The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison

Vincent Stephen Hughes

PhD

University of York

Language and Linguistic Science

October 2014

Abstract

Within the field of forensic speech science there is increasing acceptance of the likelihood ratio (LR) as the logically and legally correct framework for evaluating forensic voice comparison (FVC) evidence. However, only a small proportion of experts currently use the numerical LR in casework. This is due primarily to the difficulties involved in accounting for the inherent, and arguably unique, complexity of speech in a fully data-driven, numerical LR analysis. This thesis addresses two such issues: the definition of the relevant population and the amount of data required for system testing. Firstly, experiments are presented which explore the extent to which LRs are affected by different definitions of the relevant population with regard to sources of systematic sociolinguistic between-speaker variation (regional background, socio-economic class and age) using both linguistic-phonetic and ASR variables. Results show that different definitions of the relevant population can have a substantial effect on the magnitude of LRs, depending on the input variable. However, system validity results suggest that narrow controls over sociolinguistic sources of variation should be preferred to general controls. Secondly, experiments are presented which evaluate the effects of development, test and reference sample size on LRs. Consistent with general principles in statistics, more precise results are found using more data across all experiments. There is also considerable evidence of a relationship between sample size sensitivity and the dimensionality and speaker discriminatory power of the input variable. Further, there are potential trade-offs in the size of each set depending on which element of LR output the analyst is interested in. The results in this thesis will contribute towards improving the extent to which LR methods account for the linguistic-phonetic complexity of speech evidence. In accounting for this complexity, this work will also increase the practical viability of applying the numerical LR to FVC casework.

Contents

List of Tables

List of Figures

Acknowledgements

I am not sure I can express how thankful I am to my supervisor Professor Paul Foulkes. Thank you for introducing me to forensic phonetics as an undergraduate. There is no way I would have started (let alone finished) a PhD without your initial inspiration and encouragement. Thank you for your advice and guidance which have been instrumental in shaping this thesis. Thank you for your continual support and friendship. Thank you for all of your time, patience and energy - I truly could not have asked for anything more from a supervisor.

I wish to thank my Thesis Advisory Panel of Professor Peter French and Dr Dominic Watt. Without their insightful comments, questions and ideas this thesis would not be the same. I also want to thank my fellow LR-researcher Erica Gold for all of our chats and discussions. This work would not be the same without you. Thanks to George Brown for teaching me how to use HTK and indulging my questions about techy things I didn't understand. Thanks to Phil Harrison for your scripts, Matlab chat and proof reading, and to Richard Rhodes for advice about ageing (in linguistic terms, not physical) - but much more importantly thanks for the numerous curry outings. I'm not sure we've quite worked out the best Indian in York, but we've certainly done plenty of research.

A massive thank you must go to my best friend, (now Dr) Ashley Brereton. Ash's mathematical influence is evident throughout this thesis. Thank you for writing scripts, thank you for your advice about coding, thank you for troubleshooting and thank you for being a great teacher. It is fair to say that this thesis would not have been possible without you.

Thanks to Bill Haddican for employing me to work on the Northern Englishes grant and allowing me to use the data in this thesis. Thanks also for convincing me to invest time into learning how to use R! A debt of gratitude is also owed to Dr Frantz Clermont, Dr Kirsty McDougall and Dr Geoffrey Morrison for their direction, insight and support over the last few years.

During my PhD I was lucky enough to spend three months at the New Zealand Institute of Language Brain and Behaviour at the University of Canterbury. I found the experience inspiring and would like to thank everyone at NZILBB for making it so special. Specific thanks go to Professor Jen Hay, Robert Fromont and Dr Pat LaShell for all your time, advice, data and statistical expertise. Thanks also to Paul, Ghada and Maya for looking after me and taking me on days out!

I am indebted to Dr Bernard Guillemin, Esam Alzqhoul and Balu Nair at the University of Auckland for their time, hospitality and willingness to share scripts and knowledge which have ultimately improved the work in this thesis. I wish to thank Professor Henning Reetz and the Phonetics lab in Frankfurt for their hospitality, as well as a big thank you to Dr Michael Jessen for his time and for all of his encouragement regarding my work.

I am also grateful to the Economic and Social Research Council (ESRC) for a three year PhD scholarship and for an overseas travel grant which allowed me to visit NZILBB in NZ. I am also grateful to the Deutscher Akademischer Austausch Dienst (DAAD) for a research grant which allowed me to visit the Goethe Universiät in Frankfurt and the Bundeskriminalamt (BKA) in Wiesbaden.

Finally, a massive thank you, of course, to my family. Thank you to my mum and dad who have always shown such belief in me and allowed me to pursue the things that make me happy. Thank you for your unwavering love, support and encouragement. To my brother, Phil, thank you for the year and a half we spend living together in York. Thanks for supporting me throughout the last three years and thanks for always making me laugh. To Fiona, thank you for your unconditional love - I feel so lucky to have you. Thank you for putting up with me, especially in the last few months, and thank you for buying me treats to keep me going. You are my one and I couldn't be happier that we get to spend the rest of our lives together.

Declaration

The acoustic data used in Chapters 4 and 8 (based on /uː/) were collected as part of my Masters dissertation:

- Hughes, V. (2011). *The effects of variability on the outcome of likelihood ratios*. Unpublished MSc Dissertation, University of York.

The LR-based analysis in §4.3.1 is the same as that in my Masters dissertation. The analyses in §8.3.1.1 and §8.3.2.1 replicate the general structure of the experiments in the Masters dissertation using the same data, but were adapted to address the limitations of the earlier work.

§4.3.1, §8.3.1.1 and §8.3.2.1 have previously been published::

- Hughes, V. and P. Foulkes (in press, 2014). Variability in analyst decisions in the computation of numerical likelihood ratios. *International Journal of Speech, Language and the Law* 21(2), 279-315.

The analysis in §5.3.3 has previously been published:

- Hughes, V. (2013). Establishing typicality: a closer look at individual formants. In *Proceedings of Meetings on Acoustics (POMA)* 19. Montréal, Canada.

Elements of Chapter 6 have previously been published:

- Hughes, V. and P. Foulkes (2014). Regional variation and the definition of the relevant population in likelihood ratio-based forensic voice comparison using cepstral coefficients. In *Proceedings of 15th Australasian Conference on Speech Science and Technology*. Christchurch, New Zealand.

The entirety of Chapter 7 has previously been published:

- Hughes, V. and P. Foulkes (submitted). Defining the relevant population according to socio-economic class and age in forensic voice comparison. *Speech Communication* 66, 218-230.

