [322].}\]

\[^{20}\text{Like humans in this case who show very poor abilities at detecting the phase of audio signals [22].}\]

\[^{21}\text{Implicitly, further evidence that speaker related information is affected by prosody is demonstrated in the decreased ability to recognise speakers under stress, which manifests itself typically with increased speaking rate and elevated pitch [280].}\]
sectioned based on variables such vowel onset time, energy valleys and phone boundaries were all able to achieve alone typically around 12% EERs [226]. The mean, duration and variance of F0 have also been informative for speaker role recognition [40], which may be perceived as a high level trait itself. Empirically examined in [299] were 19 prosodic features. Pitch was found to be most informative however its inclusion with speakers having a large F0 range has been found to reduce recognition rates [429]. Attempts to incorporate prosodic information into current factor analysis modelling approaches, complementing MFCC features,\textsuperscript{22} is beginning to be considered [94, 227].

2.5.3 High-Level Behavioural Features

The learnt and behavioural traits present within speech also contain information pertaining to the speakers identity but these features are generally harder to extract and of a nature that requires different and often more complex modelling approaches than lower level features [326].\textsuperscript{23} Their discrete nature also demands we obtain greater amounts of speech to form representative models for them [299]. They are however significantly more robust to common nuisance variations [72] and have the ability however to provide independent identity information [325, 326]. High-level information considered for speaker recognition includes conversational patterns [299], language and lexical usage [29], idiolectal differences [102] and a combination of high-level traits with a system employing different features for different people inspired by how perceptually people specify the use of different features in classifying their peers [292].

Low-level features provide a baseline from which improvement may be achieved through the use of these mid and high-level sources. We note finally however that all approaches towards combining these multiple information sources within the literature are purely empirical which highlights the significant gap in scientific understanding regarding theoretical models for fusing information streams and quantifiers for indicating which specific approaches are optimal. Some attempts have been made in recent years but with little gain [5, 122] and this remains as a significant gap in scientific understanding.

\textsuperscript{22}The error rates of the best stand alone prosodic feature systems are typically an order of magnitude worse than the current best speaker recognition methods based on low-level spectral information [138]. Several studies have promisingly demonstrated however the complementary nature of such prosodic features with MFCC vectors in other contexts [65, 129, 199, 417].

\textsuperscript{23}Binary trees, neural nets, hidden Markov models and N-grams have been used for example to quantify observed patterns of high-level features. See [65, 325, 326] and references therein for greater details. Language and grammar models are also often required in parallel.
2.6 Variability Compensation and Robustness

Automatic speaker recognition systems must be robust enough to handle real world data if they are to be of significant practical use for society. In real world applications, outside of the controllable conditions of academic research, several extra sources of variation are introduced to the speech signal, all of which contribute towards masking information pertaining to speaker identity or creating a mismatch between training and testing speech. A list of common such sources and effects is given in Table 2.3.

<table>
<thead>
<tr>
<th>Source of Nuisance Variation</th>
<th>Typical Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environmental Noise</td>
<td>Additive Time Domain Blurring</td>
</tr>
<tr>
<td>Transmission Channel Differences</td>
<td>Phantom Formants &amp; Other Spectral Distortions</td>
</tr>
<tr>
<td>Microphone Differences</td>
<td>Spectral Distortions</td>
</tr>
<tr>
<td>Incomplete Phonetic Coverage</td>
<td>Incomplete Estimation of Feature Variance</td>
</tr>
</tbody>
</table>

Table 2.3: Sources of variation commonly responsible for the well understood training-testing data mismatch problem which degrades the performance of real world systems.

The importance of techniques to alleviate these effect has been well understood for a long time. In 1995 by members of the Massachusetts Institute of Technology (MIT) Lincoln Laboratories speech science group presented a GMM-UBM system able to achieve an identification rate of 99% on the TIMIT corpus which dropped significantly to only 60% on the same data transmitted over a telephone line [323, 329]. Several techniques at the feature, model and score levels have been proposed within the research literature for compensating for the various sources of acoustic smearing [420].

**Feature-level** methods describe the procedural manipulation of the speech feature vectors before they are modelled or scored against speaker models. The zeroth cepstral coefficient is generally disregarded also as a form of normalisation by ignoring the base level energy of the signal [315]. Weighting certain cepstral coefficients, known as liftering [49], can alleviate the sensitivity of the higher order coefficients to noise and the lower order coefficients to the overall spectral slope, thereby improving robustness [315]. These methods are important as common low level spectral features are very non robust to typical distortions [34]. It was concluded in [322] that compensation for training/testing acoustic differences had considerably larger affect than variations in methods of determining cepstral features. Attempts to reduce environmental noise have included the use of Wiener and Kalman filtering [38] and autocorrelation methods [221] which assumed that noise could be treated as a stationary signal. Studies with a more realistic experimental treatment of additive noise (albeit with current less than state of the art
modelling) demonstrated the significance of this problem [320, 340] where recognition accuracies where observed to differ by a whole order of magnitude with the inclusion of background noise at various signal to noise ratios.

Common feature level robustness measures include cepstral mean subtraction (CMS), relative spectral filtering (RASTA), and spectral subtraction. A common technique for spectral features is to map their distribution to a standard distribution or learnt set of parameters [297]. This feature warping approach increases robustness to channel mismatch, additive noise and slightly to non linear handset transducer effects. Mapping the distribution of cepstral vectors to a Gaussian density was shown to improve recognition rates compared to CMS and CMS with variance normalisation (where the second moment of the cepstral vectors is also normalised, unlike just the first in standard CMS) on the NIST 1999 SRE data [297]. An affine feature transform is also shown to reduce channel and noise induced effects affecting the recognition task [249].

Cepstral coefficients derived from weighted linear prediction (first applied in speech recognition) rather than the more common Fourier transform have been shown empirically to have a greater robustness to sources of additive noise [340]. A comparative evaluation of several feature normalisation methods is presented in [9].

One of the virtues of attempting to compensate at the score level is that it can be applied to any system, independent of feature and modelling choice. Score-level robustness measures involve a transformation process being applied to the raw scores coming from the classification system and these are standard even in factor analysis [216], i-vector [351] or support-vectors machine [60] systems despite all of which having

---

24CMS, one of earliest proposed compensation measures, uses the homomorphic property that cepstra of time domain convolutions are additive in the quefrency domain, allowing any consist channel induced distortions to be partially eliminated from the speech signal via subtraction of the signals long temporal term cepstral mean vector from each feature vector [29, 151]. Empirical studies focusing on the benefit of CMS alone have demonstrated its ability to lower EERs by as much as one third [354]. Implicit in the CMS method is the assumption that the long term cepstral mean of the speech signal alone (without the channel distortions) is zero, so that the mean taken over the whole utterance represents only the channel convolution effect. This results in significantly worse performance on recognition tasks when no channel or other convolutional effects induce differences between the training and testing data [249]. Further methods are needed to negate non-linear traducer effects introduced by various telephony handsets [323]. Similar methods were proposed earlier in the literature that didn’t make use of the homomorphic transform but were able to cleverly obtain an estimate of the spectral noise present within a speech signal by determining the typical spectrum present in the non-speech segments of the waveform. This then forms a representation of the constant background noise, that can be removed from the speech signal by spectral subtraction [51].

25Relative spectral (RASTA) processing techniques [182] are based on the capitalising on the differing spectral qualities of speech and noise arising from the differences in the typical rates of change between sources of noise and the vocal tract. By considering the rates of spectral change to be expected from the vocal tract motions alone, the slower and faster changing elements of noise on either side of this are able to be removed.
modelling methods designed to compensate for the nuisance variations that partially introduce the requirement for normalisation.\(^{26}\)

A score-level method based on a learnt handset dependent mapping that addresses the non-linear changes made to the spectral content of the speech waveform is proposed in [310]. Indeed microphone transducers can introduce spurious resonances at sums, differences and multiples of the true formants leading to so called ‘phantom formants’, as well as widening or flattening the spectrum, all of which generally result in reduced recognition performance. This method however requires prior knowledge of transmission handsets. See also the handset normalisation (\(h\)-norm) approach proposed in [324].

Most score normalisation methods however perform a mapping of scores to a predetermined distribution [170]. Typically these are all based on learning the parameters of a Gaussian distribution on different data and performing a standard normal mapping to obtain a Gaussian \(z\) score. Such methods include \(z\)-norm, \(t\)-norm [31], \(c\)-norm and \(d\)-norm [41]. These methods are discussed in more detail in Section 4.3 where they are contrasted with the regression based score modelling approach introduced in Chapter 4 termed \(r\)-norm.

The affect of feature vector (CMS, feature warping) and score normalisation (\(t\)-norm, \(z\)-norm, cohort method) techniques on the NIST 2002 mobile phone data are examined in [34], where it was found that feature warping and \(t\)-norm in combination gave the greatest increase in performance. It is evident that these methods do not completely account for the distortions introduced by the mobile telephony channels however as the system still falls short of the results obtained on the control landline telephony data.

Whilst feature and score level methods have been proposed in greatest numbers, recent work has also focused on model-level methods, particularly attempting to isolate speaker variability from channel variations. In an early model compensation approach affine and fixed target transformations trained on the variation present within databases containing non-variable recording conditions were able to uniformly improve performance for each cohort testing size [276]. Another study demonstrated that channel compensation in the model and feature domains are not exclusive and that each may remove separate channel effects and several procedures in combination may be beneficial [390]. However several channel compensation techniques are applied in sequence in [62] concluding that the belief that the “more boxes in the scheme, the better the system” is false. Model level attempts are addressed in greater detail in the state of the art modelling discussion of Section 2.8.

\(^{26}\)E.g. \(z\)-norm followed by \(t\)-norm has been found useful in factor analysis modelling [407, 408].
2.7 Statistically Modelling the Human Voice

Creating a concise description of the patterns present within a speaker’s collection of training feature vectors is required in order to quantify the similarity of uttered test speech against a speaker model. The earliest attempts were generally discriminative methods based on codebook and template models [41, 100] however statistical quantification of measured speech features quickly improved upon these [223].

In the Bayesian statistical paradigm the likelihood ratio is the optimal test for deciding between two competing hypotheses [42]. Applying this to the speaker verification problem, the question then becomes how to model the likelihood functions of the same and alternative speaker hypotheses. The first considerably successful paradigm for modelling speech features (again typically MFCC) came from the use of generative Gaussian Mixture Models (GMM) [328] which unlike earlier discriminative template models possessed the flexibility to suitably describe the variation of speech features and at least partially handle noise and other corruptions [328].

GMM are used to model both hypotheses, with the alternative hypothesis model referred to as an Universal Background Model (UBM) [327], which quantifies the variation of the speech features over the population. The UBM is fitted to background speaker training data via the expectation-maximisation (EM) algorithm [99] whilst greater recognition accuracy is obtained when client speaker models are obtained from the UBM via a maximum a-posteriori (MAP) Bayesian adaptation [158, 324]. Other methods of adapting client models from the UBM [97, 202, 244, 356] or forming the UBM [324] have been proposed and their effect on recognition accuracy empirically studied [125, 178, 251]. Studies examining the requisite amount of training data for the UBM in order to form an accurate understanding of the variation of speech features have observed that recognition rates stabilised once an hour of background speaker training speech was reached [178] although commonly NIST SRE [281] submissions employ several databases running to multiple hours of speech data [60, 216].

---

Optimal when the likelihood functions are known exactly.

28 With a suitable number of mixtures a GMM may be fitted to an arbitrary distribution [42], and increasing the number of mixtures can account for correlations allowing the use of diagonal covariance matrices [327]. Applied to speaker recognition, an approximate physical interpretation is that the individual multivariate Gaussians mixtures model broad individual acoustic properties which are correlated to unique vocal tract configurations. No temporal properties are captured by GMM, however these may be discarded without loss of accuracy in the text-independent task.

29 Many terms exist for the concept of a universal background model. These include population model, background cohort model, speaker independent model and world model.

30 Regarding efficiently capturing the population variation with minimal data decimation and random feature selection were explored in [35] whilst an informed selection of features for modelling based on a
Other attempts at quantitatively capturing speaker identity prior and up to the GMM-UBM approach included dynamic time warping and template matching approaches derived from speech recognition [64], codebook and vector quantisation approaches [87], k-means methods [41], polynomial classifiers [66], fractional Brownian motion modelling of a Hurst parameter [343], metric scoring on a low dimensional manifold found via graph embedding which aims to focus on the dimensions of the typically large GMM supervector spaces that constrain the greatest variability [207], creating a ‘structured score vector’ via a Fisher discriminant that describes efficiently how components of a GMM describe observed data [238] and finally using the GMM-UBM model to derive a speaker specific ‘binary key model’ which is scored via a logical AND operator [26] in a process loosely analogous to the state of the art iris recognition method [88].

2.8 State of the Art Speaker Recognition Systems

Advances in the performance of text independent automatic speaker recognition systems over the last 10 years have come from improved modelling techniques which successfully manage nuisance variations and background variability. Over this period feature selection has not varied as demonstrated by examining the systems submitted in recent years NIST Speaker Recognition Evaluations [281] where magnitude spectral information remain almost uniformly employed. The results of these evaluations demonstrate that state-of-the-art systems are approaching the 1% EER on challenging data collected from a variety of real world situations [59, 216, 287].

The central precept of this new speaker modelling baseline builds of the previously optimal GMM-UBM framework, doing so by representing speakers utterances by a supervector derived from the concatenation of the mixture means of a GMM trained on it. The dimensionality of such supervectors, typically of the order of $\sim 1000 \times 20$, meant that discriminative approaches for classification were initially examined such as Support Vector Machines\(^{31}\) (SVM) [69]. Another motivation is that whilst GMM enables accurate probabilistic modelling of the intricate distribution of speech features it does not focus on the central issue of discriminating speakers [91].

\(^{31}\)Claiming that the binary nature of SVM was well suited to the speaker verification task, SVM had also been applied to model the scores output by another kernel machine classifier [414] in a score
SVM can also be applied directly to the lower dimensional acoustic features [92] however the focus then still becomes finding channel independent representations and compensation. These issues are reduced by using the more speaker dependent supervectors but these still require the alleviation of problem perturbations as attempted to be provided by the eigenvalue Nuisance Attribute Projection (NAP) [360, 361] method for example which looks to remove supervector dimensions most associated with uninformative signals via projections to lower subspaces [70].

The issues of dimensionality of supervectors, compensating for nuisance variations and model training when presented with minimal speech have all been significant motivations in the development of generative supervector methods which have lead to the current state-of-the-art approaches. The central element of these methods is to consider the supervectors to be generated by lower dimensional latent variables which stipulate the supervectors position in a much smaller subspace. This approach is “forced upon us” [213] given the practical impossibility with limited data of estimating a supervector covariance matrix of sufficient rank to be useful. A supervector $S$ may typically be represented as:

$$S = m + Vy + Ux + Dz$$

being decomposed into a speaker and utterance independent mean $m$ and an identity dependent ($Vy$), channel/utterance dependent ($Ux$) and residual ($Dz$) components. This is referred to as the Joint Factor Analysis (JFA) model [209]. The factors $y, x$ and $z$ are assumed to have standard normal prior distributions, such that $S$ is multivariate Gaussian, and the speaker factors $y$ are assumed to be depend only on the speaker whilst the channel factors $y$ (more accurately session factors) vary with each utterance [209].

Rather than simply negating channel and nuisance effects this approach attempts to model these [62, 407] and separate variations out into speaker identity and channel dependent sources located within the subspaces defined by $V$ and $U$ respectively [217]. These hyperparameters are estimated via maximum likelihood methods with a modelling approach akin to the proposed regression technique r-norm outlined in Chapter 4. Another earlier and interesting use of SVM in the score domain had been to learn a decision function for analysing LLR rather than thresholding [37], motivated by the claim that the Bayes decision rule is not optimal in most practical cases due to imperfect likelihood estimates.

 Attempts to derive a full posterior distribution over the supervector rather than using a point estimate by latent variable methods were investigated in [217]. Given even ten seconds of speech the posterior distribution was found to be sharply peaked so that integrating over the speaker factors will give much the same result as the modal point estimate.

This unimodal distribution is shown to be superior to multimodal formulations in [214].
minimum divergence procedure designed to avoid saddle points during the optimisation process [209]. Generally the unions of large databases are used to train the JFA model hyperparameters as it is required to capture all sources of variation anticipated to be contained within testing data [215]. Typically it is observed that the eigenvalues of the \( V \) and \( U \) matrices trail off exponentially and \( D \) is small which provides some empirical justification for the supervectors to be reasonably constrained within the low dimensional subspaces.

The hyperparameters \( V \) and \( U \) provide a prior distribution on the GMM supervector \( S \), acting as important constraints when estimating the GMM model with minimal speech data [209]. This is a significant difference to the GMM-UBM approach. The JFA model avoids the non ideal adaptation behaviour of classical or relevance MAP [158, 327] which becomes asymptotic to maximum likelihood training with sufficient data [219] whereby GMM speaker models adapt to not only speaker specific characteristics but to any other signal present within the feature data. The model of (2.1) instead makes possible the combination of priors of classical MAP [158] with eigenvoice MAP [211] and eigenchannel MAP [218] into a single prior on the speaker and channel dependent supervector components.\(^{34}\) All of these methods are based on the idea using a well constructed prior distribution to overcome the fact that there is generally insufficient training data for the enrolling speaker in order to perform speaker-dependent training.\(^{35}\) Still, limited enrolment data can lead to poor estimates of the latent variables \( y, z \) and \( x \) [211].

Several approaches exist for forming the likelihood function for comparing utterances with these factor analysis based models which trade off accuracy, computational complexity and speed [160, 213].

\(^{34}\)Important research preceding the JFA model had explored eigenvector methods of finding speaker or channel dependent subspaces separately. PCA had been used to derive a set of vectors called eigenvoices and only eigenvoice coefficients were estimated during speaker enrolment, thereby constraining the GMM model representations to lower dimensional subspaces [231]. This method allowed rapid adaptation of speaker models and was especially useful with limited training data, with subsequent speaker recognition experiments with eigenvoice adapted models outperforming relevance MAP estimated models under this condition [380]. A modification of eigenvoice MAP generated a new method termed eigenchannel MAP which resulted in significant increases in identification accuracy in a speaker recognition experiment where training and testing speech were transmitted via differing channels [218]. Eigenchannel modelling assumes that the difference between the speaker and channel dependent supervectors differ only by a vector which represents the channel effects alone (the same assumption as speaker model synthesis [375]).

\(^{35}\)By ‘speaker dependent training’ it is meant that the GMM parameters are determined by Expectation-Maximisation or Maximum-Likelihood type methods using only the speakers training data and without adaptation from any prior model. Other methods have also been explored for overcoming this small data problem such as cluster adaptive training [153] and building upon the Extended MAP idea proposed in [428] by making use of the correlations between speaker pairs data [212]. This last method demonstrates that the common assumption across nearly all speech and speaker recognition tasks of the statistical independence of speakers is sub-optimal, something that is also exploited by the score post-processing method proposed in Chapter 4.
Despite the success in terms of improving recognition accuracy in trying to separate out channel effects from speaker information it was found that the channel factors \( y \) actually contained speaker dependent information [95]. This led to a reformation of the latent variable modelling of GMM supervectors as:

\[
S = m + Tv + \epsilon
\]  

(2.2)

where rather than training separate speaker and channel spaces all training data is pooled and a Total Variability subspace \( T \) is trained by similar maximum likelihood methods. The coefficients of a speakers supervector represented within this space \( v \) is termed an identity vector or i-vector and may be found as the mode of the posterior distribution of \( v \) given the training data where it is assumed to have a prior standard normal distribution. Typically \( v \) is set to have \( \sim 400 \) dimensions. This approach is currently the state-of-the-art in text independent automatic speaker recognition. It allows for very fast scoring methods based on cosine distances between i-vectors that are much faster than evaluating JFA likelihood expressions [93].

Compensation for channel and intersession variation is carried out in the i-vector space rather than the considerably larger GMM supervector space as is done in JFA where a channel supervector is estimated and subtracted. Further normalisation approaches such as LDA, NAP and Within-Class Covariance Normalisation [179] which creates a mapping that preserves directions in space, unlike LDA and NAP, scaling the i-vectors to attenuate high within class variability [9]. These compensation techniques for i-vectors are all borrowed from earlier SVM approaches.

The distribution of i-vectors is typically non-Gaussian and methods of replacing the standard normal prior distribution with a students t-distribution are being investigated which allows for better handling of outlying data and more severe channel distortions [210, 352]. A form of length normalisation of i-vectors has also been proposed to reduce this behaviour [155]. Further modelling of estimated i-vectors by probabilistic linear discriminant analysis (PLDA) [305] is also being investigated to further mitigate nuisance variations [256].

Almost solely these models have been applied to cepstral features which have additive channel effects (assuming they were introduced via convolution with the speech signal). The effectiveness of the additive channel compensation approach of these methods remains to be explored with non-cepstral features where channel effects are not additive. An overview of the evolution of text-independent speaker recognition is presented in [36] and [223].
2.9 Cautions and the Future of Speaker Recognition

One of the primary considerations in implementing practical systems is dealing with the simultaneously improving technological threats of voice transformation and synthesised speech. Systems which do not account for temporal information, as most NIST SRE state of the art systems do not, have been found to be most vulnerable to such attacks along with simpler playback type attacks [201].

Countermeasures for such activity are beginning to be investigated [10]. The inclusion of multiple levels of feature information is also aimed at increasing the practical strength of recognition systems [223]. Research efforts are also beginning to be diverted to improving text-dependent speaker recognition using state of the art text-independent techniques [365].

Another issue is the public perception and media portrayal of biometric technologies and speaker recognition in general [52] and it is important where possible to communicate the realities of the speaker recognition performance by machines! Despite the thematic depiction of miraculous feats of identification, in fact speech recognition remains the only prevalent speech based technology within society. The increasing robustness and accuracy of speaker recognition systems is likely approaching the level where its practical use begins to square with the costs of any potential errors it can make.

Prediction of the future is always precarious of course and it is interesting to note the tone of several review papers published prior to the paradigm changing breakthroughs of speaker supervectors and factor analysis modelling in that all recommended striving to incorporate additional information from extra features rather than improving speaker (feature) modelling techniques [100, 135, 326].

From the 1960s researchers could not see the introduction of speaker recognition technology given the tremendous costs and volume of signal processing machinery required [100]! By the 1980s their view had changed: “It may be that voice verification will never catch on as a premium method of identity validation. However, there are indications, as our society edges forward into the abyss of total automation and computerization, that the need for personal identity validation will become more acute” [100]

Nearly 30 years later, with the scientific understanding and necessary technology at our finger tips coupled with the greater and increasing demand for verifying yourself, especially from a distance, the dawn of the emergence of speaker recognition technology within society must be near!
Chapter 3

The Glottal Waveform

3.1 Introduction

In this chapter we provide a description of the glottal waveform, the signal processing related to it and its use as a source of information inferring speaker identity. A novel feature termed a source-frame is also introduced.

We begin by describing in section 3.2 the human physiology used during the speech production process, introducing the vocal folds and their associated glottal airflow waveforms that result from their vibratory motion. A review of the scientific literature regarding methods for estimating the glottal waveform from the digitally sampled speech signal is presented in section 3.3 which describes the source-filter theory of speech production, linear prediction of speech and inverse filtering.

In section 3.4 we review the various temporal and spectral parameters calculated from glottal waveform estimates for representing them concisely. In section 3.5 we introduce a normalised temporal glottal flow waveform feature termed a source-frame that enables visualisation of inter speaker differences and is employed frequently in the speaker recognition experiments reported in Chapter 5.

Lastly in section 3.6 we present a literature review regarding the use of the glottal waveform for the task of automatic speaker recognition.

3.2 Speech Production and the Glottal Waveform

We describe in 3.2.1 the physiology of the vocal tract, larynx and vocal folds before discussing their manipulation during the speech production process in 3.2.2. Phonation and the important role played by the vocal folds is emphasised. The section ends with an
overview of invasive methods employed for observing or inferring the dynamic motions of the vocal folds presented in 3.2.3.

3.2.1 Physiology of the Human Larynx and the Vocal Folds

The larynx is an organ located anteriorly at the top of the trachea and commonly referred to informally as the ‘voice box’. In infants the larynx is located at the level of the C2-C3 vertebrae but recedes and descends during growth to adulthood to the level of the C3-C6 vertebrae [74]. A schematic of the human larynx is shown in the left of Figure 3.1.

![Figure 3.1: An anatomy of the human vocal tract showing the larynx, articulators and resonance cavities as well as the location of the true and false vocal folds. Shown on the right is a traverse planar view of the vocal folds as might be obtained by a nasal or oral laryngoscopic camera.](http://www.nathanclarkecommunication/Anatomy-of-the-larynx.jpg)

Humans, like reptiles, amphibians and other mammals, use the larynx to control the production of sound. The larynx also performs important tasks related to both breathing and the prevention of pulmonary aspiration (choking caused by foreign objects entering the trachea).

Within the larynx are nine cartilages, three paired and three unpaired. The paired arytenoid cartilages are the most important with respect to speech because of their influence on the position and tension of the vocal folds which are the two triangular folds of ligament consisting primarily of hyaline cartilage plus other tissues that are stretched horizontally across the larynx extending from the thyroid cartilage in the front to the

---

1 Image amended from [http://www.nathanclarkecommunication/Anatomy-of-the-larynx.jpg](http://www.nathanclarkecommunication/Anatomy-of-the-larynx.jpg)
2 In contrast avian sounds are produced by the resonances of a bony structure located at the bottom of the birds trachea. This analogous vocal organ is called the syrinx.
arytenoid cartilages at the back [74]. The place where the two vocal folds\footnote{The vocal folds are also referred to as vocal cords and less commonly as vocal reeds due to their vibratory behaviour analogous to the thin strip of material in woodwind instruments, harmonicas or accordions.} meet and the gap between them during abduction is termed the glottis. The arytenoid cartilages are responsible for controlling the degree of glottal opening through which air flows during breathing and the generation of speech.

Human vocal folds are pearly white due to minimal blood flow and consist of three layers: the epithelium, the vocal ligament and the vocalis muscle. They are shown from above in the right of Figure 3.1. The folds are protected from above by the epiglottis, which prevents food irritating or lodging in the folds and from entering the trachea. The vocal folds are also referred to as the true folds to differentiate them from the more durable false folds. These pair of thick mucous membranes are located directly superior to the true folds and also play a protective role. No muscle is contained within these false vocal folds in contrast to the presence of skeletal muscle in the true folds [74]. The false folds are also called vestibular or ventricular folds and contrary to the true folds play a very minor resonance shaping role in the production of modal speech,\footnote{With certain exceptions such as the Tuvan style of throat singing, Tibetan Chant and the false cord screaming techniques common to modern styles of metal music, all of which make use of the false vocal folds to create an undertone.} a process described in Section 3.2.2. In evolutionary terms the true folds have been adapted for speech whilst the false folds have retrogressed from their initial role as air-trapping valves for respiration [242]. Subsequently all further discussion pertains exclusively to the true vocal folds.

The muscles of the larynx are divided into intrinsic and extrinsic muscles with the intrinsic muscles further divided into the respiratory and the phonatory muscles (the muscles of speech production). The respiratory and phonatory muscles abduct and adduct the vocal cords respectively and are controlled by the recurrent laryngeal nerve, a branch of the vagus nerve pair [74]. These are responsible for controlling the vocalis muscle which regulates the glottal opening by forcing the anterior portion of the vocal fold ligament located near the thyroid cartilage to contract.

The differences in laryngeal sizes with gender and age, along with genetic variations, results in different vocal fold densities and sizes. Measured along the anterior to posterior line of the body, male vocal folds are generally within 17.5-25 mm long and females 12.5-17.5 mm long, while their thickness varies between 3-5 mm [386]. At birth boys’ and girls’ vocal folds are both typically 2 mm long. Girls’ folds however grow at 0.4 mm per year and boys’ at 0.7 mm per year [177]. Testosterone also lengthens the cartilage of the
larynx and thickens the folds of boys during puberty, changing the timbre of the boys’ voice. These growth differences result in adult males, with the longer, larger and heavier folds, having an average pitch of 110 Hz and females 200 Hz compared to children who typically have a higher 300 Hz pitch [421].

The other key physical components of the human speech production system that operate with the larynx are the lungs, diaphragm, trachea, laryngopharynx, velum, oropharynx, nasopharynx, tongue, hard palate, teeth, cheeks and lips.5 Taken together, the laryngeal, oral and nasal cavities are referred to as the pharynx. The tubular segment from the lips through the oral cavity and down to the vocal folds is called the vocal tract with an average length in adult male humans of 16.9 cm and 14.1 cm in adult females [161]. In the context of speech production, the movable and flexible elements of this system (velum, tongue, cheeks, lips) are referred to as the articulators and are placed into certain configurations in order to articulate specific sounds. We now describe how this physiology is used to generate sounds and purposefully manipulated to produce speech.

3.2.2 Phonation and the Speech Production Process

Sound is a longitudinal pressure wave generated by an energy source and transmitted through some compressible medium, typically air.6 The speech pressure wave is generated from the energy provided by the lungs and diaphragm to drive air into the larynx, through the vocal folds and on beyond the oral and nasal cavities and past the lips. The term phonation is used to describe any vibration within the larynx that modifies this airflow and subsequently influences the produced sound or speech. Voiced phonation is the phenomena of vibration of the vocal folds that produces a quasi periodic airflow that acts as an energy source to the vocal tract for speech production. During normal phonation the entire length of the vocal folds vibrate, snapping open and closed. This is also referred to as modal phonation or modal voice. A glottal space of 3 mm and a minimum airflow is sufficient to initiate phonation [391]. Feedback from the vocal tract has little influence on the vibrations of the vocal folds [130].

The vocal fold vibration cycle that produces this excitation signal/energy is described in Figure 3.2. This waveform is also called the mucosal wave, as when the vocal folds

5See [421] for a functional description of these regarding speech production.
6Humans detect sound with their ears via the pressure wave entering the ear canal and vibrating the ear drum. The basilar membrane on cochlea in the inner ear translates these motions into an electrical signal that is then sent to and interpreted by the brain. The basilar membrane is narrow and rigid at the basal end (stapes bone location) but malleable and broad at the apex and as a result the incident sound wave energy peaks at different locations allowing the membrane to perform a filterbank style decomposition [348]. Psychoacoustics tells us that like other sensory systems (smell, vision) the auditory response increases logarithmically with the intensity of the stimulus [142].
oscillate, the superficial tissue of the vocal fold is displaced in a wave-like fashion. The production of this quasi periodic vibratory motion of the vocal folds is described by the myoelastic-aerodynamic theory of phonation [392]. The theory describes the following processes, where references are made to the numbered stages depicted in Figure 3.2:

- To begin the vocal folds are abducted and the glottis is open. The diaphragm is then lowered and the chest cage expanded drawing air into the lungs.
- The glottis is closed and the laryngeal muscles tensed to produce the desired pitch.
- Air is forced up through the trachea by muscular contraction of the lungs.
- The airflow slows under the vocal folds due to the glottal closure. The subglottal pressure increases (Step 1).
- At some point the airflow generating the increased subglottal pressure is enough to overcome the muscular force holding the vocal folds together and they are blown apart (Step 3). The resultant glottal opening allows air to flow into the pharynx.

An earlier theory proposed in 1950 called the Neurochronaxic Theory which was based on a neuromuscular hypothesis that stated that the frequency of the vocal fold vibration was determined by the recurrent nerve, not by air pressure or muscular tension [416]. Specifically it was believed that every single vocal fold vibration was the result of a firing of the laryngeal nerves controlled by an acoustic centre in the brain. The theory was disproved as the muscles were shown to be incapable of contracting at the observed vibration rates along with the fact that people with paralysed vocal folds are able to produce voiced phonation.

Image amended from http://www.phy.duke.edu/~dtl/136126/restrict/Voice/foldwave.gif

---

7 An earlier theory proposed in 1950 called the Neurochronaxic Theory which was based on a neuromuscular hypothesis that stated that the frequency of the vocal fold vibration was determined by the recurrent nerve, not by air pressure or muscular tension [416]. Specifically it was believed that every single vocal fold vibration was the result of a firing of the laryngeal nerves controlled by an acoustic centre in the brain. The theory was disproved as the muscles were shown to be incapable of contracting at the observed vibration rates along with the fact that people with paralysed vocal folds are able to produce voiced phonation.

8 Image amended from http://www.phy.duke.edu/~dtl/136126/restrict/Voice/foldwave.gif
The subglottal pressure is reduced by the release of this puff of air. The natural elasticity of the tissues and muscles of the larynx and vocal folds in collaboration with the Bernoulli effect then pulls the folds rapidly back into place, beginning at the lower edge (Step 5) and resulting in glottal closure (Step 7). The Bernoulli effect occurs when subglottal airflow velocity increases when passing through the constricted glottal opening (the so called Venturi tube effect) which creates a negative pressure immediately below and also between the medial edges of the vocal folds which also provides an adducting force.

The process is then repeated so long as air is continually exhaled, with the pulses of air released through the glottis providing the source of energy for the production of voiced sounds. The cycle is enabled by the symmetry in weight, mass and shape of the vocal folds. Glottal cycles also called pitch periods.

Other phonation types derived from different laryngeal configurations are possible and these result in perceptively distinct speaking styles. Sound produced by voiced phonation is labelled voiced. Conversely sounds produced without any vibration of the vocal folds are labelled as unvoiced or voiceless and typically carry less energy than voiced sounds [132]. If the vocal cords are held apart, air can flow between them without being obstructed, so that no noise is produced by the larynx. These sounds have no associated glottal waveform, as the vocal folds are stationary, and the temporal distribution of the resulting airflow through the glottis is characterised by white noise.

Voice produced by only a partial closure of the vocal folds along their anterior-posterior length during phonation results in the addition of whisper to the vibrations, a phenomena that we perceptually label as breathy voice or murmur [421]. A secondary glottal pulse can be produced from such phonation [309]. In contrast voicing produced by vocal folds stiffened through the tensing of the laryngeal muscles such that the folds are rigid and only a portion vibrates, is perceived as a creaky voice, also called vocal fry or laryngealisation . Tightly closing the glottis is also required to generate glottal stops.

A single cycle of opening and closing the vocal folds takes an average male 10 ms meaning the cycle repeats at an approximate rate of 100 times per second. This rate is called the fundamental frequency (F0) and its perceptual correlate is referred to as ‘pitch’. The cycle rate is dependent upon the stiffness, mass and size of the vocal folds as well as the subglottal air pressure generated by the diaphragm and lungs [383]. Changing certain of these factors allows a speaker to raise or lower the pitch of the voiced sound.

The resulting waveform passing through the glottis during this voiced phonation process is a sequence of air pressure variations that form a quasi periodic flow that
can be measured in cubic centimetres per second [146]. The glottal flow has a coarse structure representing the general flow and a fine structure derived from perturbations in the flow such as from partial fold closure or aspiration. An example typical of this airflow waveform is shown in Figure 3.3 where two glottal cycles are shown.9

![Figure 3.3: Waveform of the airflow through the glottis during voiced phonation.](Image amended from http://www.feilding.net/sfuad/musi3012-01/images/lectures)

This flow produced by the manipulation of the larynx, has a simple decaying spectrum with a clear F0. Energy at integer multiples of F0 called harmonics are contributed to the speech signal by the pharyngeal, oral and nasal cavities. The vocal tract, excited by this glottal energy, also imposes its own resonances onto this harmonic structure of the glottal waveform, with harmonics near or on the natural resonances of the vocal tract instantaneous shape being emphasised. These resonances from the vocal tract are called formants.11 The formant locations can be shifted by adjusting the articulators, which perceptually results in a wide range of possible sounds and is vital to conveying emotional and linguistic meaning.

### 3.2.3 Invasive Measurements of the Glottal Waveform

Glottography describes the recording of measurements of the temporal variation of the glottis during phonation.12 A common proxy for the glottal flow waveform is provided by electroglostography (EGG) which is a semi-invasive method of measuring vocal fold activity whereby electrodes are placed on the outer skin of the throat directly over the vocal folds and measure changes in the electrical resistance across the larynx which is an indicator for the vocal fold contact area. An increasing EGG waveform (resistance) is derived from increasing contact area corresponds to the glottal closed phase. This

---

9We also refer to this signal as the glottal or voice-source waveform, or volume-velocity (V-V) airflow being careful to differentiate between the V-V flow at the glottis with that radiated from the lips. Frequently, as discussed in 3.3, we work with the derivative of this signal.

10Image amended from [http://www.feilding.net/sfuad/musi3012-01/images/lectures](http://www.feilding.net/sfuad/musi3012-01/images/lectures)

11Derived from the Latin verb ‘formeare’ meaning “To shape”.

---

30
relationship has been confirmed through simultaneous laryngoscopy [80]. Different neck electrical resistances can make comparisons between speakers of EGG amplitudes problematic.

In the 1960s and 1970s mechanical inverse filters were produced for the purpose of estimating the V-V flow at the glottis [186]. These filters were designed to have a transfer characteristic the inverse of that of the vocal tract, having a antiresonance (zero) for every resonance (pole). This filter was first applied to the speech pressure wave radiated from the mouth, where a pole at zero frequency was also used to compensate for the high frequency emphasis introduced at the lips. Problems with inverse filtering the pressure wave included the resultant sensitivity to low frequency noise from this zero frequency pole increasing the low frequency response of the filter as well as the fact that the dc level of the glottal flow could not be determined, being lost by the imperfect integrator.

For these reasons, methods of inverse filtering the V-V flow at the mouth were introduced which neatly bypassed these problems introduced by the zero frequency lip radiation compensating pole, such as the wire screen pneumotachograph mask [337].

3.3 Literature Review: Estimating the Glottal Waveform

In this section we review approaches for estimating the glottal flow using digital signal processing methods applied to the digitally sampled speech pressure wave. These methods are less accurate and more vulnerable to environmental distortions (than mechanical, invasive approaches) but are non-invasive and are required tools for any use of the glottal signal for practical speaker recognition.