The entirety of Chapter 9 has previously been published:

- Hughes, V., A. Brereton and E. Gold (2013). Sample size and the computation of numerical likelihood ratios using articulation rate. *York Papers in Linguistics* 13, 22-46.

In all cases, I was the lead author and the data were collected (with the exception of Chapter 9) and analysed by myself.

Some of the arguments presented in this thesis have previously been published in a paper which I co-authored. In this paper, I was responsible for the arguments relevant to this thesis:

- Gold, E. and V. Hughes (2014). Issues and opportunities: the application of the numerical likelihood ratio framework to forensic speaker comparison. *Science and Justice* 54(4), 292-299.

Vincent Stephen Hughes

October 2014

Chapter 1

Introduction

Forensic speech science (FSS) is the application of linguistics, phonetics and acoustics to criminal investigations and legal casework (for an overview see Foulkes and French 2001; Nolan 2001; Jessen 2008). Speech is an increasingly common form of expert forensic evidence. This is due, in part, to the increased availability of speech recorded during crimes, the development in the technology used to record speech and a more advanced understanding of the principles underlying the identification of individuals from their voice. Speech evidence can be divided into two broad categories. Before the apprehension of a suspect, an expert may be instructed to conduct a *speaker profile* of an unknown offender. Based on the observed speech patterns, the expert will attempt to determine socio-indexical information about the speaker's background (e.g. regional background, class, age), thus narrowing the population of which the offender is a member (Foulkes and French 2001, 2012; see Ellis 1994 for a case report).

More commonly, forensic speech scientists become involved in legal casework after the apprehension of a suspect. This is referred as speaker identification (Nolan 1997), which itself takes two forms: naïve and technical. Naïve speaker identification involves cases where a lay (i.e. untrained) listener hears the voice of an offender but does not see their face (e.g. in a masked bank robbery). Since no recording of the offender exists the *ear-witness* may be required to demonstrate their ability to identify the voice using the aural equivalent of a visual line-up (a *voice parade*) (Nolan and Grabe 1996; Nolan 2003). Issues relating to the ability of naïve listeners to identify voices in forensic contexts and the salient perceptual cues which listeners attend to are discussed in Bull

and Clifford (1984, 1999), Nolan *et al.* (2009) and Nolan, McDougall and Hudson (2013).

Technical speaker identification involves analysis by a forensic speech scientist, although it is now more commonly referred to as forensic voice comparison (Rose 2002; French and Harrison 2007; Rose and Morrison 2009; Jessen 2012).

1.1 Forensic voice comparison

Forensic voice comparison (FVC) accounts for the majority of casework undertaken by forensic speech scientists (c. 70%; Foulkes and French 2012). FVC typically involves a recording of the voice of an unknown offender (e.g. in a covertly recorded drug deal) and a recording of the voice of a known suspect (from a police interview in the UK; PACE 1984). The expert is instructed to conduct a comparison of the speech patterns in the suspect and offender recordings to aid the court in establishing whether the voices in the two recordings belong to the same or different individual(s). This evidence is then used by the trier-of-fact (judge and/or jury), along with all other evidence, to make a decision regarding the defendant's innocence or guilt.

There is currently no clear consensus amongst forensic speech scientists as to the most appropriate way of analysing FVC evidence. Gold and French (2011) present the results of an international survey, conducted in 2010, on current practises in FVC. The aim of the survey was to make available current working practises in FVC casework around the world. Participants consisted of 34 practising forensic speech scientists from a range of academic institutions and forensic laboratories (both private and governmental) in 13 countries. Those surveyed were asked about their methods of analysis in FVC cases, the variables considered the best for speaker discrimination and the frameworks used to express conclusions. According to Gold and French (2011), there are four common methodological approaches for the analysis of speech in FVC casework: auditory-phonetic analysis, acoustic-phonetic analysis, combined auditory and acoustic analysis, and automatic speaker recognition (ASR).

1.1.1 Auditory phonetic analysis

Auditory phonetic analysis involves making auditory judgements about a range of linguistic-phonetic variables (see §1.1.3) without the aid of spectrographic-acoustic analysis. Auditory judgements are commonly qualitative, involving detailed phonetic transcription following the protocols of the International Phonetic Alphabet (IPA). For certain variables, auditory-only analysis may be quantified using counts based on the frequency of occurrence (e.g. allophonic realisations of a phoneme, frequency of lexical items). As highlighted by Baldwin and French (1990), historically, auditory analysis was the only available method for performing a linguistic-phonetic comparison of speech samples in FVC cases. However, given the advancements in techniques for performing acoustic analysis, the use of the auditory-only approach is now relatively rare. Gold and French (2011) report that this approach is used by just three of the 34 experts (9%) surveyed. Further discussion of auditory-only FVC analysis is found in Baldwin and French (1990) and French (1994).

1.1.2 Acoustic phonetic analysis

Acoustic phonetic analysis involves making observations of linguistic-phonetic variables without listening to the suspect and offender samples. An early form of acoustic-only analysis is *voice printing* (see Kersta 1962). Voice printing involves qualitative, largely text-dependent, visual comparison of spectrograms of the same utterance from a pair of suspect and offender recordings. The approach was initially claimed to achieve 100% recognition rates. Although challenging the early claims of infallibility, impressive accuracy rates are also reported in Tosi (1979) based on carefully controlled sections of recordings. However, voice printing has been largely dismissed by the FSS community[1] as unscientific and unreliable (Gruber and Poza 1995; Hollien 2002), and in *United States v Robert N Angleton* [2003] was ruled inadmissible as a form of forensic analysis in Texas (Morrison 2014). Despite this, voice printing is still admissible as expert evidence in other states of America (Tiersma and Solan 2012).

[1] International Association of Forensic Phonetics and Acoustics (IAFPA) Resolution - Voiceprints: `http://iafpa.net/voiceprintsres.htm` (accessed: 30th April 2014)

Modern acoustic-only analysis has a more principled linguistic-phonetic basis involving the extraction of acoustic-phonetic variables (e.g. formant frequencies). Despite this, there is still scepticism regarding the use of acoustic analysis without auditory analysis. Of the experts in Gold and French's (2011) survey, just one uses the acoustic-only approach.

1.1.3 Combined auditory and acoustic analysis

The majority of FVC evidence presented and admitted in courts (including in the UK, Germany, Turkey, Brazil and China) (Gold and French 2011) is based on a combination of auditory and acoustic analysis. This involves making qualitative judgements using auditory-analysis and where possible quantifying observations using spectrographic-acoustic analysis. A range of linguistic-phonetic variables is typically analysed and an overall conclusion provided to the court. For this reason auditory-acoustic analysis may also be referred to as the *componential* approach. The linguistic-phonetic variables analysed include segmental variables (vowels and consonants), suprasegmental variables (e.g. voice quality, prosody (incl. articulation rate, rhythm)), speech pathologies (e.g. stuttering, hyper-nasality), higher-order linguistic variables (e.g. lexical choice, syntax) and non-linguistic variables (e.g. hesitation phenomena, clicks) (see French *et al.* 2010: 146-147).

Gold and French (2011) report that all experts using the combined approach analyse mean fundamental frequency (f0). Voice quality (VQ) was considered the best speaker discriminant, although it is not routinely examined by all experts. Further, of those experts who do conduct VQ analysis, it is far from clear how such analyses are conducted and how methods differ between analysts. Gold and French (2011) also report that 97% of experts analyse vowel formants in FVC casework. The use of acoustic analysis (and specifically the extraction of vowel formants) as part of FVC was, in effect, made obligatory following the Northern Irish Court of Appeal decision in *R v O'Docherty* [2002] which is persuasive in England and Wales. Although the England and Wales Court of Appeal in *R v Flynn* [2008] re-affirmed the judgement in *R v Robb* [1991] that the use of acoustic analysis should be determined on a case-by-case basis, experts have largely continued to follow the guidelines in *O'Docherty* [2002].

1.1.4 Automatic speaker recognition (ASR)

An alternative to the analysis of linguistic-phonetic variables is the use of ASR systems. ASR systems can consist of a piece of stand-alone, commercial software which performs signal processing, speaker selection and statistical modelling (e.g. BATVOX[2]). ASR systems may also be built manually using widely available speech processing and statistical software (e.g. MATLAB). ASRs differ from linguistic-phonetic approaches in three key areas: how the speech signal is processed and analysed, the variables extracted, and the procedures for statistical modelling.

ASR typically involves treating the speech-active portion (i.e. with silences removed) of a recording holistically, by analysing the signal at equally spaced intervals (called frames). Following this *global* approach, the signal is not analysed as a series of discrete linguistic units as in §1.1.3 (although segmental ASR analysis is possible; Rose 2011a, 2013a). ASRs typically extract cepstral coefficients (CCs) (although many other variables are extractable in automatic analyses; e.g. PLPs) from each frame, which provide information about the power spectrum of the signal, capturing properties of the size, shape and short-term configuration of the supralaryngeal vocal tract. CCs can also be used to calculate derivatives which capture information about the dynamic properties of spectral change. Finally, the performance of ASR systems is typically analysed statistically using Gaussian Mixture Models (GMMs) (Reynolds *et al.* 2000). A detailed explanation of ASR variables and GMMs is presented in Chapter 3.

The benefit of the ASR approach is that a considerable amount of speaker discriminatory information can be extracted from speech samples without requiring labour intensive procedures for preparing or segmenting samples. ASR data are also continuous, allowing for efficient statistical modelling to generate probabilistic, numerical output. However, CCs are highly sensitive to noise in recordings, technical quality (e.g. sampling rate) and channel mismatch (commonly in FVC the offender sample is recorded via telephone transmission, while the suspect sample is recorded directly), although procedures for compensating for these factors are available (Alexander *et al.* 2004; Botti *et al.* 2004; Alexander 2005). Gold and French (2011) report that eight

[2] `http://www.agnitio-corp.com/products/government/batvox` (accessed: 29th April 2014)

experts (24%) currently use ASR for FVC. In all cases, the analysis includes some element of human supervision, although the role of the human-supervisor was not made explicit.

1.2 Expressing conclusions in forensic science

A number of frameworks are used for evaluating evidence and expressing expert conclusions in FVC cases. These include a binary decision (the suspect and offender are the same or different speaker(s); §2.1.1), classical probability scales (involving a gradient assessment of the likelihood of the suspect and offender being the same or different speaker(s); §2.1.2) and the UK position statement (two stage evaluation of the consistency and distinctiveness of the suspect and offender samples; §2.1.3). However, there have been increasing cross-disciplinary demands for changes in the way such forensic comparison evidence is evaluated and presented to the courts. This has led to claims that the field of expert evidence provision is undergoing a *paradigm shift* (Saks and Koehler 2005; Morrison 2009a). This shift involves a move away from expert judgements based on the probability (or likelihood) of the suspect and offender being the same or different individual(s), and towards the evaluation of the evidence using the likelihood ratio (LR) framework (§2.1.4).

Across forensic sciences, the LR is now widely accepted as the "logically and legally correct" (Rose and Morrison 2009: 143) approach for evaluating the strength of expert comparison evidence (Aitken and Stoney 1991; Robertson and Vignaux 1995b). The LR provides a gradient assessment of the strength, or weight, of the evidence, indicating the degree to which the it supports both the prosecution and defence. Applied to FVC, the LR involves analysing the similarity of the suspect and offender samples to each other, as well as the typicality of the offender sample (i.e. the evidence) with respect to the wider (relevant) population. Considerable support for the LR as the appropriate framework for the evaluation of expert evidence has also developed within the field of FSS (Rose 2002; Morrison 2010), and since 2001 there has been extensive quantitative research applying the numerical LR to FVC (Kinoshita 2001; see §2.2.2).

Despite this, worldwide very little FVC casework is performed using the numerical

LR framework (four of the 34 experts surveyed in Gold and French 2011). This is due, primarily (but not exclusively), to theoretical and practical difficulties in generating a single numerical estimate of the strength of speech evidence. Such difficulties derive, primarily, from the inherent complexity of speech as a form of forensic evidence; difficulties which are commonly overlooked in much of the current LR-based FVC research and casework.

1.3 Research aims and implications

This thesis explores some of the difficulties in applying the data-driven, numerical LR framework to FVC, by considering and accounting for the complexity of speech evidence from a linguistic-phonetic perspective. Numerical LR output in a given FVC case is necessarily dependent on decisions made by the analyst: the initial sample of suspect and offender speech, methods of analysis (§1.1) and choice of variables for comparison, as well as method-internal factors such as the definition of the relevant population, collection of representative data for testing, amount of data used, formula for LR computation, procedure for calibration and means of combining LRs from individual variables. Therefore, it is essential to understand the extent to which such dimensions of variability affect the numerical estimate of the strength of evidence and the performance of LR-based FVC systems.

Two specific issues for numerical LR computation are considered in this thesis. The first relates to the definition of the relevant population, against which the typicality element of the LR is quantified. In particular, consideration is given to varying dimensions of sociolinguistic sources of between-speaker variation (e.g. regional background, age and socio-economic class), and the extent to which such factors should be controlled in LR-based FVC using both linguistic-phonetic and ASR input variables. The second issue is the effects of different sources of sample size variation on numerical LR output. This involves analysing both how many and which speakers are used as development (or training), test and reference data (§3.2.1) in LR-based system testing. The experiments in this thesis also consider how the amount of data per reference speaker affects the resulting LRs.