3.3.1 Source-Filter Theory and Linear Prediction of Speech

Much development and research has lead to the source-filter model of speech production, which is today the most common approximation to the speech production process, used extensively for speech analysis and synthesis. Introduced primarily in the 1960s
the source-filter theory posits that speech is produced from a voice source signal with a decaying harmonic spectrum then modulated by a time-varying linear filter representative of the changing vocal tract. The filter transfer characteristics of the vocal tract can be recreated via physical tube model approximations and advances in practical modelling began with observations that specifically shaped cylindrical tubes joined piecewise together could be used to generate the formants of speech [253].

The source-filter theory of speech production specified by \( z \)-transforms, where \( P(z) \) is the transform of the pressure wave at the lips assumes:

\[
P(z) = G(z)V(z)L(z)
\]

where \( G(z) \), \( V(z) \) and \( L(z) \) are the \( z \)-transforms of the impulse responses of the linear filters representing the action of the glottis, vocal tract and the lips respectively [317].

The source filter model can be represented by grouping these linear filters into a single linear filter \( H(z) = G(z)V(z)L(z) \) excited by an impulse train for voiced speech and \( H(z) = V(z)L(z) \) excited by white noise for unvoiced speech. \( H \) is said to colour the source excitation signal having uniformly flat spectrum (i.e. white). A mixed excitation model can be used for voiced fricatives or breathy voice.

This lead to the application of linear prediction\(^ {\text{16}} \) to speech as the strong interrelationships allowed the acoustic tube model parameters to be estimated directly from the digitised speech waveform and thus to analytically estimate the filter of the vocal tract. Indeed it was shown that the acoustic tube representation can be equivalently represented by the inverse vocal tract filter obtained from linear prediction (LP) of the speech signal [412].

\(^ {\text{15}} \)In the simplest approach the vocal tract is considered as a single tube extending from the vocal folds to the lips. Though the exact shape of the vocal tract is quite complex, many of its most prominent features can be recreated with such simple models. For example the resonances of a 17 cm cylinder closed at one end occur around 500, 1500 and 2500 Hz, which are close to the formant frequencies of the vowel /e/ [421]. Increasing the complexity, the transfer function of a concatenated model with \( N \) tubes possesses \( N/2 \) complex conjugate roots which allows modelling of \( N/2 \) vocal tract resonances [30]. Indeed a unique concatenation of equal length tubes can be found having resonance frequencies matching those of any given order transfer function polynomial [30]. These tube models may be analysed with a system of partial differential equations to describe the sound pressure and V-V flow over time and space or in the frequency domain with \( z \)-transforms to determine the filtering they perform on incident signals.

\(^ {\text{16}} \)In European languages voiced sounds are much more frequent than unvoiced sound with 78% of UK English speech sounds being voiced for example [75]. One result is that linear prediction has been used extensively in many domains of speech science including speech coding [288], speech recognition [196] and speech synthesis [191]. It is a fundamental signal processing technique however used in a wide range or engineering applications.
The coefficients of the vocal tract filter may be found from LP analysis [316], which posits that the speech signal can be recreated from a linear sum of previous samples, predicting \( x[n] \) by:\(^{17}\)

\[
\hat{x}[n] = \sum_{k=1}^{P} a_k x[n - k]
\] (3.2)

The three common approaches to solving this linear system for the predictor coefficients are the autocorrelation, covariance and lattice methods [253, 315] and evolve from different specifications of the interval of speech to minimising the squared prediction error over.\(^ {18}\) The vocal tract acts as a time-varying linear acoustic filter during speech production, changing configuration at the order of \(10^{-2}\) seconds. As such linear prediction is performed on 20 to 30 ms segments of speech called frames.

A frequency domain schematic of this source-filter model is shown in Figure 3.4 which shows the spectrum of the speech signal \( X(n) \) (top left) which is seen to be formed from the combination of the vocal tract filter spectrum \( V(z) \) (top right) with the glottal derivative spectrum \( G(z)L(z) \) (bottom left). The time domain waveform of the glottal derivative is also shown. The formant structure of the vocal tract is clearly imposed upon the glottal source signal.

![Figure 3.4: The source-filter theory of speech production. Amended from [388].](image)

This source-filter model of speech production is suitable for non-nasal speech. To represent nasal sounds\(^ {19}\) a side branch can be added to the single cylinder model which

---

\(^{17}\)This is a \(P^{th}\) order linear predictor as it uses the previous \(P\) samples to estimate the current sample \(x[n]\). Setting \(P = \text{Fs}/1000 + 2\) where \(\text{Fs}\) is the sampling frequency is generally sufficient to capture the formant locations and bandwidths, the most defining characteristics of voiced sounds [333].

\(^{18}\)Autocorrelation LP solves the linear system over all time \([-\infty, +\infty]\), applying a windowing function to the interval of interest. The Levinson-Durbin recursive algorithm provides an efficient solution to this formulation by making use of the resulting symmetric and Toeplitz properties of the autocorrelation matrix [315]. In contrast the covariance solution is formed from analysis of a specific interval. The resulting matrix is not Toeplitz but more efficient solutions than Gaussian elimination still exist. Typically a Cholesky decomposition or square-root method is used to obtain the prediction coefficients [253].
introduces zeros into the transfer function however (modelling the anti-resonances where speech energy is lost), making it less analytically tractable than the all pole model.\textsuperscript{20}

Other considerations in applying the source-filter model include that the vocal tract acts as a linear acoustic system, that the source excitation is independent and does not interact with the vocal tract and to differentiate between acoustic and physiological interactions. Note that these are dissociate assumptions on the speech signal \textsuperscript{337}.

Linearity is a simplifying assumption with many practical benefits. Neglecting energy losses due to absorption at the soft palate, cheeks and velum, the vocal tract is quite linear, although several studies have demonstrated it is not perfectly so \textsuperscript{[24, 374, 391]}\textsuperscript{21}.

Regarding independence, ripple in the glottal flow estimates can occur due to source filter interaction when the first formant is close to the fundamental \textsuperscript{[13]}. The degree of source-filter interaction was found to be speaker dependent in \textsuperscript{[338]} where simultaneous EGG and voice recordings were made whilst moving in a small tube near the lips so as to extend the vocal tract length and lower formants.

Acoustic properties of the vocal tract have the ability to affect the airflow at the glottis without necessarily altering the vibratory pattern of the vocal folds. Perturbations from formants present in the flow and skew\textsuperscript{22} of the glottal pulse can indicate interaction between the vocal tract acoustics and the airflow at the glottis \textsuperscript{[339]}.

### 3.3.2 Inverse-Filtering

The speech signal can by passed through the inverse of the linear filter for the vocal tract found via LP analysis in order to remove the spectral shaping effects of the vocal tract and obtain an estimate of the glottal flow in a process called inverse-filtering (IF).

Most modern\textsuperscript{23} state of the art IF methods adopt the approach of estimating the vocal tract transfer function over only the speech samples corresponding to the closed phase (CP) of the pitch cycle. Theoretically over this region the signal represents only a freely decaying oscillation shaped by the vocal tract resonances and with no or minimal interaction with the glottis, thus best adhering to the assumptions of the independent source-filter theory assumption.

\textsuperscript{20}One can still assume an all pole model for nasals (and fricatives) as including more poles can model the zeros \textsuperscript{[316, 130]} although increasing the LP model order to much results in instability regarding noise (overfitting of the spectrum) and difficulty in estimating the larger set of polynomial coefficients for the transfer function. Nasals can automatically be detected from the speech waveform \textsuperscript{[306]}.

\textsuperscript{21}An acoustic system is linear if and only if it can be characterized by its frequency response \textsuperscript{[253]}. The estimated glottal flow should be zero over closed phases in a perfectly linear system \textsuperscript{[186]}.

\textsuperscript{22}Skew causes greater high frequency energy and is caused by non zero vocal tract impedance \textsuperscript{[336]}.

\textsuperscript{23}IF the speech signal to strip away the vocal tract effects and estimate the glottal flow was first proposed in 1959 \textsuperscript{[261]}. These early methods began by removing only the first formant and were mechanical.
The covariance specification of the LP normal equations is used for CP analysis as it minimises the prediction error only over the region of interest and whilst CP analysis generally provides a good vocal tract estimate, it is the nature of the covariance solution that the determined filter but may have unstable poles [253]. At the cost of computational complexity the locations of the poles in the z-plane are generally checked and reflected if necessary about the unit circle [316]. Another issue with CP analysis is having sufficient samples over the closed region to be able to define a well posed problem for the appropriate prediction order. For this reason the method has been shown to estimate the glottal waveform consistently well for low pitched voices with well defined closed phase [230, 403], but may be unsuited to certain voices having high pitch or phonation styles with minimal to no glottal closure period [302, 413]. Better formant tracking is also obtained via CP covariance LP [76].

The speech signal as we have seen can be considered to be produced via the action of convolving the excitation signal produced at the glottis with the transfer characteristic of the vocal tract (including lips). Without having information about either of these components iterative methods have been proposed similar to maximum likelihood estimators for unknown parameters/class membership problems. The iterative algorithm suggested in [20] initially estimates the glottal spectrum via performing a small order LP (essentially estimating the low-pass glottal filter characteristic typically specified as a 6 dB roll off [248]). The resulting all-zero filter is then used to remove the glottal component from the speech spectrum by IF. Higher order LP is then performed on the resulting time domain signal, ideally now possessing no glottal effects, in order to obtain an accurate estimate of the vocal tract filter which finally is then used to inverse filter the original speech signal obtaining the final estimate of the glottal flow. Integration

\[ \text{Integration} \]

with filters being manually tuned until zero flow over glottal closure periods was obtained [261]. More sophisticated analogue systems were developed that were capable of removing the first four formants [186, 73]. The first pitch-synchronous techniques, estimating the glottal flow over each vocal fold cycle, were also proposed [257]. Most of these methods had significant drawbacks aside from the mechanical apparatuses required that relate to the researchers having a prior belief of what the glottal flow should resemble and adjusting the filters formant locations and bandwidths to achieve specifically this. Another interesting early non-digital technique relied on placing a reflectionless tube at the lips to create acoustic conditions whereby the glottal waveform can be recorded directly with minimal distortion via a microphone placed within the tube [362]. Digital filtering was introduced in 1970 [279] beginning the transition towards application and estimation in practical situations where there is an unquestionable restriction on obtaining the source waveform from the digitised speech signal alone [190].

Notes:

24 Autocorrelation LP is computationally easier and guarantees stable poles but requires a +6 dB per octave pre-emphasis [347] and is typically biased by vocal tract effects.

25 Note that stability is with respect to a causal linear filter; indeed the poles must lie outside the unit circle for a stable anticausal filter [253].

26 The 6 decibel per octave fall off in pressure due to radiation at the lips is found to be well approximated by the derivative of the V-V flow there, especially at low frequencies [253]. Thus a single pole is
may also then be performed to remove the radiation effects of the lips. This method was shown to estimate glottal flow well even from breathy voices. Further improvements to this method have since been obtained via the addition of extra iterations in order to perform pitch synchronous so called iterative-adaptive IF (IAIF), whereby the flow at the glottis is estimated precisely over each single pitch cycle [12]. An adaptive method that integrates the Liljencrants-Fant glottal flow model [132] into the estimation procedure and claims to be capable of also modelling anti-resonances was proposed in [147]. Recently a Markov-Chain Monte-Carlo algorithm has been used to improve the estimation of the first few poles of the LP model within an IAIF process [32].

Another technique for obtaining accurate glottal flow estimates is to impose constraints upon the equations for the LP coefficients with the goal of only producing plausible vocal tract filter representations. LP coefficients are typically found by minimising the prediction error over whole residual even though it is known that the error is much larger over the period of glottal closure that produces the most energetic impulse to the vocal tract for most speakers [421]. Various other error forms have been proposed but none have been found suitable for all noise and environmental conditions [317]. Such attempts include minimising the error between the LP residual and an ideal, theoretic glottal pulse which naturally suffers from the problem of prior conception, and substituting the L-2 norm of least squares for the L-1 norm. The LP coefficients solutions have also been constrained to explicitly come from certain prior specified sets [98].

Regarding CP IF, the closed region is typically short such that incorrectly specifying it by only a few samples can have significant effects. This prompted the development of DC-constrained LP where by regularisation is used, requiring that the sum of filter coefficients remains constant over frames or pitch periods [16].

The method of weighted LP proposed applying a temporal weighting of the square of the residual signal based on the short-time energy (STE) function [245]. In essence this method aimed to emphasis LP estimation only over those samples best fitting the implicitly used source-filter theory, but stability of the all-pole filter was an ever present

27 These methods are based on the discrete all-pole modelling process for obtaining well fitting parametric models to the speech spectrum that are suitable for accurate formant estimation [121].

28 Aiming to overcome the fragility of CP analysis and to also make use of all of the speech samples per pitch period a method of jointly estimating a source model and vocal tract filter was proposed in [259]. Similar joint optimisation based methods derived from this study were tested [148, 149] but since ignored due to its complexity and non-robustness. Similar issues regarding robustness and suitability typically only for synthetic speech occur in [145], where discrete all-pole modelling was combined with the Liljencrants-Fant (LF) glottal flow model as a source in a process whereby multidimensional optimisation performed with simultaneous IF was used in order to determine the best LF parameters.
issue. A modification to the STE was proposed in [247] that was demonstrated to result in stable all-pole models and resulted in increased robustness to noise. This method was analysed on a synthetic speech database generated using a glottal model proposed in [385] in [205] whereby it was found to perform superior to IF with autocorrelation LP estimated vocal tract models and comparable to CP covariance IF.

Predefined values of the gain of the estimated vocal tract filter at specific frequencies have also been imposed during the determination of the LP coefficients [17] and evidence presented for the claim that this method produces filters with less spurious low frequency formants, improving the robustness of CP IF to the location of the analysis window [18].

3.3.3 Epoch Estimation

A necessity for the application of the CP techniques is to make an accurate determination of the segment of glottal closure within each pitch period with several studies showing the variability of the glottal estimate is significant with changes in the CP analysis window [235, 330, 426]. Good determination of the vocal tract filter requires specification of the closure instant to within 0.25 ms [284].

Determining GCI from the speech signal alone was first explored in 1974 where closed phase was found by LP and a sliding window [369]. Estimating the closure instants has also been attempted via analysis of the formant frequency modulation that is predicted to occur with the time varying degree of source/tract coupling over each glottal cycle [25], by using the minimum of the least squares LP prediction error [419], by performing a singular value decomposition of the speech signal via the Frobenius norm [246], by a sliding covariance window with reference to formant modulations [302], by the group delay function [345, 358], by unwrapping the phase [425], by using the Hilbert envelope of the group delay [319]. Empirical studies of group delay methods is presented in [56] and a review of many of these methods is presented in [278]. Location based on the maximum phase characteristics of the glottal pulse prior to glottal closure is discussed in [86] and [116].

The DYPSA algorithm improves on the group delay methods and uses a dynamic programming cost function to break the estimation problem down into computationally efficient sub problems with factors relating to pitch deviation, Frobenius norm amplitude consistency, ideal phase slope function deviation and speech waveform similarity [229].

29 Early methods of developing algorithms for the determination of the glottal closed phase from the speech signal alone used a two-channel analysis with speech and a simultaneously recorded EGG signal as a ground truth [230]. Estimation of CP is difficult even with EGG signals when the voice is high pitched or the voice source is breathy [426] and GCI are often not even present in the voices of speakers under stress or with vocal disorders [265].
The main components of the algorithm are to detect GCI candidates based on zero crossings of the phase-slope function then use a phase-slope projection technique to determine more candidates that may have been missed by not quite crossing zero. Falsely detected GCI are then determined by the dynamic programming cost function. Two-channel analysis of the DYPSA algorithm found it to be state of the art in 2007 with a GCI location standard deviation of 0.29 ms about the true EGG value [229, 284].

In the speech of most speakers greatest excitation is found at the point of glottal closure, making the determination of the more gradual glottal opening instant (GOI) comparatively harder with little literature published on the task. CP IF estimates are also much more tolerant of imprecise specification of the opening instants and the glottal closure period is often taken as a fixed percentage of the pitch period (ie 30% to 45% in length) starting from the found GCI [302].

One algorithm that estimates both the GCI and GOI in each pitch period is presented in [115], where a low-pass filtered version of the speech signal is used to demarcate containers (small segments of continuous samples) within which lie the true GCI and GOI. GCI are said to lie between the minimum and the positive going zero crossing and GOI to lie between the maximum and the negative going zero crossing, with an extra 0.25 ms added on each side for leniency. Appropriate low pass filtering of the speech signal is a vital stage of the process. Two-channel analysis of the method with the small CMU ARCTIC database found that the method was more accurate than DYPSA whilst also more robust to additive noise [115].

The authors of DYPSA also recently proposed the YAGA algorithm for epoch detection where GOI are also estimated and which also uses the group delay function and dynamic programming [379]. The CP in this algorithm is defined as the region over which LP analysis results in minimum deviation from an all-pole model for speech. The GOI detection method is based on the idea that the CP measured from the estimated GCI should be locally consistent. Two-channel analysis is used to report high GCI and GOI identification rates with 0.3 ms and 0.5 ms accuracies respectively.

An algorithm has also been proposed that estimates a collection of glottal flows per pitch period with sliding windows, some of which should accurately cover the true CP (should it have existed), before selecting the best estimated glottal waveform from these with reference to characteristic properties of idealised glottal flow [270]. This algorithm was claimed to produce glottal waveforms very alike those obtained from simultaneous EGG measurements.
3.3.4 Exploiting the Mixed Phase Properties of Glottal Flow

The remaining main category of glottal flow estimation methods aim to achieve source-filter separation by exploiting the phase properties of the glottal pulse. Considering the source-filter model as an excitation/filter model where the filter comprises all of the properties of the glottal flow, vocal tract and lip radiation, it was shown that the glottal flow can be modelled as an anticausal (maximum phase) filter before the glottal closing and a causal (minimum phase) filter after the glottal closing [105]. Given this capability questions are introduced regarding IF methods about the possibility of the ‘vocal tract’ filter capturing glottal information.

For such reasons methods aiming to perform the source-filter separation in the phase space using zeros of the z-transform (ZZT) have been investigated. Such methods perform no inverse filtering and typically provide accurate flow estimates (two channel analysis has confirmed this) when the recording conditions are optimal, meaning that the phase\(^{30}\) properties of the speech signal are captured with full fidelity [86]. These methods are computationally intensive however and perform significantly worse than IF approaches in the presence of almost any distortions to the speech signal.

The computational issue of the ZZT decomposition method has inspired phase based separation using homomorphic analysis via the complex cepstrum [110] within which similar flow estimates are obtained with significant reductions in processing time. These increases are relative however as the method is still based on the mixed phase properties of speech and a large number of points must be taken during Fourier analysis in order to be able have a fine scale on the unit circle for unwrapping the phase. Note also that the choice of windowing function for speech frames plays an important role due to the importance of the phase signal in these methods.

Comparative studies of differences in glottal flows estimated from a range of methods typically conclude that, of the approaches discussed, IAIF techniques are most robust and more suitable for estimation in a wider variety of recording conditions but that CP IF and decompositions based upon the mixed phase properties of speech are best for high fidelity estimation [112, 114, 300, 413].

\(^{30}\)Note that ‘phase’ has many meanings across digital signal processing literature and indeed the both the instantaneous phase and Fourier decomposition phase component have been used [86], neither of which is robust.
3.3.5 Objective Measures for Glottal Estimates

The glottal flow pulse is a low frequency signal and is thus sensitive to recording conditions and phase responses of microphones and transmission channels [14] and the fidelity of glottal flow estimates can be considerably affected by any phase distortions introduced during the recording or transmission processes [413, 419]. Given these sensitivities in performing the already difficult blind separation problem there exists a need for tools to quantify broadly the quality of obtained flow estimates.

While EGG and laryngographs both provide information on glottal V-V flow and vocal fold motion, they obtain these by invasive methods which are not feasible in nearly all non-research situations, meaning that often blind source-filter separation is attempted from the speech signal alone without having reference to any ground truth. This is a fundamental problem in developing new glottal flow estimation methods and in testing existing methods. Synthetic speech can be used to overcome this problem however the synthetic speech is generally produced explicitly by the source-filter model which in practice varies in applicability when examining truly generated human speech.

This has meant that several quantitative measures have been proposed for assessing the quality and fidelity of estimated glottal flow waveforms. Phase space plots, first suggested in [120], have been used to check the validity of IF glottal estimates with theory suggesting that their phase plane plots should form a circle [316] (when estimated from modal voice) and that any extra circles indicate the remaining presence of vocal tract effects, typically formant ripple. Several objective heuristics plus statistical measures for assessing the quality of the obtained IF glottal model are considered in [33] which also includes an automatic implementation of [120]. Estimates are considered against expected or plausible flow waveforms with the degree of formant ripple quantified based on spectral calculations of the determined flow.

Other common indicators include the signal-to-reconstruction-error-ratio calculated from ratios of standard deviations of the true and resynthesised signals [205], the H1-H2 measure of the difference in amplitudes of the first and second harmonics of the source spectrum [205], calculating the group delay function of the estimated waveform [14] and spectral roll-off measures [270]. An empirical review of several glottal quality measures is presented in [268].

A comprehensive review of glottal waveform estimation methods can be found in [13, 108] and [413].
3.4 Literature Review: Parameterising the Estimated Glottal Waveform

Modelling or quantifying the glottal waveform has applications in the fields of speech pathology [113], speech synthesis [81], speaker recognition (Section 3.6) and affect classification (Section 7.2). Yet modelling of the glottal excitation is still not well established and significantly less is known about how to parameterise the voice source signals in comparison to the vocal tract filter [378]. The three main approaches for extracting information relevant to these tasks has been (1) fitting theoretical models of modal glottal flow, (2) calculation of temporal or spectral statistics and (3) data driven approaches whereby characteristic flows are derived from the actual dataset of glottal estimates.\(^\text{31}\)

Importantly, the parameterisation method must reference the practical aims however [13]. Classification tasks like identity and affect prediction require primarily discriminative features, possibly that are also complimentary to baseline features. In such cases low fidelity estimates may be acceptable provided that they are informative and consistently reproducible. This engineering perspective contrasts with the requirements of speech pathology for example where high fidelity estimates may be essential to accurately infer physiology and dynamics.

3.4.1 Temporal Domain Models

Functional forms for the glottal flow are of particular importance to speech synthesis where their variations can result in significant perceptual differences [79, 200, 336]. The glottal pulse signal has a coarse structure whose shape is typified by the EGG measures [311]. These temporal function form models capture this coarse component and are able to describe the open, closed and return phases as well as the pulse shape and peak glottal flow. The finer structure\(^\text{32}\) of the glottal flow is often discarded when fitting these synthesis type models.

The Rosenberg-Klatt (RK) model is one of the earliest and simplest such models, describing the pulse as a quadratic \(\sim bn(2n_c - 3n)\) over the samples \(n \in [0, n_c]\) from the GOI to the GCI and identically zero over the closed phase [225]. The parameter \(b\) controls the flow amplitude. A review of glottal models proposed before 1986 is presented in [148] where the Fujisaki-Ljungqvist (FL) polynomial model is also proposed which

\(^{31}\text{Outside of the signal processing domain physical models of the vocal folds capable of reproducing their dynamic behaviour have also been developed [384].}\)

\(^{32}\text{The fine structure derives from ripple (related to the coupling of source and vocal tract) and aspiration (the result of turbulent airflow through the glottis which varies with area shape and degree of closure) [302].}\)
encompasses all their properties. The KLGLOTT88 model proposed in [225] aims to improve the RK model via appending a first order low-pass filter with the aim of creating a smoother glottal closure which is strongly related to the speech’s spectral slope [308].

The most frequently employed of these models however is the Liljencrants-Fant (LF) model [132] for the derivative of the glottal flow that models the opening and return phases with an exponentially increasing sinusoid followed by a decreasing exponential respectively while the closed phase is identically zero. It has been frequently used for speech synthesis [413]. The model is described by Equation (5.11) in Section 5.7 where its parameters are used in a speaker verification experiment.

The model is suited to modal phonation with non-interactive flow such that the linear source-filter independence assumption holds [132]. It has been shown to be functionally equivalent to other parametric models [106] and is in fact a combination of an earlier introduced three parameter model of Liljencrants (L) with an earlier Fant (F) model [132]. The combination reduced the discontinuity at the flow peak of the F model.

One issue with functional time domain models is that their spectral properties (phase response in particular) can be considerably different to true glottal flow signals [86], although it is possible however to formulate the LF model as a combination of anti-causal and causal linear filters [105] with well matching spectral properties. Frequency domain analysis of the LF model is presented in [131].

### 3.4.2 Statistics and Quantifiers of Glottal Flow

In non-synthesis domains such as classification tasks often the estimated glottal flow is decomposed into a set of descriptive statistical measures [413]. Such descriptors in the temporal domain have included ratio measures of the phases of the glottal cycle [197, 372] relative to the instantaneous pitch such as open quotients first proposed in [381], or closing quotients [264]. The normalised amplitude quotient is another attempt to parameterise the closing phase and has been found to relate well to laryngeal tension [15]. The ratio of the open to closed phases (the speed quotient) has been shown to describe well the asymmetry of the glottal pulse [180]. The return quotient measured on the derivative flow signal was suggested in [303].

Measures of differences in the cycle to cycle variation of the glottal flow in duration and amplitude known respectively as jitter and shimmer are also used as quantifiers [134].

---

33 A computationally efficient alternative to the LF model termed the R++ model (being derived from the Rosenberg (R) model [336]) is presented in [404].

34 Given the noted difficulty in accurately detecting epochs, often a fixed percentage is assumed and these features can thus relate more then to the instantaneous pitch [344]. The crossing of a fixed threshold on an EGG signal is another approach to measuring these durations [107].
Amplitude statistics such as the maxima and minima of the V-V flow at the glottis and its derivative are also used [13]. The Teager energy operator [373] is also used to define the energy of the speech signal in terms of the voice source input.

It has been shown that the mentioned common glottal flow models are not able to spectrally represent the voice during speech nor singing [180] and in many speech processing fields it is often preferable to have a frequency domain representation of the glottal flow [13] which may be obtained via autoregressive (LP) or parametric (Fourier)\(^{35}\) means. For example voice quality measures are generally described by spectral parameters such as the harmonic richness factor [81], the spectral slope of the glottal flow [267] and the parabolic spectral parameter [19]. The harmonic richness factor [78] calculated as the ratio of the sum of harmonics above the fundamental to the fundamental has been used to quantify the spectral decay (slope) also.

Having accounted for formant peaks [53] amplitude difference between the first two harmonics denoted \(H_1 - H_2\) has been shown to correlate to the open quotient of the glottal waveform [175]. The peak in the spectrum of the flow derivative is known as the glottal formant. Using two-channel EGG analysis it has been shown to correlate strongly with the open quotient[55]. Many other possible spectral quantifiers of the glottal flow signal are discussed in [388] as well as in the documentation for the Aparat IAIF toolkit [7]. A review of the spectral features of voice source models in both the amplitude and phase domains is presented in [105] and of glottal flow estimates in general in [112].

### 3.4.3 Data-Driven Representations

Motivated by the observation that many typically finer but idiosyncratic characteristics of real voice source waveforms are not accounted for by function fitting methods, data based approaches that aim to extract salient feature from the estimated glottal flow waveforms rather than perform a top down approach of fitting a preconceived analytic form to the data have been proposed for improved speaker recognition and for the generation of more perceptually natural synthetic speech.

The Deterministic plus Stochastic model (DSM)\(^{36}\) is one such attempt, proposed initially for improved speech synthesis in [119] while subsequent investigations of its potential for speaker identification were presented in [117]. This model fits an average waveform to the source signal below a cut-off frequency (4 kHz), and models the

\(^{35}\)Some works have looked to obtain glottal pulse spectral information from the magnitude spectrum of the speech signal, avoiding complications with IF methods for instance. See [174, 175] and [388].

\(^{36}\)The DSM model is based on a similar approach to the Harmonic plus Noise model [370] where the speech signal (not the LP residual/glottal flow) is modelled as the sum of sinusoids of the fundamental frequency and its harmonics up to a maximum voiced frequency and as noise above this.
higher frequencies with a Hilbert envelope of the signal. This coarse component average waveform is given by the first principal component vector of a PCA decomposition of the collection of prosody normalised single pitch period glottal flow waveforms and is referred to as the ‘eigenresidual’. It is found on the CMU-Arctic database that the eigenresidual accounts for 46% of the observed variation [112]. By retaining only the first eigenvector an implicit filtering of the glottal waveforms is performed, the claim being that subsequent eigenvectors contain primarily higher spectral information, including vocal tract resonances remaining from imperfect IF for example [119].

It was found in [378] that only 16 ‘prototype’ voice-source waveforms, selected via a Fisher discriminant ratio, were suitable for use as a pseudo basis for the representation of estimated glottal flow curves from the 10 speakers of the APLAWD database. Further analysis in terms of signal to noise ratio and Bark spectral distortion (a psychoacoustic measure) were presented in [172] demonstrating the improved synthesis results stemming from this data motivated codebook. Further the waveform representations were shown to be independent of pitch or phonetic content in correspondence with theory [204].

The source-frame feature introduced in Section 3.5 is another such data-driven representation and its utility for the speaker recognition task is investigated in several experiments described in Chapter 5.

3.5 Source-Frames: A Normalised Glottal Flow Feature

The time domain features which represent the derivative of the glottal V-V airflow are extracted from the speech signal via the process outlined below. They are fundamentally CP IF estimates that are then normalised to capture only waveform shape characteristics and are inspired by the data based methods discussed. We refer to each of these normalised, two pitch period, GCI centred glottal flow curves as a source-frame.

Preprocessing: The speech signal is first segmented into 30 ms frames with a 10 ms shift and pre-emphasised with an alpha of 0.95 [316]. A voiced/unvoiced decision is then made on each speech frame. Frames are labelled voiced if the frames energy and short term autocorrelation measures exceed some dynamic or empirically predetermined thresholds. We set the thresholds quite high and only retain the top 30-50% of data by these measures in order to retain highly voiced speech only. The pitch estimation algorithm described in [127] was also used and speech sections found to be voiced by both of these methods only were labelled voiced.

One thousand pitch periods are sufficient to reliably estimate the DSM model [112].
In order to estimate the glottal waveform corresponding to each of these voiced frames we then perform CP LP analysis, where the assumptions of the source-filter model of speech production [130, 252] are most valid due to the maximal separation of interaction between the vocal tract and vocal folds.

**Determination of Glottal Closure:** Stepping through voiced sections of speech the CP of each glottal cycle is then estimated. An implementation of the algorithm described in [115] is used for determining the GCI and GOI. This glottal instant detection algorithm was reviewed in Section 3.3.3 and is described in detail in [115] where empirical evidence is also presented for the claim that it is superior to alternative glottal instant detection methods such as DYPSA [284].

**Closed Phase Linear Predictive Analysis & Inverse-Filtering:** CP LP, using the covariance solution to the Yule-Walker equations, is then performed over the detected glottal-closure regions of voiced speech in order to determine as accurately as possible the all-pole linear filter representing the vocal tract at each frame length moment of voicing of the speech signal. Any roots of the filters transfer function found to fall outside the complex unit circle are reflected for stability of the filter, which is then used to IF the speech frame, determining the pitch-synchronous error signal representing the glottal flow derivative waveform.

**Pitch Period Extraction & Prosody Normalisation:**

Having estimated the glottal wave over each voiced pitch period of the speech signal we then segment the glottal signal into consecutive two pitch periods frames, with glottal closure points at the start, centre and end of each extracted segment. This creates a set of vectors of different lengths.

Each of these two-pitch period segments is standardised by a process called prosody normalisation whereby the data is scaled in both length $x$ (duration/pitch) and amplitude $y$ (energy/flow-volume) dimensions. The amplitude scaling is done by normalising by the standard deviation of the source-frame data. In the pitch normalisation process the frame length is mapped to a pre-specified constant number of samples by interpolation or decimation as necessary, along with the required low pass filtering to negate aliasing effects. Data were normalised to a length of 256 samples.

Finally these now uniform length vectors are Hamming windowed to emphasise the shape of the signal around the central glottal-closure instant. Each such vector is called a *source-frame* and with this windowing each source-frame vector contains information

---

38 It also makes an estimate of the more difficult to determine point of glottal opening, avoiding the common assumption that a fixed portion of each pitch period is closed.

39 The length depends upon the F0 of the pitch period they are extracted from and the sampling frequency of the digitised speech.
centrally describing the glottal flow shape over a single pitch period. It retains information pertaining to the overall periodic motion of the vocal folds at the cost of shimmer and jitter information.

This normalisation enables the signals to be analysed statistically with discriminative and generative models and for fitting functional forms to the data via regularised methods. Without normalising, only scale invariant methods like non-regularised least squares can be used in exploring parameterised forms for establishing a concise representation of the data.\textsuperscript{40}

To overcome some imperfections with the inverse filtering process the mean of small blocks of consecutive source-frames is often taken providing an averaged estimate of the shape of the glottal waveform over a short period of voicing, reducing the effects of imperfect IF\textsuperscript{285}. These features are based upon the DSM model proposed for speech synthesis in \cite{119} and aim to capture both coarse and fine structure features of general flow in a single representation.

3.6 Literature Review: Glottal Waveforms for Speaker Recognition

As we have seen in Chapter 2 the currently prevailing automatic speaker recognition paradigm involves modelling short-term spectral magnitude characteristics, which primarily reflect vocal tract information \cite{223}. Perceptual and invasive studies however indicate physical reasons for expecting speaker’s glottal flow signals to aid in discriminating between speakers.\textsuperscript{41} Indeed whilst the largest source of variation between speakers voices

\textsuperscript{40}Multiple examples of such source-frame vectors are shown in the appendices, Section A.1. A Matlab implementation of this algorithm is available from \url{http://staff.estem-uc.edu.au/davidv/}.

\textsuperscript{41}Several studies were conducted in the 20th century, particularly within the fields of pathology and speech synthesis, which alluded to the potential speaker discriminative abilities of the glottal flow waveform. As early as 1940 video footage of vocal fold motions was captured which demonstrated their variable nature over different speakers \cite{133}. Differences in vibration patterns between genders were inferred from direct measures of airflow at the mouth \cite{185}. A Rosenberg airflow mask was used in \cite{208} to determine how perceptual qualities (‘breathy’, ‘creaky’, ‘husky’) correlated to glottal flow characteristics. Evidence of the glottal flow resulting in perceptual differences between speakers, as adjudged by an expert listener, were obtained via synthesis experiments on time aligned TIMIT \cite{157} utterances whereby switching the glottal source waveform between speakers resulted perceptually in a perceived dialect more indicative of that of the owner of the glottal source waveform \cite{423}. Perceptual differences arising from synthesis with varying LF voice source model parameters are reported in \cite{393}. Yet another perceptual study concluded that speakers have certain phonation termination habits, maintaining periodicity until the end or transitioning into an aperiodic phase, a factor which human listeners remember and use to inform their identity decisions \cite{50}. EGG signals have also shown speakers to have differing source waveforms as a result of differences in their vocal fold vibration patterns \cite{82}. Another study demonstrated that people could recognise others voices from listening to only the linear-prediction residual \cite{139}.
originates from their different vocal tract and articulator physiologies, speaker identifying characteristics have been found in several parameterisations of the voice-source waveform which are now discussed. Further, based on human physiology and the general speech production theory, this glottal flow information should be complementary to standard spectral features, in the sense that it should hold independent information.

These two points are fundamental and the glottal flow signal is typically considered with a view towards increasing the accuracy from a baseline information source, typically MFCC features. Many of the studies within the research literature adopt this approach and their reporting focusses on increases rather than the accuracy of isolated voice-source systems alone. In point of fact almost all studies report some level of increased recognition performance upon including glottal flow information in their systems [413].

In one of the earliest studies several spectral parameters extracted from the LP residual were combined with LP filter parameters to reduce the EER from 5.7% to 4% on a small set of French speakers collected over three days of radio broadcasts [377]. Jitter and shimmer measurements were able to lower the EER of a GMM-UBM spectral system from 10.1% to 8.6% when scores with a k-means classifier and combined by weighted score fusion in [134]. The Fourier decomposition of IF estimated glottal flows from vowel centres of four different words from the TI46 database alone were able to produce EERs of 20% and 31% for eight female and male speakers respectively [410]. Another small study [254] reported improvements in EER and identification rates over a variety of GMM modelling configurations when combining the mel-cepstral coefficients of the LP residual with LP spectral envelope cepstral features. The data for this study was recorded in a sound proof room in contrast to another examination of the use of cepstra to parameterise the LP residual performed on telephone speech where the IAIF estimates were found to be “not too useful for recognizing speakers” [222]. Voice pathology software implementing IF to obtain glottal flow estimates was used in [163]. Various GMM-UBM models of output spectral domain parameters in combination with standard MFCC features resulted in an EER of $\sim 0.5\%$ in comparison to the MFCC baselines of $\sim 1\%$ for a clean database of 240 speakers.42

Larger and more recent studies include the following. The four parameters of the LF model were used in [302] to model the coarse temporal structure of CP IF glottal flow estimates along with energy and perturbation measures to capture finer grained details. Speech data was provided by the TIMIT database. Using a GMM-UBM framework identification rates were able to be improved via feature fusion with MFCC from 91% to

---

42Another study by the same researchers shows that PCA coefficients of pitch normalised IF estimates effectively separate genders [162].
93.7% for the 112 male speakers however they were lowered for the 56 female speakers from 93.6% to 92.6%. It was observed that the female voice-source features contained many outliers (low GMM likelihoods of the data) and the LF parameters could likely have been estimated with greater accuracy via a non-linear least squares algorithm, the likes of which are better suited to fitting piecewise functions [42]. The use however of mel-cepstral parameterisation of the glottal flow signal in combination with the standard spectral MFCC resulted in identification rates of 95.1% and 95.5% for the male and female groups respectively. Further studies on the telephone degraded speech (NTIMIT) showed increases of 56.7% to 59.4% and 66.3% to 69.0% for the male and female systems respectively when combining MFCC spectral magnitude features with MFCC of the glottal flow derivative.\footnote{We note also that these accuracies are improved by the absence of intersession variation within the TIMIT databases.}

A unique parameterisation and modelling approach on challenging data was presented in [277]. The residual phase, derived from the cosine of the phase of the analytic signal found from the Hilbert transform on the LP residual, was modelled via a neural network producing an EER of 26% on the NIST 2003 SRE data. This was reduced to 19% via the application of $t$-norm [31]. Linear weighted score fusion with a similarly modelled MFCC spectral system (baseline 14%) achieved an EER of 10.5%. Using the same database CP IF estimates were parameterised by the Fujisaki-Ljungqvist (FL) [148] model in [357], adjusting the time parameters to percentages of the local pitch period. An interesting LP process was performed whereby the LP parameters were estimated over a smoothed CP where the smoothing was achieved via the addition of the autocorrelation matrices from the previous and following pitch periods. Of note also was the novel approach of estimating the glottal flow model parameters in the frequency domain. A codebook was built linking the five parameters of the FL model with their frequency domain transform coefficients, and a metric of least squares on these frequency domain coefficients was used to find the parameters from the codebook. Using neural net modelling very similar EERs are observed for the raw and score fused (with magnitude spectral MFCC) systems with these methods.