The findings of this thesis have a number of implications for LR-based FVC, and potentially other areas of forensic science by extension. The results of these studies will allow analysts to understand and acknowledge the effects of the wide-range of different sources of variation encountered throughout LR-based analyses. The results of these experiments will also help analysts determine which sources of variation to control, based on the magnitude of their potential effects on the resulting LRs. More generally, these findings will contribute towards increasing the extent to which LR-based FVC accounts for the linguistic-phonetic complexity of speech evidence. In accounting for this complexity, the quality of FVC evidence will be improved in terms of the underlying, fundamental linguistic principles involved in the analysis. The studies will therefore help make the numerical LR more practically viable for the analysis of FVC evidence in casework. Further, from a theoretical perspective, the analysis of particular linguistic-phonetic and ASR variables will expand our understanding of how *group* and *individual* (Garvin and Ladefoged 1963; see further §3.3.1.1) information is encoded in FVC variables and how this information affects LR output.

The analysis of both linguistic-phonetic and ASR variables will contribute towards the integration of linguistic-phonetic and automatic methods of speech analysis. This is particularly important for two reasons. Firstly, there is a general consensus within the field of FSS that an integrated approach based on linguistic-phonetic and ASR analysis will provide the best method for successful speaker discrimination (as shown in the evaluation of human assisted speaker recognition (HASR) systems in NIST 2010; Greenberg *et al.* 2010). Secondly, ASR research very rarely considers the sociolinguistic dimensions of variability known to affect the distributions of linguistic-phonetic variables. ASR systems are commonly viewed as 'black boxes' and are treated with suspicion by the courts. The integration of techniques from linguistics and ASR therefore helps to improve the understanding of ASRs, in turn addressing recent calls for the improvement in the quality and transparency of forensic evidence presented to the courts (National Research Council 2009; Law Commission of England and Wales 2011).

1.4 Overview of the thesis

The Research Review in Chapter 2 discusses different approaches to the expression of conclusions in FVC cases. Further, it provides an overview of the position of the paradigm shift in FVC and the development, and application, of the LR in research and casework. The complexity of speech evidence, and the specific problems this causes for the definition of the relevant population and the collection of data for system testing, are also considered in detail. Finally, Chapter 2 provides a critical review of current approaches for dealing with these issues and outlines the specific research questions addressed in the thesis.

Chapter 3 presents the general methods applied throughout the experiments in the thesis. These include the speech corpora used, the structure of LR-based experiments, methods for LR computation, the linguistic-phonetic and ASR variables analysed, and the procedures used for extracting quantitative data. The general limitations of the experiments are also outlined.

Chapters 4, 5, 6 and 7 provide empirical data relating to theoretical issues of the definition of the relevant population. Chapters 4 and 5 explore how regional variation affects LR output using linguistic-phonetic variables (namely the formant trajectories of /uː/ and /aɪ/). Chapter 6 considers the effects of regional variation on LR output using ASR variables (CCs and derivatives). Chapter 7 examines the role of socio-economic class and age in defining the relevant population using the formant trajectories of /eɪ/.

Chapters 8, 9 and 10 provide empirical analysis relating to the practical issue of the amount of data required in LR-based system testing. Chapter 7 presents preliminary studies into the number of reference speakers and tokens per reference speaker using the raw data from Chapters 4 and 5. Chapter 9 considers the upper limit of the number of reference speakers and tokens required for LR testing based on Monte Carlo simulations (MCS) using articulation rate (AR) data. Chapter 10 expands the methods in Chapter 9 by using MCS to investigate the number of development, test and reference speakers (see §3.2.1) required in LR-based FVC using formant data from the hesitation marker UM (*erm*).

Finally, in Chapter 11 there is a discussion of the findings of the experiments with

suggestions for alternative approaches to the definition of the relevant population for FVC. This chapter also presents implications for future research and casework, and a series of general conclusions.

Chapter 2

Research Review

Across the forensic sciences there has long been debate about the most appropriate methods for analysing and presenting forensic evidence to the courts. Over time the criteria for the admissibility of expert evidence has changed. In *Frye v United States* [1923], the court ruled that expert testimony was admissible if the method used had received general acceptance within the relevant scientific community. Prior to *Frye*, the admissibility of expert evidence had been based on the expertise and experience of the analyst. The *Frye* ruling therefore shifted the focus for admissibility away from the expert and onto the widespread professional acceptance of the methods themselves.

However, the Supreme Court in *Daubert v Merrell Dow Pharmaceuticals* [1993] ruled that *Frye* was superseded by Federal Rule of Evidence (FRE) 702 (1975) which stated that "if scientific, technical, or other specialized knowledge will assist the trier-of-fact to understand the evidence or determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise."[3] FRE 702 therefore does not include the *Frye* requirement for general acceptance within the relevant field in determining admissibility. The court in *Daubert* also produced a series of guidelines for determining what constitutes admissible scientific evidence in US courts. Amongst other requirements, the court determined that valid scientific methodology should be based on empirical testing and peer review and that the error rates of the method should also be known.

[3] `http://www.law.harvard.edu/publications/evidenceiii/rules/702.htm` (accessed: 20th September 2014).

Saks and Koehler (2005) identified *Daubert* as the "driving force" (Morrison 2009a: 299) behind what they describe as a *paradigm shift* in the methods applied to evaluating forensic evidence. According to Saks and Koehler (2005), the new paradigm is based on "empirically grounded science" (p. 892) consistent with practices in forensic DNA analysis. Similarly, in 2009 the National Research Council (NRC) produced a report calling for the improvement in the quality of forensic evidence presented to the courts in line with the paradigm advocated by Saks and Koehler (2005). Morrison (2009a: 299) claims that implicit within both Saks and Koehler (2005) and the NRC report (2009) is the proposition that forensic evidence (of all kinds) should be evaluated using the likelihood ratio (LR) framework.

There has also been much debate within the field of FSS as to the most appropriate methods for analysing samples in FVC and the reliability of such evidence (for an overview see Nolan 2001; Foulkes and French 2001; Eriksson 2011). Central to this debate is the issue of how experts express their conclusions, since as Nolan (2001) states "the expression of the opinion is ... an outward sign of the way (an expert) conceptualises the task in which they are engaged" (p. 12). Within the field of FVC there is considerable acceptance of the LR as the logically and legally correct framework for evaluating evidence, at least in principle. Yet the complexity of speech as a form of forensic evidence introduces issues with the application of a fully numerical, data-driven LR approach, such as that advocated in Morrison (2014). Therefore, it is fair to say that the paradigm outlined in Morrison (2014) is not consistent with the Frye ruling that methods of forensic evaluation should be generally accepted within the field.

This chapter considers the frameworks currently used for evaluating FVC evidence, the development of the LR in FVC and the place of the *paradigm shift* within FVC. The theoretical and practical issues with the application of the numerical LR to FVC are explored in light of the complexity of speech evidence. Attention is then given to the specific issues relating to the experiments in this thesis: the definition of the relevant population and the collection of data for LR-based system testing. Finally, the research questions are detailed.

2.1 Conclusion frameworks in FVC

In their survey, Gold and French (2011) found that several frameworks are currently used worldwide for evaluating evidence and expressing conclusions in FVC casework. The use of different frameworks is determined by legal rulings in different countries, employers and governments, as well as by individual experts themselves. This section provides an overview of these frameworks. For each approach, the acceptance within the FSS community is discussed together with the logical, legal and practical issues surrounding its use.

To contextualise these issues, it is useful to define the elements of FVC analysis in terms of conditional probability or likelihood (p) based on propositions (or hypotheses) (H) and evidence (E). Propositions relate to the statements offered to the court by the prosecution and defence to explain the evidence. As is typical in the forensic statistics literature, the term *proposition* is preferred here since the term *hypothesis* has implications of frequentist hypothesis testing (Aitken and Taroni 2004: 6-7). Applied to FVC, the prosecution proposition is typically that the suspect and offender are the same speaker, while the defence proposition, in general terms, is that the suspect and offender are different speakers. The evidence is the data extracted from the offender sample (i.e. the unknown source).

2.1.1 Binary decision

Following the binary decision framework, the expert is restricted to a two-way, categorical decision: either the samples contain the voice(s) of the same or different speaker(s). A limitation of this approach is that it prohibits a gradient assessment of the degree of consistency between the samples. The expert is therefore forced to make illogical *cliff-edge* decisions about the identity of the offender (Robertson and Vignaux 1995b: 118). The *cliff-edge* effect refers to the arbitrary turning point between the two potential conclusions and the evidence required to move from one to the other. Given the multidimensionality of speech variables analysed as evidence and the inherent sources of variability (see §2.2.5), the binary decision approach has largely been rejected by the FSS community. This is reflected in the fact that just two of the experts surveyed in

Gold and French (2011) currently use this framework.

2.1.2 Classical probability scales

Some of the limitations of the binary decision framework are resolved by classical probability scales, in which the expert expresses conclusions in terms of the gradient probability of the samples containing the voice(s) of the same or different speaker(s) given the evidence. An example of such a scale is in Table 2.1. Gold and French (2011) report that classical probability scales are the most commonly used framework for FVC evidence, accounting for 13 of the 34 (38%) practitioners surveyed. This approach is used worldwide (including Europe, USA, Brazil, South Korea and Australia) and is typically employed by experts using auditory and acoustic analysis (§1.1.3).

Table 2.1: Example of a classical probability scale for FVC conclusions (Broeders 1999: 129; equivalent to that in Baldwin and French 1990: 10)

Positive identification	**Negative identification**
sure beyond reasonable doubt	*probable*
there can be very little doubt	*quite probable*
highly likely	*likely*
very probable	*highly likely*
probable	
quite possible	
possible	
… that they are the same person	… that they are different people

2.1.2.1 Issues with posterior probability

Both the binary decision and the classical probability scale frameworks have been criticised within the field of FSS (Broeders 1999, 2001; Champod and Evett 2000; Champod and Meuwly 2000) and the wider forensic community (Robertson and Vignaux 1995b). The primary criticism is that these frameworks are based on posterior probability involving an assessment of the probability of the propositions given the

evidence $p(\mathrm{H}|\mathrm{E})$. However, posterior probability is ultimately an issue for the trier-of-fact, as it is equivalent to an assessment of the probability of the innocence or guilt of the suspect based on the evidence. The overlap between the expert and trier-of-fact when expressing $p(\mathrm{H}|\mathrm{E})$ conclusions is most evident where the offender sample constitutes the crime, meaning that propositions are formulated at the offence level (Lucy 2005: 118). Labov (1988; see also Labov and Harris 1994) reports a case in which a baggage handler was accused of making threatening telephone calls to Los Angeles airport. Based on auditory and acoustic analysis, Labov concluded that the voices in the samples belonged to different speakers, and the suspect was subsequently found innocent. However, such a categorical decision is directly equivalent to the trier-of-fact's assessment of the innocence of the accused.

Furthermore, in order to determine posterior probability the expert requires access to information "from sources other than an objective scientific evaluation of the (suspect) and (offender) samples" (Morrison 2009c: 4). That is, to assess the likelihood of the suspect and offender being the same or different individual(s), it is necessary to have access to all of the evidence presented to the court, such as whether the suspect was in the country at the time or whether they had an alibi. Such information should theoretically only be available to and assessed by the trier-of-fact. Even if such knowledge is available to the expert, it is not the expert's role to evaluate it. It is also essential that the other evidence in the case does not influence the expert's conclusion, even subconsciously or inadvertently.

Finally, conclusions expressed as a binary decision or using a classical probability scale only account for the probability of one proposition (usually the prosecution proposition). However, only with an assessment of the likelihood of the evidence under both the prosecution and defence propositions is the trier-of-fact able to evaluate its strength with regard to innocence and guilt. To consider only one proposition is also inconsistent with the objective responsibility of the expert to aid the court. Therefore, it is preferable to use a framework which considers the strength of the evidence under the competing propositions rather than the probability of the propositions themselves. This is emphasised by the ruling in *R v Doheny and Adams* [1996], which states that "the scientist should not be asked his opinion on the likelihood that it was the defendant

who left the crime stain" (Rose 2007b).

2.1.3 UK Position Statement (UKPS)

To address these issues, French and Harrison (2007) present an alternative model for evaluating FVC evidence, now often referred to as the UK Position Statement (UKPS). UKPS is the result of debate within a sub-section of the FSS community (French 2005; French and Harrison 2006) regarding the appropriateness of classical probability scales, which until 2007 had been the dominant framework for expressing conclusions in UK casework.