An exploratory study of several temporal and spectral quantifiers of IF glottal flow estimates for speaker recognition was performed in [389]. Performing feature fusion, frame-level identification rates in differentiating pairs of 50 TIMIT speakers with various sized GMM models demonstrated that the glottal information was complementary to spectral magnitude mel-cepstra.

Representing speakers via their DSM model [119] features (eigenresidual and Hilbert...
energy envelope of spectral content above 4 kHz) discriminative modelling with a distance metric achieved an impressive identification rate of 96.35% on the 630 speakers of the TIMIT database and 70.7% on the YOHO database [117]. The reduction over databases highlighting the information loss and corruption that occurs with practical environmental recording conditions and these rates are similar to previous glottal studies on these data [378, 302]. Weighted score fusion of systems using exclusively the eigenresidual or the second eigenvector were also presented with no positive fusion results observed at any weighting, providing empirical justification for the retention of the first eigenvector only.

Finally a novel method which uses cepstral subtraction rather than the equivalent frequency domain IF for inferring glottal flow information is presented in [171] along with promising recognition results. Without employing IF the method is also claimed to be robust to low frequency phase distortions that often affect LP analysis. The algorithm consists of calculating standard magnitude spectral MFCC over the glottal CP (as determined by the DYPSA algorithm [284]) which are then subtracted from MFCC calculated over the entire pitch period, resulting in what are termed vocal-source cepstral coefficients (VSCC). Identification experiments on the TIMIT and YOHO databases with GMM modelling (no background model) demonstrate that with weighted score fusion the VSCC features are complementary to the magnitude spectral MFCC.44 With no further mention in the literature this interesting approach is replicated in Section 5.2.

Despite these results the voice-source waveform remains to be regularly and efficiently utilised in automatic speaker recognition systems. Further studies are required to determine the extend to which glottal information can assist the speaker recognition task and should these provide sufficient encouragement several difficulties likely have to be overcome to enable practical uses. These include the difficulty of reconstructing time domain glottal signals given speech enhancement and nuisance compensation algorithms for additive and convolutional noise do not reduce phase distortions [309] and the bandwidth limiting effect of most current narrowband telephony standards which filters much of the low frequency content of the speech signal which relates strongly to the glottal flow [302]. Almost no research exists regarding attempts to estimate high fidelity glottal flow waveforms in such real world conditions.

The difficulty of this task may be implicitly indicated by the lack of further publications following the influential work from 1999 in [302] which concluded with the statement that “we are currently investigating source features in other databases such as switchboard.”

44For comparison with the DSM model reported in [117] an identification rate on YOHO of 63.7% was achieved.
Chapter 4

Score Post-Processing: r-Norm

4.1 Introduction

In this chapter a versatile and novel technique is introduced for increasing the performance of any speaker recognition system by making use of previously undervalued information held within the scores output by a given classifier. Specifically, we hypothesise that there exists a relationship between the scores of a test probe against all enrolled client models that has not been attempted to be captured yet, and we propose a flexible regression based normalisation/post-processing method for adjusting these scores to alleviate predictable biases. A structured learning method called Twin Gaussian Process Regression [47] is used by the method to capture these inter-speaker (inter-model) relationships hypothesised to be present within the scores.

We name the approach r-norm for regression-normalisation and, depending upon the choice of data used in learning the r-norm model, it may be viewed as a normalisation method and/or purely as a performance boosting approach. Indeed we have come to generally view the method as another stage of speaker modelling that takes place at the score level. Note that although our focus is on automatic speaker recognition, the r-norm algorithm is a general technique applicable to any classification task that generates a quantitative assessment (score) when comparing a test input against a given class.

The hypothesis regarding these inter-score relationships and the r-norm algorithm proposed for exploiting them are described in detail in the following Section 4.2. In Section 4.3 the r-norm method is discussed in relation to and contrasted with existing score normalisation or post-processing methods. In Sections 4.4, 4.5, 4.6 and 4.7 the experimental results of applying the r-norm method to various speaker verification systems and databases are reported. These results suggest that the r-norm method is able to
perform very well with respect to increasing a systems’ recognition accuracy. Finally, in
the chapter summary found within Section 4.8 the experimental results are condensed,
conclusions are drawn from the gathered empirical evidence, and areas not covered along
with future research directions are commented upon.

4.2 Theory of Regression Score Post-Processing: r-Norm

We now describe in detail the proposed regression score post-processing/normalisation
technique r-norm. The method aims to increase the recognition accuracy of any classifi-
cation system that outputs a quantitative assessment (i.e. a score) regarding the question
of how likely it is that a presented test probe belongs to a specific class. For example it
may be an object classification problem where a typical assessment for the system may
be to determine whether a given image (test probe) is of a table (class 1), chair (class 2)
or waste bin (class 3). The r-norm method that we propose was developed after observ-
ing certain relationships between the scores of a speaker recognition system; important
relationships that we hypothesise to be present across many classification systems for
many problems and that the method aims to exploit. These relationships are described
presently, but first we note that despite the generality of the r-norm method we couch its
description and further discussion in the context of our classification problem of primary
interest, namely automatic speaker recognition.

Now to the issue of what it is within the scores of a system that we are trying to
capture and profit from. Our enrolled speakers (classes) will have a complex and certainly
multivariate set of dimensions along which they vary. Similarly a given utterance (test
probe) presented for verifying a claimed identity will have several different characteristics
present in various quantities. Given that the feature extraction, modelling and scoring
systems accurately capture this information, we hypothesise that the score of this certain
test probe against each enrolled model will reflect the characteristics of the probe that
each of the models shared and differed in. Actually of greater importance (and what
occurs in practice with sufficient but never perfect modelling) is that, irrespective of any
biases or approximations in our classifier, its scores consistently reflect to some degree
this mutual information shared between the probe and model training utterances. The
r-norm method aims to learn these consistent biases and then correct for or infer from
them.

The method can also be viewed as a speaker modelling process that occurs at the
score level. This is seen when we consider the score of a probe against a model as
a scalar feature quantifying model-probe similarity but acknowledging the presence of inherent errors due to approximations and imperfect representations of both the speaker characteristics (in the model) and probe utterance (in the features). A probes’ score vector then contains information about the inter-speaker variation, that is about how the speakers, as approximated by their models, vary and compare to one another. During the training of the \( r \)-norm model, we aim to learn any such patterns within the scores, particularly the ones that result in frequent and repeatable sources of biases or errors, and make adjustments given this information to enable improved classification decisions.

By way of a simplistic example, we may for instance observe a pattern where utterances from speaker A are scored well against model A but typically always higher against model C and say quite low against model D. Having captured this and other such relationships evident within the scores used to train our \( r \)-norm model we can then apply the model to compensate for these effects, and in this simplistic example, then perhaps increase the score against model A when we observe this pattern. A visualisation of this simplistic example is presented in Figure 4.1 and is designed to partially illuminate the kinds of patterns or trends we hypothesise exist within the scores of a classification system and how we aim to exploit them via the proposed method. In this example during the training process of the \( r \)-norm model we learn that the target trials for model A are consistently scored too highly against model C and typically very low against model D. This is shown in the left ‘Before’ pane. In testing, we then aim to adjust the scores, increasing the score against model A whenever this pattern is observed as shown in the right ‘After’ pane. This is of course a highly simplistic example. Indeed here only the score against model A is adjusted whereas in the true \( r \)-norm process in fact the scores against each model would be modified. It is provided simply to demonstrate the types of inter-model relationships that we hypothesise are present within the scores of a given classification system and that the regression normalisation method attempts to learn and exploit.

Before proceeding to describe the regression based algorithm \( r \)-norm that we propose in order to capture and make use of these relationships, we clarify some frequently used descriptive terminology. As described, the relationships we hypothesise to be present exist across the scores of a single probe against all enrolled class models. We therefore assume an ordering of the enrolled speakers and term this collection of the score of a given probe against each of the given models a score vector. The \( i^{th} \) element of a score vector for probe \( X \) then represents the score of \( X \) against model \( i \). To introduce the method we concern our description only with closed-set recognition where any one test probe was uttered by a client for whom we have an enrolled model. A classification sys-
Figure 4.1: A simple visual example of the relationships and patterns within scores that the r-norm method aims to learn and adjust for.

tem outputs raw scores which the r-norm model is applied to. We assume that scores are organised into a matrix where client models correspond to rows and test probes to columns; that is to say that score vectors are column vectors. We work with three disjoint sets of data: a training set for estimating client models, a development set scored by the system to produce the matrix of raw scores used in learning the r-norm model, and finally an unseen testing data set of utterances which are scored by the classification system producing raw scores that the now learnt r-norm model is applied to. We are required to score each probe from our development data set against all enrolled models. If we have n target trials per model then we state that we are training on n score vectors per model. Without the symmetry of all models having the same number of target trials then we state that the development set for training the r-norm model contained α score vectors for a specific model having α target trials. This outlines the terminology we use in describing the training and implementation of a r-norm model.

We now describe the r-norm algorithm, the training of which begins by having a set of enrolled client speaker models and the scores of the development data set arranged in a score matrix which we denote by \( D \). The key observation for attempting to exploit the described relationships is that during the development stage we are aware of the true identity label of all utterances. With this information we may create a set of scores that represent the scores of the development data hypothetically output by an idealised, ultra recogniser. We refer to this matrix as the ideal matrix, and denote it by \( I \). Hence \( I \) has the same dimensions as \( D \).

The central concept of r-norm lies in learning a regression model from the develop-

---

1 At the expense of some additional computation time that is linearly proportional to the cost of scoring a probe against a single model. As mentioned, this is typically not an issue even when executed on modest modern computing architecture.
ment data score matrix $D$ to the ideal score matrix $I$. There is much freedom regarding
the choice of matrix $I$, (there are no constraints other than that the dimensions are
fully specified by and equal to those of $D$), but the choice does influence how the $r$-norm
process may be described. For example we may choose a highly synthetic matrix where
target trial scores are all set to +1 and non-target trial scores all set to 0 in which case
we may view the $r$-norm method as a score post-processing step for improving recogni-
tion performance where $I$ does not relate to our data or conditions and is purely a form
of optimal, error free, classification output. Alternatively if the system is to be tested
on speech that has different characteristics to that used for training or is challenging in
some sense (channels, noise, babble, microphone)\(^2\), we may choose the ideal score matrix
$I$ to be taken from the scores of clean data (or data used to train the speaker models)
and the development data should be as similar as possible to the anticipated testing
speech. In this way the $r$-norm process may be viewed as a compensation or normali-
sation method. These are just some potential choices regarding the specification of $I$.
The resolute factors always guiding its choice are however that it should correspond to a
set of scores with zero or nearly zero errors for the development data set. The empirical
results presented in the final sections of this chapter demonstrate that the specification
of $I$ is very important. Deductions are made from observations of these experimental
results in the chapter summary of Section 4.8 regarding specifying the ideal matrix $I$ in
a sensible and likely optimal manner.

Having obtained our development data set scores $D$ and specified their ideal equivalent
$I$, the question is how can we establish an algorithm to adjust future test set scores,
cores that should hopefully display the same patterns and relationships that are present
within our development data scores, so that they provide improved recognition accuracy?
To do this we learn a regression function that maps $D$ to $I$. This regression function we
denote by $r$.

We use Twin Gaussian Process Regression (TGPR) [47] as the regression model for
learning this mapping between $D$ and $I$. TGPR is a structured prediction\(^3\) method that
firstly builds models for the relationships found within $D$ and within $I$ separately, before
learning a regression function $r$ between these preliminary models. This type of struc-
tured learning is a suitable choice for the regression model in the $r$-norm method given
the discussed relationships we are attempting to capture. This training phase of the $r$-

\(^2\)Different transmission channels and/or recording microphones may introduce acoustic mismatches
derived from their own additive or convolutional effects. The recording environments of the training
and testing speech may also differ. Such differences could derive from the additive presence of traffic or
industrial noise or the so-called babble noise created by a group of simultaneous talkers.

54
norm method is shown in Step 1 of the schematic for the whole r-norm process presented in Figure 4.2. In step 1 the Twin Gaussian Process Regression (TGPR) function r is learnt; the arrow here implies capturing the relationship between the development data matrix D and the ideal score matrix I. In step 2 the function r is used to map the raw test scores to adjusted r-norm versions; the arrow here implies a mathematical mapping of scores under the function r. Note that the r-norm model maps the vector of scores of a probe against all enrolled models to a vector of equal length representing the r-norm score of the probe against each enrolled model. In our terminology r maps score vectors to score vectors.

Figure 4.2: Schematic outlining the steps of the r-norm method.

By performing this structured prediction we aim to capture any relationships found within the raw scores D between client models and the scores of a test probe that we postulate exist due to correlations between client models (derived from true similarities in actual speakers voices pending accurate speaker modelling). We then aim to make use of these discovered relationships, held within the regression function r, by mapping the raw scores of unseen test set data under r where these inter-speaker correlations have been accounted for by accentuating target scores and diminishing incorrectly high impostor scores. This mapping is the second and final key stage of the r-norm process and is shown in Step 2 of Figure 4.2.

In machine learning the term structured prediction is used to convey that the method outputs a structured object as opposed to a label or real value. Here, as used in the r-norm process, the TGPR method outputs a function for mapping a raw score vector to a modified or r-normed version.

55
Note that in applying the r-norm model to a certain test probe, we require each probe to be scored against all client models in order to produce a score vector that is mapped under the regression function r. The r-norm process adjusts the score of an test utterance against a model with reference to how the test probe scores against all other client models of the system. This of course increases online computational time during the verification process in direct proportion to the number of enrolled clients, but in most modern automatic systems implemented by average CPUs the scoring of a single utterance against one model is sufficiently quick that this should not be of large concern. The r-norm score of a specific probe against the enrolled model i is then found by reading the i\textsuperscript{th} element of the score vector output by the regression function r.

This describes the novel regression score-post processing idea proposed to learn and compensate for inter-speaker or strictly inter-model relationships or biases. A summary of the method is presented next in subsection 4.2.1. Later, speaker verification experiments are performed and the empirical evidence found regarding the strength of the r-norm model with respect to its ability to improve recognition accuracy is presented in Sections 4.4, 4.5, 4.6 and 4.7.
4.2.1 Outline of the r-Norm method

A summary of the proposed r-norm process is presented below:

1. **Obtain the raw scores** → These are the scores to which we wish to eventually apply the r-norm model we shall learn in the following steps. This implies having specified the set of classes we wish to assign probes to having a model representation to score against for each. We call the vector of scores of a probe against all enrolled models the score vector for the specific probe. As such an ordering of enrolled models is maintained and corresponds to the indexing or elements of the score vector.

2. **Specify the development score matrix** $D$ → This requires the determination of what data from each class will form the development set and scoring each of these data against all of the class models.

3. **Select an ideal score matrix** $I$ → This typically involves either specifying distinct values $i_T$ & $i_{NT}$ or distributions for target and non-target scores or else selecting some scores output from a specific classifier on some relevant data. The size of $I$ is fixed by and equal to that of $D$ as the ideal matrix is in 1-to-1 correspondence element by element with the development score matrix $D$. This is due to the fact that $I$ is abstractly supposed to represent the scores of the development data as assessed not by the actual classifier we are working with, but by some hypothesised, error free super-classifier.

4. **Learn the TGPR regression function** → This function, which we denote by $r$, is a learnt mapping from the development score matrix $D$ to the ideal score matrix $I$. $r$ maps raw score vectors to r-norm score vectors, by which we mean each element represents the r-norm score of the probe against the specific model.

5. **Apply the r-norm model to the raw scores of the test set** → Under the learnt relation $r$, map the raw score vectors to their r-norm versions. To determine the r-norm score of a single probe against a single model, the probe must first be scored against all enrolled models in order to produce a score vector which may then be mapped under the TGPR regression function $r$. Determining the r-norm score for this probe then against any model of interest amounts to reading out the element of the resulting vector output from the mapping at the index corresponding to the desired model/class.
4.3 Contrasting r-Norm with Standard Normalisation Techniques

In this section the r-norm method is compared to and contrasted with existing standard score post-processing techniques from the speaker recognition literature.

Most speaker recognition systems assume an equal (uniform) prior on client speakers and adopt a Bayesian approach for obtaining the posterior probability for the test speech against a client model, as such outputting a likelihood ratio where the numerator is a similarity measure (likelihood of speech data against a client model) and is measured with respect to the expected variability of the speech features over the entire population of (relevant) speakers, that is by the typicality value of the denominator (likelihood of speech data against a world model).

The typical score from an automatic speaker recognition system is a log likelihood ratio, denoted by $\varphi$ in Equation (4.1) for the score of a test probe $X$ between a client model $\lambda_{\text{client}}$ and a Universal Background Model (UBM) $\lambda_{\text{UBM}}$ [327]:

$$
\varphi(\lambda_{\text{client}}, X) = \frac{P(X | \lambda_{\text{client}})}{P(X | \lambda_{\text{UBM}})}
$$

This implicit normalisation by a world model is a standard and logically rigorous process for generating a score. In real world systems further adjustments are often made to this score with thought towards compensating for specific sources of nuisance variation or mismatches present between training and testing utterances. These normalisations are different and involve an explicit distribution scaling of the output scores\(^4\). The sources of nuisance variations may be due to telephony effects from handset microphones or transmission channels, additive noise effects such as babble or environmental noise or from large differences in phonetic content between the utterances.

Whatever the obstacle in each case may be, the aim with all of these score normalisation methods is to increase the recognition accuracy of the classifier. What this explicitly means is increasing the separation between the score distributions of target and non-target trials, as it is the overlaps of these distributions that define the Type I and Type II errors [42, 140]. Approximating the target and non-target (impostor) score distributions with $N(\mu_t, \sigma_t)$ and $N(\mu_i, \sigma_i)$ Gaussians respectively, the system Equal-Error Rate

\(^4\)As discussed in Section 2.6 these differences may be addressed at earlier stages of the statistical comparison process than the score level. Here we are only concerned with what manipulations can be performed to the set of scores output from our classifier for the purpose of increasing accuracy.
(EER) is given by the cumulative standard normal $\Phi (\text{Score}_{EER})$ where

$$\text{Score}_{EER} = \frac{\mu_i - \mu_t}{\sigma_i + \sigma_t} \quad (4.2)$$

The aim then is to minimise the systems EER, and addressing this at the score level the approach of standard score normalisation methods is fundamentally about changing the relative distributions of impostor and target scores. This framework is shown in Figure 4.3. Shown are typical score distributions for target and non-target trials along with an operating point (threshold) marked as ‘Criterion’. With all score post-processing or normalisation methods the aim is to increase the separation between the target and non-target score distributions. An optimal system that makes neither any Type I nor Type II errors necessarily has zero overlap between these two distributions.

Note that this Gaussian approximation on the scores, one that is also made in Figure 4.3, is common and well validated experimentally [31, 324], although long tailed distributions are more common now with state of the art factor based systems [256].

In attempting to adjust the score distributions, common score normalisation methods apply a standard normal $N(0,1)$ transform using a priori parameters that estimate the raw impostor scores curve. Further as it requires much score data to accurately estimate mean and variance statistics, most normalisation methods are impostor centric. Nor-

![Figure 4.3: The “decision landscape” of all biometric systems. The image and phrase are taken from [88] by the pioneer of iris identification John Daugman.](image-url)
malising scores in this approach, working under this assumption that the impostor and target scores are normally distributed, was first proposed in [236] (zero normalisation or z-norm) and is now standard in speech processing. This approach was designed to compensate for inter-speaker variation and was followed by other similar transforms that have proved useful such as test-normalisation (t-norm) [31], and handset-normalisation (h-norm) [324]. The development data used to learn these a priori parameters is dependent on the aims of the normalisation. z-norm compensates for inter-speaker variation by using estimates of the mean and variance of \( \varphi(\lambda_{\text{client}}, \cdot) \) to normalise all probe scores against \( \lambda_{\text{client}} \). t-norm aims to compensate for inter-session differences by performing a standard normal mapping of \( \varphi(\cdot, X) \) that is based on an a priori approximation of the distribution of \( \varphi(\cdot, X) \). These two common approaches are conveyed graphically in Figure 4.4. The purpose of this diagram is to contrast the nature of z-norm and t-norm with the proposed r-norm method which considers the relationships of scores over the whole matrix.

h-norm performs this mapping based on parameters learnt from observations of how the different telephony handsets used in recording the test utterances systematically change the score distributions. Others have been suggested for text-dependent speaker recognition such as utterance normalisation u-norm[156], but unlike r-norm all of these may be grouped under the theory of a \( N(0, 1) \) standard normal mapping.

Figure 4.4: Schematic of the z-norm and t-norm score adjustment processes which operate via a standard normal type mapping of scores on a model-by-model and probe-by-probe basis respectively. LLR is Log-likelihood ratio.

60
The proposed $r$-norm method contrasts with these approaches in that it uses the relations found between client models and development probes through analysing scores of a development data set to then adjust the scores of a test utterance against all client models. It is both different in implementation and also in purpose as it aims to increase performance in a wider set of circumstances by modelling deeper relationships than the aforementioned normalisation methods. A comparison between the $r$-norm process shown in Figure 4.2 and the score matrix of Figure 4.3 shows that while $z$ and $t$-norm act on a model-by-model or probe-by-probe basis respectively, the $r$-norm method makes use of the inter-relationships between probes and models over the entire matrix.

We anticipate that the pure normalisation methods $z$-norm and $t$-norm are still beneficial if implementing $r$-norm when there is any significant and specific mismatch between the development data used for learning the regression function $r$ and the anticipated testing data. In such a situation we predict benefits in applying a $z$-norm mapping (before applying $r$-norm) using parameters for each client model estimated on data that is similar to that used for regression model development. This remains to be validated experimentally. However it is our hypothesis that the $r$-norm method is of use as a modelling tool and is able to improve recognition accuracy independent of the existence of these sources of variation that inspired those earlier normalisation methods. As such we often use the term regression score post-processing and avoid the use of the word normalisation. Also, more so than the other methods, $r$-norm should be useful in a wider range of other pattern recognition domains outside of speaker recognition.

The reasons for the requirement to normalise the scores output by a classifier are varied; it may be to achieve speaker and system independent thresholds, to compensate for nuisance variations that are present within the training and testing speech sets, or to adjust for a mismatch of acoustic conditions between these two sets. These are the typical reasons for the invention and application of pure score normalisation methods discussed.

It may also be however that a clever mapping of scores can reliably increase performance across many situations. This better describes the paradigm of the proposed regression score post-processing method $r$-norm that we now investigate empirically.
4.4 Experiment 1: NIST 2003 SRE Data

The proposed regression score post-processing method $r$-norm is applied to the scores of a GMM-UBM speaker verification experiment on female speakers of the NIST 2003 SRE data. This first empirical test of $r$-norm provides strong support for the hypothesis that it is able to increase recognition accuracy, with a poor initial baseline EER of 19% considerably reduced to 7%. The work in this section was presented at Interspeech in 2013 [399].

4.4.1 Experimental Design

To begin empirically testing the $r$-norm idea we generate scores from an automatic speaker recognition system. Specifically we performed a text-independent speaker verification experiment using a Gaussian Mixture Model (GMM)- Universal Background Model (UBM) system [327] on the 1speaker Female portion of the NIST-2003 Speaker Recognition Evaluation (SRE) data [281]. We now describe first the experimental procedure of obtaining scores from the speaker verification experiment and then the procedure used in applying the $r$-norm model to the scores.

Mel-frequency cepstral coefficients (MFCC) were used as features. MFCC were extracted with 28 filters mel-spaced over the frequency range up to the Nyquist frequency of 4 kHz. We retained the first 12 coefficients, appending the log energy and first order deltas for a 26 dimensional feature vector. Speech frames were Hamming windowed at a length of 25ms and shifted by 10ms. A decibel measure of frame energy was used as a simple voice activity detector and an empirically determined threshold used to remove non-speech frames.

The UBM, comprising 1024 mixtures with diagonal covariance matrices, was trained by the standard Expectation-Maximisation (EM) [99] algorithm for the chicken and egg problem of initially having neither information on the assignment of MFCC frames to mixtures (the latent variable information of the GMM model) nor any reasonable initial estimate on mixture centres or variances. The EM algorithm was started with uniform weights, random diagonal covariance matrices and centroids estimated from a fast implementation of the k-means algorithm [402] with k=1024. The UBM was trained
on the data of all female speakers from the NIST 2000 and 2001 SRE [281] databases. It was initially desired to also use the Switchboard 2 - Phase II [169] data however available computation resources limited us to performing only 10 iterations of EM on this data alone. This is the most significant reason for the weak overall performance of the baseline system reported in Table 4.1.\footnote{The EM algorithm was implemented here in Matlab. In the following chapters training of GMMs is performed via a much faster C implementation [262] which in turn allows model estimation to be performed on larger training sets.}

We deemed this baseline acceptable for the aims of this investigation; namely to explore how well the proposed $r$-norm technique could improve recognition accuracy post obtaining the raw scores\footnote{In preparation for this experiment a similar but preliminary experiment was performed on the small and clean ANDOSL speaker recognition corpus [260]. The ANDOSL data was partitioned into 30 background speakers and 24 enrol/test speakers. An EER of $\ll 1\%$ was achieved with this GMM-UBM system alone and the $r$-norm EER was $\sim 0\%$.}. These results presented here at a minimum demonstrate evidence regarding the benefit of using $r$-norm in circumstances where the modelling has been substandard due to the training data or otherwise.

Speaker models for all 207 female NIST 2003 speakers were adapted from the UBM in the partial Bayesian manner of the maximum a posteriori (MAP) [158] algorithm using the single training utterance for each speaker within the corpus. We considered only closed-set speaker verification and thus removed the test utterances not attributed to any of the 207 clients. This left 1899 testing utterances from which the first 1000 where used for development data (learning the $r$-norm regression model), and the remaining 899 utterances were used for testing. All 207 clients had at a minimum of one utterance (target trial) in both of these development and test sets.

We denote by $\text{score}(\lambda_{\text{client}}, X)$ the score of test trial $X$ having $T$ MFCC feature vectors $\{x_t\}$ for $t = \{1, \ldots, T\}$ against client model $\lambda_{\text{client}}$. This is calculated, where $\lambda_{\text{UBM}}$ is the UBM, as as the base 10 log of the likelihood ratio of Equation (4.1):

$$\text{score}(\lambda_{\text{client}}, X) = \log_{10} \left( \frac{P(X | \lambda_{\text{client}})}{P(X | \lambda_{\text{UBM}})} \right)$$

(4.3)

where the likelihood $P(X|\lambda)$ of $X$ against a generic Gaussian mixture model $\lambda$ of $M$ mixtures with mixture means, covariances and weights ($\mu_i, \Sigma_i, \pi_i$) is given by:

$$P(X|\lambda) = \left( \prod_{t=1}^{T} \sum_{i=1}^{M} \pi_i N(x_t; \mu_i, \Sigma_i) \right)^{1/T}$$

(4.4)

where $N(\cdot; \mu, \Sigma)$ is the multivariate normal density with mean $\mu$ and covariance $\Sigma$.

The geometric mean provides a normalisation for the utterance length $T$, and as
noted in [327], some partial compensation for the patently untrue but highly useful independence assumption on the speech frames $x_t$ that allows $P(X)$ to be calculated as $P(X) = \prod P(x_t)$. The log likelihood ratio of Equation (4.3) is abbreviated as LLR.

We now describe how the TGPR function of the $r$-norm model was learnt, using the MATLAB implementation supplied by the authors of the TGPR method [47]. In this examination of the $r$-norm idea we did not perform any parameter search to optimise the TGPR model, employing only the default TGPR parameters given in the code. The authors knowledge of the use of TGPR for image problems (pose estimation and occlusion detection) in computer vision suggests that a parameter search could be beneficial with respect to the performance of $r$-norm.

We explore two $r$-norm implementations by learning a regression onto two separate ideal score matrices $I$. The first, which we denote as $I_1$, consisted of only $i_{NT} = 0$ impostor scores and $i_T = 1$ target scores. These regression targets may be thought of as zero-variance distributions, and at higher level of abstraction as the output of some hypothesised super classifier that makes zero errors and has zero uncertainty.

In the second exploration (denoted $I_2$), the ideal matrix $I_2$ was based on the actual raw target and non-target score data from the development utterances. The target distribution and non-target distribution means $\mu_T$ and $\mu_{NT}$ were first calculated and $I_2$ set to exactly the development score matrix. Then each target score $s_t$ in $I_2$ was adjusted to $s_t + \mu_t$ and each non-target score $s_{NT}$ in $I$ adjusted to $s_{NT} + \mu_{NT}$. That is shifted target and non-target score distributions were used for $I_2$ and as $\mu_T > 0$ and $\mu_{NT} < 0$, the result of this shift was to reduce the overlap of the two distributions. Indeed this shift of one mean was sufficient to completely separate the target and non-target scores so that again we are regressing towards the output of some hypothesised system that makes no errors, but this time it has some uncertainty or variation in the values that it outputs.

As mentioned, we used the scores from the first 1000 test utterances for learning the TGPR function of the $r$-norm model and validate this model on the unseen scores from the remaining 899 test utterances.

In this first investigation for comparison we also perform test-normalisation ($t$-norm) and zero-normalisation ($z$-norm). The disjoint data used for $t$-norm utterances and $z$-norm GMM model building was taken from Female NIST 2000 SRE speakers. We use 110 utterances for $t$-norm and train 60 speaker models for $z$-norm. We would expect better $t$-norm and $z$-norm results with a larger number of utterances and models respectively [216], however computational resources restricted us to these numbers.

We report EER and detection cost minima. Detection cost minimums are abbreviated
at minDCF and for this experiment are with respect to the detection cost function (DCF) parameters specified in the NIST-2003 SRE, namely $C_{\text{miss}} = 10$, $C_{\text{fa}} = 1$ and prior on detecting a target speaker as $P_{\text{target}} = 0.01$. Of course, system decisions are determined by accepting (rejecting) the identity claims corresponding to scores above (below) some set threshold. Specification of this threshold is often referred to as the operating point as adjusting the threshold makes a trade-off between the FRR and the FAR. For a specific threshold, the verification system will possess equal FAR and FRR; this is the EER that we report.

4.4.2 Results and Discussion

We now report the performance of the GMM-UBM system without any score post-processing and after the application of $z$-norm, $t$-norm and the two $r$-norm instances which we denote $I_1$ and $I_2$. The EER and minDCF values are given for each of these conditions in Table 4.1.

<table>
<thead>
<tr>
<th>Normalisation Method</th>
<th>EER</th>
<th>minDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>18.8%</td>
<td>0.061</td>
</tr>
<tr>
<td>$z$-norm</td>
<td>18.2%</td>
<td>0.068</td>
</tr>
<tr>
<td>$t$-norm</td>
<td>19.6%</td>
<td>0.069</td>
</tr>
<tr>
<td>$r$-norm: $I_1$</td>
<td>7.01%</td>
<td>0.030</td>
</tr>
<tr>
<td>$r$-norm: $I_2$</td>
<td>9.3%</td>
<td>0.036</td>
</tr>
</tbody>
</table>

Table 4.1: EER and minDCF for each normalisation method. $I_1$ and $I_2$ refer to the $r$-norm models learnt on $i_T$ & $i_{NT}$ with zero and non-zero variance respectively.

The proposed $r$-norm score post-processing step has been shown to perform very strongly on the Female data of the NIST-2003 SRE. Using the regression function $r$ learnt on the $I_1$ matrix, which contained only two values, $i_T = 1$ for target scores and $i_{NT} = 0$ for non-target scores, reduced the EER to 7.01%.

Learning the TGPR regression function $r$ on the $I_2$ matrix which represented well separated target and impostor score distributions, but with non-zero variances reduced the EER to 9.3%.\(^7\)

Both of these results were considerably better than the compared normalisation methods $z$-norm and $t$-norm which displayed mixed performance on this task with $z$-norm (18.2%) slightly improving the EER but $t$-norm slightly reducing it (19.6%). As mentioned, the $z$ and $t$ normalisation parameters were estimated using a relatively small number of speaker models and test probes. However although their performance could
likely be improved upon in this specific instance, nowhere in the literature are resulting improvements in recognition performance observed on the same order as that observed here with either r-norm method.

Whilst this comparison between r-norm and other common normalisation methods was performed in this first empirical exploration and reported at Interspeech in [399], the author now believes them to be redundant. Both z-norm and t-norm aim to compensate for specific and precise forms of nuisance variation. This is in contrast with the evolved view of the regression score post-processing method r-norm, which is to view it as an extra modelling step that occurs at the score level.

Detection Error Trade-off (DET) curves are shown in Figure 4.5 for the raw (original GMM-UBM LL ratios) and post-processed scores.

![Figure 4.5: NIST-2003 DET curves for the raw scores of the female-speaker GMM-UBM experiment (red curve) and for each normalisation z-norm (blue, dashed), t-norm (black, dashed) and r-norm with $I_1$ (blue) and $I_2$ (cyan).](image)

We must mention that, with what has been learnt from experimental work performed subsequent to that reported in this section regarding an optimal selection of the $i_T$ & $i_{NT}$ values for r-norm performance, we believe these results could likely be improved upon. The used $i_T$ & $i_{NT}$ values in $I_1$ of 1 and 0 respectively do not match either of the two patterns observed in subsequent experiments and noted in the summary content given in Section 4.8 for this chapter. See that section to note that the $i_T$ & $i_{NT}$ values used in the only two tried formulations of r-norm here (namely conditions $I_1$ and $I_2$) do not fit into either Pattern 1 or Pattern 2 described there. Of course we note again too that we were trying to patch a very leaky vessel in starting from a baseline EER of 18.8% in this experiment and that many proposed algorithms may act as bungs here.
The effect of $r$-norm (with $I_1$) on the target and non-target score distributions is shown in Figure 4.6. A significant factor in $r$-norm improving the recognition performance here is in removing the left skew of the target score distribution. The range of the score values is not significantly altered in this application of $r$-norm.\footnote{This is in contrast to the values resulting from its application to the G.711 condition scores of Section 4.7 shown in Figure 4.11.}

![Figure 4.6: Relative frequency histograms are shown in the top panel for the distribution of raw LLR scores. The effect of $r$-norm ($I_1$) on these distributions is shown below.](image)

It remains to be tested how the proposed $r$-norm method improves the accuracy of a recognition system when the baseline starts from a more acceptable level of performance (EER $< 10\%$) and when applied to more sophisticated, state of the art automatic speaker verification systems. To address this first case, the $r$-norm method is applied to the scores of GMM-UBM systems with $\sim 5\%$ EERs next in Section 4.5.

To address the second point, the $r$-norm method is applied to an i-vector system on NIST-2006 SRE data in Section 4.6. This experiment also addresses the suggestion made in [59] that score normalisation is not a factor in the performance of advanced speaker recognition systems (despite contrary positions [407]). This comment is made from a pure normalisation perspective however, as these systems have modelling methods designed to cope and adjust for nuisance variations that give reason to the requirement for score normalisation. The $r$-norm approach, viewing it as a post-score modelling methodology by using a synthetic ideal score matrix that is designed to leverage inter-speaker differences, should still have a purpose here. The experiments performed in Section 4.6 with an i-vector system on recent and challenging NIST SRE corpora enable conclusions to be drawn regarding this claim.
4.5 Experiment 2: AusTalk

Further empirical evidence for the performance of $r$-norm is provided by gender dependent speaker verification experiments on subsets of 100 speakers of the AusTalk corpora [411, 85]. The GMM-UBM MFCC system baseline EER of 5.26% is reduced to 0.06% in the male experiment, and from 5.60% to 0.07% in the female experiment.

4.5.1 Experimental Design

The new AusTalk database contains a large and growing collection of multi-session, read and spontaneous speech of native Australian speakers [85, 411] recorded over three sessions with a minimum separation of one week.\footnote{AusTalk is an Australia wide Australian Research Council funded project with data collected from at least 14 universities from all over Australia. Note that it is also an audio-visual database.} We performed gender dependent, GMM-UBM speaker verification experiments on subsets of AusTalk participants in order to provide scores to further test the hypothesis that $r$-norm is able to increase the recognition accuracy of our verification system. One hundred (100) speakers of each gender were taken as clients with their ’story’ data from session-1 used for training and their ’interview’ data from session-2 used for testing. The ’story’ speech is non-spontaneous and read from a computer screen, on average running to 4 minutes. The ’interview’ speech was recorded a minimum of one week later and is spontaneous speech recorded from a dialogue between the participant and a research assistant (RA) who simply provided prompts. Ideally, and for the majority of cases, very little (< 10%) RA speech is present within the recordings and these run to 11 minutes on average. All data was downsampled from $f_s = 44.1$ kHz/32 bit to 16 bits at $f_s = 16$ kHz. Table 4.2 provides a breakdown of the recording locations of AusTalk that these speakers were taken from.

<table>
<thead>
<tr>
<th>Recording Location</th>
<th>University</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canberra</td>
<td>ANU</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>Sydney</td>
<td>UNSW</td>
<td>18</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>USYD</td>
<td>28</td>
<td>39</td>
</tr>
<tr>
<td>Melbourne</td>
<td>UMELB</td>
<td>34</td>
<td>26</td>
</tr>
</tbody>
</table>

Table 4.2: Breakdown of the used 100 male and 100 female AusTalk participants.
MFCC features of dimension 32 were used comprising the first 15 cepstral-coefficients + log energy along with the first order deltas of these. These MFCC were warped [297] to target distributions learnt from the ‘story’ speech of 20 disjoint AusTalk speakers\textsuperscript{10} from each gender. Frames were 25ms long with a 10ms overlap and Hamming windowed. Mel-cepstra were extracted with \(\lceil 3 \times \ln (f_s) \rceil\) mel-spaced filters over the bandwidth of \([0, \frac{f_s}{2}]\) Hz.