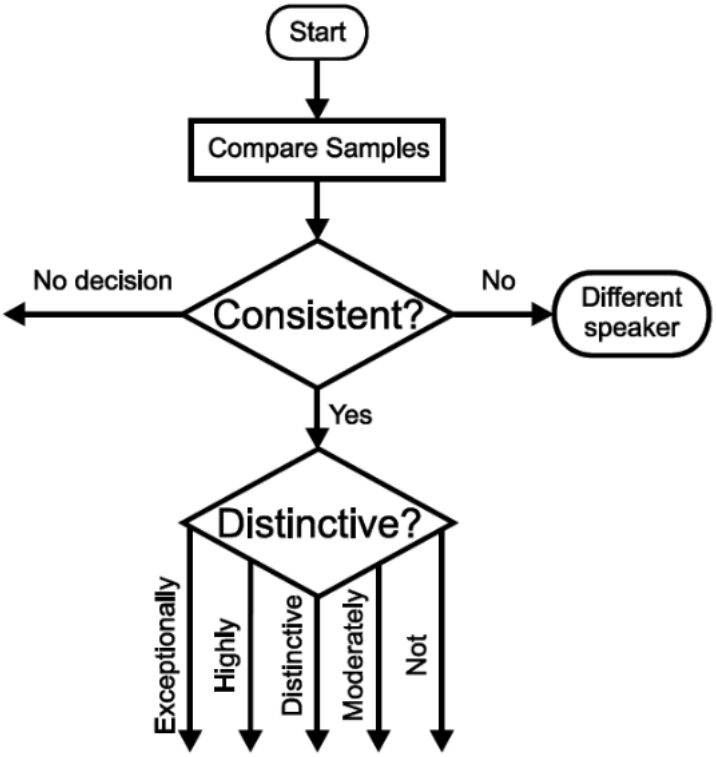

Figure 2.1: Flow chart of the UK Position Statement framework for FVC evidence (from Rose and Morrison 2009: 143)

UKPS consists of a two-stage evaluation (Figure 2.1). The first stage requires an assessment of the similarity between the suspect and offender samples, termed the consistency judgement. It allows experts to reach one of three mutually exclusive conclusions: *consistent*, *not consistent* or *no decision*. According to French and Harrison (2007), a *not consistent* verdict should be preferred unless the differences

between the samples can be explained by "established models of acoustic, phonetic or linguistic variation" (p. 141). If the two samples are judged to be consistent, the expert moves to the second stage, termed the distinctiveness judgement. This is an assessment of the typicality of the shared features across the samples within the wider population since, as Nolan (2001) states, strength of evidence is dependent on "whether the values found matching … are vanishingly rare, or sporadic, or near universal in the general (relevant) population" (p. 16). Distinctiveness is classified using the following five-point scale:

5. Exceptionally distinctive - the possibility of this combination of features being shared by other speakers is considered to be remote
4. Highly distinctive
3. Moderately distinctive
2. Distinctive
1. Not distinctive

from French and Harrison (2007: 141)

Distinctiveness is, for the majority of variables, assessed qualitatively. That is, while the analysis of the samples may involve quantification of acoustic variables, their typicality is assessed based on the expert's knowledge and professional experience, or with reference to published studies of sociolinguistic variation. When applying the UKPS, the "general (relevant) population" (Nolan 2001: 16) used to assess distinctiveness is defined according to the regional and social groups to which the expert believes the offender belongs.

UKPS has been signed by 25 forensic practitioners and interested academics. According to Gold and French (2011), UKPS is currently employed by 11 (32%) of the 34 practitioners surveyed and has largely replaced classical probability scales in the UK. With the exception of one expert, the combined auditory and acoustic approach is the preferred method of analysis for those using UKPS.

2.1.3.1 Limitations of UKPS

Although UKPS represents a shift away from posterior probability, there remain logical shortcomings with the approach, as raised in Rose and Morrison's (2009) response to French and Harrison (2007). Firstly, consistency and distinctiveness are analysed on different scales, meaning that it is difficult to interpret the relative similarity and typicality of the suspect and offender samples. Secondly, the scales are categorical with a finite number of potential outcomes and are serially ordered such that distinctiveness is only assessed if the samples are judged to be consistent with each other (issues with similar two-stage approaches are discussed in Evett 1991: 10-11). This is problematic since it prohibits the gradient assessment of the strength of the evidence under the two competing propositions in all cases. Thirdly, the categorical, binary outcome of the consistency judgement introduces *cliff-edge* effects into the analysis. A *not consistent* judgement is also equivalent to an assessment of the propositions given the evidence (i.e. the samples contain the voices of different speakers). Finally, Rose and Morrison (2009) state that it is not clear how the analysis of multiple variables should be combined using UKPS.

However, the overarching criticism of UKPS in Rose and Morrison (2009) is that it falls short of either a conceptual or numerical implementation of the Bayesian LR (discussion of French *et al.*'s 2010 rejoinder to Rose and Morrison 2009 is at §2.2.4).

2.1.4 The Bayesian approach

The LR is the probabilistic framework used for the evaluation of forensic DNA evidence, and the move towards the LR reflects the role of DNA in "setting the standard" (Balding 2005: 55) across forensic sciences. The LR forms a component of Bayes' theorem which may be applied to the entire criminal trial. This section discusses the application of Bayes' theorem and Bayesian inference for reasoning under uncertainty in criminal trials. The LR as an independent component for estimating the strength of forensic evidence is then discussed.

2.1.4.1 Bayes' theorem and Bayesian reasoning

Bayes' theorem (Bayes 1763) provides the theoretical foundation for a branch of statistics based on conditional probability, in which probability is considered as a measure of belief in an event or series of events. It is conditional in that it is dependent on available information (Redmayne 2001: 55). According to Aitken and Taroni (2004: 22), Bayes' theorem is defined by two fundamental elements: (1) that belief can be modified as new information emerges or existing information changes, and (2) that different individuals' beliefs in the same event will vary due to differences in the weights attached to each piece of information. Bayes' theorem may be expressed as:

$$p(\mathrm{H}_n|\mathrm{E}) = p(\mathrm{H}_n)p(\mathrm{E}|\mathrm{H}_n) \tag{2.1}$$

from Lee (2004: 8)

where p is probability, E is evidence, H_n is a sequence of events and $|$ is given. According to Bayes' theorem, the probability of a sequence of events given the evidence $p(\mathrm{H}_n|\mathrm{E})$ is equivalent to the product of the prior probability of that sequence events $p(\mathrm{H}_n)$ and the probability of the evidence assuming that series of events $p(\mathrm{E}|\mathrm{H}_n)$ (Iversen 1984: 12).