The ‘story’ data of session one (∼ 3 minutes of speech) was used for client training and a classic GMM-UBM experiment \([327]\) was performed with client GMM consisting of 1024 Gaussians each with diagonal covariance. These were learnt by MAP adaptation of the mixture means only from a gender dependent UBM trained on all 200 of the phonetically diverse sentences recorded by all 54 Australian speakers of each gender present within the ANDOSL corpus \([260]\).

The AusTalk ‘interview’ data was used for testing and consisted of spontaneous speech primarily from the participant but also including prompt questions from the RA. AusTalk participants are recorded via a head-mounted microphone typically located only a few centimetres from their mouth. The vast majority of RA speech is clearly of a lower intensity and was removed via an energy based threshold on the frames. This also implicitly acted as a voice activity detector (VAD) to remove silence frames also. The same method with a lower threshold was used purely as a VAD to remove all non-speech from the ‘story’ data used for client training.

The remaining ‘interview’ data, representing speech only from the participant, were divided into 20 equal length sections for use as target and non-target trials, producing a matrix of scores of dimension 100 × 2000 for each gender, with 2000 target scores and 198000 non-target scores. This was done to create a sufficient number of target scores against each of the 100 client models in order that the raw scores could be partitioned to train an r-norm model and subsequently test it on unseen data. Scores were obtained as the log likelihood ratios of the data between the client and UBM generating hypotheses\([327]\), as per Equation 4.3.

The r-norm equal-error rates (EER) and minDCF in all experiments were found by performing 5 fold cross-validation (CV) on the raw scores, training the TGPR regression model r on 4 of the disjoint folds and testing on the remaining 5th at each of the 5 possible such permutations. All resulting EER and minDCF values are means taken over the 5 folds of the cross-validation process. The ‘pre’ r-norm values in the given tables were taken over the r-norm testing sets, not the original full raw score matrix. Typically the values are slightly worse than those of the full matrix.

\textsuperscript{10}Taken from the University of Canberra portion of the database recordings.
A grid search over a subset of $\mathcal{R}^3$ was performed in order to determine the parameters of the regression target matrix $\mathcal{I}$, namely $i_T$, $i_{NT}$ and a scaling factor $\kappa$ that controls the amount of Gaussian $\mathcal{N}(0,1)$ noise added to these 2 values. Optimal parameters were selected with respect to the development set EER, and this r-norm model was then applied to the held out testing fold of scores at each permutation. In all cases, as in the NIST-2003 experiment of Section 4.4, the training EER was found to be maximised when the r-norm model was trained by regressing from the raw development scores to an ideal matrix $\mathcal{I}$ having only two values $i_T$ and $i_{NT}$ representing ‘ideal’ target and non-target scores respectively. That is the ideal target and non-target score distributions had zero variance.

In all three experiments the detection cost was calculated using the thresholds specified in the NIST-2006 Speaker Recognition Evaluation [4], namely $C_{miss} = 10$, $C_{fa} = 1$ and $P_{target} = 0.01$. In Table 4.3 the minimum of this decision cost function is given under the abbreviation minDCF.

Finally, note that no comparisons are made here between r-norm and t-norm or z-norm as were done in Section 4.4 and published in [399]. This is because, with a greater understanding of the method, we now expect r-norm to consistently outperform those measures and believe that it is not a fair or valid comparison when viewing r-norm as a score level modelling tool of inter-speaker (or inter-class in the general application) variability that aims to increase recognition performance irrespective of the presence of nuisance factors or mismatch conditions.

### 4.5.2 Results and Discussion

Results are given in Table 4.3 in the form of equal-error rates and detection cost minima pre and post r-norm. The r-norm process is able to nearly completely separate the target and non-target score distributions, for both genders, reducing the EERs to almost 0%. The post r-norm equal-error rates of approximately 0.1% in both experiments meant that at this operating point and with 20000 verifications performed (as 2000 target trials and 18000 non-target trials) our r-norm system would have made 2 type 1 errors (false rejects) and 18 type 2 errors (false accepts).\(^{11}\)

The values used for the ideal score matrix are given in the final two columns of Table 4.4, with ideal target scores set to $i_T = 0.6$ and non-target scores to $i_{NT} = 0.4$ for the female experiment and $i_T = 0.55$ and $i_{NT} = 0.35$ for the male experiment. In all experiments these values were found by a coarse grid search over suitable candidate

\(^{11}\) This is only an estimation for illustrative purposes as the r-norm EER reported is the averaged value of the post EER of r-norm on each of the five test folds from the cross-validation process.
Table 4.3: *AusTalk*: EER and minDCF are shown pre and post $r$-norm.

ranges suggested by the range of the values within the raw scores. Optimal values were determined with respect to the resulting five-fold CV-averaged EER. Note that initially a third parameter was searched over, namely a scaling of some additive $\mathcal{N}(0, 1)$ noise about each of the $i_T$ and $i_{NT}$ values. In both experiments however the EER was minimised with this noise scaling parameter equal to 0. That is to say then that both $r$-norm models were learnt on an ideal set of scores having just two distinct values, namely $i_T$ and $i_{NT}$.

Also shown are the values of the EER threshold $\hat{x}$ prior to $r$-norm.

The remaining columns of Table 4.4 provide information on the distributions of the raw target and non-target trials (that is prior to $r$-norm). The mean ($\mu$) and median ($\tilde{x}$) measures of centre are given along with the standard deviation ($\sigma$) as a measure of spread and the skewness ($\gamma_1$), calculated as the third standardised moment. These statistics are apparently important in informing the optimal choice of values for the ideal matrix $I$ in the $r$-norm training process, more about which is said in the chapter summary of Section 4.8.

As is typical of log likelihood ratios coming from GMM-UBM models, the target and non-target score distributions are both well approximated by Gaussians, being unimodal and having very little skew. This has a predictable effect upon the locations of the $i_T$ & $i_{NT}$ values, about which observations are made regarding $r$-norm in a general nature in the summary statements of Section 4.8.

<table>
<thead>
<tr>
<th></th>
<th>Pre $r$-norm</th>
<th>Post $r$-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>minDCF</td>
</tr>
<tr>
<td>Male</td>
<td>5.26%</td>
<td>0.021</td>
</tr>
<tr>
<td>Female</td>
<td>5.60%</td>
<td>0.029</td>
</tr>
</tbody>
</table>

Table 4.4: *AusTalk* experiment summary statistics for the scores of target ($T$) and non-target ($NT$) trials. Given are the mean $\mu$, median $\tilde{x}$, standard deviation $\sigma$ and skewness $\gamma_1$. The last three columns give the values for the location of the pre $r$-norm EER threshold $\hat{x}$ as well as the found optimal ideal matrix parameters $i_{NT}$ & $i_T$ used in the obtaining the $r$-norm results reported in Table 4.3.

<table>
<thead>
<tr>
<th>Scores</th>
<th>$\mu$</th>
<th>$\tilde{x}$</th>
<th>$\sigma$</th>
<th>$\gamma_1$</th>
<th>$\hat{x}$</th>
<th>$i_{NT}$</th>
<th>$i_T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target</td>
<td>0.667</td>
<td>0.662</td>
<td>0.169</td>
<td>0.265</td>
<td>0.39</td>
<td>-</td>
<td>0.55</td>
</tr>
<tr>
<td>Non-Target</td>
<td>0.189</td>
<td>0.192</td>
<td>0.128</td>
<td>0.128</td>
<td>0.35</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target</td>
<td>0.660</td>
<td>0.657</td>
<td>0.144</td>
<td>-0.077</td>
<td>0.43</td>
<td>-</td>
<td>0.60</td>
</tr>
<tr>
<td>Non-Target</td>
<td>0.231</td>
<td>0.230</td>
<td>0.128</td>
<td>0.063</td>
<td>0.40</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Shown in Figure 4.7 is a surface plot of resultant $r$-norm EER values for different ideal score matrices $I$ as specified by the parameter pair $(i_T, i_{NT})$ when applied to the MFCC scores of the male speakers. EER values are shown only for a subset of the halfspace $i_T > i_{NT}$. EER values are seen to be approximately equal at all slices parallel to the line $i_T = i_{NT}$ and to increase as these slices move away from this line. This suggests that searching along the line $i_T = -i_{NT}$ is sufficient for finding near optimal $I$ parameters for the $r$-norm method when both target and non-target distributions are well approximated by Gaussians.

![Surface plot of $r$-norm EER values](image)

Figure 4.7: Post $r$-norm EER surface over the parameter space of $i_T$ & $i_{NT}$ values for the male MFCC system.

The strength of these results was anticipated due to several favourable factors which include the amount of speech available for both model training and testing and the quality of the recordings. Indeed a state of the art factor analysis system, if additional Australian speech data was procured to appropriately learn the speaker and channel or total variability subspaces, would likely achieve similarly low EER and detection cost values without requiring any application of an $r$-norm model. What has been supported with evidence here however is the claim that $r$-norm, when given a significant number of target score vectors per enrolled model, is able to considerably increase recognition accuracy.
4.6 Experiment 3: NIST 2006 SRE Data

An r-norm model is applied to the scores of a state-of-the-art i-vector system on a selection of the NIST-2006 SRE core condition data. A baseline EER for the i-vector system modelling MFCC features of 6.64% was achieved. By segmenting the test trials in order to generate enough scores to suitably train an r-norm model, the EER was able to be reduced to 2.44%.

4.6.1 Experimental Design

In this experiment, an i-vector model [95] with cepstral features was used with clients from the male core-condition data of the NIST-2006 Speaker Recognition Evaluation (SRE) [4]. The experimental design for this speaker verification experiment, performed to generate more score data to further investigate the r-norm model, is described now.

Mel-frequency cepstral coefficients (MFCC) were used as speaker features with 27 mel-spaced filters. The log energy + leading 20 cepstral coefficients were retained along with their Δ and ΔΔ resulting in a 63 dimensional feature vector. Frames were Hamming windowed at 25ms and incremented by 10ms shifts. An energy threshold was used as a voice activity detector to discard non-speech frames. Feature warping [297] was also performed to cepstral target distributions learnt on the CMU Arctic database [228]. A Universal-Background-Model (UBM) was trained as a Gaussian-Mixture-Model (GMM) having 1024 mixtures each with a diagonal covariance.

The total variability space $T$ of the i-vector extractor had 400 factors and $T$ and the UBM were trained on the male speakers of the following collection of databases:

- NIST-2005 SRE [281]
- Switchboard 2 - phase 2 [169]
- Switchboard 2 - phase 3 [169]

With consideration towards applying r-norm the following deviations were taken from the NIST-2006 core condition: the test probes of the NIST-2006 core condition not belonging to one of the enrolled speakers were removed in order to perform closed-set recognition. The speech file of each test trial was then divided into 5 and 10 chunks such that each enrolled model had enough target scores for five-fold cross-validation (CV) testing of the r-norm process. This left 3445 and 6890 trials in the two cases respectively.
All test utterances were scored against all enrolled models. Without performing this segmentation of test data files there would not have been enough target trials to train an r-norm model and then validate it on unseen data. For brevity, we will refer to these two experiments as the 5-chunk and 10-chunk conditions.

Cosine similarity scoring [93] was used in evaluating (enrol, probe) i-vector comparison pairs.

In order to test the performance of the r-norm model, five fold CV was used in the following process: the score vectors (again, meaning the vector of scores of each test utterance against all enrolled models) were divided into five disjoint groups each having the same number of target trials for each model. Then at each of the five permutations, four of these groups (‘the development set’) were used to train an r-norm model and the held out fifth group (‘the test set’) was used to validate this trained model. The parameters of the ideal matrix providing the regression target in training the r-norm model were optimised with respect to the EER on the development set itself. With our understanding of the past performance of the r-norm model, the ideal matrix was chosen to have only two distinct values, namely $i_T$ & $i_{NT}$; no measure of spread was incorporated into the ideal scores forming the regression target in learning the twin-GPR function r.

Average equal-error rates and detection cost minima pre and post the application of the learnt r-norm model are reported, with these values calculated as the average over the five cross validation test sets. The detection cost was calculated using the thresholds specified in the NIST-2006 SRE [4], namely $C_{miss} = 10$, $C_{fa} = 1$ and $P_{target} = 0.01$. The abbreviation minDCF is used for the minimum of the detection cost function.

4.6.2 Results and Discussion

In Table 4.5 equal-error rates (EER) and minDCF are given for the raw i-vector system (pre) and for the r-norm method applied to these scores (post). Both of these performance measures degrade as the test utterances are chunked into smaller pieces, in line with the decreasing information content of each probe. For reference, the EER of the same i-vector system without chunking the test probes was 6.6%. Thus, in some sense chunking the data into 5 and 10 pieces in order to train the r-norm model was validated by the resulting EERs of 2.4% and 3.8% respectively. It remains a hypothesis given the available data here that the post r-norm EER could be lowered further if more test utterances of a suitable length were able to produce more score vectors from the i-vector system in order to train a superior r-norm model. Detection cost minima are shown to improve similarly.
Table 4.5: NIST-2006: EER and minDCF values pre and post r-norm are shown for the i-vector system on the NIST-2006 SRE data. As described in 4.6.1, the test utterances were divided into 5 and 10 chunks in two separate experiments in order to provide enough score data to train the r-norm model.

Summary statistics of the pre r-norm target and non-target score distributions are given in Table 4.6 for both the 5 and 10 chunk conditions. Also shown for both conditions is the location of the EER threshold \( \hat{x} \) and the optimal values found for \( i_T \) & \( i_{NT} \). The shape of the target and non-target score distributions is central to determining the locations of \( i_T \) & \( i_{NT} \), which are in turn central to the performance of the r-norm method. Observations are made regarding this relationship in the concluding remarks of Section 4.8.

<table>
<thead>
<tr>
<th>Scores</th>
<th>( \mu )</th>
<th>( \hat{x} )</th>
<th>( \sigma )</th>
<th>( \gamma_1 )</th>
<th>( \hat{x} )</th>
<th>( i_{NT} )</th>
<th>( i_T )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 chunks Target</td>
<td>0.274</td>
<td>0.267</td>
<td>0.110</td>
<td>0.307</td>
<td>0.13</td>
<td>-</td>
<td>0.30</td>
</tr>
<tr>
<td>5 chunks Non-Target</td>
<td>0.038</td>
<td>0.034</td>
<td>0.069</td>
<td>0.417</td>
<td>0.10</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>10 chunks Target</td>
<td>0.221</td>
<td>0.214</td>
<td>0.098</td>
<td>0.329</td>
<td>0.11</td>
<td>-</td>
<td>0.40</td>
</tr>
<tr>
<td>10 chunks Non-Target</td>
<td>0.031</td>
<td>0.028</td>
<td>0.063</td>
<td>0.305</td>
<td>0.10</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.6: NIST-2006: Summary statistics of the scores of target and non-target trials for the segmentation of test probes into 5 and 10 pieces: mean(\( \mu \)), median(\( \hat{x} \)), standard deviation(\( \sigma \)) and skewness(\( \gamma_1 \)) for target and non-target scores. Also given are the EER threshold \( \hat{x} \) and optimal values found for \( i_T \) & \( i_{NT} \).

A graph of these distributions is presented in Figure 4.8, plotting relative frequency histograms for the target and non-target scores pre (top) and post r-norm. The range of values of the scores post r-norm is greatly reduced in this instance. The optimal values found for \( i_T \) and \( i_{NT} \) are plotted as black vertical lines on either side of a green line marking the r EER threshold value \( \hat{x} \).

Detection error trade-off (DET) curves are shown in Figure 4.9; there are 5 curves for both pre & post r-norm corresponding to the 5 folds of cross validation and these all show consistent improvements at all operating points with the red post r-norm curves displaying little variability. Both Figures 4.8 & 4.9 relate to the 10-chunk condition.
Figure 4.8: NIST-2006 Relative frequency histograms pre & post $r$-norm for the target and non-target scores. Marked on the pre curves is the location of EER threshold $\hat{x}$ prior to $r$-norm and on either side are the optimal values $i_T$ & $i_{NT}$ found for the ideal matrix.

Figure 4.9: NIST-2006 DET curves pre(blue) & post(red) $r$-norm for each of the test sets across the 5 folds of cross-validation.
4.7 Experiment 4: Wall Street Journal - Phase II

An i-vector system was used to model MFCC features in a speaker verification experiment on the Wall Street Journal - Phase II database. Four sub-experiments were performed with both training and testing speech (1) in original 16kHz wideband format, (2) downsampled to 8kHz, (3) passed through a AMR-WB wideband mobile codec and (4) passed through a narrowband G.711 landline codec. Cross-validation was used to apply $\mathbf{r}$-norm to the scores from these experiments. Original EERs of 3.6%, 1.6%, 4.4% and 3.9% are reduced to 1.4%, 0.5%, 0.5% and 0.1% respectively.

Disclaimer:
This section involves work done by my colleague Laura Fernández Gallardo, who was studying the effect of communication channels on speaker recognition. The division of labour was as follows: Laura performed all work to the point of generating scores for these 4 sub-experiments at which point the scores were then passed onto the author of this thesis in order to apply the score post-processing technique $\mathbf{r}$-norm. Everything written here in describing the process of obtaining these scores is in the author’s own words. These 4 sub-experiments, without any mention of the $\mathbf{r}$-norm process, will contribute towards Laura’s doctoral studies.

Laura Fernández Gallardo’s doctoral studies are jointly supervised by Professor Michael Wagner at the University of Canberra and Professor Sebastian Möller of Telekom Innovation Laboratories, TU Berlin, Germany.

Her email address is: Laura.FernandezGallardo@canberra.edu.au
4.7.1 Experimental Design

An i-vector system [95] was used to perform a speaker verification experiment on the Wall Street Journal Continuous Speech Recognition Phase II (WSJ1) database [296, 1, 3]. Frames were Hamming windowed at a length of 25ms with a 10ms shift. From these, 63 dimensional MFCC features were extracted by taking the first 20 cepstral coefficients and log energy along with their Δ and ΔΔ. ⌈3 × ln (fs)⌉ mel-spaced spectral filters were used, where $f_s$ is the sampling frequency, and the MFCC were not warped at all [297].

A UBM comprising 1024 mixtures, and a total variability matrix $T$ with 400 factors were trained on the collection of the following corpora whose combination amounted to 648 male speakers and approximately 50 hours of speech:

- TIMIT Acoustic-Phonetic Continuous Speech Corpus, [157]
- North American Business News Corpus, [3, 168]
- Resource Management Corpus 2.0 Part 1, [3, 304]
- WSJ Continuous Speech Recognition Phase I (WSJ0), [3, 296]

Enrol and test data came from the WSJ1 database [1] which contains 134 male speakers with 10 sentences per speaker. Five sentences were employed for enrolment and the remaining five for testing. Experiments were performed with the data at different sampling rates or having been passed through certain codecs designed to simulate mobile or landline telephony transmission conditions. There were four conditions and these are shown in Table 4.7.

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>NB</td>
<td>Narrowband</td>
<td>8 kHz</td>
</tr>
<tr>
<td>2.</td>
<td>WB</td>
<td>Wideband</td>
<td>16 kHz</td>
</tr>
<tr>
<td>3.</td>
<td>AMR-WB</td>
<td>Wideband</td>
<td>16 kHz</td>
</tr>
<tr>
<td>4.</td>
<td>G.711</td>
<td>Narrowband</td>
<td>8 kHz</td>
</tr>
</tbody>
</table>

Table 4.7: Conditions under which the Wall Street Journal - Phase II data was used. Enrol and test data were always matched with respect to these conditions.

Note that the WB condition corresponds to the natural state of the WSJ1 database. The codecs were applied via software simulation, which implemented standard International Telecommunication Union (ITU) and Third Generation Partnership Project (3GPP) tools. Before the coding processes, the signal was band-limited to either narrowband or wideband by applying channel filters complying with the ITU-T Recommendations G.712 and P.341 respectively. In each of the four cases the enrolment and testing
sentences were matched with respect to these four conditions. As mentioned, these procedures were initially performed in order to investigate hypotheses outside of this thesis but, with multiple target trials, allows \( r \)-norm to be investigated here in more conditions, providing more empirical evidence regarding its performance in different circumstances.

Enrol and probe i-vector pairs were scored via cosine similarity scoring [93], where smaller angles between i-vector pairs residing in identity space provide stronger evidence for the hypothesis that the probe supervector derived from the observed feature data was generated from the same latent identity variable; that is the same-speaker hypothesis.

Taking the scores from each of the four experiments, the following five fold cross-validation process was used to investigate the performance of \( r \)-norm. The raw cosine similarity scores from the i-vector system were divided into a development set and a testing set, with the development set containing four of each client’s probes scored against all enrolled models, and the test set holding the remaining fifth probe for each client scored against all models. This selection was permuted over each of the five possibilities.

For each cross-validation fold the \( r \)-norm model was learnt between the development data and a constructed ideal score matrix \( I \), and then applied to the disjoint test set. Equal-error rates (EER) reported as 'pre' \( r \)-norm values are the EER of the raw i-vector test score set, averaged over the five folds. The reported 'post' \( r \)-norm EER is the EER of the scores mapped under the \( r \)-norm regression function, averaged over the five distinct test sets coming from the cross-validation process. Minimum detection costs values are also calculated, using the thresholds specified in the NIST-2006 SRE [4], namely \( C_{\text{miss}} = 10, C_{\text{fa}} = 1 \) and \( P_{\text{target}} = 0.01 \). The abbreviation \( \text{minDCF} \) is used to refer to this.

The parameter selection of ideal score matrix \( I \) values was achieved via a grid search minimising the EER of the post \( r \)-norm development scores. Values are fixed over the 5 folds of cross-validation performed, and for a specified value pair \( i_T \& i_{NT} \), the averaged EER is calculated on the development scores at each fold and averaged. This is the value minimised in performing the optimisation for the parameter search. Again, as implied, the ideal matrix \( I \) is constructed to contain only 2 values, \( i_T \) for target trial scores and \( i_{NT} \) for non-target trial scores; our regression target in learning the \( r \)-norm TGPR model had zero noise about the target and non-target values.

### 4.7.2 Results and Discussion

Table 4.8 shows the pre and post \( r \)-norm EER and \( \text{minDCF} \) values (averaged across the five cross-validation folds) for each of the 4 speech data formats within the WSJ1 exper-
ments. The r-norm method is again observed to perform very strongly with respect to both EER and detection cost minima. Both of these measures were reduced considerably in all four data conditions.

<table>
<thead>
<tr>
<th></th>
<th>Pre r-norm</th>
<th>Post r-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>minDCF</td>
</tr>
<tr>
<td>NB</td>
<td>3.60%</td>
<td>0.026</td>
</tr>
<tr>
<td>WB</td>
<td>1.58%</td>
<td>0.015</td>
</tr>
<tr>
<td>AMR-WB</td>
<td>4.37%</td>
<td>0.013</td>
</tr>
<tr>
<td>G.711</td>
<td>3.88%</td>
<td>0.015</td>
</tr>
</tbody>
</table>

Table 4.8: WSJ1: EER and minDCF pre and post r-norm for the 4 experiments performed on the WSJ1 data.

Score distribution statistics, which provide some insight into the location of the optimal values for $i_{NT}$ & $i_T$ are presented in Table 4.9. Given are the statistics for target and non-target trials of the full raw score matrix output by the i-vector system.

<table>
<thead>
<tr>
<th>Scores</th>
<th>Type</th>
<th>$\mu$</th>
<th>$\bar{x}$</th>
<th>$\sigma$</th>
<th>$\gamma_1$</th>
<th>$\hat{x}$</th>
<th>$i_{NT}$</th>
<th>$i_T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Narrowband</td>
<td>Target</td>
<td>0.649</td>
<td>0.646</td>
<td>0.110</td>
<td>-0.035</td>
<td>0.45</td>
<td>- 0.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-Target</td>
<td>0.198</td>
<td>0.188</td>
<td>0.138</td>
<td>0.305</td>
<td>- 0.42</td>
<td>- 0.47</td>
</tr>
<tr>
<td>2.</td>
<td>Wideband</td>
<td>Target</td>
<td>0.699</td>
<td>0.699</td>
<td>0.093</td>
<td>-0.226</td>
<td>0.49</td>
<td>- 0.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-Target</td>
<td>0.188</td>
<td>0.182</td>
<td>0.125</td>
<td>0.454</td>
<td>- 0.45</td>
<td>- 0.50</td>
</tr>
<tr>
<td>3.</td>
<td>AMR-WB</td>
<td>Target</td>
<td>0.889</td>
<td>0.898</td>
<td>0.043</td>
<td>-3.279</td>
<td>0.82</td>
<td>- 0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-Target</td>
<td>0.714</td>
<td>0.727</td>
<td>0.077</td>
<td>-0.790</td>
<td>- 0.82</td>
<td>- 0.87</td>
</tr>
<tr>
<td>4.</td>
<td>G.711</td>
<td>Target</td>
<td>0.671</td>
<td>0.689</td>
<td>0.103</td>
<td>-2.120</td>
<td>0.48</td>
<td>- 0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-Target</td>
<td>0.220</td>
<td>0.210</td>
<td>0.150</td>
<td>0.178</td>
<td>- 0.20</td>
<td>- 0.90</td>
</tr>
</tbody>
</table>

Table 4.9: WSJ1: Summary statistics of the scores of Target and Non-Target trials for each sub-experiment. These values are taken prior to r-norm and relate to the selection of the optimal ideal score matrix values. Given are the mean($\mu$), median($\bar{x}$), standard deviation($\sigma$) and skewness($\gamma_1$) values for target and non-target scores. Also given are the EER threshold $\hat{x}$ and found optimal values for $i_T$ & $i_{NT}$.

Figures 4.10 and 4.11 provide a graph of the summary statistics for the wideband and G.711 data conditions, which corresponds to lines two and four of Table 4.9. Shown are the relative frequency histograms of the target and non-target scores before and after r-norm. Marked with a vertical green line on the before curves is the location of the EER threshold $\hat{x} = 0.484$. Also marked are the optimal ideal matrix values $i_T$ & $i_{NT}$ as found over the development sets in the cross-validation process. The post r-norm distributions shown in the lower panel of each figure plot the results from the first cross-validation.
fold; their shape and location is typical of the result from each of the five folds.

Figure 4.10: WSJ1 Relative frequency histograms pre & post r-norm for the target and non-target scores of the wideband (WB) condition. Marked on the pre curves is the location of EER threshold $\hat{x} = 0.49$ prior to r-norm and on either side are the optimal values $i_T$ & $i_{NT}$ found for the ideal matrix. Note that the i-vector distributions are reasonably symmetric.

The range of the scores after the application of r-norm in the G.711 condition is increased greatly, from $[-1, 1]$, as constrained by the cosine angular measure, to $\sim [-20, 120]$. This is demonstrated in the lower panel of Figure 4.11. The interpretation of the resulting r-norm score is simply that it is just a score which may be compared against a threshold by some classifier in performing a given recognition task. Calibration [61, 58] of these scores must be performed if it is to be interpreted in any probabilistic terms, e.g. as a likelihood ratio\(^\text{12}\).

The distribution of the scores of non-target trials from the i-vector system was found to be approximately symmetric across all four experiments, with the skewness ranging

\(^\text{12}\)As it must from a cosine similarity score in any case. Further even, most systems which actually calculate notionally a score as the ratio of two likelihoods often produce output that doesn’t display the properties of a likelihood ratio and must then be calibrated if this behaviour is desired [398]. This is of vital importance in a forensic context, but typically not in a speaker verification system. The point is, in this context, we are concerned only with classification accuracy.
Figure 4.11: WSJ1 Relative frequency histograms pre & post \( r \)-norm for the target and non-target scores of the G.711 condition. Marked on the pre curves is the location of EER threshold \( \hat{x} = 0.48 \) prior to \( r \)-norm and on either side are the optimal values \( i_T \) & \( i_{NT} \) found for the ideal matrix. Note the left skew of the distribution of i-vector target scores.

from \( \gamma_1 = -0.79 \) to \( \gamma_1 = 0.45 \). The distribution of scores from target trials was found to be either approximately symmetric (narrowband \( \gamma_1 = -0.04 \) and wideband conditions \( \gamma_1 = -0.23 \)) or left skewed (AMR-WB \( \gamma_1 = -3.28 \) and G.711 \( \gamma_1 = -2.12 \)). This was empirically observed to have an affect on the location of the found optimal (with respect to development EER) \( i_T \) & \( i_{NT} \) values, patterns which were evident over all experiments reported in this chapter. As it relates to the \( r \)-norm method in general, it is addressed in the chapter summary found in Section 4.8.
4.8 Chapter Summary

Primarily we note that in all cases the proposed regression score post-processing method \( r \)-norm was able to lower equal-error rates and detection cost minima, often considerably. Table 4.10 provides an overview of the experimental results regarding the application of the \( r \)-norm method to the scores of several speaker recognition systems on various corpora as outlined in this chapter.

<table>
<thead>
<tr>
<th>Database</th>
<th>System</th>
<th>Experiment</th>
<th>pre EER %</th>
<th>post EER %</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIST-2003 SRE</td>
<td>GMM-UBM</td>
<td>Female</td>
<td>18.8</td>
<td>7.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male</td>
<td>5.26</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Female</td>
<td>5.60</td>
<td>0.07</td>
</tr>
<tr>
<td>AusTalk</td>
<td>GMM-UBM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIST-2006 SRE</td>
<td>i-vector</td>
<td>5-chunks</td>
<td>9.06</td>
<td>2.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10-chunks</td>
<td>11.47</td>
<td>3.80</td>
</tr>
<tr>
<td>Wall Street Journal</td>
<td>i-vector</td>
<td>NB</td>
<td>3.60</td>
<td>1.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WB</td>
<td>1.58</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AMR-WB</td>
<td>4.37</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G.711</td>
<td>3.88</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Table 4.10: *Summary of the performance of \( r \)-norm over all experiments performed in Chapter 4. Equal-error rates are given for the base classification system and post the application of \( r \)-norm.*

Performance, with the EER as a metric, was improved on both classic Gaussian mixture model systems and state of the art factor analysis systems such as the i-Vector models employed in the NIST-2006 and Wall Street Journal experiments. Performance gains were similar in respect to lowering the detection cost minima, as can be seen in Tables 4.1, 4.3, 4.5 and 4.8.

Some observations may be made regarding the behaviour of the location of the found optimal ideal matrix values \( \mu_{\text{non-target}} \), \( i_{NT} \), \( \hat{x} \), \( i_{T} \), and \( \mu_{\text{target}} \). These values empirically were always found in one of two patterns. The first pattern is that they are located on either side of the EER threshold location \( \hat{x} \) and well inside the modes of the non-target and target distributions. The upper panel of Figure 4.10 is a good example of this first pattern, which is frequently observed when both target and non-target score distributions are symmetric, or occasionally when both are non-symmetric with the same direction of skewness. The template for this pattern, with respect to ordering, may be described as:

**Pattern 1:** \([\mu_{\text{non-target}}, i_{NT}, \hat{x}, i_{T}, \mu_{\text{target}}] \)

The second observed pattern regarding the location of the found optimal ideal matrix parameters is that they are located beyond the modes of the non-target and target score
distributions. This is evidenced in Figure 4.11 for the G.711 data condition of the Wall Street Journal-Phase II i-vector experiments. Here the target scores are highly left skewed ($\gamma_1 = -2.120$) and this pattern regarding the location of the found optimal ideal score matrix values is generally observed when one of the score distributions is symmetric but the other is highly skewed. This pattern may be categorised by the following ordering:

**Pattern 2:** $[i_{NT}, \mu_{\text{non-target}}, \hat{x}, \mu_{\text{target}}, i_T]$

Knowledge of these patterns may be informative in restricting the parameter space within which we expect to find some optimal values for the parameters $i_T$ & $i_{NT}$. Often however the skew of the target distribution is the only important factor. This is evidenced in the NIST-2006 experiments of Section 4.6 where a Pattern 1 style is found for $i_T$ & $i_{NT}$ although the non-target distribution is slightly right skewed. This skew is minimal however.

One may also backtrack in order to give fuller consideration to the choices to be made in implementing an $r$-norm model with regard to what development data is used for creating the raw score matrix $D$ and what the ideal score matrix $I$ should be. These should be informed by both the nature of the testing data and what the aims in applying $r$-norm are. In some circumstances, typically when various data is readily available, it may be beneficial to create $D$ and $I$ from the scores of certain classifiers on data of specific conditions or degradations. Our empirical results suggest that creating matrix $I$ with only the two distinct values $i_T$ & $i_{NT}$ works well when we believe that the raw scores generated from our classifier will be distributed in the same manner as those used for $D$ in learning the $r$-norm model. When we do use only the two distinct values $i_T$ & $i_{NT}$ in $I$ we have found it useful to think of such a construction as being the score output by a hypothesised super-classifier in the sense of making no errors and having zero uncertainty.

The experiments performed in this chapter have used the $r$-norm method from the viewpoint of score post-processing to improve recognition rates, which we view as a type of speaker modelling that uses scores as pseudo features. The $r$-norm method may also focus on normalisation alone, where the emphasis is not on boosting system performance by capturing correlations between client models and test probe scores that relate to interspeaker variability, but on compensating and overcoming specific mismatch conditions between training and testing, more like $z$-norm and $t$-norm. A potential configuration of the $r$-norm system for dealing with large differences between training and testing speech could be selecting the development data used in forming the raw score matrix $D$
to match as well as possible the anticipated testing data type and basing the ideal score matrix on scores derived from clean data (or data well matching that used to train client models). This and other conceivable permutations remain to be tested.

As another example, the testing data in all experiments in this chapter used to validate the $r$-norm model (coming from a partition or cross-validation on the same database), whilst completely disjoint from speaker model training and $r$-norm development data, presumably shared many acoustic characteristics with the development data that was used to generate the raw score matrix that the TGPR function $r$ was learnt on. Other future work in developing further the $r$-norm method and demonstrating it experimentally should focus on cases where there is no a priori information as to what the characteristics of the testing speech will be, necessitating that the development data set should be large and acoustically varied, and/or that the matrix $I$ should be representative of a $z$-norm mapped score matrix and that the test scores should undergo $z$-norm before applying $r$-norm. There are many mismatch scenarios that have several choices for combinations of development data and ideal matrix and in each case there exist theoretically justifiable reasons for the choices of $D$ and $I$ that remain to be tested experimentally.

It is important to remember, however, that the scores and whatever patterns, trends or relationships that exist within them across the enrolled models are the only central considerations for the $r$-norm method. The class modelling methodology, evaluation procedure for trials or properties of the data are irrelevant for the $r$-norm method beyond the effect these and other variables have on the distribution and inter-relationships of their pre-$r$-norm scores. With the empirical evidence presented in this chapter, so long as the scores contain some information regarding the inter-relationships of class models we expect the $r$-norm method to perform well.

To close we briefly address some further issues regarding the $r$-norm method that remain to be tested or developed.

- Firstly, this has been an empirical validation of the proposed algorithm and we have not made any theoretic statements regarding conditions for optimal or expected performance. Bounds on $r$-norm error rates, an analytic specification of the ideal matrix $I$ for minimum classification error, an optimal relationship between the number of classes and the required amount of development score data for training the $r$-norm model are just some of the conceivable theoretical insights that would be of obvious practical value. Monte Carlo simulation with synthetic scores and
an analysis of the twin-Gaussian process regression model as used in the r-norm algorithm may be beneficial for developing such theory.

○ It would be of interest to the automatic speaker recognition community to observe the performance of the r-norm method when applied to PLDA compensated i-vector comparisons [305, 256, 220], as such a system represents the absolute current state of the art in the field.

○ No optimisation was performed in any of the experiments reported in this chapter with respect to the parameter selection of the twin-Gaussian process regression (TGPR) model that we employ in the r-norm method. The author’s knowledge of the use of the TGPR model in pose estimation problems within computer vision suggests that this would be a beneficial additional stage in learning the TGPR model.

○ The proposed r-norm method is of course applicable to all classification tasks that make the determination of class membership of given test input based upon some quantitative comparison between its features and the class models. How the r-norm method performs in domains other than speaker recognition remains to be established\(^{13}\).

○ The r-norm method remains to be extended to open set recognition tasks. Potential methods for achieving this that remain to be explored could be including a world model within our class labels that out-of-set probes can be assigned to or developing a threshold on the r-norm scores below which the probe is labelled out-of-set.

○ In certain domains the interpretation of the magnitude of the probe-model evaluation is of interest or even vital importance. Tasks involving the quantitative assessment of forensic evidence are (or certainly should be) quintessential examples of this latter case. This is not an easy task with regard to the values coming from an r-norm system. The key to overcoming it likely lies in the process of calibration [58, 57], and in the field of speaker recognition is a common final step for aiding the interpretation of the typical modern scoring system [398].

\(^{13}\)An informal experiment was performed with positive results on the task of identifying file formats (.wav/.txt/.pdf etc) from given file fragment data. Positive results meaning a pre r-norm misidentification rate of \(\sim 3\%\) was reduced to \(\sim 2\%\).
Chapter 5

Glottal Waveforms: Text-Independent Speaker Recognition

5.1 Introduction

Having established in the speaker recognition literature review of Chapter 2 that any speaker information from the action at the larynx is generally an underutilised signal in automatic speaker recognition, in this chapter we report on various speaker recognition experiments making use of the speaker’s estimated glottal flow waveforms.

Glottal estimates are derived in general from speech recorded in clean environmental conditions, acknowledging that estimation methods are not robust to noise and phase distortions to the speech signal. Under such circumstances we aim to explore the extent of speaker dependent information contained within the glottal flow signal and to then demonstrate that such information is beneficial in the sense that it can be used to improve upon the accuracies of systems employing magnitude spectral information of the speech signal alone (i.e. MFCC) which primarily relate to the vocal-tract configuration.

Used are data-driven parameterisations of the voice-source waveform under the hypothesis\(^1\) that these approaches better capture the speaker dependent idiosyncrasies that are essential to the speaker recognition task. This includes several studies of the source-frame feature introduced in Section 3.5, and to begin an investigation into the promising but since unmentioned cepstral coefficient representation of the voice-source [171].

\(^{1}\)And guided by existing literature [117, 171, 389].
5.2 Experiment 1: Replication of Voice-Source Cepstrum Coefficients for Speaker Identification

In this initial experiment the speaker identification results presented in [171] were replicated on the YOHO speech corpus. The scores are also interpreted in a verification paradigm and EER reported. This paper had introduced a promising and novel means of representing the glottal signal in the cepstral domain that has not been reported on further by the authors or elsewhere in the literature. With 108 speakers a misidentification rate of 6.76% is achieved for the MFCC system alone, 10.99% by the voice-source cepstrum coefficient system alone, and 5.42% for a combined system using score fusion. These misidentification rates obtained by the method are lower than those reported in the original paper.

5.2.1 Introduction

A method of inferring a cepstral representation of the voice source waveform was presented in [171] where the features were then used in a speaker identification experiment on the YOHO corpus [71]. Our motivation for this initial exploration of the use of voice source information for speaker recognition is derived from the absence of any further literature published regarding the method by the authors or others, and it was felt that replicating the method would be informative and of use to several scientific sub-communities.

The paper proposes a new feature termed a voice-source cepstrum coefficient (VSCC) that is found via making use of the convolutional combination of voice-source and vocal-tract properties being transformed to an additive relationship in the cepstral domain. A summary of the method presented in the paper [171] is now provided:

1. Enframe the speech signal and make a voiced/unvoiced decision on each frame.
2. Extract mel-frequency cepstral coefficients (MFCC) [366, 49] from the frames of the speech signal.
3. Determine the linear predictive coding (LPC)/auto-regressive coefficients. For unvoiced frames perform covariance LPC over the entire frame. For voiced frames determine the LPC coefficients over the closed phase of the glottal period only. In [171] the authors use the DYPSA algorithm [229, 284] to determine an estimate of
the glottal closure instant and then estimate the closed phase of the pitch period as beginning at this point and extending for 35% of the pitch period.

4. Using these LPC coefficients, determine the spectral envelope of the frame. A mel-filter bank is then applied to the envelope and the cepstrum is calculated from the cosine transform of the log filterbank outputs. This forms what is called the vocal-tract cepstrum coefficients (VTCC).

5. Finally we calculate for each frame the voice source cepstrum coefficients, in vector notation, as VSCC = MFCC - VTCC.

Note that this method requires no inverse filtering in order to obtain an implicit estimation of the glottal signal, represented here in the cepstral domain. Of course, like MFCC, all phase information is lost during this process. We now describe using the VSCC in a speaker identification experiment performed on the YOHO corpus.

5.2.2 Experimental Design

As done in [171] we used the YOHO corpus [71] to perform a speaker identification experiment making use of VSCC in combination with MFCC. 80 speakers of the YOHO database were used and their training and testing data came from the original ‘Enroll’ and ‘Verify’ labelling of the database respectively. A GMM-UBM system [327] was used with the UBM being trained on the combination of all remaining 28 YOHO speakers training data by the expectation-maximisation (EM) algorithm before client models were adapted from the UBM via the Bayesian MAP process [158].

Shown in Table 5.1 are various parameters used in the experiment. The extraction of the VSCC was performed with a Matlab script written from the author’s understanding of the original specification of the algorithm in [171]. The DYPSA algorithm was not used for making a determination of the closed phase of each period. Instead a Matlab implementation by the author of the algorithm presented in [115] was used. Voiced/unvoiced (V/UV) decisions on each frame were made with a short term autocorrelation measure with lags bounded by specified maximum and minimum fundamental frequency (F0) values given in Table 5.1. A simple voice activity detection measure was implemented via an energy metric given by $E = 10 \times \log_{10}(\|X\|)$ where $X$ is a speech frame vector. Frames in the lower 30% of frame energies were discarded. Scores were obtained as a standard GMM-UBM log-likelihood ratio per (4.1).

MFCC and VSCC scores were fused via a convex combination per (5.1), where $w \in [0, 1]$.

$$\text{Score}_{MFCC+VSCC} = w \times \text{Score}_{VSCC} + (1 - w) \times \text{Score}_{MFCC}$$ (5.1)
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cepstral coeffs</td>
<td>12</td>
<td>GMM mixtures</td>
<td>32</td>
</tr>
<tr>
<td>Linear prediction order</td>
<td>12</td>
<td>Frame size</td>
<td>32 ms</td>
</tr>
<tr>
<td>Max. F0</td>
<td>250 Hz</td>
<td>Frame shift</td>
<td>10 ms</td>
</tr>
<tr>
<td>Min. F0</td>
<td>70 Hz</td>
<td>Frame threshold</td>
<td>30%</td>
</tr>
<tr>
<td>Pre-emphasis</td>
<td>0.95</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: Parameter-Value pairs as used in replicating the VSCC speaker identification results of [171] on the YOHO corpus.

5.2.3 Results and Discussion

Results are now reported for the individual MFCC and VSCC GMM-UBM systems and their score fused combination. These are presented in Table 5.2 where both misidentification and equal-error rates are given.\(^2\) Note that the minimum values for each of the EER and misidentification rates specified here were found independently. They occurred at different weightings in the score fusion process: \(w = 0.3\) for the misidentification rate which reached a minimum of 5.42\%, and \(w = 0.15\) for the EER whose minimum was 14.64\%.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Misidentification Rate</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>6.76 %</td>
<td>14.81 %</td>
</tr>
<tr>
<td>VSCC</td>
<td>10.99 %</td>
<td>19.61 %</td>
</tr>
<tr>
<td>MFCC + VSCC</td>
<td>5.42 %</td>
<td>14.64 %</td>
</tr>
</tbody>
</table>

Table 5.2: Misidentification and Equal-Error Rates (EER) for the MFCC and VSCC feature systems on the YOHO corpus. Note that the chance misidentification rate with 80 speakers is \(\sim 99\%\).

The VSCC features are highly informative for the task of speaker identification and are able to increase the identification accuracy when combined by score fusion with the MFCC baseline system. Regarding the verification paradigm they are also seen to be individually informative, although they are not observed to provide any complementary information to the vocal-tract features as no real improvement is evident over the MFCC baseline at any weighted score fusion combination of the MFCC and VSCC scores.

The misidentification rates of the MFCC, VSCC and score fused systems of \(\sim 7, 11\)\(^2\) Misidentification rates are based on assignment of identity to the maximum LLR score of the probe against all models. EER are obtained by hard acceptance decisions on the same scores above a common threshold.

90
and 5% respectively are all lower than those of the original paper of \( \sim 14, 36 \) and 10%.

This is likely to be attributable to the use of fewer speakers, the use of the (disjoint) YOHO data to also fit the UBM model and the use of the more accurate \cite{115} glottal closure detection algorithm compared to the DYPSA algorithm used in \cite{171}. Within some margin of variation however we have seen that the VSCC feature and more importantly the information available from the voice-source waveform of the speaker is indeed beneficial to the speaker recognition task.

In Figure 5.1 the variation of the EER and misidentification rate over the combinations defined by each choice of \( w \in [0,1] \) are plotted respectively. In Figure 5.2 detection error trade off curves are plotted for the individual and fused systems.

![Weighted combination of MFCC and VSCC classifiers](image)

**Figure 5.1:** The EER and misidentification rate is plotted for each convex combination of the MFCC and VSCC scores. A minimum of 14.65\% is achieved with \( w = 0.15 \) for the EER and 5.42\% is achieved with \( w = 0.3 \) for the misidentification rate.
Figure 5.2: DET curves for the MFCC, VSCC and fused scores. A minimum EER of 14.65% is achieved by the fused system with $w = 0.15$ (red curve). The MFCC and fused curves display considerable overlap.

We now move on to explore the normalised glottal waveform parameters termed source-frames as introduced in Section 3.5. In Section 5.3 we investigate whether this representation is dependent upon the spoken phonetic content from which the source-frame is estimated, then in Section 5.4 we propose a simple but novel speaker verification framework for comparing these features in Euclidean space.

---

3From which we can also infer a partial answer regarding whether the true glottal waveform in general inherits characteristics based on the uttered phonetic content.
5.3 Experiment 2: Phonetic Independence of the Source-Frame

Experiments were performed on the TI-46 database to investigate the question of whether there exists any evidence of dependence of the shape of the glottal waveform either within or between speakers on the phonetic content. Groupings were formed from the voiced letters of the database based on their phonetic similarities. Glottal waveforms were parameterised as source-frames and compared by the Euclidean distance between them. The distributions of these distances from comparisons both between and within phonetic groups were compared via Kolmogorov-Smirnov hypothesis tests. At the 5% significance level the null hypothesis, that the shape of the glottal waveform has no dependence on phonetic content, was not able to be rejected in any experiment.

5.3.1 Introduction

In this experiment we are concerned with the source-frame representation of the derivative glottal waveform of a speaker as described in subsection 3.5 of Section 3.4 which dealt with parameterising glottal waveform estimates. Specifically we test whether the shape of the source-frame has any dependence on the phonetic content of the utterance from which it was estimated. Our null hypothesis was that there is no phonetic dependence and the alternative hypothesis was that the phonetic content influences the source-frame.

Ideally for text-independent speaker recognition the source-frame features, and glottal features in general, should be independent of the phonetic content being uttered. This has indeed been found in [117, 172] for different representations of the glottal signal, and is now investigated on the source-frames that will be used as features in speaker identification experiments in the following sections of this chapter.

5.3.2 Experimental Design

We use the TI 46-Word database [239] which has 16 speakers, evenly divided by gender, recording 10 instances of each letter of the English alphabet for training. The testing portion of the database consists of 8 sessions with each speaker recording 2 utterances of the alphabet each time. TI-46 is sampled at 12.5 kHz with 12 bit resolution.
The voiced letters of the English alphabet are grouped according to their phonetic similarities, as shown below in Table 5.3. These form the phonetic classes that we compared glottal waveforms from in order to investigate whether any statistically significant differences were present that may be attributable to the uttered content that they are estimated from. Group labels are based on the International Phonetic Alphabet. Source-frame features, as described in Section 3.5, are used to represent the speakers glottal waveform information; stated explicitly, we are testing for any phonetically derived variations present in source-frames and not the speakers true glottal flow waveform for which the source-frame is only a surrogate.

<table>
<thead>
<tr>
<th>Phonetic Group</th>
<th>Letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>ei</td>
<td>A, H, J, K</td>
</tr>
<tr>
<td>i</td>
<td>B, C, D, E, G, P, T, V, Z</td>
</tr>
<tr>
<td>ai</td>
<td>I, Y</td>
</tr>
<tr>
<td>ou</td>
<td>O</td>
</tr>
<tr>
<td>u</td>
<td>Q, U</td>
</tr>
<tr>
<td>e</td>
<td>F, S, X</td>
</tr>
</tbody>
</table>

Table 5.3: Groupings of 21 English letters as created for phoneme dependence testing on the TI-46 database of the source-frame parameterisation of the derivative glottal flow.

Two experiments were performed with the aim of discerning the phonetic dependence of:

1. **INTER-SPEAKER VARIATION**: How the source-frames vary BETWEEN SPEAKERS over the voiced letter groups

2. **INTRA-SPEAKER VARIATION**: What variation exists WITHIN SPEAKERS source-frame features over the voiced letter groups

We now describe how the TI-46 data was used to perform each of the two experiments. Note that in performing a comparison of any two source-frames $\vec{X}, \vec{Y}$, where $\vec{X}, \vec{Y} \in \mathbb{R}^N$ and $N$ is the length the source-frames are normalised to, a scaled Euclidean distance metric was used, as given by:

$$d(\vec{X}, \vec{Y}) = \frac{1}{N} \|\vec{X} - \vec{Y}\|$$  \hspace{1cm} (5.2)

This is a simple measure for detecting differences in the shapes of the prosody-normalised source-frames. We refer to this as an arithmetic Euclidean distance.

94
5.3.2.1 Inter-speaker phonetic variation

The following method was used for determining whether there existed any statistically significant, phonetically derived differences within the glottal waveform features independent of the generating speaker. First, we combined all training and testing data of TI-46, thereby also eliminating any clear temporal/sessional variations. For each letter in our vowel groups we calculated a mean glottal waveform from all of the source-frames from this letter, done per speaker, per training/testing set. This gave us a collection of 672 mean glottal waveform features: number of speakers (16) × training/testing (2) × number of letters in our vowel groups (21). We then calculated arithmetic Euclidean distances by (5.2) between vowels of the same-group, and between different-groups. This formed our two sets of score data which we compared: a distribution of same group distances and a distribution of different group distances. These distributions were compared in the same manner as for the intra-speaker phonetic variation tests: by two-sample Kolmogorov-Smirnov tests which we describe below after the description of intra-speaker method. We know from earlier experimental work including [401] that the distributions of scores generated in this way from comparisons of source-frames from same and different speakers are highly separated. Thus we did not compare any two mean source waveforms calculated from utterances of the same speaker.

5.3.2.2 Intra-speaker phonetic variation

To test for phonetically derived differences in the source-frames within individual speakers the 16 speakers of the TI-46 database were processed as follows. Using only the training portion, source-frames were extracted from each letter in a vowel group, and the collection of source-frames from each letter were divided into four groups and mean source waveforms were calculated for each of these. Euclidean distances (5.2) were then calculated between letters from the same phonetic groups and between letters from different phonetic groups. This produces two distributions of scores which were again compared with non-parametric Kolmogorov-Smirnov tests (K-S test) in order to determine if any statistically significant differences existed. These results are given in Table 5.5 below.

In each of these two experiments, our confirmatory data analysis by using a two-sample K-S test to compare the relative-frequency histograms (via their cumulative density functions) of the scores of the same group and different group comparisons. This K-S Test quantifies distance based differences between the empirical distribution functions (cumulative functions) of the two sets of data to infer whether they share the
same distribution. As stated, our null hypothesis was that there are no differences in the source-frame representation of the derivative glottal waveform shapes due to the uttered phonetic content, and we considered the p-values at the $\alpha = 5\%$ significance level to determine whether this null hypothesis could be rejected.

Results for these two experiments are now presented in 5.3.3. Two notes of caution. These glottal waveforms are only compared via their source-frame representations so any conclusions we draw must be limited to statements regarding the source-frame parameterisation only. Secondly, we are comparing time series data when we compare two source-frames and this time series factor is not accounted for in our analysis. As such we must be cautious regarding any use of the word ‘independence’.

5.3.3 Results

5.3.3.1 Inter-speaker phonetic results

Based on the p-values from the K-S tests, for none of our phonetic groups could we conclude a distinctive glottal waveform existed. At the $\alpha = 5\%$ significance level we could not reject the null hypothesis for any of the vowel groups. In fact no p-value was close to allowing the rejection of the null hypothesis, with p-values for all comparisons greater than 65% as shown in Table 5.4.

This agreed qualitatively with the observed relative frequency histogram distributions of same-vowel and different-vowel scores, all of which displayed nearly complete overlap. Figure 5.3 shows plots of the two score distributions of comparisons from same and different phonetic groups for all six phonetic groupings.

<table>
<thead>
<tr>
<th>Phonetic Group</th>
<th>K-S test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon i$</td>
<td>0.975</td>
</tr>
<tr>
<td>$i$</td>
<td>0.975</td>
</tr>
<tr>
<td>$ai$</td>
<td>0.975</td>
</tr>
<tr>
<td>$ou$</td>
<td>0.675</td>
</tr>
<tr>
<td>$u$</td>
<td>0.675</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>0.675</td>
</tr>
<tr>
<td>All</td>
<td>0.975</td>
</tr>
</tbody>
</table>

Table 5.4: Inter-speaker Kolmogorov-Smirnov test p-values for TI-46 phonetic groups. The empirical distributions of scores from both same and different vowel group comparisons displayed significant overlap as captured by the p-values of the K-S tests. The p-value is also given in the final row for the combination of same and different scores from all phonetic groups.
Figure 5.3: Inter-speaker relative-frequency histograms are plotted for each of the six phonetic groups. Abscissa shows the arithmetic Euclidean distance scores on mean source waveforms between letters from the same and from different vowel groups.
5.3.3.2 Intra-speaker phonetic results

Having found no phonetically dependent, distinct source-frame shapes across speakers, we then investigated whether there existed any significant differences within speakers. Shown in Table 5.5 are the p-values from the two-sample K-S tests. The empirical distributions of scores for same-vowel group and different-vowel group comparisons for each speaker displayed significant overlap and this is captured by the p-values of the K-S tests. Speaker F7 in post analysis was found to have considerably less source-frame data extracted from her utterances and this is the primary reason for her outlying behaviour, which is however still not statistically significant.

<table>
<thead>
<tr>
<th>Female Speaker</th>
<th>p-value</th>
<th>Male Speaker</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>0.6751</td>
<td>M1</td>
<td>0.9748</td>
</tr>
<tr>
<td>F2</td>
<td>0.6751</td>
<td>M2</td>
<td>0.9748</td>
</tr>
<tr>
<td>F3</td>
<td>0.9748</td>
<td>M3</td>
<td>0.9748</td>
</tr>
<tr>
<td>F4</td>
<td>0.9748</td>
<td>M4</td>
<td>0.6751</td>
</tr>
<tr>
<td>F5</td>
<td>0.9748</td>
<td>M5</td>
<td>0.9748</td>
</tr>
<tr>
<td>F6</td>
<td>0.6751</td>
<td>M6</td>
<td>0.3129</td>
</tr>
<tr>
<td>F7</td>
<td>0.1108</td>
<td>M7</td>
<td>0.9748</td>
</tr>
<tr>
<td>F8</td>
<td>0.9748</td>
<td>M8</td>
<td>0.6751</td>
</tr>
</tbody>
</table>

Table 5.5: Intra-speaker Kolmogorov-Smirnov test p-values for TI-46 speakers.

The distribution of the arithmetic Euclidean distance scores from the same-group and different-group comparisons are plotted in Figures 5.4 and 5.5 for all female and male speakers respectively. In comparison to the inter-speaker distributions of comparison scores in Figure 5.3 for all of the phonetic groups, for each of which the two distributions displayed nearly complete overlap, the intra-speaker scores show much greater variation. Based on the K-S test statistics however, these differences over the phonetic groups in the shape of the individual speakers source-frames as measured by the Euclidean distance between them, are not statistically significant at the $\alpha = 5\%$ significance level for any of the 16 YOHO speakers.

The use of the Kolmogorov-Smirnov test allows us to conclude that for none of the 16 speakers can we reject the null hypothesis, since there is no significant difference between the source-frames of the different phonetic groups. The closest we come to rejecting the null hypothesis is for female speaker F7 where the calculated p-value was 0.11.
In neither the inter-speaker nor intra-speaker experiments was any evidence found for rejecting the null hypothesis of there existing no significant phonetically based differences in source-frame shapes.

The two experiments are necessary and complement each other as the pathological case is conceivable where no phonetically based differences are observed between speakers due to an unfortunate distribution of within-speaker differences that cancel out when combined. The results of the K-S tests in comparing the distribution of distances between source-frames from individual speakers over the phonetic groups (intra) and from comparing different speakers over the phonetic groups (inter) provide no evidence for the alternative hypothesis. This is of course a positive conclusion when considering employing source-frame glottal features as an information source for text-independent speaker recognition. If strong evidence had been established for source-frames being being dependent upon the uttered phonetic content then either a restriction to text-dependent speaker recognition or else compensation measures to remove phonetic effects would be necessary before attempting to compare speakers based on their source-frame features. With a suitable collection of databases the dependence of the source-frame on language may also be investigated.

Finally we recall points made in [45] among other places, that the absence of evidence does not imply evidence of absence. To be able to conclude with statistical significance that the two distributions are statistically similar we would require further exploration with equivalence testing rather than statistical difference testing. Our weaker conclusion is sufficiently encouraging to begin text-independent speaker recognition experiments using source-frame features for the task of inferring speaker identity.

We commence this exploration next in Section 5.4 where a speaker verification experiment on the YOHO and ANDOSL corpora employing the source-frame glottal features is reported.
Figure 5.4: The distribution of arithmetic Euclidean distance scores within and between the phonetic groups for all 8 TI-46 female speakers.
Figure 5.5: The distribution of arithmetic Euclidean distance scores within and between the phonetic groups for all 8 TI-46 male speakers.
5.4 Experiment 3: Distance Metric on Mean Source-Frames

In this section a novel speaker verification framework is introduced that uses a distance metric to compare speakers’ average glottal waveforms. Preliminary experiments were performed on the ANDOSL and YOHO speech corpora. The source-frame feature representation of the derivative glottal waveform, found by closed-phase inverse filtering and prosody normalised as described in Section 3.5 is used in both experiments. It was found that the mean of several hundred source-frames were highly discriminating with each speaker having a characteristic glottal waveform. An equal-error rate of 9.43% was achieved on the YOHO corpus in a speaker verification experiment while the ANDOSL speakers were completely separated when sufficiently large amounts of training and testing data were employed.

5.4.1 Introduction

Having found in the previous Section 5.3 no evidence for the source-frame features to be dependent upon the phonetic content of the utterance they are extracted from we now begin to explore their use for the task of automatic, text-independent speaker recognition.

A simple modelling approach is taken whereby averages of source-frames extracted from the training speech are compared to an average of those extracted from the testing utterance. This comparison is once again performed by an arithmetic Euclidean distance. A method for evaluating these comparisons against the expected population variance is provided through the introduction of a background model suitable for speaker verification.

The remainder of this experimental section is structured as follows: in 5.4.2 we describe in detail these proposed modelling and scoring methods before illustrating the experimental design employing these techniques for two experiments performed on the ANDOSL and YOHO corpora. In 5.4.3 the results are then presented for each database and discussed.
5.4.2 Experimental Design

5.4.2.1 Glottal Waveform Modelling & Scoring via a Distance Metric

We now describe the speaker verification system employing source-frames features as used in the experiments of this section. Source-frames are denoted by capital Latin letters \( \vec{X}, \vec{Y} \) whilst the mean of a collection of source-frames is denoted as \( \hat{X} \) and \( \hat{Y} \).

Consecutive source-frames are combined into blocks of size \( \beta \), \( \{ \vec{X}_i \mid i = 1, \ldots, \beta \} \), which are represented by their mean vectors in \( \mathbb{R}^n \):

\[
\hat{X} = \frac{1}{\beta} \sum_{i=1}^{\beta} \vec{X}_i \quad (5.3)
\]

A speaker model \( \lambda \) is represented by a set of mean vectors

\[
\lambda = \{ \hat{X}_k \mid k = 1, \ldots, \kappa_\lambda \} \quad (5.4)
\]

where the number of means \( \kappa_\lambda \) is determined by the number of blocks of size \( \beta \) present within the speakers training data. Figure 5.6 shows a mean source-frame waveform resulting from taking the mean of a block of \( \beta = 1000 \) source-frames.

Similarly a Universal Background Model \( \lambda_U \) is represented by a set of mean vectors

\[
\lambda_U = \{ \hat{U}_k \mid k = 1, \ldots, \kappa_{\lambda_U} \} \quad (5.5)
\]

with the number of means \( \kappa_{\lambda_U} \) determined by the number of blocks of \( \beta \) source-frames within the collection of all the background speakers’ data.

Scoring is done in the Euclidean space \( \mathbb{R}^n \) that the source-frames reside in, by a distance between training and test blocks. Several distance measures on \( \mathbb{R}^n \) were investigated in preliminary studies for this experiment, including the non-symmetric relative time squared error (RSTE) measure defined in [117] where, if the score of a comparison between \( \hat{X} \) and \( \hat{Y} \) is denoted by \( s(\hat{X}, \hat{Y}) \), the distance is:

\[
s(\hat{X}, \hat{Y}) \sim \frac{\| \hat{X} - \hat{Y} \|}{\| \hat{X} \|} \quad (5.6)
\]

It was found that the arithmetic distance measure (5.7) below achieved the best system performance. This arithmetic distance measure \( d \) used is the length-normalised
Euclidean distance in $\mathbb{R}^n$, such that the distance between mean vectors $\hat{X}, \hat{Y} \in \mathbb{R}^n$ is:

$$d(\hat{X}, \hat{Y}) = \frac{1}{n} \| \hat{X} - \hat{Y} \|$$  \hspace{1cm} (5.7)

Testing was performed as follows. For a system with block size $\beta$, test speech is recorded until $\beta$ source-frames are obtained, and a test mean $\hat{Y}$ is calculated from these per (5.3). The system produces a score against this test mean, and makes a recognition decision on this score based on a threshold found during the development phase. The score between the test utterance block mean $\hat{Y}$ and a speaker model $\lambda$ is given by:

$$d(\hat{Y}, \lambda) = \frac{1}{K} \sum_{k=1}^{K} d(\hat{Y}, \hat{X}_k)$$  \hspace{1cm} (5.8)

The score between the test utterance block mean $\hat{Y}$ and the background model $\lambda_U$ is similarly given by

$$d(\hat{Y}, \lambda_U) = \frac{1}{K} \sum_{k=1}^{K} d(\hat{Y}, \hat{U}_k)$$  \hspace{1cm} (5.9)

Other ways of scoring the test sample against the UBM were investigated including taking the minimum rather than the average. The average score was found to give the best system performance, superior to the similar approach used in [117] of taking the first principal component of the collection of normalised glottal signals and comparing those with the RTSE metric.

The score of the test utterance with block mean $\hat{Y}$ against the speaker model $\lambda$ is then given by the ratio of these:

$$score(\hat{Y}; \lambda, \lambda_U) = \frac{d(\hat{Y}, \lambda)}{d(\hat{Y}, \lambda_U)}$$  \hspace{1cm} (5.10)

This proposed modelling and scoring method is applied in two separate experiments performed on the ANDOSL [260] and YOHO [71] databases, which are described now.

### 5.4.2.2 Preliminary investigation on the ANDOSL database

Preliminary investigations were done on the Australian National Database Of Spoken Language (ANDOSL) [260]. A subsection consisting of the first 30 male speakers from the native English portion of the database was used. Each spoke the same 200 sentences. The recordings were downsampled from 20 to 16 kHz for processing. The waveforms were not downsampled to 8 kHz as done in [119], so that the full spectrum could be modelled.
by the source-frame, and we do not investigate the stochastic (fine structure) element of the deterministic plus stochastic model proposed there. The first 100 sentences are used for training and the remaining 100 for testing. The ANDOSL database is such that all data for each participant was recorded over one session, and the recording equipment and environment were the same for all speakers. Thus ANDOSL is not suitable for testing of automatic speaker recognition systems in realistic circumstances, but was used here only for preliminary studies to determine under close to ideal conditions whether the proposed method works well.

A disjoint group of 10 male speakers from ANDOSL were used for the UBM, with 3000 source-frames extracted for each speaker. The other 20 speakers were used as clients and impostors. Feature extraction of the source-frames was performed as described in Section 3.5 based on 30 ms frames with 10 ms shifts and an LP order of 18 for the autoregressive analysis. We examine mean source-frames calculated from block sizes of \( \beta = 500, 1000, 2000, 3000 \) and all available source-frames. Source-frames are normalised to \( n = 256 \) samples long, and pseudo-likelihood-ratio scores are calculated as per (5.10).
5.4.2.3 YOHO: A more realistic database

We then explored this proposed glottal based speaker verification system on the YOHO speech corpus [71] which contains American English speech of ‘combination lock’ phrases (e.g. 14-39-85) recorded over several sessions and sampled at 8 kHz. We used the 108 male speakers within the database, all having recorded 4 sessions for enrolment, speaking 24 combination lock phrases per session, and 10 verification sessions where 4 combination phrases were recorded each time. The background model is formed from the first 54 speakers, while the remaining 54 are used for client and impostor testing. Source-frame extraction was performed in the same manner with the same parameters, normalised to \( n = 256 \) samples but using a 10\textsuperscript{th} order linear prediction. At the 8 kHz sampling frequency, 128 samples for a pitch period corresponds to a 63 Hz fundamental, meaning essentially all glottal waveforms were interpolated during the prosody normalisation process of the source-frame extraction. Means were calculated on block sizes of \( \beta = 500, 1000, 2000 \) and 3000 source-frames.

5.4.3 Results and Discussion

5.4.3.1 ANDOSL

The equal-error rates for the mean source-frame/distance modelling method are given in Table 5.6 for the subset of 20 ANDOSL speakers that this preliminary investigation was performed on.

<table>
<thead>
<tr>
<th>Block Size ( \beta )</th>
<th>Estimated Speech Duration</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>~ 8 seconds</td>
<td>12.27%</td>
</tr>
<tr>
<td>1000</td>
<td>~ 16 seconds</td>
<td>5.77%</td>
</tr>
<tr>
<td>2000</td>
<td>~ 30 seconds</td>
<td>1.84%</td>
</tr>
<tr>
<td>3000</td>
<td>~ 45 seconds</td>
<td>0.0%</td>
</tr>
<tr>
<td>ALL</td>
<td>&gt; 5 minutes</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Table 5.6: Equal-error rates for each block size \( \beta \) and corresponding speech durations for the verification experiment on the 20 ANDOSL speakers.

It was found that the variation within speaker means is reduced as the block size increases, and indeed that speaker verification performance improves, with no found limit as the block size grows for this small collection of 20 speakers recorded under optimal recording conditions. Indeed complete separation of the target and non-target scores (EER = 0\%) is achieved from comparisons between the mean of \( \beta \geq 3000 \) training data source-frames against the mean of \( \beta \geq 3000 \) testing data source-frames.
For comparison using a Gaussian mixture model and UBM with 64 mixtures, modelling 20 MFCC extracted from those source-frames, produced an EER of 3.96%.

The implied duration of speech required to obtain $\beta$ source-frames is given in column two of the table. Based on an estimate that $\sim 60\%$ of speech frames are voiced and with a frame shift of 10ms we can approximately obtain $\beta = 500$ source-frames from a amount of speech of: $500 \times 1/0.6 \times 10$ ms or $\simeq 8$ seconds.

### 5.4.3.2 YOHO

Figure 5.7 shows detection error trade-off (DET) curves for the system with various block sizes. Again the EER of the system decreased with each increase in the number of source-frames used to obtain a mean from. The EER of the system taking block sizes of $\beta = 3000$ source-frames was found to be 9.43% for the 54 client speakers taken from the YOHO database.

In both the preliminary ANDOSL study and the investigation on the YOHO corpus we find that the speaker verification system improves as means are compared from increasing block sizes. Our interpretation of this is that a speaker has a characteristic glottal waveform, independent of the voiced sound being uttered, and that by taking the mean of larger block sizes we remove the speakers natural variation about this characteristic waveform, as well as any errors, such as from poor linear prediction or incorrect determination of glottal closure periods.

Plots of multiple source frames from the same speaker and from different speakers are shown in section A.1 of the Appendix. Significant overlap and similarity in waveform shape are evident for the same-speaker plots presented for YOHO speakers s7 (Figures A.1, A.2 and A.3) and s8 (Figures A.4, A.5 and A.6). This is typical of the observed intra-speaker behaviour of the mean source-frames. In contrast, much greater variation is observed in the inter-speaker plots of Figures A.7 to A.11 where single mean source-frames from five different speakers are plotted.

For a male with a 100Hz fundamental, (having a pitch period of 10ms), a block size of 1000 frames corresponds to 10 seconds of purely voiced speech, which implies the need for approximately 20 or 30 seconds of natural speech. The achieved results show that

---

4During development several other methods were explored for modelling these prosody normalised source signals. Spectral parameterisations including FFT coefficients, linear cepstral coefficients, PCA and LDA representations and MFCC were also investigated in a GMM-UBM framework [327]. See Section A.2 of the appendix for some preliminary results regarding separating source-frames by LDA. Despite the physical and perceptual motivations for the mel warped spectral range not being as applicable to glottal waveforms, the MFCC parameterisation achieved the best EER of these methods, but as noted this was worse than the distance measure approach.
when this much speech is available, and recorded under favourable circumstances, the speaker’s source waveform is highly discriminating. Our goal is now to find approaches to achieve these recognition rates without the requirement for as much testing data. Improving the feature extraction process and introducing a variance measure on top of the mean with a view towards forming a generative probability model would be of practical value in regard to reducing the amount of test speech required for low EER speaker verification.

Next, in Section 5.5, we begin exploring the ability of discriminative support vector machine models to separate these source-frame parameterisations of the speakers derivative glottal waveform.
5.5 Experiment 4: Frame Level Identification with Support Vector Machines

A preliminary speaker identification experiment was performed on the YOHO corpus using the source-frame features representing a speaker’s derivative glottal waveform. Reported are lower level correct source-frame, rather than utterance level, identification rates. Using a support vector machine model to compare source-frames, a 65% correct frame level identification rate was obtained with a cohort of 20 speakers.

5.5.1 Introduction

We now begin to investigate the ability of support vector machine (SVM) classifiers to differentiate speaker’s source-frame features.

While most successful applications of SVM models to speaker recognition have been based on supervectors constructed from the parameters of generative models [92, 96, 216], and because of the positive results obtained using a distance measure alone to compare source-frames in their native $\mathbb{R}^n$ feature space, we hypothesise that a well trained SVM will exhibit superior performance.

In order to focus on optimising discriminative performance we report correct identification rates at the individual source-frame level. In the next experiments reported in Section 5.6 we build upon the results presented in this experiment to combine the SVM models individual frame classifications to make utterance level speaker assessments.

5.5.2 Experimental Design

We use the YOHO speaker identification database consisting of multiple-session, real-world office environment recordings of combination-lock phrases sampled at 8 kHz [71] to report source-frame identification rates for different sized cohorts of male speakers. Source-frames were extracted per the description in Section 3.5 from the male YOHO speakers and these two pitch period, variable length waveforms were normalised in length to $n = 256$ samples.

Primarily for the purpose of dimension reduction, the source-frame training data for each closed cohort of speakers is taken and a common basis is derived via a principal component analysis (PCA). This basis is used then to project all training and testing
source-frames into. The percentage of variation within the different cohort data explained by retaining increasing number of principal components is shown in Figure 5.8. We see that 95% of the variation within the data is covered by the first 50 principal components, independent of cohort size.

![PCA Coverage of Variance for Each Cohort Size](image)

**Figure 5.8: Proportion of source-frame variation covered by retaining increasing numbers of PCA dimensions. For all cohort sizes the first 30 principal component dimensions retain at least 90% of the observed variation of the source-frame training data.**

Support vector machines are then applied to the principal component representation of these source-frames with the intention of training speaker discriminating hyperplanes. The principal component projection of the source-frame data was not scaled or normalised in any way. Empirically it was found that radial basis function (RBF) kernels achieved the greatest separation of the speaker data, performing slightly better than 3rd degree polynomials. A disjoint set of male YOHO speakers was used with a coarse to fine scale grid search to determine the best parameters for the RBF kernel, (gamma = 0.007 & cost = 32).\(^5\)

The number of source-frames used for training the SVM hyperplane was increased along with the cohort size as shown in Table 5.7, where for example, for each speaker

\(^5\)We used the common C++ SVM library LIBSVM [77] to implement these experiments.
Table 5.7: Number of source-frames used to train SVM models for speakers in each cohort size. For example 1000 source-frames (projected into the PCA basis for dimension reduction) were used to train the speakers SVM hyperplanes in the cohort of size 20.

<table>
<thead>
<tr>
<th>Cohort Size</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. Source-Frames</td>
<td>300</td>
<td>500</td>
<td>800</td>
<td>1000</td>
</tr>
</tbody>
</table>

in the group of 5 speakers, 300 source-frames were used for SVM training. Due to computational limitations, results were only obtained from cohort sizes of 5, 10, 15 and 20 speakers. The ‘Enroll’ and ‘Verify’ divisions of the YOHO speakers data were ignored, pooling all data and performing 10-fold cross-validation, training the SVM model on a mutually disjoint selection each time and testing on the remaining data. The reported identification rates are averages over these. A 1-against-all method was used to train a SVM hyperplane for each individual speaker.

5.5.3 Results and Discussion

More than 95% of the variation within the source-frames was captured by the first 30 basis vectors of a PCA for all cohort sizes. Using this PCA basis for dimension reduction, SVM with radial basis function kernels were used to separate the speaker’s source-frames with an average source-frame identification rate of 66% obtained over the 5, 10, 15 and 20 speaker cohort sizes. Figure 5.9 shows identification rates on a per source-frame level. The highest of these identification rates for each speaker cohort size tested is given in Table 5.8.

Table 5.8: Best source-frame identification rates over all number of retained PCA dimensions for each speaker cohort size.

<table>
<thead>
<tr>
<th>Cohort Size</th>
<th>Correct Identification Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>70.8 %</td>
</tr>
<tr>
<td>10</td>
<td>64.3 %</td>
</tr>
<tr>
<td>15</td>
<td>65.0 %</td>
</tr>
<tr>
<td>20</td>
<td>64.7 %</td>
</tr>
</tbody>
</table>

These results are considerably better than the chance correct identification rates for each cohort size. As mentioned, due to computational reasons the largest cohort size explored was 20 speakers, although interestingly the results for the cohorts of size 15 and of 20 speakers are almost identical.

Despite the first 30 principal components seemingly capturing the majority of the
variation present within the source-frame data, the identification rates for each cohort size plotted against the number of retained PCA dimensions display a flattening trend from approximately 50 PCA dimensions on.

Figure 5.9: Source-frame correct-identification rates as a function of PCA dimension size. The identification rate curves are formed by averaging the results over each of the 10 cross-validation folds. Group sizes 15 and 20 are nearly identical in correct identification rates at all retained PCA dimensions.

Whilst we initially investigated how best to maximally separate speakers with these source-frame features, results were reported at the lowest level, being on a per source-frame basis rather than a per utterance basis. In the same way as low phoneme recognition rates translate into acceptable word recognition systems, these identification rates are expected to translate to strong utterance based speaker recognition systems.