The conceptual application of Bayes' theorem is commonly referred to as Bayesian inference or reasoning. Bayesian inference plays an important and natural role in daily life, since our beliefs and opinions relating to uncertainty change as we come into contact with relevant information. Since the real world truths of the events of a crime are inherently uncertain, Bayesian inference provides the probabilistic model for making judgements in criminal trials. On the basis of the combined weight of the evidence presented to the court, the trier-of-fact assesses the likelihood of the defendant's innocence or guilt (Good 1991: 89-90). Since the burden of proof lies with the prosecution, a guilty verdict may only be reached when the likelihood of guilt assigned by the trier-of-fact is greater than the *beyond reasonable doubt* threshold (i.e. where likelihood of guilt approaches one).

The odds form of Bayes' theorem as applied to criminal trials is given as:

$$\frac{p(\mathrm{H}_p)}{p(\mathrm{H}_d)} \times \frac{p(\mathrm{E}|\mathrm{H}_p)}{p(\mathrm{E}|\mathrm{H}_d)} = \frac{p(\mathrm{H}_p|\mathrm{E})}{p(\mathrm{H}_d|\mathrm{E})} \tag{2.2}$$

where H_p is the prosecution proposition (guilty), H_d is the defence proposition (innocent) and E is evidence. The prior odds reflect the trier-of-fact's assessment of the probability of the competing propositions before the introduction of (new) evidence. The weight or strength of each piece of evidence is expressed as the ratio of $p(\mathrm{E}|\mathrm{H}_p)$ and $p(\mathrm{E}|\mathrm{H}_d)$ (the LR or Bayes Factor) which modifies the prior odds to establish the posterior odds. The posterior odds concern the "ultimate issue" (Lynch and McNally 2003: 96) of innocence or guilt; an assessment of the probability of the competing propositions given the combined weight of the evidence.

There are a number of advantages to using Bayes' theorem in criminal trials. The theorem is flexible, allowing the trier-of-fact's belief in the competing propositions to be modified as new evidence is introduced. Further, in Bayes' theorem conditional probability is subjective (Redmayne 2001: 54). Therefore, where a jury is entrusted with interpreting the evidence, Bayes' theorem allows each individual to assign different weights to that evidence and thus potentially generate different posterior probabilities. For forensic evidence, where the trier-of-fact cannot reasonably be expected to interpret the evidence, the expert is responsible for assessing the weight of the evidence and, following Bayes' theorem, can do this using the LR. The trier-of-fact can then use this to generate posterior probability.

However, there are a number of issues with the practical application of Bayes' theorem in criminal trials. The first is the appropriate definition of the prior odds, to reflect the initial assumption that the suspect is innocent until proven guilty. Cohen (1982) highlights that the presumption of innocence requires the prior probability to be zero, meaning posterior probability would also necessarily be zero, irrespective of the evidence. It is preferable, therefore, to think about the prior odds in terms of the *island problem* (Aitken and Taroni 2004: 117-118). A crime is committed on an island with a population N, of which the suspect is a member. Without any evidence, each member of the population is assumed to be equally likely to have committed the crime. Robertson and Vignaux (1995b) argue that the prior odds can then be thought of as the ratio of the probability of choosing the suspect at random from the population ($p(\mathrm{H}_p)$) divided by

the probability of choosing any other member of the population at random ($p(\mathrm{H}_d)$):

$$\frac{\left(\frac{1}{N}\right)}{\left(\frac{N-1}{N}\right)} = \frac{1}{N-1} \tag{2.3}$$

Secondly, the assessment of posterior probability requires an arbitrary division between innocence and guilt. $p(\mathrm{H}|\mathrm{E})$ is therefore necessarily susceptible to the *cliff-edge* effect since the trier-of-fact needs to make a categorical decision as to whether the defendant is guilty beyond a reasonable doubt (Evett 1991: 12). It is not clear how the threshold for determining innocence and guilt is determined by juries. Finally, the formal quantification of Bayes' theorem in criminal trials (i.e. assigning numerical values to each piece of evidence to generate an overall probability of innocence and guilt) has largely been rejected by the courts in England and Wales (see §2.1.4.5.

2.1.4.2 Assessing strength of evidence using the LR

The LR provides a gradient estimation of the strength of the evidence (E) based on the ratio of its probability given the prosecution proposition (H_p) and its probability given the defence proposition (H_d):

$$\frac{p(\mathrm{E}|\mathrm{H}_p)}{p(\mathrm{E}|\mathrm{H}_d)} \tag{2.4}$$

As a ratio, the outcome is a value centred on one, such that LRs of greater than one offer support for H_p while LRs of less than one offer support for H_d. The magnitude of the LR determines how much more likely the evidence is given one proposition over the other (Evett *et al.* 2000). A LR of ten, for instance, means that the evidence is ten times more likely assuming the proposition that the samples contain the voice of the same speaker than assuming the proposition that the samples contain the voices of different speakers.

The numerator of the LR is equivalent to the similarity between the suspect and offender samples (i.e. the probability of the offender values assuming they were produced by the suspect). The denominator is equivalent to the typicality of the offender sample with respect to the relevant population (i.e. the probability of the offender values assuming they were produced by another member of the relevant population) (Aitken and Taroni 2004: 206; see §2.3). Using the numerical data-driven approach, typicality

is quantified using statistical models generated from a sample of the relevant population (the forensic scientist may also estimate typicality based on experience and expertise, although there is considerable debate over how to implement this approach scientifically, with suggestions including testing the performance of experts in controlled experiments; see Evett 1991: 21). Such a sample is termed the background or reference data, and the statistical model of these data is called the background or reference model.