Next in Section 5.6 we test this hypothesis by combining frame level classifications to create utterance level predictions of speaker identity.
5.6 Experiment 5: Support Vector Machine Approach at the Utterance Level

Speaker identification experiments were performed on the YOHO corpus, reporting correct identification rates at the utterance level. Source-frames were modelled via support vector machines using both multiclass and regression methods. The multiclass SVM model correctly identified 85.3% of test utterances with a cohort of 20 speakers, while the regression SVM model with the same closed set of 20 speakers correctly identified 72.5% of the test utterances. These results are similar to previous investigations of the voice-source waveform for speaker identification using the YOHO corpus [171, 302]

5.6.1 Introduction

In this section we introduce a temporal divide between training and testing speech (maintaining the existent one present in the YOHO corpus), and report correct identification rates at the utterance (.wav file) level, testing the hypothesis that these source-frame level correct identification rates from our initial exploration in Section 5.5 translate into strong utterance level systems. Indeed this is the behaviour observed in moving from the micro level (speech frame, visual frame) to the macro level (utterance recording, visual sequence) in the majority of automatic recognition systems, and logically so for any non-trivial distributions of micro level recognition rates.

Two SVM modelling approaches are taken in this experiment: the first is a multiclass implementation and the second is a regression model where class labels were directly approximated.

5.6.2 Experimental Design

Again our feature of interest is the source-frame. We examine the ability of both multiclass SVM and single class SVM regression to discriminate between speakers based on these source-frame features in closed-set speaker identification experiments.

In both experiments we used male speakers from the YOHO corpus [71]. YOHO contains multisession recordings divided between ‘Enroll’ and ‘Verify’ database labels and in all experiments training and testing speech is taken from these respectively. YOHO, whilst non-challenging for current state-of-the-art automatic speaker recognition systems
such as those derived from factor analysis [209, 95, 220] or even more traditional GMM-UBM modelling [327], permits the voice-source waveform to be initially examined in the absence of channel and noise variations that can impact negatively on the linear prediction and inverse filtering processes. This is the approach of several significant papers in their preliminary investigations of the voice-source [302, 171, 172, 118]. Estimating information regarding the glottal waveform, particularly temporal domain information, in the presence of convolutional and phase distorting effects is infeasible with current algorithms and perhaps fundamentally so.

Source-frames were normalised to $n = 256$ samples. This dimensionality is an issue for computation considerations and principal component analysis (PCA) was used purely for dimension reduction. A disjoint set of 10 male speakers from the YOHO dataset, not used in any identification experiments as clients or impostors, was selected and source-frames extracted from all of their enrol data. We shall refer to this set as the background set. Using this background set, a basis of principal components was determined onto which experimental source-frames could be projected for dimension reduction.

The percentage of variation retained from the background data by increasing the number of principal components is shown in Figure 5.10. We see that more than 90% of the variation within the data is covered by retaining the first 50 principal components. This was expected as the windowing of the source-frame produces many near zero samples shared at each end of all source-frame vectors, meaning that almost certainly there are at most $\sim 180$ dimensions of variation (refer for example to the plotted mean source-frame of Figure 5.6).

The C++ release of the LIBSVM package [77] was used to implement all SVM experiments. Cross validation on a further disjoint set of male YOHO speakers was used to determine the optimal kernel function (radial basis function) and kernel parameters ($\gamma = 0.0325$, $c = 32$) prior to these experiments, but with the same PCA basis used as derived from the background set. Data was not scaled further than the prosody normalisation step during feature extraction.

5.6.2.1 Design: Multi-Class SVM Modelling

Multi-class support vector modelling was explored with closed cohorts of 5, 10, 15, 20 and 30 speakers. For each experiment, hyperplanes (SVM models) were trained on speech coming only from the 'Enroll' partition of YOHO, and test probes were taken only from the 'Verify' partition. Identification rates are given at both the source-frame level and the utterance level. For all source-frames, from all probe utterances, probabilities
measuring class membership are output. Frame level identification rates were calculated based upon assignment of source-frames to a speaker/class whose model generates the maximal probability. Utterance level scores were determined by calculating the mean probability value over all source-frames from the utterance. Utterance level identification decisions were then made by assigning the utterance to the model/speaker with the maximum score.

Training and testing source-frame data is projected against the background data PCA basis for dimension reduction. Experiments are performed whilst retaining 30, 40, 50 and 60 principal component. Table 5.9 reports average (across the four PCA dimension results) utterance and frame level correct identification rates.

5.6.2.2 Design: Regression SVM Modelling

We also examined the ability of binary SVM regression models in closed set speaker identification experiments on the YOHO corpus. One-versus-all models were created as follows. For each speaker within the cohort, a regression SVM model was trained on the pooled training data of all speakers. Training data was assigned the class +1 for PCA projected source-frames belonging to the target speaker, and −1 for all other speakers present within the training set. Source-frames presented at testing time were assigned a predicted label by the regression model with a value on the continuum between these training class labels [−1, +1].

For frame level identification rates, source-frames were assigned to the speaker whose
regression model output the largest score. For utterance level identification, the mean of all the regression outputs of all the frames of the test utterance was taken to create an utterance score against the speaker SVM model. The speaker whose model outputs the largest utterance level score was identified as the speaker of the test utterance.

Taking the mean of the frame scores was empirically determined to achieve higher identification rates than other statistical fusion possibilities such as maximums or product.

This experimental process was performed for cohorts of size 5, 10, 15 and 20 speakers. In all regression SVM experiments we retain only the first 30 principal component dimensions, informed by the results of the multiclass SVM experiments which provided strong evidence for the proposal that there was little benefit in retaining more principal component coefficients. The speakers and their training and testing data remained the same for each cohort size, as done for the multiclass SVM experiments, which is to say that the increasing cohorts of speakers formed supersets.

5.6.3 Results

5.6.3.1 Results: Multi-Class SVM Modelling

Correct identification rates for the one-versus-all multiclass SVM model are presented in Table 5.9 for identifications at both the frame level and utterance level.

<table>
<thead>
<tr>
<th>Cohort Size</th>
<th># Source-Frames</th>
<th>Frame %</th>
<th>Utterance %</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1100</td>
<td>38.3%</td>
<td>71.74%</td>
</tr>
<tr>
<td>10</td>
<td>2200</td>
<td>30.9%</td>
<td>86.7%</td>
</tr>
<tr>
<td>15</td>
<td>6000</td>
<td>39.4%</td>
<td>89.5%</td>
</tr>
<tr>
<td>20</td>
<td>6000</td>
<td>25.6%</td>
<td>85.3%</td>
</tr>
<tr>
<td>30</td>
<td>6000</td>
<td>20.5%</td>
<td>80.8%</td>
</tr>
</tbody>
</table>

Table 5.9: Identification rates at the frame level and utterance level for the multiclass SVM modelling as a function of cohort size and number of source-frames.

Identification rates do not evolve in relation to the chance identification rate of \(1/\text{CohortSize}\). Instead the maximum utterance identification rate of 89.5% is obtained for 15 speakers. This is likely due to the amount of data available for training the SVM model, where overtraining and undertraining are likely occurring on either side of 15 speakers. Limitations in available computational power necessitated training on 6000 source-frames for the 20 and 30 speaker cohorts, the same number as used for the 15 speaker group, whereas the number of training source-frames had been increased
from cohorts of 5 to 10 to 15 respectively. The method for combining the frame level probabilities in order to make an utterance level assessment may also be a factor as this behaviour is not observed in the frame level rates as shown in column 3 of Table 5.9.

The frame level identification rates for each cohort and each PCA dimension size are plotted in Figure 5.11. Utterance level identification rates, again for each cohort and PCA dimension, are shown in Figure 5.12. The influence of the PCA dimension is shown to be minimal over these two plots. The identification rates are promising, especially under the working assumption that the voice-source information is orthogonal or complementary to common spectral magnitude features (mel-cepstra).

![Frame Level Correct Identification Rates](image)

**Figure 5.11:** *Frame level correct identification rates for each closed set speaker size and PCA dimension. Little variation is observed with increasing PCA dimensions.*

The number of principal components used was varied from 30 to 60 in the multiclass SVM experiments. It can be seen from the statistically non-significant increase in correct identification rates in the multiclass SVM results as the PCA dimension was increased (see Figure 5.12) that there is also a large amount of noise variation in the source-frame data not related to speaker identity, and that retaining larger numbers of principal components is not beneficial to recognition accuracy.

Misidentifications are found to be approximately uniform across speakers in all cohorts. Figure 5.13 demonstrates how the frames from test utterances of Speaker 2 are assigned when scored against the multiclass SVM model for the size 15 cohort. While the majority are correctly assigned to Speaker 2, the misclassifications are approximately
5.6.3.2 Results: Regression SVM Modelling

Regression SVM correct identification rates are reported in Table 5.10 for all cohort sizes, based on retaining 30 principal component dimensions. Note that the number of source-frames available per speaker (Column 2) ideally should increase with the cohort size for reliable training of the SVM regression model. We were limited to using the quantities given due to constraints on computational resources. The maximum cohort size was limited to 20 speakers for similar reasons.

Figures 5.14 and 5.15 show utterance scores for the cohort group of 5 speakers, where there were 184 test utterances (roughly split uniformly between speakers). Figure 5.14 shows the test probes from the 5 speakers scored against the regression model for speaker 1 whilst Figure 5.15 shows the same utterances scored against the regression model of speaker 2.

A verification style threshold is drawn on each figure along with points demarcating the continuous section of utterances coming from the speaker whose model is being tested against. This threshold line is drawn only to indicate the typical distribution of scores observed in all experiments; we perform speaker identification experiments which make
Figure 5.13: Histogram of frame level assignments to each speaker in the cohort of size 15 using the multiclass C-SVM model where all test source-frames belong to Speaker 2.

<table>
<thead>
<tr>
<th>Cohort Size</th>
<th># Source-Frames per Speaker</th>
<th>Frame %</th>
<th>Utterance %</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1100</td>
<td>37.4%</td>
<td>90.0%</td>
</tr>
<tr>
<td>10</td>
<td>2000</td>
<td>30.4%</td>
<td>89.3%</td>
</tr>
<tr>
<td>15</td>
<td>1000</td>
<td>21.7%</td>
<td>80.2%</td>
</tr>
<tr>
<td>20</td>
<td>1000</td>
<td>17.8%</td>
<td>72.5%</td>
</tr>
</tbody>
</table>

Table 5.10: Mean correct identification rates at frame and utterance level using SVM regression as a function of the average number of source-frames used per speaker for SVM regression training and of cohort size.

Correct identification rates were reasonably consistent for each individual speaker in all experiments. Figure 5.16 shows the breakdown of correct identification rates for each speaker of the cohort of size 20. For the majority this rate is above or near 70%, however for speakers 16 and 20 the performance is well below the average for the cohort. Upon inspection these identification rates are strongly correlated with the total number of source-frames extracted from each speaker’s training utterances. Whilst the training utterances were the same in number for each speaker, the number of pitch periods in total over these utterances differed for each speaker. We believe this affected the SVM regression model accuracy for those speakers.

Identification results in both experiments have shown further evidence that the voice-source waveform obtained by inverse filtering of the speech signal contains significant
information pertaining to speaker identity. Using a multiclass SVM model 85.3% of test utterances, all disjoint to those used for training, were correctly identified for the cohort of 20 speakers. Using a binary SVM regression model with the same closed set of 20 speakers, 72.5% of test utterances were correctly identified. These results are similar to previous investigations of the voice-source waveform for speaker identification using the YOHO corpus [302, 171].

The identification rates of the regression model are inferior to the multiclass support-vector model, and there are logical reasons for this. Each single regression model attempts only to differentiate between binary classes, and doesn’t model the variations of non-client speakers when training a client model. Such a model structure is better suited to a speaker verification paradigm where accept and reject decisions are required, and not selection from a group. Such a paradigm would require the introduction of some measure of typicality, that is a measure of how the speakers source-features are distributed over the speaker population of interest for the system [327].

For the cohort of 20 speakers (the largest used in both experiments), correct identification rates of 85.3% for the multiclass system and 72.5% for the regression system
compare well to previous investigations of voice-source features using the YOHO [71] corpus, although it must be noted that we use smaller cohort sizes here. An analytic model of the glottal wave based on parameterising its opening, closing and return phases obtained a classification rate across all of YOHO (averaged over male and females) of 71.4% [302]. A classification rate on all of YOHO of 64% using a cepstral parameterisation of the spectrum of the voice-source was achieved in [171].

To implement such identification systems as explored here, especially on large or open cohorts, would required significant computational power. Clients would also be required to give more enrolment speech than would be convenient. These points are acknowledged, however the focus of this voice-source investigation using discriminating models has been on exploring the identity information content of the source-frame features. To this end our aims have been achieved.

Finally we believe these results further support the hypothesis that data driven models of the voice-source [172, 378] are more useful for speaker recognition than analytic models parameterising the sections of a pitch period of the voice-source waveform (opening, closing, returning), such as those proposed earlier originally for speech synthesis such as Liljencrats-Fant [132, 131] and Rosenberg [336]. These analytic models of the glott-
tal waveform are suitable for speech synthesis but we believe they do not capture the nuanced variations that differentiate speakers.

We continue testing this point over the next experimental Section 5.7 where generative models for the source-frames or parametric representations derived from them are explored. There are several significant reasons for developing generative models for these features. This would alleviate the requirement for excessive amounts of enrolment speech and allow constrained adaptation of distribution models using Bayesian methods which are particularly beneficial when presented with limited data. Further advantages include employing scoring based on probabilistic measures which quantify a systems assessments in a more logically rigorous manner than discriminative methods. This point particularly holds for the use of such features in a forensic context where the reporting methodology should be consistent across practitioners and cases, the most logically rigours approach to achieving this being application of a Bayesian framework to update beliefs based upon the presented evidence [272], and this is better adhered to by generative/probabilistic models. In the experiments of the next section the r-norm score post-processing method introduced in Chapter 4 is also applied to the classifications of the generative glottal systems and other approaches explored there.

Figure 5.16: Percentages of the utterances of each speaker of the cohort of size 20 correctly identified as belonging to that speaker.
5.7 Experiment 6: Glottal Information with r-Norm

A large speaker verification experiment was performed on the AusTalk corpus where glottal information was paired with a MFCC baseline, and the r-norm score post-processing method was applied to the scores of both the individual and score fusion systems. Source-frames were fitted to glottal waveforms estimated in two different manners, and these were in turn parameterised and modelled by four different methods. Score fusion demonstrates that the glottal features typically provided some complementary information to the MFCC baseline. However no improvements in either EER or minDCF were observed of the same magnitude as those resulting from the application of r-norm.

5.7.1 Introduction

We test in this section the performance of several glottal estimation and modelling methods for the task of speaker recognition. Still used are the normalised time domain waveform features (source-frames) representing the derivative of a speakers volume-velocity airflow through the glottis during phonation. The derivative glottal waveform estimates were obtained via inverse-filtering as in the previous sections as well as by a decomposition of the speech signal in the complex cepstrum domain. These are then parameterised in four different approaches which included the use of the well known Liljencrants-Fant model for the glottal flow.

These experiments allow us to draw several inferences in regard to which estimation, parameterisation and modelling methods for the glottal signal are most beneficial for the task of speaker recognition.

5.7.2 Experimental Design

A subset of the AusTalk [85, 411] dataset was used to perform gender dependent speaker verification experiments. 100 males and 100 females were obtained from the AusTalk recording locations as shown in Table 4.2. The session 1 reading of a short story was used as training data whilst the testing data was taken from the session 2 interview speech. Only speakers with completed story and interview components were used. Session 2 was recorded a minimum of one week after session 1 and the interview generated approximately 10 minutes of speech per subject once the research assistants prompting
questions and non-speech segments had been removed via an empirically determined frame energy threshold. In order to apply the $r$-norm model to the various systems that we examine, each subject's interview data was evenly divided into 20 segments, each of which was then treated as a single trial. All data was downsampled from 44.1 kHz, 32 bit depth to 16 kHz, 16 bit depth.

Background models we trained in all experiments on the relevant features extracted from the ANDOSL corpus [260], using all 200 of the recorded phonetically diverse sentences and using all 54 Australian speakers of each gender. ANDOSL was downsampled to match the used AusTalk sampling rates.

A baseline mel-frequency cepstral coefficient (MFCC) system was used against which we compare the glottal waveform information streams as well as quantifying their complementary nature. The scores from this MFCC system were used in Chapter 4 to empirically validate the proposed score post-processing model $r$-norm and a full description of this baseline is given in 4.5.1. In summary a GMM-UBM [327] system was used to model 32 dimensional MFCC features extracted from 25 ms frames shifted by 10 ms.

In tandem to this baseline system we extracted 2 different estimates of the derivative volume-velocity flow waveforms. The first estimate was obtained through a closed-phase inverse filtering using the glottal closure instant detection algorithm presented in [115]. This is the common method based on linear prediction used throughout this thesis and as detailed of the description of Section 3.5 on extracting the source-frame features. The second method used for obtaining estimates of the speakers glottal waveform is based on the complex cepstrum [110, 111]. This algorithm is reviewed in section 3.3 on methods for estimating glottal waveforms from digitised speech and the Matlab implementation provided by the original authors and available in the Glottal Analysis Toolbox (GLOAT) package [109] was used to implement it. We denote the inverse-filtered estimates by IF and the estimates obtained by the complex cepstrum decomposition by CC. From both of these time-domain glottal waveforms the normalised source-frame representation as described in Section 3.5 were then extracted. Source-frames were normalised to 256 samples.

Figure 5.17 shows the mean source-frame taken over a single utterance resulting from inverse-filtering and from the complex-cepstrum decomposition. The resulting waveform from the complex cepstrum estimation shows less closed phase ripple (around the 150th sample) as well as a more pronounced negative peak located at the boundary of the opening to return phases (approximately sample 125). This is typical of the differences observed between the two methods of estimation.

Four different methods for parameterising and scoring the set of IF and CC estimated
source-frame vectors were explored:

1. Parameters of a fitted Liljencrants-Fant model [131] in a GMM-UBM classifier.
2. GMM-UBM modelling of the coefficients of fitted piecewise polynomials.
3. A Probabilistic Linear Discriminant Analysis (PLDA) classifier.
4. The Euclidean distance method on mean source-frames as described in Section 5.4.

We now describe each of these four systems in turn. In the first system the four-parameter Liljencrants-Fant (LF) model [131, 132] for the derivative glottal flow waveform was fitted to each source-frame. Like the source-frame, the LF model describes only the coarse structure of the glottal flow. It is a piecewise model that describes the open, closed and return phases while capturing the pulse shape and the peak glottal flow. Proposed for speech synthesis, it has previously been used for speaker identification to reduce an MFCC EER by 5% relative [302]. The LF glottal flow model for the derivative
of the volume-velocity airflow of a single glottal cycle \( v_{LF}(t) \) is given by (5.11):

\[
v_{LF}(t) = \begin{cases} 
0 & 0 \leq t < T_o \\
E_0 e^{\alpha(t-T_o)} \sin[\omega_0(t - T_o)] & T_o \leq t < T_e \\
E_1 \left[ e^{-\beta(T_e-T_c)} - e^{-\beta(t-T_c)} \right] & T_e \leq t < T_c 
\end{cases}
\]

Closed Phase  
Opening Phase  
Return Phase  \hspace{1cm} (5.11)

This model was fitted to the central 128 samples (single glottal cycle) of each source-frame by minimising the sum of residual square errors via the LevenbergMarquardt algorithm by using the \textit{lsqcurvefit} function from the Matlab Optimisation Toolbox. Figure 5.18 shows a LF model fitted to a single source-frame estimated by inverse-filtering.

Figure 5.18: \textit{Liljencrants-Fant glottal model fitted by least-squares to a single source-frame. The opening phase occurs over samples [25,87] and closed phase over [87,105].}  

The four LF model parameters \( E_0, \alpha, \beta \) and \( \omega_0 \) were modelled via Gaussian mixture models (GMM). The same features extracted from the 54 ANDOSL speakers were used to train a GMM universal background model (UBM) which formed a prior density from which the client models were adapted in a MAP process [158].

The timing parameters \( T_o, T_e, T_c \) representing the opening instant, instant of derivative flow minima and closing instant respectively were found during the glottal closure.
The instant detection step of the inverse-filtering process and used for both the IF and CC estimates. As described in [132], the full set of timing parameters are sufficient to specify the whole LF model. Here these selected timing parameters are used to demarcate the piecewise opening and returning regions over which the LF model segments are fitted.

The GMM used to model these features comprised 32 mixtures with diagonal covariances. Only the means were adapted from the UBM during the MAP process. Log likelihood ratios were obtained from test trials per (4.3).

The second system fitted piecewise polynomials of degrees 3, 7, 3 over the sections of source-frame samples [51, 100], [101, 150], [151, 200] respectively. Also included in the second section was the sample indexing the residual maxima. This was included to capture information pertaining to the sharp negative peak demarcating the commencement of the return phase that could only be captured with a polynomial of such high degree that considerable overfitting resulted. This piecewise fitting then resulted in a feature vector of length 17 coefficients. The concatenating of this index and the polynomial coefficients parameterisation was also used in the exploration of the voice-source waveform for forensic voice comparison reported in Section 6.4. See Figure 6.5 for an example of these piecewise polynomials fitted to a mean of source-frames found by inverse filtering.

These 17 dimensional vectors were modelled via GMM also with 32 mixtures and mean only MAP adapted from a UBM trained on the 54 ANDOSL background speakers. Again scores were calculated as log likelihood ratios for each target or non-target trial.

The third method of parameterising and modelling the extracted source-frames was to use a probabilistic version of linear discriminant analysis (PLDA) as proposed in [305] for face recognition. With this approach we assume that source-frames themselves are produced from a generative model that encompasses the variation both within and between speakers source-frames and also includes a Gaussian noise component. This model is defined as:

\[ \bar{X}_{ij} = \mu + S_u_i + N_v_{ij} + \epsilon_{ij} \]  

(5.12)

where the \( j^{th} \) training source-frame of the \( i^{th} \) speaker \( \bar{X}_{ij} \) is decomposed into a signal component \( \mu + S_u_i \) that depends only on the speaker and a noise component \( N_v_{ij} + \epsilon_{ij} \) which captures the noise or variation within individual speakers’ source-frames. The columns of the signal matrix \( S \) describe a basis for the inter-speaker source-frame variation and the position in this subspace of an individual source-frame \( \bar{X}_{ij} \) is specified by the \( u_i \) vector. All source-frames originating from a the speech of a single speaker
are assumed to derive from this latent identity variable \( u_i \). Similarly the noise matrix \( N \) contains in its columns a basis for the intra-speaker source-frame variation and the vector \( v_{ij} \) represents the position of source-frame \( \hat{X}_{ij} \) within this subspace. Both vectors \( u_i \) and \( v_{ij} \) are distributed as \( N(0, I) \), where \( N(a, b) \) is the multivariate normal with mean \( a \) and covariance matrix \( b \). Remaining residual variation is captured by the pure noise term \( \epsilon_{ij} \) which is distributed as \( N(0, \Sigma) \) where \( \Sigma \) is diagonal. The speaker and source-frame independent factor \( \mu \) is the mean of all training speakers source-frames about which observed variation is described. Thus the PLDA model consists of the parameters \( \Theta = (\mu, S, N, \Sigma) \).

Training of this PLDA model for each gender was achieved as follows. Source-frames extracted from all 54 ANDOSL speakers were used with the expectation-maximisation (EM) algorithm [99] to iteratively estimate point estimates of the PLDA model parameters \( \Theta \) (E-step) and then the full posterior distributions over the latent variables \( u_i \) and \( v_{ij} \) (M-step) in such a way that the likelihood of the ANDOSL training source-frames under the PLDA model is monotonically non-decreasing.

We performed 5 iterations of the EM algorithm to train the PLDA model \( \Theta \) for each gender on the ANDOSL data. Only the first 40 sentences of each of the 54 ANDOSL users were used and the median their source-frames used, meaning the PLDA model was trained on \( 54 \times 40 \) source-frames for each gender. Only the central section of samples \([40, 110]\) of the source-frames were used. The signal subspace described by matrix \( S \) had 32 dimensions while the noise subspace \( N \) had 30 dimensions.

Having specified the PLDA model \( \Theta \), there is no enrolment or training of speaker models whereby any point estimate of the latent variable \( u_i \) implied by speaker \( i \)'s training source-frames is estimated. Instead, at the speaker verification stage we calculate a posterior probability via Bayes’ rule that two source-frames shared the same latent identity variable, irrespective of what that exact vector \( u_i \) was. This is done by marginalising over the unknown latent identity variables. Models then in fact represent a relationship between all of the observed source-frame data (from a single speakers training data and from the test probe data we are evaluating) and the hidden identity variables. See [305] for an example regarding calculating this posterior probability.

The final and forth system for modelling and scoring the source-frame data was a non-generative approach with the arithmetic Euclidean distance metric method used to compare source-frames in their native \( \mathbb{R}^n \) as described in Section 5.4.

The complementary potential of each of these 8 glottal information systems (2 different glottal waveform estimates IF and CC × 4 different modelling & scoring methods)
with the MFCC baseline was then explored through 2 separate score fusion approaches. The first approach is based on a convex combination of the MFCC log likelihood ratio score $S_{MFCC}$ and the score from the particular glottal system being fused $S_G$:

$$S_{FUSED} = w \times S_{MFCC} + (1 - w) \times S_G \quad \text{where} \quad w \in [0, 1]$$

(5.13)

The scores from the discriminative distance modelling $S_{Distance}$ of the forth glottal system were transformed per (5.14) so that like the generative MFCC log likelihood scores a greater value for a target trial was optimal. This transformation was sufficient given $0 < S_{Distance} \ll 1$ for all trials.

$$S_{Distance} \mapsto 1 - S_{Distance}$$

(5.14)

The second score fusion approach was to also use a linear fusion but with weights trained by logistic regression. This was achieved via Niko Brummer’s FOCAL toolbox [57].

Finally the score post-processing algorithm r-norm introduced in Chapter 4 were then applied to both the scores of each individual system and to scores from the fused systems. The r-norm results are based on 5 fold cross-validation over the 20 probes per speaker with 16 score vectors used as the development set for training the TGPR function of r-norm and the remaining held out 4 score vectors used to validate it at each fold.

In all cases the ideal matrix $I$ specified in learning the TGPR regression function described zero variance distributions, meaning that it contained only two distinct values $i_T$ and $i_{NT}$ representing ideal target scores and ideal non-target scores respectively. No parameter search was performed over these two values other than for the MFCC baseline system as reported in Section 4.5. Rather they were all specified with reference to the distribution of target and non-target trial scores in the format of Pattern 1 as noted in 4.8 in the summary of Chapter 4. Recall that this pattern was defined by the ordering: $[\mu_{non-target}, i_{NT}, \hat{x}, i_T, \mu_{target}]$, where $\hat{x}$ is the location of the EER threshold.

We now report the results of these methods, providing the equal-error rates (EER) and minima of the NIST 2006 detection cost function (minDCF) for each individual system and for each score fusion system as well as for the r-norm algorithm applied of each of these.
5.7.3  Results

Figure 5.19 shows the log-likelihood (LL) of the training ANDOSL MFCC data for the UBM as the number of mixtures is increased in powers of 2 for the male and female data and was used to determine the number of GMM mixtures for the MFCC systems. From the training LL data it was determined that overfitting began to occur after 1024 mixtures, beyond which certain components were responsible for only a single cepstral frame observation⁶. By this method it was also determined that 32 mixtures were optimal for all GMM modelling of glottal features.

![Figure 5.19: Log-likelihood of the ANDOSL MFCC UBM female training data as a function of the increasing number of GMM mixtures.](image)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Pre r-norm EER %</th>
<th>Pre r-norm minDCF</th>
<th>Post r-norm EER %</th>
<th>Post r-norm minDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>5.63</td>
<td>0.0293</td>
<td>0.07</td>
<td>0.023</td>
</tr>
<tr>
<td>SF(IF) - LF</td>
<td>29.25</td>
<td>0.0985</td>
<td>11.94</td>
<td>0.0713</td>
</tr>
<tr>
<td>SF(IF) - Poly</td>
<td>39.05</td>
<td>0.0993</td>
<td>43.00</td>
<td>0.1</td>
</tr>
<tr>
<td>SF(CC) - LF</td>
<td>27.75</td>
<td>0.098</td>
<td>11.49</td>
<td>0.0586</td>
</tr>
<tr>
<td>SF(CC) - Poly</td>
<td>46.10</td>
<td>0.1</td>
<td>48.12</td>
<td>0.1</td>
</tr>
<tr>
<td>SF(IF) - PLDA</td>
<td>26.22</td>
<td>0.0964</td>
<td>8.77</td>
<td>0.0617</td>
</tr>
<tr>
<td>SF(CC) - PLDA</td>
<td>33.80</td>
<td>0.0965</td>
<td>10.97</td>
<td>0.067</td>
</tr>
<tr>
<td>SF(IF) - Distance</td>
<td>26.33</td>
<td>0.0954</td>
<td>7.54</td>
<td>0.0592</td>
</tr>
<tr>
<td>SF(CC) - Distance</td>
<td>27.62</td>
<td>0.0949</td>
<td>9.65</td>
<td>0.0718</td>
</tr>
</tbody>
</table>

Table 5.11: EER and minDCF for the MFCC baseline and each of the 8 glottal systems for the 100 female AusTalk speakers. SF=source-frame, IF=inverse-filtered, CC=complex-cepstrum decomposition, LF=Liljencrants-Fant parameters, Poly=Polynomial coefficients.

⁶“responsible” in the sense of cepstral frames being assigned to mixture centroids whose individual Gaussian mixture had maximum likelihood.
The results of each individual system for the 100 female AusTalk speakers are shown in Table 5.11 along with the post r-norm EER and minDCF for each system. As for the males, the polynomial features are not found to be discriminating and the r-norm post-processing also fails, increasing the already near chance level EER for both the IF and CC polynomial systems. The base results of the distance and PLDA systems are weak in comparison to the MFCC baseline but the r-norm method again finds enough information within their scores to reduce all of their EER values to \( \sim 10\% \), as occurred with the male speakers.

Shown in Table 5.12 are the EER and minDCF for all of the male individual systems as well as the cross-validated r-norm results for each of them. The performance of the individual glottal systems is weak although the r-norm method finds enough information in the scores of the PLDA and distance metric systems to be able to reduce the EER to \( \sim 10\% \) for both of these features fitted to each of the IF and CC estimated source-frames. The GMM modelling of piecewise fitted polynomial coefficients was particularly weak for both IF and CC estimated glottal waveforms.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Pre r-norm EER</th>
<th>Pre r-norm minDCF</th>
<th>Post r-norm EER</th>
<th>Post r-norm minDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>5.26</td>
<td>0.0214</td>
<td>0.06</td>
<td>0.00019</td>
</tr>
<tr>
<td>SF(IF) - LF</td>
<td>27.98</td>
<td>0.0992</td>
<td>12.09</td>
<td>0.0569</td>
</tr>
<tr>
<td>SF(IF) - Poly</td>
<td>33.67</td>
<td>0.0998</td>
<td>20.83</td>
<td>0.0995</td>
</tr>
<tr>
<td>SF(CC) - LF</td>
<td>28.87</td>
<td>0.10</td>
<td>16.06</td>
<td>0.0755</td>
</tr>
<tr>
<td>SF(CC) - Poly</td>
<td>39.14</td>
<td>0.0996</td>
<td>44.2</td>
<td>0.0994</td>
</tr>
<tr>
<td>SF(IF) - PLDA</td>
<td>26.2</td>
<td>0.092</td>
<td>9.53</td>
<td>0.092</td>
</tr>
<tr>
<td>SF(CC) - PLDA</td>
<td>31.3</td>
<td>0.0989</td>
<td>11.97</td>
<td>0.0684</td>
</tr>
<tr>
<td>SF(IF) - Distance</td>
<td>29.24</td>
<td>0.0960</td>
<td>8.44</td>
<td>0.0606</td>
</tr>
<tr>
<td>SF(CC) - Distance</td>
<td>28.59</td>
<td>0.0961</td>
<td>9.85</td>
<td>0.0714</td>
</tr>
</tbody>
</table>

Table 5.12: EER and minDCF for the MFCC baseline and each of the 8 glottal systems for the 100 male AusTalk speakers. SF=source-frame, IF=inverse-filtered, CC=complex-cepstrum decomposition, LF=Liljencrants-Fant parameters system, Poly=Polynomial coefficients system.

Tables 5.13 and 5.14 show the female AusTalk speakers’ score fusion results for the weighted and logistic regression approaches respectively. None of either the IF and CC polynomial systems, the LF parameters fitted to the CC estimated source-frames nor the distance metric on the inverse-filtered source-frames showed any benefit with the MFCC baseline by weighted fusion at any point over the range \( w \in [0,1] \). This meant that the weighted fused scores for these systems were identically the MFCC baseline system
and thus \( r \)-norm results are not reported for these in the final two columns of Table 5.13. The logistic regression fusions all resulted in minor decreases in the EER relative to the MFCC baseline, although the maximum relative improvement was only 11% as achieved by the distance metric classifier on the IF estimated source-frames. The \( r \)-norm method was able to improve the performance of each individual fused system. However for no system was the \( r \)-norm EER lower than the \( r \)-norm method applied directly to the MFCC baseline scores alone.

<table>
<thead>
<tr>
<th>Weighted Fusion</th>
<th>Weight w</th>
<th>Pre ( r )-norm</th>
<th>Post ( r )-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER %</td>
<td>minDCF</td>
<td></td>
</tr>
<tr>
<td>MFCC baseline</td>
<td>5.63</td>
<td>0.0293</td>
<td>0.07 0.023</td>
</tr>
<tr>
<td>MFCC + SF(IF) - LF</td>
<td>0.75</td>
<td>5.40 0.0287</td>
<td>0.63 0.0064</td>
</tr>
<tr>
<td>MFCC + SF(IF) - Poly</td>
<td>1</td>
<td>5.63 0.0293</td>
<td>-   -</td>
</tr>
<tr>
<td>MFCC + SF(CC) - LF</td>
<td>1</td>
<td>5.63 0.0293</td>
<td>-   -</td>
</tr>
<tr>
<td>MFCC + SF(CC) - Poly</td>
<td>1</td>
<td>5.63 0.0293</td>
<td>-   -</td>
</tr>
<tr>
<td>MFCC + SF(IF) - PLDA</td>
<td>0.7</td>
<td>5.24 0.0275</td>
<td>0.50 0.0050</td>
</tr>
<tr>
<td>MFCC + SF(CC) - PLDA</td>
<td>0.7</td>
<td>5.42 0.0276</td>
<td>0.62 0.0066</td>
</tr>
<tr>
<td>MFCC + SF(IF) - Distance</td>
<td>1</td>
<td>5.63 0.0293</td>
<td>-   -</td>
</tr>
<tr>
<td>MFCC + SF(CC) - Distance</td>
<td>0.85</td>
<td>5.30 0.0268</td>
<td>1.07 0.0108</td>
</tr>
</tbody>
</table>

Table 5.13: Weighted fusion results for the AusTalk female speakers and the post \( r \)-norm results for each combination. \( r \)-norm results are not given for the four systems for which no improvement upon the MFCC baseline was observed (being equal to the MFCC).

<table>
<thead>
<tr>
<th>Logistic Regression Fusion</th>
<th>Pre ( r )-norm</th>
<th>Post ( r )-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER %</td>
<td>minDCF</td>
</tr>
<tr>
<td>MFCC baseline</td>
<td>5.63 0.0293</td>
<td>0.07 0.023</td>
</tr>
<tr>
<td>MFCC + SF(IF) - LF</td>
<td>5.40 0.0287</td>
<td>0.43 0.0016</td>
</tr>
<tr>
<td>MFCC + SF(IF) - Poly</td>
<td>5.42 0.0284</td>
<td>0.44 0.0017</td>
</tr>
<tr>
<td>MFCC + SF(CC) - LF</td>
<td>5.35 0.0276</td>
<td>3.54 0.0091</td>
</tr>
<tr>
<td>MFCC + SF(CC) - Poly</td>
<td>5.35 0.0276</td>
<td>0.14 0.0033</td>
</tr>
<tr>
<td>MFCC + SF(IF) - PLDA</td>
<td>5.30 0.0275</td>
<td>0.49 0.0017</td>
</tr>
<tr>
<td>MFCC + SF(CC) - PLDA</td>
<td>5.59 0.0290</td>
<td>0.56 0.0018</td>
</tr>
<tr>
<td>MFCC + SF(IF) - Distance</td>
<td>5.00 0.0258</td>
<td>0.89 0.0090</td>
</tr>
<tr>
<td>MFCC + SF(CC) - Distance</td>
<td>5.33 0.0276</td>
<td>0.13 0.00023</td>
</tr>
</tbody>
</table>

Table 5.14: Logistic regression fusion results for the 100 AusTalk female speakers and the post \( r \)-norm results for each combination.
The male score fusion results from combining each individual glottal system in turn with the MFCC baseline are reported in Tables 5.15 and 5.16 for the weighted and logistic regression score fusion approaches respectively. In all cases, except the FOCAL fusion of the inverse-filtered source-frame distance system, were the glottal systems seen to be complementary to the baseline MFCC system. No considerable reductions of either EER or minDCF are observed, however, with the best relative improvement of EER seen to be 20% for the weighted fusion of the LF parameters fitted to the CC estimated source-frames.

<table>
<thead>
<tr>
<th>Weighted Fusion</th>
<th>Weight w</th>
<th>Pre r-norm EER</th>
<th>minDCF</th>
<th>Post r-norm EER</th>
<th>minDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MFCC baseline</strong></td>
<td>1</td>
<td>5.26</td>
<td>0.0214</td>
<td>0.06</td>
<td>0.00019</td>
</tr>
<tr>
<td>MFCC + SF(IF) - LF</td>
<td>0.5</td>
<td>5.01</td>
<td>0.0211</td>
<td>0.17</td>
<td>0.00062</td>
</tr>
<tr>
<td>MFCC + SF(IF) - Poly</td>
<td>0.9</td>
<td>5.10</td>
<td>0.0209</td>
<td>0.49</td>
<td>0.005</td>
</tr>
<tr>
<td>MFCC + SF(CC) - LF</td>
<td>0.4</td>
<td>4.25</td>
<td>0.0208</td>
<td>0.2</td>
<td>0.00036</td>
</tr>
<tr>
<td>MFCC + SF(CC) - Poly</td>
<td>0.95</td>
<td>5.15</td>
<td>0.0211</td>
<td>1.99</td>
<td>0.0207</td>
</tr>
<tr>
<td>MFCC + SF(IF) - PLDA</td>
<td>0.75</td>
<td>4.71</td>
<td>0.0184</td>
<td>0.50</td>
<td>0.0051</td>
</tr>
<tr>
<td>MFCC + SF(CC) - PLDA</td>
<td>0.9</td>
<td>5.03</td>
<td>0.02</td>
<td>0.91</td>
<td>0.0092</td>
</tr>
<tr>
<td>MFCC + SF(IF) - Distance</td>
<td>0.95</td>
<td>5.05</td>
<td>0.0196</td>
<td>1.03</td>
<td>0.0085</td>
</tr>
<tr>
<td>MFCC + SF(CC) - Distance</td>
<td>0.85</td>
<td>5.00</td>
<td>0.0202</td>
<td>0.32</td>
<td>0.0032</td>
</tr>
</tbody>
</table>

Table 5.15: Weighted fusion results for the 100 AusTalk male speakers as well as the post r-norm results for each combination.

<table>
<thead>
<tr>
<th>Logistic Regression Fusion</th>
<th>EER</th>
<th>minDCF</th>
<th>Post r-norm EER</th>
<th>minDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MFCC baseline</strong></td>
<td>5.26</td>
<td>0.0214</td>
<td>0.06</td>
<td>0.00019</td>
</tr>
<tr>
<td>MFCC + SF(IF) - LF</td>
<td>5.05</td>
<td>0.0207</td>
<td>0.05</td>
<td>0.00041</td>
</tr>
<tr>
<td>MFCC + SF(IF) - Poly</td>
<td>5.10</td>
<td>0.0215</td>
<td>0.2</td>
<td>0.00073</td>
</tr>
<tr>
<td>MFCC + SF(CC) - LF</td>
<td>4.75</td>
<td>0.0219</td>
<td>0.027</td>
<td>0.00035</td>
</tr>
<tr>
<td>MFCC + SF(CC) - Poly</td>
<td>5.15</td>
<td>0.0214</td>
<td>0.061</td>
<td>0.00023</td>
</tr>
<tr>
<td>MFCC + SF(IF) - PLDA</td>
<td>4.69</td>
<td>0.0183</td>
<td>0.23</td>
<td>0.00037</td>
</tr>
<tr>
<td>MFCC + SF(CC) - PLDA</td>
<td>5.10</td>
<td>0.0209</td>
<td>0.06</td>
<td>0.00017</td>
</tr>
<tr>
<td>MFCC + SF(IF) - Distance</td>
<td>5.38</td>
<td>0.0199</td>
<td>0.29</td>
<td>0.0011</td>
</tr>
<tr>
<td>MFCC + SF(CC) - Distance</td>
<td>5.00</td>
<td>0.0203</td>
<td>0.11</td>
<td>0.00054</td>
</tr>
</tbody>
</table>

Table 5.16: Logistic regression fusion results for the 100 AusTalk male speakers as well as the post r-norm results for each combination.
We also report speaker identification results based on these scores for comparison with our earlier work in Sections 5.5 and 5.6 and with other literature on the use of the glottal waveform for speaker recognition [302, 117]. Shown in Tables 5.17 and 5.18 are the correct-speaker identification rates\(^7\) for the 9 female and males systems respectively.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Identification Rate %</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>93.30</td>
</tr>
<tr>
<td>IF</td>
<td></td>
</tr>
<tr>
<td>LF</td>
<td>14.40</td>
</tr>
<tr>
<td>Poly</td>
<td>9.95</td>
</tr>
<tr>
<td>PLDA</td>
<td>12.50</td>
</tr>
<tr>
<td>Distance</td>
<td>21.45</td>
</tr>
<tr>
<td>CC</td>
<td></td>
</tr>
<tr>
<td>LF</td>
<td>18.25</td>
</tr>
<tr>
<td>Poly</td>
<td>6.95</td>
</tr>
<tr>
<td>PLDA</td>
<td>14.45</td>
</tr>
<tr>
<td>Distance</td>
<td>16.30</td>
</tr>
</tbody>
</table>

Table 5.17: Identification rates for the MFCC baseline and each of the 8 glottal systems for the 100 female AusTalk speakers.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Identification Rate %</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>97.55</td>
</tr>
<tr>
<td>IF</td>
<td></td>
</tr>
<tr>
<td>LF</td>
<td>12.90</td>
</tr>
<tr>
<td>Poly</td>
<td>10.10</td>
</tr>
<tr>
<td>PLDA</td>
<td>24.00</td>
</tr>
<tr>
<td>Distance</td>
<td>18.55</td>
</tr>
<tr>
<td>CC</td>
<td></td>
</tr>
<tr>
<td>LF</td>
<td>15.75</td>
</tr>
<tr>
<td>Poly</td>
<td>8.55</td>
</tr>
<tr>
<td>PLDA</td>
<td>12.35</td>
</tr>
<tr>
<td>Distance</td>
<td>14.80</td>
</tr>
</tbody>
</table>

Table 5.18: Identification rates for the MFCC baseline and each of the 8 glottal systems for the 100 male AusTalk speakers. IF=inverse-filtered, CC=complex-cepstrum decomposition

Shown in Figures 5.20 and 5.21 are the progressive decrease in identification rates as the closed cohorts of speakers are increased for females and males respectively with each of the 4 glottal modelling systems (on the inverse-filtered estimated glottal waveforms). Also shown is the MFCC rate and the chance identification rate for each cohort size from 5 to 100 speakers. These were produced from the scores of the full 100 speakers by randomly sampling each cohort of size \(n\) speakers and removing the models and trials necessary.

\(^7\)This is the correct average recall for all speakers based on assignment to maximum scores without reference to thresholds.
of the remaining $100 - n$ speakers from the scores to maintain a closed set identification paradigm. The distance and PLDA methods are seen to perform superior to the function fitting approaches.

![Female Identification rates](image1)

Figure 5.20: Identification rates against cohort sizes of 5 to 100 randomly selected female speakers for the 4 IF glottal systems and the MFCC system.

![Male Identification rates](image2)

Figure 5.21: Identification rates against cohort sizes of 5 to 100 randomly selected male speakers for the 4 glottal systems (on the IF estimated waveforms) and the MFCC system. The chance identification rate for each cohort size is also shown.

Figures 5.20 and 5.21 were generated in an attempt to gain some quantitative insight into the chance population hit rate regarding similar glottal waveforms as based on the various glottal parameterisations examined. There is a large difference observed in the
identification rates for each cohort size between the MFCC and glottal systems. Whilst
the MFCC system displays an absolute decrease in correct identification rate of ~ 5%
in both genders in moving from cohorts of 5 to 100 speakers, the glottal systems identi-
fication rates all decrease considerably. Of interest however is the stabilising behaviour
of the identification rates begun from approximately cohorts of size 60 speakers which
was observed in both male and female graphs and all 4 glottal systems. This provides
limited evidence that these time domain glottal waveform parameterisations are able to
correctly identify with approximately 10 to 20% accuracy in large (> 100) groups of
speakers. Equivalently stated, this is weak evidence for the claim that the time domain
glottal waveforms of two speakers chosen at random from the population (having the
same gender) will match significantly enough to be misidentified by these methods with
an 80 to 90% chance. These results are also useful given the fact that many speaker
recognition experiments with the glottal waveform [302, 389], including those in earlier
sections of this chapter, are reported on cohorts of less than 100 speakers.

The identification rates given are well below the ~ 60% rates for the male and female
speaker systems consisting of 112 males and 56 females reported in [302] on the TIMIT
database. TIMIT however was also recorded in optimal environment conditions (minimal
phase distortions or spurious noise) but consisted of no intersession variation. It is likely
that glottal information over multiple sessions recorded in practical environments would
be minimally informative given the identification rates in [302] decreased to ~ 10% when
using speech transmitted via telephone.

Of the several variables explored in this experiment the results (EER and identifi-
cation rates) provide evidence that data-driven parameterisations of estimated glottal
information (PLDA modelling, distance metric on mean source-frames) are superior8 to
function fitting approaches using piecewise polynomial estimates or the LF glottal flow
model.

Information from the glottal flow was also shown in almost all cases to complement
the MFCC spectral magnitude features. Therefore in circumstances where estimates of
the glottal flow waveform may be obtained with confidence, this signal may be used to
boost the recognition performance.

Strong evidence for the ability of the r-norm method to improve classification accu-

cracy was also again demonstrated with the largest improvements in recognition derived
from its application.

---

8This includes taking the MFCC of the LP residual, not done here, for which strong empirical evidence
exists [254] and the VSCE parameter introduced in [171].

136
5.8 Chapter Summary

The glottal waveform has been shown over several text-independent speaker identification and verification experiments to contain speaker discriminating information. The focus of the majority of these experiments was the proposed source-frame feature that captures the shape of the glottal flow waveform in a length and amplitude normalised pitch synchronous manner.\(^9\)

Evidence was presented for the beneficial property for text-independent speaker recognition that the source-frame feature, and implicitly the glottal flow waveform shape, possessed no dependence upon the phonetic content for the utterance.

Discriminative modelling with the length normalised Euclidean distance metric and SVM models were then investigated. The source-frame feature is a time domain waveform, such that unlike most speaker recognition features it can be plotted and speaker differences and natural variation visualised.\(^10\) From this perspective, our investigations began with the use of the distance metric to compare source-frames in their native Euclidean space, a method which was shown to strongly distinguish speaker identity. A key component of this approach, as for the Euclidean distance metric method, was to form an average source-frame from a collection (or ‘block’) of greater than 100 source-frames. Through this process stable waveforms for individual speakers were able to be obtained through diminishing the effects of imperfect inverse-filtering and other nuisance effects that may be considered to create variations distributed around high fidelity estimates.

SVM classification with both multiclass and regression models on source-frames, reduced in dimension by a PCA, was shown to be able to differentiate well speaker’s individual source-frames and subsequently speaker’s full utterances.

GMM modelling of fitted function parameters (polynomial curves and LF model) were also explored but found to be typically less informative than data-driven approaches which included probabilistic linear discriminant analysis modelling of the source-frame data. Estimation by closed phase inverse-filtering was seen to be typically slightly beneficial for identifying speakers than estimates made by the decomposition of the complex cepstrum, using the mixed phase properties of the glottal flow.

The glottal flow information was also shown to be beneficial in improving the recognition rates of systems using the most informative speaker recognition features, namely MFCC. Although only small improvements were typically observed; the extra complexity of implementing such systems may only be beneficial in certain high risk environments.

\(^9\)The rare VSCC feature representation [171] was also successfully replicated.

\(^{10}\)Observe and compare the multiple plots of same-speaker and different-speaker source-frames presented in the Appendix, Section A.1.
Chapter 6

Glottal Waveforms: Forensic Voice Comparison

6.1 Introduction

In this chapter we examine briefly the use of the glottal waveform as an information source for forensic voice comparison. The task of forensic voice comparison (FVC) is to quantitatively evaluate some given speech evidence with respect to the likelihood of the given sample having been uttered by a known suspect against it having been uttered by someone else from the population of potential offenders.

A fuller description of this task is given in Section 6.2 along with a review of the existing literature regarding the modern development of FVC and the considerably limited use of a speakers’ glottal waveform for aiding in this evaluation. Section 6.3 presents the results of a small human listening task on the new YAFM database [332]. This experiment was performed to better understand the extent to which the YAFM speakers were perceptually similar and gauge the suitability and difficulty of using the database for carrying out a FVC experiment. In Section 6.4 we then report on a likelihood ratio FVC experiment using statistical modelling that is performed on YAFM, where the information from the estimated glottal waveforms of each speaker is used to provide complementary information to a cepstral system.
6.2 Literature Review: Forensic Voice Comparison and the Glottal Waveform

In section 6.2.1 we describe the forensic voice comparison task and the evolution towards a more rigorous and scientific treatment of it. With this context provided, in section 6.2.2 we then review the sparse literature addressing the use of the glottal flow waveform as a source of information for accurately performing a given forensic voice comparison.

6.2.1 The Evolving Paradigm of Forensic Voice Comparison

Forensic voice comparison is the application of scientific methods to the analysis of voice samples where the identity of one or both of the generating speakers is unknown or disputed. A FVC may be performed to aid investigators in pursuit of specific theories or leads, as a form of evidence presented to the trier of fact in a court of law (i.e. judge, jury), or even in matters of private investigation [333].

Arguably forensic voice comparison (FVC) is the least understood speaker recognition task and over the last decade has been undergoing a process of evolution, moving away from the auditory and visual spectrographic analysis carried out by ‘experts’ towards a more objective, repeatable and scientifically logical foundation based on statistical acoustic-phonetic analysis and to a certain extent automatic systems in the vein of speaker recognition [272].

Central attempts to adhere to this new standard for the evaluation of evidence is the framework of Bayesian probabilistic reasoning [198, 331], named after the 18th century English philosopher Thomas Bayes, which prescribes the use of a Likelihood Ratio for updating probabilistic formulations of beliefs given the analysis of pertinent new information as: Posterior = Likelihood Ratio × Prior. In the context of FVC where we have two samples of speech and two competing hypotheses, namely $H_{ss}$ that the two samples were spoken by the same speaker against $H_{ds}$ that two different speakers respectively produced the two samples, then expressed slightly more formally the odds form of Bayes’ Theorem states that:

$$\frac{\text{Prob} (H_{ss} \mid E)}{\text{Prob} (H_{ds} \mid E)} = \frac{p(E \mid H_{ss})}{p(E \mid H_{ds})} \times \frac{\text{Prob} (H_{ss})}{\text{Prob} (H_{ds})}$$

(6.1)

where the speech evidence $E$ is typically a crime scene recording and the other speech

1This gradually occurring paradigm shift was motivated by the 1993 Daubert ruling [2] which formed part of a series of rulings from the United States Supreme Court regarding the admissibility of expert witness testimony and the quality of evidence presented to the court.
sample is a recording of a suspect's voice used to build the model representing $H_{ss}$. The likelihood-ratio (LR) is the ratio of the evaluation of the evidence $E$ under probability densities representing the competing hypotheses: $LR = \frac{p(E | H_{ss})}{p(E | H_{ds})}$.

In adhering to this framework for the analysis of the speech evidence, a forensic scientist calculates the LR and makes the strict departure from the old paradigm by presenting only this and not reporting any assessment of the probability of either hypothesis [334]. The logical reasoning for this is that specifying the posterior probability of either hypothesis requires knowledge of the prior probability, which is dependent upon every other evidential factor being considered in the overall analysis, all of which fall outside of the domain of the forensic scientist [331, 333]. It is the job of the forensic scientist solely to evaluate the forensic sample evidence and not to calculate the probability of the hypothesis!

The actual calculation of the strength of the evidence, the LR, is dependent upon the chosen methods for modelling the selected features of the suspect's speech (captured in $H_{ss}$) and estimating their variation over the population (captured in $H_{ds}$) [165, 166]. LR are compared when calculated via the common GMM-UBM [327] and multivariate kernel density (MVKD) [8] approaches to model the feature space in [273] where features derived from the formant trajectories of 27 male speakers were used to conclude that the GMM-UBM approach was more accurate and precise. Methods of likelihood-ratio calculation and alternative approaches to evaluating trace evidence are also explored in [8] where glass fragment evidence is considered using elemental composition as a feature.

It is the process which is fundamental however: “There can be no general recipe [for a LR formula], only the principle of calculating the [Bayesian LR] factor to assess the evidence is universal” [241]. The LR only conforms to a general probability theory within which any generative model can be used to describe the distributions of the chosen features and to then make calculations from.

The LR contains two components. The numerator which measures similarity, namely how plausible is it that the observed evidence could have been generated by the suspect speaker ($H_{ss}$) and the denominator which measures the typicality, quantifying the likelihood that the observed speech could have been produced by a person chosen at random from the population of potential offenders (PPO). A key element of the LR calculation is this typicality, which requires accurately learning how the features of interest are

---

2 Some authors choose to make this explicit by claiming to work within the 'likelihood-ratio framework' rather than performing 'Bayesian analysis' so as not to imply that they report a posterior [272]. This of course highlights the central difference between FVC and automatic speaker recognition. It is this reporting of LR only and not of an identification or 'match' that is the reason for the use of the term FVC and not forensic speaker recognition which would be a misnomer if adhering the LR paradigm.
distributed over the individuals of the PPO represented by null hypothesis $H_{ds}$.

The LR only provides insight into the specific question implied by the prosecution and defence hypotheses and a constantly reoccurring practical problem arises in determining what this implies regarding defining the PPO. By specifying an inappropriate background the desired question may not truly be probed. Two issues are present here in general: (1) disagreement upon a definition for the PPO, and (2) actually building a model to represent it.

Regarding (1), agreement is not present within the forensic research community and this has even been cited as a reason to avoid the data driven LR approach entirely [143, 144]. One proposal is that the background speakers should be “sufficiently similar” to the offender recording that a hypothetical investigator would require FVC performed to differentiate them, perhaps with consideration of other evidence also [274]. Empirical results of selecting such a background set by human listeners did result in a superior LR than that produced from randomised selection of background speakers [274]. This method is questionable however, with logical concerns regarding the use of using naive listeners discussed in [274] itself, along with the fact that the LR will be returned to the trier of fact who is expected to update their prior probability regarding the competing hypotheses which has been based on other evidence presented and any other indisputable facts of the case which narrow down the global population of potential offenders to some specific group of which it is likely the case that many of the adjudged similar speakers are not members. These are important considerations given that certain specifications have the ability to result in findings partial to either competing parties hypothesis.

Regarding point (2), if it can be agreed upon as to what the background population should represent, there is often an issue with obtaining sufficient data to accurately model the variability of the FVC features being used as this generally means having to sampling data from the specific sub-population which is both expensive and time consuming. Like in all statistical modelling, the general rule is that more precise calculations may be made given larger data sets and it becomes a question of obtaining acceptable accuracy. Using a parametric representation of fundamental frequency trajectories it was concluded that the LR begins to stabilise once the background model training dataset contains speech

---

3. An example demonstrates that in some cases the specification of this population can be fundamental: regarding DNA evidence the alternative hypothesis that the DNA sample came from the suspects identical twin would be sufficient to deem the evidence worthless [334]. More pragmatically different specifications of the PPO can certainly alter the calculated LR [195, 274].

4. It is in general agreement that one should try to match as well as possible recording conditions between suspect and background speakers recordings, as well as their speaking style. This can be done when suspect speech is recorded in controlled circumstances, typically a police interview, although cooperation is often a significant issue.
from at least 30 speakers [195]. The calibration of the LR also improves with the growing cohort size [58]. Monte-Carlo sampling has been used to draw very similar conclusions [188, 335], although the transferability of the methods insights are predicated on knowing the distribution of real world features where we can easily understand the diminishing effects of sampling larger numbers. Monte-Carlo can’t break out of the distribution it is sampling from!

Another issue with FVC is that given the potential implications of the analysis one way or the other it is important to be able to quantify the confidence in ones findings. Indeed this is part of the requirements of the Daubert ruling [2] that is motivating the development of this scientifically rigorous LR paradigm in many countries besides just the USA. The National Research Council of the USA has also published its desire for forensic science to report measures of accuracy and reliability [283].

One proposed measure for assessing the accuracy of FVC methods is the log likelihood-ratio cost $C_{llr}$ which penalises LR values in proportion to the amount they incorrectly favour one hypothesis over the other [58]. If errors are being made it is important that their effects are minimal. The general discrimination ability of a system can also be reported by measures which average performance over various applications [397].

The precision of LR values may be inferred via statistical credible intervals (the Bayesian analogue of frequentist statisticians confidence intervals based on sampling from a normal distributions) estimated by in lab repetition of calculations [271]. Measuring the precision and accuracy of the LR is also not a widely agreed upon process and a summary of various objections may be found in [367], which concludes that improving the modelling methods is the fundamental action required: “An excellent research goal would be to devise procedures for calculating likelihood-ratios which include better models of within-speaker and between-speaker variability and which therefore result in more accurate and precise likelihood-ratio estimates”.

Due to the practical scarcity of suspect data and the mismatched conditions between traces and reference populations common in daily casework, significant errors appear in

---

5 Although specifying a methods classification accuracy was another such demand of the report, something not applicable to the LR paradigm. The publication was also concerned about bias of forensic scientists, unconscious or otherwise, and this is limited by the LR approach of advocating that the scientist analyses the relevant evidence only and is not privy to extraneous information.

6 Note that there are two types of accuracy in LR systems. One is with respect to the ground truth of the comparisons (known during in lab experiments) and the other is with respect to the distribution of the variables being measured. The $C_{llr}$ function measures the first type of accuracy and is therefore independent of the features the system employs. See [61] and [68] for other potential measures for establishing LR accuracy.

7 Often individual costs are specified for each type of error that may reflect a desire to incorrectly pass over a guilty subject rather than detain an innocent one.
LR estimation if specific robust techniques are not applied [165]. Gaussian mapping of LR values in the vein of test normalisation [31] has been investigated to overcome poor quality recording conditions and mismatch between speech samples in FVC [54] as well as the migration of automatic speaker recognition modelling methodologies to deal with the common case of working with limited quantities of data [250].

Understanding the properties of nominal LR values from certain modelling methodologies has been studied in [67], where agreement was reached with [398] that most such values do not display the behaviours expected from true likelihood-ratios and that, whilst these systems may be discriminative, they require calibration [397] before being valid for forensic situations.

Another important consideration with this evolving paradigm is that communicating the results of correct analysis to the relevant parties accurately and without misunderstanding is difficult and yet essential in order to be correctly applying or acting on the science.

People have a bias to focus on the evidence at hand and report a posterior probability without knowing the prior probability and in most cases it is highly abstract thinking to assume that the forensic scientist will present the calculated LR value and the interested parties will actually update their prior beliefs in any Bayesian way, if such a thing is even possible given various dependencies on the relevant pieces of evidence and the non-quantifiable nature of circumspection. It is important to communicate however that the LR at its foundation means that the trier of fact, should believe the ‘prosecution’ hypothesis LR times more than before considering this piece of speech evidence. Communicating this essence of the logical theory has been made more difficult given the use of terms in popular culture such as ‘voiceprint’ and ‘a match!’ which have lead to the so called CSI effect whereby the triers of fact have the wrong expectations of FVC and believe that it can over achieve [164].

With regard to features commonly used within FVC traditionally F0 and formant centres and common automatic features like cepstral coefficient have been chosen [334]. Many features have been proposed but any real world application relies on linguistic, phonetic and acoustic analysis of the speech samples to determine what is applicable [333]. In theory any discriminative and quantifiable aspect of the speech signal, including everything covered in the literature review of automatic speaker recognition in Section 2.5 is a potential candidate for modelling. In general, analysis of multiple information sources can also be combined by naive Bayes independence assumptions or other fusion methods [57, 59]. Commercial software packages sold as FVC tool kits, such as Agnitio’s BATVOX [6] and Oxford Wave Research’s Vocalise [290] typically allow such analysis.
over multiple features. In the next section a review of the limited use of the glottal flow waveform for FVC is presented.

6.2.2 The Glottal Waveform in Forensics

Despite there having been various parameterisations proposed within automatic speaker recognition, the voice-source remains to be regularly employed as a information source for FVC although differences in vocal fold vibration patterns have been used to inform subjective opinions of voice quality in auditory-phonetic and auditory-spectrographic contexts outside of the LR paradigm [183].

The few investigations into the discriminative ability of glottal flow features within the LR paradigm have to date found little to no benefit. The commercial voice pathology GLOTTEX package has been used by [124], with catastrophic fusion (decreases in both $C_{5tr}$ accuracy and precision measured by the 95% credible interval) observed with a MFCC GMM-UBM system. GLOTTEX decomposed the inverse-filtered glottal flow estimates into low and high frequency components related to the vibration of the folds and the motion of their mucosal covering respectively. It then produces descriptive statistics from the spectral domain, measures which are perhaps useful for voice pathology but seemingly miss many of the characteristics of identity. The authors also suspected the performed all-pole modelling was not suitable for much of the analysed speech [124].

Another small study of the voices of twins using glottal flow parameters determined that despite the very similar physiology differences were present in the acoustics, although larger prosodic differences were found [342].

To the authors’ knowledge this is the extend of literature on the topic of FVC with glottal flow information. Although not couched in a forensic context, the glottal features reviewed in 3.6 are also of relevance to FVC whenever they can reliably be estimated. Indeed one of the key considerations in forensic cases, namely the quality of speech data, may be the primary reason for the scant use of the glottal signal in FVC. Band limiting by telephone transmission and phase distortions introduced by transmission channels and microphones may indeed make the use of glottal information in forensic science limited. Automatic speaker recognition literature demonstrates that it is a useful signal whenever it may be estimated however.

\[8^8\text{In fact due to perinatal hardships being more often experienced by monozygotic twins than dizogotic twins, there is generally a larger linguistic difference between the former [368].}\]
6.3 Experiment 1: YAFM Database Naive Listener Task

We first introduce the Young Australian Female Map-task (YAFM) database [332] before describing a small preliminary experiment performed on it. The experiment involved human listeners undertaking a naive forensic voice comparison task and was carried out to develop an initial quantitative assessment regarding the degree of perceptual acoustic similarity of the YAFM speakers, which was hypothesised to be significant. The small group of 39 volunteer listeners were able to correctly make the ‘same speaker’ or ‘different speaker’ assessment on the 10 comparisons of approximately 5 seconds of speech per speaker side presented to them with 61% accuracy.

6.3.1 Introduction

For our investigation into the use of a data-driven parameterisation of the voice-source waveform as a useful feature for forensic voice comparison (FVC) we used the Young Australian Female Map-task database [332], which is available upon request. In this section this new database which is suitable for performing FVC experiments is introduced and the results of a naive listening task are reported in order to develop some understanding of the corpora. With this understanding in the proceeding Section 6.4 a statistical FVC experiment calculating likelihood ratios based on glottal and MFCC information is reported.

The current section however is organised as follows: first the YAFM database is introduced in 6.3.2 before describing the experimental design of the naive listening task in 6.3.3. The results are reported in 6.3.4 along with a brief discussion.

6.3.2 The YAFM Database

The Young Australian Female Map-task (YAFM) consists of 26 female speakers with Australian English as their first language (L1), all within their twenties and all from the same university-attending social group. There exists two sessions for 24 of speakers separated by at least one week, where each time the same guide (also a member of the social group) lead the participant through a verbal repetition of place names and locations, before proceeding to a map-task. The map used was synthetic and the place names and the map-task route were designed to elicit several tokens for a range of phonemes, typically including extended vowels. Examples of these tokens included “Eden
Railway Station”, “Burns Freeway” and “Lovers Lookout”. The map-task involved a scenario whereby the guide assumed the role of a nervous groom (Tim) needing to get to the church on time by following the verbal instructions given by the participant. With this task there is only a single map, held by the participant, and thus minimal confusion between guide and participant as the participant is simply required to trace and narrate the route from Tim’s house to the church, with the guide in the role of Tim typically just periodically stating an affirmative “yes/ok”. This does however elicit semi-spontaneous speech from the participant and more instances of the above mentioned location and street names.

General convergence of certain acoustic properties of speech is present within the YAFM data and is hypothesised to be signifying social group membership. The most prominent example being the significant vocal creak possessed by several speakers. Qualitatively, the speakers are very similar to both the authors ears and to the ears of the database instigator, the linguist and phonetician Phil Rose. This convergence and the multisession recordings make YAFM a forensically realistic database where speakers are also challenging to differentiate. To gain some quantitative insight into this last statement a preliminary human listening experiment was conducted, with our null hypothesis being that we would observe authentication rates at the chance level. We now describe the organisation of this listening task.

6.3.3 Experimental Design

Responses to the naive listening task were collected as follows. Volunteers replied to an online survey, first providing the following meta data: age bracket, gender and a yes/no response as to whether English was their L1 and whether they were aware of any hearing impairments.

They were then able to begin the naive listening task, making a determination on each of ten comparisons of two files as to whether the two files were recordings of the same speaker or different speakers. The same 10 comparisons were used for all participants. The twenty files all contained the short utterance “A5 - Eden Railway Station” only and involved sixteen different speakers (6 different speaker trials, 4 same speaker trials).

The term naive in ‘naive listening task’ is used to imply that the same or different speaker decision for each comparison was made by the volunteer without the aid of any tools such as those provided by a statistical acoustic or phonetic analysis: the same speaker or different speaker assessment was made only by listening (through headphones or speakers at the volunteers discretion) to the two samples per comparison. Samples
could be listened to multiple times, however in the description of the task given to the volunteers they were encouraged to come to their decision within a “couple” of plays of the speech segments in each comparison.

6.3.4 Results and Discussion

Results of the naive listening task are reported first for all responses and then for the subset of native English speaking responders only. Only one of the volunteer respondents to the naive listening task reported having any known hearing damage, stating the presence of mild tinnitus. Figure 6.1 shows the average response accuracies across all 39 participants for each of the 10 given voice comparisons. The truth of each comparison, whether the two audio files being listened to were spoken by the same speaker or by different speakers, is marked with an ‘S’ or a ‘D’ respectively.

The random selection of the 10 comparison tokens generated a range of response accuracies. An indication of the individual variation in performance across the 10 comparisons is given by Figure 6.2 which displays a histogram of response accuracies (count of correct responses out of 10). Summary statistics are shown in Table 6.1. Based on these sample statistics a 95% confidence interval for the mean population response accuracy is 60.5 ± 9.84. This result is statistically significant and we could reject the null hypothesis that the population would perform at chance level accuracy. It is not though indicative of accurate performance.

![Figure 6.1: Percentage of correct responses for the 10 randomly paired YAFM “Eden Railway Station” token comparisons. Comparisons 1, 5, 7, 8 are the same speaker (S); remaining 6 are different speakers (D). Responses given by 39 volunteer participants.](image)

The results for the sub category of participants who were native English speakers, num-
Table 6.1: Summary statistics from all YAFM naive listener responses. The standard error is calculated using all 39 responses to the task.

<table>
<thead>
<tr>
<th>Mean Response Accuracy</th>
<th>60.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Error</td>
<td>5.02%</td>
</tr>
</tbody>
</table>

Table 6.2: Summary statistics from the responses of English L1 speakers only. The standard error is calculated based on the sample size of 21.

<table>
<thead>
<tr>
<th>Mean Response Accuracy</th>
<th>64.3%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Error</td>
<td>7.07%</td>
</tr>
</tbody>
</table>

Figure 6.2: Histogram of YAFM listening task accuracies from all 39 participants.

Comparing Figure 6.4 with Figure 6.1 we see that the English L1 participants are typically more accurate on each of the 10 comparisons, as may have been expected. This increase is not statistically significant however. A histogram of the accuracies of the English L1 only participants is given in Figure 6.3 and their mean and standard error are shown in Table 6.2.
Figure 6.3: Histogram of YAFM listening task accuracies from the subgroup of 21 English L1 participants.

Figure 6.4: Percentage of correct responses for the 10 randomly paired YAFM “Eden Railway Station” token comparisons for just the 21 English L1 participants. Comparisons 1,5,7,8 are the same speaker (S); remaining 6 are different speakers (D).

by ear (‘naively’) of just better than 5 from 10 or chance levels. Large differences were observed between the accuracies of individual comparisons. For example all 39 participants correctly claimed the voices of comparison #2 were different, while only 1 of the 39 claimed correctly to hear the same speaker in the 2 audio segments presented in comparison #8. Whilst not of relevance here, one suspects one could given time and effort
test other possible comparison pairs and corral the YAFM speakers into pens on Dodd-
dington’s farm of sheep, goats, lambs and wolves [103]. Of primary note here however
is the fact that the comparisons were only correctly recognised as the same or different
speaker with slightly above chance accuracy.

To summarise the results of the naive listening task we conclude that encouraged
to take just a short amount of time and to decide without the assistance of any tech-
nology (bar the options of headphones) or quantitative measures that the participants
found the given YAFM audio comparisons a challenging task, with their collective re-
sponses providing some evidence for the claim that the YAFM speakers are not easily
differentiated.

These results are also especially interesting in light of the fact that forensic voice
comparison practitioners for a long time relied solely on a listening test to make a
determination of identity [333]. Of course such comparisons are typically less naive in
the sense that the listener has garnered from experience a degree of specialised talents
for the task, although it is fundamentally unscientific.

Having obtained some small insight into the collection of speakers in the YAFM
database we now in Section 6.4 analyse the results of quantitative methods, in particular
via making use of the glottal waveform, to distinguish speakers in a simulated forensic
voice comparison scenario.
6.4 Experiment 2: YAFM Database Statistical Forensic Voice Comparison

An acoustic forensic voice comparison experiment was performed on the YAFM database using both MFCC and glottal source-frame features. These two information sources were fused via logistic regression. Tippet plots were produced for each of the systems and it is seen that both speaker information is present within the glottal information and that this information complements the cepstral system.

6.4.1 Introduction

In this section we performed a forensic voice comparison experiment on the YAFM database introduced previously in 6.3.2. Statistical modelling of acoustic features is used to produce likelihood ratios for each voice comparison. These acoustic features comprise information from the speakers glottal waveform parameterised by the source-frame representation used in tandem with a set of MFCC vectors.

We describe in detail the design of this experiment now in 6.4.2. Results are then reported and discussed in 6.4.3.

6.4.2 Experimental Design

The YAFM data was prepared by demarcating the guide and participant speech into labelled sections and diarisation achieved by using a Praat script to extract the labels of relevance. Blocks of speech that were considered noisy due to the presence of both voices at the same time, or laughter from one party masking the others speech, were removed. This left on average approximately 1.5 minutes of participant speech per session.

The automatic system was based on a traditional speaker recognition GMM-UBM framework [327], combining a mel-cepstral parameterisation with parameters coming from an estimation of the speakers glottal waveform. We considered the MFCC system a baseline, although with forensic speech work any feature that is informative regarding identity is useful and may provide contributing evidence.

YAFM speech was down-sampled from a 44.1kHz sampling frequency to 8kHz for both cepstral and glottal parameter estimation. The features for the mel-cepstral baseline experiment were comprised of 12 MFCC + log energy with first order deltas ap-
pended. This 26 dimensional vector was then feature warped [297] to target distributions learnt from the cepstra of 30 female ANDOSL [260] speakers. These data were then modelled with a 32 component GMM with diagonal covariances MAP adapted [158] from a UBM. The UBM used was trained on all 54 female speakers from the ANDOSL non-accented corpus, which contains microphone recordings of 200 phonetically varied sentences spoken by speakers evenly distributed among the dialect/sociolect categories of ‘Broad’, ‘General’ and ‘Cultivated’ Australian English. The intersession variation was maintained by MAP adapting suspect models from the UBM by using the first YAFM sessions data and withholding the second session for offender speech. A fixed relevance factor of $r = 16$ was used in MAP adapting the means and weights only [327].

The glottal waveform features representative of the speakers voice-source are the normalised source-frames obtained by closed-phase inverse filtering as described in detail in Section 3.5.

As detailed there, they provide an estimate of the time-domain waveform of the the derivative of the voice-source waveform (volume-velocity of air through the glottis during speech production). They represent the derivative due to the modelling of radiation at the lips in the source-filter theory of speech production.

A frame length of 20 ms was used with a 10ms frame increment and source-frames were normalised to a length of 256 samples. To overcome some imperfections with the inverse filtering process, the most common of which being the obvious presence of vocal-tract formants in the glottal waveform estimate (ripple), the mean of groups of 5 consecutive source-frames was taken providing an averaged estimate of the shape of the glottal waveform over a short period of produced voiced speech.

The following process was then employed in order to parameterise this time domain waveform. Remembering that this mean source-frame was a windowed signal, we note that the shape of one glottal pulse is evident centrally, with the windowing tapering the signal off to zero at each boundary. We use a polynomial fitting process in order to concisely represent the shape observed in this central glottal pulse. We created 3 sections of interest: samples $[51 – 100]$, $[101 – 150]$ and $[151 – 200]$ which we shall respectively refer to as sections A, B and C. Over sections A and C we fitted by least-squares a 3rd order polynomial, giving 4 coefficients for each section. Over section B which we observed typically contained more variation, we fitted a 7th order polynomial, giving another 8 coefficients. To add to this we also included the index of the sample over section B that had the greatest absolute difference between our polynomial approximation and the mean source-frame. This was done in an attempt to capture information about where the mean source-frame reached a negative peak, which was empirically observed to be reasonably
consistent within speakers source-frame estimates but was not typically captured by the polynomial estimates unless the polynomial order was made so excessively large that overfitting resulted.

Thus we transformed the time based glottal waveform in the form of a mean source-frame into a 17 dimensional parameter vector comprising of 4 section A coefficients, 8 section B coefficients plus the difference index and finally 4 section C coefficients. This parameterisation of the source-frame data is fundamentally a dimension reduction technique. Figure 6.5 shows the 3 polynomials estimated over their respective sections along with the index of maximum difference within section B as estimated over a the mean source-frame.

Figure 6.5: Polynomials whose coefficients are used to parameterise the mean source-frame are shown for some data of YAFM speaker 1. Also shown with a black mark at sample 126 is the index of greatest difference between section B and its polynomial estimate. The 3 sections over which each polynomial was fitted are labelled A, B and C. The unlabelled starting and ending sections were almost uniformly the same for all speakers due to the normalisation and windowing process of representing the glottal waveform in the source-frame manner and were ignored. In this waveform there remains significant energy over the closed-phase, an example of imperfect inverse filtering.

The reason for not simply looking to fit existing synthetic glottal flow models such
as the LF [131, 132] or alternatives [404] to our source-frame glottal pulse representation or directly to the estimated glottal waveforms were two fold. First, earlier investigations [400, 401] had suggested that speakers reliably reproduced distinguishing traits in their source-frame estimations of their glottal pulse that were often small and non-continuous. These characteristics were not able to be captured by the smooth synthetic flow models which capture the broader quality of the waveform. Secondly, and more pertinently, our final source-frame representation that we parameterised with the polynomial coefficients over segments is a manipulation of a glottal pulse that involved a windowing operation and an averaging of a small number of consecutive frames. As such using a synthetic flow model to represent this data was not appropriate.

Once these polynomial coefficients plus the index of the greatest difference between the polynomial prediction and original source-frame were concatenated these 17 dimensional vectors were modelled via a GMM-UBM approach, with the same dimensions for the Gaussian mixture models as were used for the cepstral system (32 mixture components, diagonal covariances, all parameters MAP adapted from the UBM for client models). The background data for the UBM was again provided by the same ANDOSL [260] females as used for the MFCC baseline. Scores were produced then in the standard manner of obtaining the likelihood of the testing/offender feature vectors against both the suspect GMM and the background population Gaussian mixture model [327]. The ratio of these two likelihoods is then taken producing a likelihood ratio that we use as a score and can be verbally interpreted in terms of probabilities of random match for any trier of fact or investigator in real forensic cases [333] given their starting information represented as prior probabilities.

Likelihood ratio scores from the ratio of data explained by client and background GMMs were also calculated for the MFCC system [327]. Having obtained likelihood ratios for both information sources (glottal and cepstral/vocal tract) these were then fused via logistic regression using the FOCAL tool kit [57].

### 6.4.3 Results and Discussion

In Table 6.3 the log likelihood ratio costs ($C_{llr}$) are shown for the MFCC, polynomial source-frame (Glottal) and fused systems. We report $C_{llr}$ values for these systems rather than just equal-error rates which relate only to the discriminative ability of the classifier. The $C_{llr}$ measures not just the discrimination ability of the classifier but also its calibration meaning how well the optimal threshold is located [61, 397]. This is of vital importance in forensic science; see the discussion in Section 6.2 for more details.
Table 6.3: Log likelihood ratio costs ($C_{llr}$) are shown for each of the individual and fused statistical acoustic systems on the YAFM database. The fused system exhibits the best discrimination and calibration qualities.

Tippet plots\(^9\) are shown for the individual and fused systems. Figure 6.6 shows a Tippet plot for the MFCC acoustic system on the YAFM data. Figure 6.7 shows the Tippet plot for the glottal system using the polynomial coefficient features. In Figure 6.8 the Tippet plot resulting from the fused cepstral and glottal systems is shown and in Figure 6.9 all three plots are overlaid. The fused system is seen to have both better discrimination (less overlap of target and non-target comparisons) and better calibration (visually the optimal threshold is seen to be $\sim 0$). This was quantified by the minimum $C_{llr}$ of 0.477 achieved by the fused system\(^10\).

Generative modelling of the source-frame information via the dimension reduction achieved by the piecewise polynomial fitting was used rather than the distance or discriminative scoring methods performed in the speaker recognition experiments of Sections 5.4, 5.5 and 5.6. This was due to the fact that in the forensic context typically significantly less data is available, especially from offender or crime scene recordings and these earlier discriminative methods required much data in order to obtain the accurate mean waveform estimates that they relied upon. This also resulted in a comparatively weaker result based on the glottal information alone than the results reported in those earlier speaker recognition experiments.

Interpretation of results and strength of evidence is particularly important in all forensic science, forensic voice comparison being no exception. This is where calibration

\(^9\)A Tippet plot is used in forensic science to graph the cumulative distribution of observed log-likelihood ratios (LLR) for each of the known target and non-target comparisons [382]. Both the performance and the calibration of the system are able to be inferred from the plot: performance is increased with diminishing overlap, whilst in a well calibrated system target and non-target curves should overlap near LLR = 0 and there should be very few target trials with LLR < 0 and similarly few different speaker trials with LLR > 0.

\(^10\)For comparison purposes we note an unpublished result obtained not by the author but by Phil Rose from an examination of the schwa vowel /er/ for the YAFM speakers where a $C_{llr}$ of 0.32 was achieved by using 26 disjoint UBM speakers and polynomial parameterisations of the first three formant trajectories. This could provide an alternate baseline to the MFCC one used and shows that formant information is also useful in differentiating these similar speakers.
of likelihood ratios becomes important. Where once calibration really meant being able to confidently threshold a score (“left of here suggests guilt”), increasingly as forensic science slowly moves towards a greater scientific foundation, calibration means being able to explain the numbers to investigators, judges and jurors and this requires having confidence in their quantitative meaning.\footnote{Indeed, typically likelihood ratios obtained from a given system strictly as the actual ratio of two likelihoods calculated from generative probabilistic models must still be calibrated to achieve these aims \cite{61, 58}. See \cite{398} for details regarding the need most systems have for calibration and some nice properties exhibited by well calibrated (‘true’) likelihood ratios.}

Figure 6.6: Tippet plot for the MFCC features and GMM-UBM model. Red curves are for different speaker comparisons, blue for same speaker comparisons.

Figure 6.7: Tippet plot for the glottal system using polynomial representations of mean source-frame waveforms. Red curves are for different speaker comparisons, blue for same speaker comparisons.
Figure 6.8: Tippet plot for the fused MFCC and Glottal systems. Red curves are for different speaker comparisons, blue for same speaker comparisons.

Figure 6.9: All three YAFM Tippet plots: MFCC (dashed), Glottal (dot-dashed) and Fused (solid) systems respectively. Red curves are for different speaker comparisons, blue for same speaker comparisons.

Whilst the quality of real world forensic recordings may often limit our ability to infer anything about the speakers voice-source waveform we conclude this experimental section by stating that, based on the empirical evidence presented here, that it is a useful signal for the task of forensic voice comparison when ever circumstances are favourable for the estimation of it.
Chapter 7

Glottal Waveforms: Depression Detection and Severity Grading

7.1 Introduction: Depressive Disorders and the Need for Quantitative Assessment Tools

In this section we introduce the significance of the problem of mental illness within society before reviewing in section 3.6 the nascent research direction of recognising depression from speech, in particular via using estimates made of the speakers glottal flow.

Major depression is an extremely debilitating condition that typically severely limits an individuals capabilities, interest levels and mood whilst also potentially causing physical health problems. Directly or indirectly the illness also affects the family and friends of the sufferer [39]. Mild depression, dysthymia, has similar but less severe effects. Unfortunately depression is not only serious, it is also common with depressive disorders ranking among the most significant reasons for disability worldwide. In the United States of America (USA) approximately 6.7% of the population (totalling nearly 15 million people) are affected each year by a severe mental illness, and it is the leading cause of disability for Americans aged between 15-44 [282]. In Australia 1 in 5 people aged 16-85 experiences a mental illness in any one year [43]. The Australian Bureau of Statistics states that over 40% of 16 to 85 years across both genders will experience some form of mental disorder during their lives [376].

The costs of the illness are large by any metric, be it of individual health or of an individuals contribution to society. Regarding health, the lifetime risk of suicide for depressive patients when left untreated is placed at 20% [167]. Encouragingly when treated
the suicide risk is reduced to below 1% [194]. These statistics convey information on only the most extreme effect depression can have on health. There are many other significant health related concerns related to the condition; people with depressive illnesses carry a higher risk of developing other serious health problems such as stroke and heart attack [275] for example.

In economic terms significant detrimental impacts are also observed, with the cost of depressive illnesses in the USA estimated at $51 billion dollars annually in lost production and absenteeism [318]. In Australia the cost is $14.9 billion dollars and over 6 million working days lost annually [39]. Across the Asia-Pacific region the untreated costs of depression are similarly significant [187]. A 2006 study concluded that within Europe, “In 28 countries with a [combined] population of 466 million, at least 21 million were affected by depression. The total annual cost of depression in Europe was estimated at Euro 118 billion in 2004, which corresponds to a cost of Euro 253 per inhabitant.” [359]. This makes depressive illnesses the most costly brain disorder in Europe, with costs totalling 1% of the European gross-domestic-product [359].

There are reasons for hope however. Results on improving peoples lives once depression is detected are positive with 70 to 80% of people successfully treated [275]. That is if depression is detected. Unfortunately it is estimated that only 20% of people with a depressive illness seek treatment [275].

Organisations such as Beyond Blue [39] and the Black Dog Institute [43] continue to educate the public to seek treatment and to look out for others, slowly removing the stigma which may be perceived to be attached to the illness. It is likely however that detection rates and rates of people seeking treatment can be improved via simpler automatic tools which are minimally invasive and which can diagnose depression earlier and in an environment of the patients choosing. With increasing bandwidths and the propagation of mobile devices, such automatic tools that can analyse information submitted via microphone and camera and quickly output a useful health related synopses including suggestions on how the patient may proceed can be expected to have large benefits in both economic and health domains.

Such tools could also assist health care providers who currently rely on patient self reporting coupled with an experts informed assessment, which can vary significantly across practitioners. Introducing automatic methods can provide general practitioners, nurses, psychologists, psychiatrists and patients themselves them with objective tools to support their decisions.
7.2 Literature Review: Glottal Flow for the Automatic Detection of Depression from Speech

Early investigations of the speech signal for depression (and affect in general) focused on prosodic and spectral magnitude features primarily relating to the vocal-tract configuration. As early as 1965, 32 hospitalised depressive patients recorded interview data which was also paired with mood assessments from two clinicians. Large adjustments in patient mood were able to be detected using spectral information [176]. Recordings of 16 people before and after having a depressive illness showed correlates of condition with fundamental frequency measures [286]. This study is interesting and unique in the detail that it is difficult and thus rare to obtain recordings of the same subject in multiple states of health. Pitch related parameters (F0-contour, F0-bandwidth and F0-amplitude) were also shown to have a strong correlation with almost two-thirds of a group of 30 depressive in-patients in [232]. In [11] energy measures of the speech waveform are shown to achieve the highest depressed/non-depressed classification scores on average across a range of modelling techniques. References 1,4-9 contained within [371] give a broad account of the research into prosodic features for depression detection. References 13-21 and 22-28 within [267] cover prosodic and spectral information for depression diagnosis respectively.

There exist reasons for exploring in detail the use of the glottal waveform for prediction of a speaker's affective\textsuperscript{1} state that are stronger than the fact that it is under researched in comparison to other information sources. Table 6.1 in Rosalind Picard's seminal book on affective computing lists several indicators of affect within the speech signal. Several of these relate strongly to the glottal flow. Anger for example is said to be characterised by being \textit{breathy} and having a \textit{chesty tone} [301]. A perceptual study of affect prediction confirms many of these characterisations [224].

Depression can cause changes in voice, perhaps even dysphonia,\textsuperscript{2} induced by emotional stress increasing laryngeal tension which presents in modified vocal fold vibrations. This is not detected by fundamental frequency measurements (including jitter and shimmer) but is contained in the waveform shape of the glottal pulse during a voiced pitch-period [266].

Neurological changes as a result of depressive illness may also affect laryngeal control, the hypothesis being that modifications of the brains basal ganglia may result in a decline

\textsuperscript{1}One of the three divisions of state of modern psychology: cognitive, conative and affective. Affective refers to the experience of emotion or feeling.

\textsuperscript{2}Dysphonia is an exhaustive medical term pertaining to any disorder of the voice.
of motor coordination similar to Parkinson’s disease [63, 291]. It was stated in 2008 that “Psychomotor disturbances are of great diagnostic significance for the depressive subtype of melancholia” and that “to date research into functional outcome and studies applying objective experimental assessment methods are lacking” [349].

Investigated in [313] were the glottal related features jitter, shimmer, degree of aspiration, F0 dynamics and velocity of energy for the purpose of inferring laryngeal control regarding depression and the psychomotor hypothesis. With 35 subjects weak correlations were found for most voice-source features with both severity (clinical and self-reported assessment scores of depressed state) and psychomotor retardation (also clinically assessed).

In a small study using groups of 15 males and 18 females approximately evenly split between control and patient groups, several statistical measures of the glottal waveform were determined to be statistically significant by ANOVA and shown to give good classification results [267]. Spectral measures of the glottal flow were found to be key features. Limitations of this study (and these are typical) are primarily the limited number of subjects, which makes drawing strong statistical conclusions difficult. A study highlighting this point is presented in [289] where an audio-visual database containing subjects without any affective disorder at the time of recording was initially collected. Within two years of this recording a small group numbering 15 were assessed by a psychologist to have developed a major depressive disorder. Via an analysis using glottal ratio based parameters it was claimed that with 69% accuracy the onset of depression could be predicted within the proceeding two years given the recording of a currently mentally healthy patient. Much further work is required to strengthen this claim.

In [291] a study with 10 high-risk for suicide depressed patients, 10 low-risk depressed patients and 10 non-depressed controls, found that the slope of the glottal flow spectrum and vocal jitter were statistically significant discriminators between all 3 classes, however again larger databases are required for confirmation and the different recording conditions used in the data collection may have played a roll in the jitter measurements. Interestingly the work was initiated by psychiatrists after reviewing their case history and realising that the patients voice, independent of linguistic content, was their primary information source for insight into near-term suicide risk.

Automated methods are not yet able to classify types of depressive illness (bi-polar, retarded and agitated depression) or accurately rate the severity of a subjects detected depressive illness from the speech signal alone, however this demonstrates a growing body of research which exhibits tentative but positive results regarding the potential accuracy of automatic depression detection from the speech waveform.
One of the key steps required to develop robust and accurate algorithms for detecting depression is the need to replicate published results that suggest promise. A core ingredient of the scientific process at any rate, replication is of particular importance to this field given the sparsity of good quality data available to the research community to develop and importantly validate proposed methods on. Difficulty often arises also given ethical and sometimes practical considerations with accessing existing collected data.

As mentioned these issues are almost universally present and generally reported in the literature. Typical of these sentiments: “Clearly, establishing stronger significance in these relationships requires a larger database” [313], “…employment of larger subject groups would have yielded statistically more accurate results” [291] and “The major limitation of this study was the sample size” [291].

With this firmly in mind, in the next section we look to replicate and build on the promising glottal feature subsystem described in [266] and [267] with a new and larger set of clinical data.
7.3 Experiment 1: Investigation on the Black Dog Institute Dataset

In this experiment a large number of glottal descriptors and their statistics were used as features in a discriminant analysis classifier for the purpose of recognising speakers from the Black Dog Institute [43] dataset as either depressed or control. This method is based on that outlined in [267] and its replication is particularly important in a field with limited and often very small datasets. Having successfully replicated the results of [267] these features are then used in a logistic regression model to investigate not just the classification of depression status but to make a prediction of the severity of depression. This is enabled by the clinical severity ratings of depression (in the Hamilton depression rating scale) that are part of the BDI dataset.

7.3.1 Introduction

As noted in the literature review of Section 7.2 the glottal waveform is an information rich signal that, beyond identification for example, may also be probed to obtain insights into a speakers affective state. In this section the use of the glottal signal for classifying the presence of the common affective disorder depression as previously reported in [267] by Moore et. al. is replicated. Typically studies reported in the literature dealing regarding this classification problem report results on small or very small groups of speakers and the data themselves are typically very difficult to obtain given legitimate privacy and ethics concerns. These trends make the task of developing robust classification systems and establishing confidence in the proposed scientific methods a difficult problem and for this reason, more so even than in many other domains of science, replication of promising methodologies and results is of great importance. The results of the methods presented in [267] on groups of 15 males (9 controls, 6 patients) and 18 females (9 controls, 9 patients) where impressive classification accuracies of $\sim 90\%$ were reported are archetypal of these types of studies warranting replication.

In 7.3.2.2 we provide an overview of the glottal feature extraction and discriminant analysis modelling methodologies used in [267] before reporting the results of our own replication of the method on the clinical dataset of 60 speakers provided by the Black Dog Institute (BDI) [43] in 7.3.3.1. Interwoven between this replication we describe our own use of the same feature set of glottal descriptors in a logistic regression classifier for not only predicting depressed/non-depressed class labels for speakers but also for pre-
dicting the severity of depression in depressed patients. This is enabled by the Hamilton Depression Rating Scale (HDRS) clinical severity ratings attached to the BDI patients. The experimental methodology for this investigation is reported in 7.3.2.3 and the results are presented and discussed in 7.3.3.2.

7.3.2 Experimental Design

Both the classification replication of [267] and the investigation into the use of these features with logistic regression for predicting patient severity ratings are performed on the Black Dog Institute (BDI) [43] dataset. We now describe this real-world clinical data.

7.3.2.1 Black Dog Institute (BDI) Dataset

The Black Dog Institute [43] is a Sydney based clinical research facility specialising in the depression and bi-polar affective disorders. With the particular purpose of fostering research into the development of machine learning style classification tools that will be able to assist clinicians and possibly patients in the detection and monitoring of these affective disorders (depression in particular) the institute has and continues to collect an audio-visual dataset. At present data has been gathered from over 40 depressed subjects and over 40 healthy controls comprising approximately an equal number of male and female data and spread over the age range of 21 to 75 years.

The data collection process begins with each subject voluntarily completing a ‘pre-assessment booklet’ that covers general information regarding their health history. They are then assessed by trained researchers following the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [23] diagnostic rules for classifying affective disorders. Only participants with a Hamilton depression rating (HAM-D) of greater than 15 but with no other mental disorders or medical conditions were selected for recording. All control participants were screened to ensure no history of mental health disorders and also were chosen with the aim of matching the depressed subjects in age and gender. All data were acquired only after having obtained informed consent from the participants and in accordance with approval from their local institutional ethics committee.

The database contains a mixture of read and semi-spontaneous interview speech. For the experiments reported in this section we used the interview speech only from male and female speakers in groups of 30 evenly divided such that they both comprised 15 control and 15 depressed patients. Only native English speakers were included in these 60 participants in order to minimise as many sources of confounding variation as possible. HAM-D ratings of the depressed subjects in this study ranged from 13-26
with a mean of 19; for reference the DSM-IV diagnostic defines subjects affective state as being “Moderate” at 11-15 points, “Severe” at 16-20 points and “Very Severe” as anything above 21 points on this Hamilton scale).

The interview speech was designed to induce an emotional response from the participants with questions typically asking participants to describe an emotionally evocative event. Such questions were for example “Can you recall some recent good news you had and how did that make you feel?” and “Can you recall news of bad or negative nature and how did you feel about it?”.

Manual labelling of the interview speech in Praat was performed to extract pure subject speech, with a resulting total duration of 290 minutes of participant speech obtained.

We now describe the design of the experiments performed on this BDI dataset. First in 7.3.2.2 we describe the glottal features and modelling we replicate from [267] before we detail in 7.3.2.3 how we used these features in our own experiment employing a logistic regression framework in an investigation of predicting severity ratings.

7.3.2.2 Replication of Moore et. al.

We only explore the glottal features described in the work we replicate presented in [267] and do not include the vocal tract or prosodic information included in their complete system. A summary of the extraction of these glottal features is now presented.

The glottal waveforms are first estimated from the voiced speech by the use of the Rank Based - Glottal Quality Assessment (RB-GQA) algorithm [265, 270] that attempts to alleviate the sensitivity of the resulting glottal waveform estimate to the determination of the closed phase of the pitch period. To do this many inverse filtered estimates are obtained based on varying by single samples the location of the estimated closed-phase and then selecting from these waveforms the ‘best’ glottal waveform as adjudged by the glottal quality assessment measures. A more detailed review of this algorithm is given in Section 3.3 on the various methods of estimating the glottal waveform.

A large number of base features were then calculated from these estimates along with their statistics. These base features were grouped into 9 glottal timing features (return phase, closed phase, max./min. of waveform, etc.), 5 glottal ratio features (open to closed quotient, closed to pitch period quotient, etc.) and 4 glottal spectral features (spectral tilt and bias over 0-1000 Hz and 0-3700 Hz). From these 18 base features statistics were calculated on their collection over individual BDI sentences (so called Direct Feature Statistics DFS) and over groupings of sentences (so called Observation Feature Statistics DFS).
Two different observation feature groupings were used comprising clusters of 5 sentences (Group 1) and of 4 sentences (Group 2). These statistics resulted in a feature vector of 1222 components; please see Tables II and III of the original paper [267] for the many statistics calculated.

A forward selection algorithm with selection based on the Fisher discriminant values pre-calculated for each individual vector component according to (7.1) was then used to iteratively grow a feature set modelled with a quadratic discriminant function for the task of classifying BDI subjects as depressed or control. The features included for modelling were grown until the accuracy of the QDA model failed to increase from one inclusion to the next. This process was performed in a leave-five-out cross-validation process where the quadratic discriminant analysis (QDA) function was repeatedly trained on a random selection of 25 subjects and tested on the remaining 5 subjects. 50 different permutations were used in this cross-validation process. All experiments are gender dependent.

\[
\text{Fisher Discriminant} = \frac{||\mu_c - \mu_p||^2}{s_c^2 + s_p^2}
\] (7.1)

The Fisher discriminant provides a measure of the differences between the distributions of two classes; here \(\mu_c\) & \(\mu_p\) and \(s_c\) & \(s_p\) are the mean and sample variances of a specific feature of the 1222 in our feature vector for the control and patient classes respectively.

Matlab code for performing the glottal estimation by the RB-GQA algorithm was provided by Dr. Elliot Moore II of the Georgia Institute of Technology, the lead author of the original paper [267]. All other steps were implemented in Matlab by the author of this thesis.

The results of this replication are presented below in section 7.3.3.1. First however we describe how this feature set of glottal descriptors was also used in a separate, non-replication experiment in order to investigate the ability of logistic regression to output depression severity ratings.

7.3.2.3 Classification & Severity Prediction via Logistic Regression

Taking only the top 30 highest Fisher discriminant features from the 1222 glottal descriptors we again performed cross-validation but with a leave-one-out process and rather than a discriminant based classifier we used the logit function to not just predict class labels but to also output an estimate for the probabilities of membership of the classified data to each of the control and patient classes.
The mean of the output class probability estimates, taken over all of a single speakers test sentences, was then calculated and the correlation with the HAM-D label measuring the severity of the BDI patients’ depression was calculated.

Results are reported as follows. Firstly the accuracy of the predicted labels based on the logistic regression model learnt on all 30 features is reported, where class labels of test data are determined by the maximum of the two class membership probabilities.

Then correlations between the given HAM-D levels and the predicted mean class probabilities are given for many logistic regression systems using growing feature sets starting from only the top feature (Fisher rated) up to the set including all 30 highest Fisher rated features.

### 7.3.3 Results and Discussion

#### 7.3.3.1 Replication with the QDA Classifier

Shown in Table 7.1 are the quadratic discriminant analysis accuracies for each gender and group based on the predictions obtained by the leave-five-out cross-validation process. The standard errors are also given along with the average sensitivity and specificity of the QDA classifier. The sensitivity is also referred to as the true positive rate (proportion of depressed subjects correctly labelled as depressed) while the specificity is also referred to as the true negative rate (percentage of control subjects correctly identified); these are the respective compliments of Type I and Type II error rates. Only 30 features were included in the final QDA model for which results are reported.

<table>
<thead>
<tr>
<th>QDA</th>
<th>Accuracy ± Standard Error</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male Group 1</td>
<td>0.651 ± 0.013</td>
<td>0.574</td>
<td>0.792</td>
</tr>
<tr>
<td>Male Group 2</td>
<td>0.684 ± 0.015</td>
<td>0.621</td>
<td>0.795</td>
</tr>
<tr>
<td>Female Group 1</td>
<td>0.748 ± 0.016</td>
<td>0.682</td>
<td>0.830</td>
</tr>
<tr>
<td>Female Group 2</td>
<td>0.748 ± 0.016</td>
<td>0.672</td>
<td>0.839</td>
</tr>
</tbody>
</table>

Table 7.1: A summary of the depression classification results from the cross-validated QDA system on the BDI dataset. This is a replication, using only the glottal features, of the method proposed in [267].

For comparison with the original paper, Table 7.2 gives the accuracies reported on their Medical College of Georgia database. Note that these values are based on the combination of glottal, prosodic and vocal tract information, not just from the glottal features alone. In Figures 7.1 to 7.4 histograms of the QDA accuracies over the cross-validation folds are shown for the male and female, group 1 and group 2 experiments. Maximum accuracies
Table 7.2: For comparison the accuracies of the original paper on their own clinical dataset collected by the Medical College of Georgia are provided. These results are for the system combining glottal information with prosodic and vocal tract streams. The male system had 9 controls and 6 patients; the female system had 9 controls and 9 patients.

<table>
<thead>
<tr>
<th>Group</th>
<th>Original Paper [267]</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.913</td>
<td>0.867</td>
<td>0.936</td>
<td>0.956</td>
<td></td>
</tr>
</tbody>
</table>

of ∼ 90% are obtained in each experiment. For comparison using a linear discriminant classifier (not modelling any covariance structure between classes) resulted in lower cross-validated maximum accuracies over the repetitions of 79.3% (Female Group 1), 81.25% (Female Group 2), 75% (Male Group 1) and 81.82% (Male Group 2). Thus the greater modelling flexibility introduced by the quadratic terms in our learnt discriminant function enabled more accurate predictions.

Figure 7.1: Female Group 1

Figure 7.2: Female Group 2
7.3.3.2 Logistic Regression for Detection and Severity Grading

Results are now presented from the logistic regression model on the same feature set. In Table 7.3 the accuracy of the logistic regression model in predicting class labels is given for the male and female, group 1 and group 2 experiments.

<table>
<thead>
<tr>
<th>Logistic Regression</th>
<th>Accuracy ± Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male Group 1</td>
<td>0.701 ± 0.008</td>
</tr>
<tr>
<td>Male Group 2</td>
<td>0.672 ± 0.005</td>
</tr>
<tr>
<td>Female Group 1</td>
<td>0.695 ± 0.011</td>
</tr>
<tr>
<td>Female Group 2</td>
<td>0.722 ± 0.007</td>
</tr>
</tbody>
</table>

Table 7.3: Average accuracies and their standard error for the logistic regression classifier using the full 30 top Fisher discriminant rated glottal features.

Comparing Table 7.3 with Table 7.1 we see that the logistic regression classifier was able to predict class labels as well on average as the QDA classifier. We observe that the logistic regression method was here able make the depressed or control classification more accurately on average for males and but less accurately for females when compared with the replicated QDA methods.
Histograms of the logistic regression depression/control classification accuracies over the performed repetitions are shown in Figures 7.5, 7.6, 7.7 and 7.8 for the female and then male, group 1 and then group 2 experiments respectively.

Figure 7.5: Group 1 Female

Figure 7.6: Group 2 Female

Figure 7.7: Group 1 Male

Figure 7.8: Group 2 Male
Shown below are plots of the resulting correlations between the logistic regression predictions for depressed speakers and the clinical HAM-D severity ratings against the number of features used in the logistic regression model; Figures 7.9 and 7.10 show the group 1 and 2 female plots while the male plots are shown in Figures 7.11 and 7.12. Whilst weak correlations of approximately 0.25 are achieved in each plot for the logistic regression systems with the first five forward selected features, the lack of trends and non-smooth nature of the curves do not suggest that the method is either predictable or reliable.

Figure 7.9: Female Group 1

Figure 7.10: Female Group 2

Figure 7.11: Male Group 1

Figure 7.12: Male Group 2
The glottal waveform has been shown to be an informative signal for the task of classifying the presence or absence of depression given a subject's speech. The glottal component of the larger method also incorporating prosodic and vocal tract features presented in [267] was shown to perform equally well on the larger clinical database provided by the Black Dog Institute. Replication is a fundamental tenet of science and is of particular importance in this domain where as stated most studies are performed on small datasets that are typically difficult to obtain. These circumstance necessitate multiple replications in order to establish confidence in the methodology and allow the development of accurate and objective clinical aids. Indeed, quoting from the conclusion of the original paper [267]:

“The authors fully acknowledge that the latitude to which these results can be widely generalized is limited due to the small sample size.”

As in the original paper, no trend is observed with respect to accuracy and the number of sentences used in the groupings from which statistics are calculated, as the maximum obtained accuracies vary between male and females over group 1 and group 2 and indeed in all cases the differences are not statistically significant.

The second part of the reported experiment, the non replication part that involved using a logistic regression model on the same set of glottal features found to be useful in the replication study, produced mixed results. Classification based on assigning to the class with maximum logistic regression output probability resulted in slightly lower on average identification accuracies, although not statistically different to the quadratic discriminant analysis model used in [267].

The main purpose however of using the logistic regression model was to determine whether there was any evidence that the output class membership probabilities for the depressed patients displayed any correlation with the Hamilton clinical severity ratings that are provided for each depressed subject of the Black Dog Institute dataset. Disappointingly little to no evidence was found for such a claim with the output class probabilities showing no correlation with the clinical severity ratings of depressed subjects and perhaps worse they displayed a very non-smooth and erratic behaviour with the varying number of glottal features used in the logistic regression model. Note that whilst these correlations behaved erratically, the accuracy tended to increase smoothly before establishing a maximum around the use of 30 features.
Chapter 8

Conclusion

8.1 Thesis Contributions

The key outcomes of this study are:

- The glottal waveform \(\rightarrow\) Features related to the volume-velocity flow of air through the glottis during phonation as estimated from digitised speech have been shown to contain significant information pertaining to the identity of the speaker. In particular the prosody normalised representation of the derivative glottal flow pulse termed a source-frame was shown to be a useful feature for speaker recognition.\(^1\) An important component of the experimental process with the source-frame feature was taking the average (mean or median) of the collection of source-frames obtained from approximately ten seconds of voiced speech. Training and testing with such ‘block’ measures was seen to improve recognition considerably. This is likely due to diminishing the effects of imperfect estimation of the glottal flow and focusing on what may be considered the prototype waveform of each speaker about which some fluctuations arise from natural variation in the glottal opening and closing.

The glottal information was shown to complement MFCC features of the speech waveform in gender dependent experiments with 100 speakers in each case. It was shown to improve the performance of a forensic voice comparison when compared to using MFCC alone also. The improvements were marginal however and coupled with the difficulty in obtaining high fidelity estimates of the glottal flow in many

\(^1\)The VSCC glottal feature proposed in [171] was also investigated with its identity information content confirmed.
practical circumstances, it is likely that the use of this signal is best employed by incorporation in high security systems where speech is able to be recorded in a controlled environment. Where ever it is possible to estimate the signal however, particularly in forensic case work, it is likely, based on both the results presented here and within the limited research literature, to be able to improve the ability of a system to differentiate or recognise speakers.

- **Depression Detection** → Glottal information was also shown to contain indicators useful for detecting clinical depression in a replication of components of the study presented in [267] but performed on a separate and larger clinical dataset provided by the Black Dog Institute [43]. Building upon this work, the use of a logistic regression classifier on these features was demonstrated to classify speakers as depressed on non-depressed comparably well but the estimated class membership probabilities displayed weak and non-consistent correlations with the clinical HAM-D severity ratings which accompany the BDI dataset.

- **Regression score post-processing** → Strong empirical evidence for the ability of the proposed score post-processing method \( r \)-norm to increase the classification accuracy of speaker recognition systems was presented. Regression to ideal scores with zero-variance distributions was observed to be most effective. Two patterns for the specification of ideal target and non-target scores (\( i_T \) and \( i_{NT} \)) for optimal post \( r \)-norm results were observed which related to the distributions of raw target and non-target scores as outlined in Section 4.8. The application of \( r \)-norm requires the test probes to be scored against all enrolled models, producing a score-vector, which necessitates a small computational increase. The largest obstacle to the implementation of the method is the demand for a large amount of training data to firstly train speaker models and then to learn the TGPR (regression) function \( r \). This new practical method is applicable to the wide range of pattern classification tasks that generate scores in assigning class labels to unknown objects and when sufficient data is available to permit its application the \( r \)-norm method could potentially benefit such diverse fields as object or instance detection in surveillance to fraud identification in finance for example.
8.2 Future Research Directions

In this final section future research opportunities based on the results presented in this study are discussed.

Whilst further evidence for the speakers glottal waveform to aid the speaker recognition task has been presented, additional study on a large corpus comparing the several ‘data-driven’ approaches to parameterising the glottal flow to fitting synthetic models would be useful in firmly establishing that these approaches are better suited for speaker recognition. This is in parallel to the speech synthesis studies demonstrating that ‘data-driven’ glottal models are perceptually superior to functional form type synthetic models [172].

Beyond this, no modelling was considered of the temporal evolution of the glottal flow waveform in the presented experiments. Variations over pitch periods likely contain much information related to the onset of voicing, its transition into unvoiced speech, the use of glottal stops as well as an inter-period differences in the vocal fold vibratory motion. All of these factors likely hold information pertaining to a speaker’s identity.

Regarding the use of this signal in practical systems, investigation of the ability of glottal information to complement MFCC features in state-of-the-art factor analysis type systems is hindered by the catch-22 that such systems typically demonstrate considerably minimal errors on clean speech, the likes of which are required to make reliable inferences of the speakers glottal flow. This thesis has not considered the difficult, and potentially infeasible, problem of estimating the glottal flow waveform in common real world conditions where significant environmental noise may be present or speech may be band-limited through transmission via telephone. Whilst the development of techniques for the accurate inference of glottal flow descriptors in such conditions would enable the use of the signal in a broader context, it may be that key information is irreversibly lost.

The ‘data-driven’ approach advocated as a result of these studies for speaker recognition whereby statistics and quantifiers of the actual glottal flow estimates are used to parameterise the signal rather than applying a synthetic flow model that misses idiosyncratic features of the flow waveform is also suspected to be optimally helpful in classifying depression and affective state in general.

Whilst a handful of studies suggest the strong potential of the glottal flow signal for the depression detection task, important research questions for future analysis include determining the optimal way to model the voice-source data for this purpose and whether it is possible to quantify the severity of detected depression with the glottal flow signal (or indeed via other speech features)? To the best of the authors knowledge no research
exists on this last question of considerable practical importance which if solved can both assist in the early detection of the onset of depression and gauge the progress of treatments.

To make progress with these issues the development of larger clinical databases are essential to develop statistical confidence in the proposed algorithms. Given the practical and ethical difficulties with achieving this, the publication of replications of proposed methods on data available to individual research groups is also essential in the development of these important tools which have the potential to improve the lives of many.

Regarding the proposed $r$-norm algorithm, which too may prove beneficial to the depression detection task among others, it remains to test the method on more state-of-the-art speaker recognition tasks (PLDA compensated i-vector systems for example) and extend its application to open-set tasks. Applying $r$-norm to other classification problems is also of interest; the author is aware of it improving slightly upon already strong recognition accuracies in the fields of human-action and file-fragment recognition. The development of a stronger theoretical framework for explaining how the method works and for predicting the optimal parameters for its application is also of interest.
Appendix A

Source-Frame Addendum

A.1 Source-Frames Plots of Same and Different Speakers

In this section means of small collections of source-frames are plotted, first from the same speaker and then from five separate speakers, using data from the YOHO corpus [71]. Means are taken over single YOHO utterances. These visualisations provide some insight into the idiosyncratic shapes individual speakers consistently reproduce in their source-frames which enable their use as features for speaker recognition. The plots of the mean source-frames are overlaid to highlight intra-speaker similarities (s7 and s8) and detect inter-speaker differences.

The variations observed over the source-frames are constrained by the fact that:

1. they are a representation of the physically constrained process shared by all humans of producing quasi-periodic speech by tensioning the laryngeal muscles to control the vibration of the vocal folds, and

2. they are both prosody normalised, removing amplitude and duration characteristics, and then windowed, leaving centrally only waveform shape traits of the derivative glottal flow.

With consideration towards their use as informative features for inferring speaker identity, we are primarily concerned with the relative degrees of inter-speaker and intra-speaker variation present within these waveforms.
A.1.0.3 Plots of Mean Source-Frames from YOHO Speaker s7
One, three and five mean source-frame waveforms are plotted for YOHO speaker s7.

Figure A.1: A single mean source-frame from is plotted from speaker s7.

Figure A.2: A second and third mean source-frame still from speaker s7 are overlaid on the original source-frame. The three waveforms display significant similarities.
Figure A.3: Lastly a fourth and fifth mean source-frame from speaker s7 are overlaid. The five waveforms display little variation. This is typical of the behaviour of most speakers and this adherence of each speaker to an individual ‘template waveform’ is what makes these features informative for speaker recognition.
A.1.0.4 Plots of Mean Source-Frames from YOHO Speaker s8

One, three and five mean source-frames are now plotted for YOHO speaker s8.

Figure A.4: A single source-frame is shown from a new speaker, identified in YOHO as s8.

Figure A.5: A second and third mean source-frame also from s8 are overlaid. Minor variation over the three waveforms is again observed.
Figure A.6: Finally a fourth and fifth mean source-frame are overlaid. This speaker’s adherence to their implicit template waveform is again evident.
A.1.0.5 Plots of Mean Source-Frames from Five Different YOHO Speakers

Now a single mean source-frame from five different speakers are plotted over five plots, adding a speaker with each additional plot. Unlike the lack of variation present within the prior plots for single speakers, the differences between speakers source-frames become increasingly visible.

Figure A.7: A mean source-frame is plotted from a single speaker.

Figure A.8: Mean source-frames from two distinct speakers are plotted. Greater variation between speakers is already suggested than was observed within individual speakers.
Figure A.9: A mean source-frame from a third distinct speaker is overlaid.

Figure A.10: A mean source-frame from a fourth distinct speaker is overlaid.
Figure A.11: Finally a mean source-frame from a fifth distinct speaker is overlaid. The plot now displays a single mean source-frame from five different speakers. The evident inter-speaker variation is significantly greater than the previously observed intra-speaker variation.
A.2 Preliminary Investigation: Linear Discriminant Analysis Classification of Source-Frames

Some preliminary results are visually reported regarding initial investigations of the use of linear discriminant analysis (LDA) [141, 42] for separating source-frame features.

Ten male ANDOSL [260] speakers source-frames were extracted and LDA dimensions determined based on the central 150 samples of the source-frame data (removing the near zero windowed samples at the start and end of each source-frame vector). Shown in Figure A.12 are the projections of 900 source-frames from 2 separate male ANDOSL speakers onto the principal LDA dimension. On the left of Figure A.2 the relative frequency histogram of these ‘scores’ is shown and on the right are shown for means of the projected values of each segment of 50 continuous source-frames.

![Figure A.12: Projections of 900 source-frames onto the first LDA dimension for two separate ANDOSL males, shown in blue and red respectively. A reasonable separation of the two speakers is evident.](image1)

Figure A.13: Left: The distribution of the LDA projections are shown in separate colours for each speaker. Right: Shown are the distribution of means of the LDA projections, where means are taken over groups of 50 consecutive source-frames. Separation of these two speakers ‘scores’ are achieved by this process.

![Figure A.13: Projections of 900 source-frames onto the first LDA dimension for two separate ANDOSL males, shown in blue and red respectively. A reasonable separation of the two speakers is evident.](image2)
Appendix B

Bibliography


N. Brummer. tinyurl.com/nbrummer.