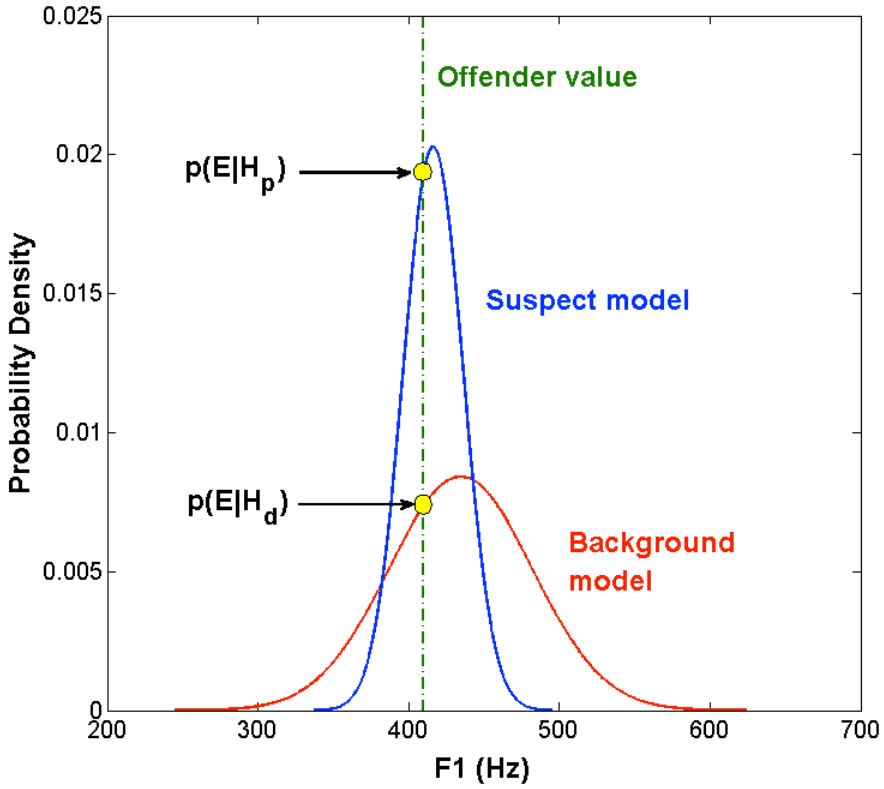

Figure 2.2: Univariate example of a LR computed using Lindley's (1977) model for a same speaker comparison based on midpoint F1 (Hz) values for New Zealand English (NZE) /u:/

An example of the computation of a numerical LR is shown in Figure 2.2. The data consist of midpoint F1 (Hz) values for /u:/ from the Origins of New Zealand English (ONZE) corpus (see further §3.1.2). The background model is generated using 50 speakers (13 tokens per speaker) and the suspect model consists of 13 tokens from a single speaker. Following Lindley (1977), these data are modelled using normal distributions. The evidence is a single offender value for F1 of 410 Hz. The evidence is firstly assessed under the assumption that it was produced by the suspect, by calculating

its probability at the intersection with the suspect model ($p(\mathrm{E}|\mathrm{H}_p) = 0.0196$). The evidence is then evaluated under the assumption that it was produced by a random speaker in the relevant population, by calculating its probability at the intersection with the background model ($p(\mathrm{E}|\mathrm{H}_d) = 0.0072$). The LR is the ratio of these probabilities. Consistent with the fact that the suspect and offender data are from the same speaker in this example, the LR is greater than one (2.72) and so offers support for the prosecution.

Bayes' theorem also allows experts to combine different pieces of evidence. Naïve (or idiot's) Bayes (Kononenko 1990) states that numerical LRs from separate sources can be combined by taking their product providing the strands of evidence are independent of each other. This is because the product of LRs from correlated pieces of evidence will overestimate their strength by weighting the same evidence more than once. Although the LR is a fundamental element of the Bayesian approach, its application in forensic contexts is independent of Bayes' theorem since it "does not (itself) make use of prior probabilities" (Morrison 2009c: 4). Therefore, the LR is free from the criticisms of Bayes' theorem outlined in §2.1.4.1.

2.1.4.3 Logical and legal correctness of the LR

Robertson and Vignaux (1995b) emphasise that "expert evidence should be restricted to the (LR) given by the test or observation of its components" (p. 21). There are a number of logical and legal arguments for this. The distinction between assessing the probability of the propositions and the probability of the evidence separates the roles of trier-of-fact and expert. This theoretically leaves the trier-of-fact free to interpret the expert forensic evidence within the context of the other evidence presented to the court. Further, by applying the LR, the expert's conclusion is not informed by priors which should be determined by the trier-of-fact or influenced by information beyond the scope of the FVC evidence itself. The LR overcomes logical shortcomings of the UKPS since it ensures that similarity is considered relative to typicality, rather than having two independent stages of analysis. Therefore, it is not the case that typicality is only assessed if there is judged to be consistency between the suspect and offender samples. Rather, the denominator of the LR ensures that the defence proposition is considered when judging typicality. Furthermore, the LR provides a gradient estimation

of the probability of the evidence, thus avoiding *cliff-edge* effects.

2.1.4.4 Logical fallacies

Despite the logical and legal correctness of the LR, there remain a number of potential fallacies in probabilistic reasoning and inference when presenting the results of expert analysis to the court. The *prosecutor's fallacy* (Thompson and Schumann 1987) involves presenting the probability of the evidence in terms of the probability of the propositions. This fallacy is also referred to as the transposed conditional since E and H are inappropriately switched such that probability is conditional on the evidence $p(\mathrm{H}|\mathrm{E})$ rather than on the propositions $p(\mathrm{E}|\mathrm{H})$. An example of this is provided in Aitken and Taroni (2004: 112) in which a bloodstain from an unknown offender is compared with the blood of the known suspect. There is a match between the blood groups across the samples meaning that $p(\mathrm{E}|\mathrm{H}_p)$ is equal to one. This blood group is also found in 1% of the population, meaning that $p(\mathrm{E}|\mathrm{H}_d)$ is 0.01 (for evidence types where $p(\mathrm{E}|\mathrm{H}_p)$ is equal to one (e.g. DNA evidence), $p(\mathrm{E}|\mathrm{H}_d)$ is referred to as the *random match probability*). The prosecutor's fallacy involves presenting this assessment of typicality as a 99% (posterior) probability that the suspect is guilty, despite the fact that the LR (strength of evidence) is $1/0.01 = 100$. Clearly a 99% chance of guilt is not the same as the evidence providing 100 times more support for the prosecution proposition than the defence proposition.

Aitken and Taroni (2004: 115) also provide an example of the *defender's fallacy* (Thompson and Schumann 1987). Considering the same case as above where the analysis of the bloodstains provided a LR of 100, assume that the relevant population consists of 200,000 people. Based on the assessment of typicality (1% of the population share that blood group), the defence claim that there are 2000 people who share that blood group. Therefore the probability that the defendant is guilty is 1/2000 and so the evidence is of little probative value. There are number of issues with such inference. Firstly, the evidence is again presented in terms of posterior probability which is inappropriate since it requires access to the prior odds. Secondly, the assumption about the probability of guilt assumes that each of the 2000 people who share that blood group are equally likely to have committed the crime. However, based on other evidence in

the case this assumption is likely to be inappropriate. Finally, as highlighted by Aitken and Taroni (2004) "evidence which increases the odds in favour of guilt from 1/200,000 to 1/2000 is surely relevant" (p. 115).

2.1.4.5 Bayes and the LR in the courts

The courts in England and Wales have displayed sensitivity to the issues of fallacious interpretation and presentation of statistical evidence. Early legal concerns over the expression of conclusions based on DNA evidence were raised by the Court of Appeal in England and Wales in *R v Deen* [1993], following Deen's conviction for rape in 1990. The court quashed the original conviction based on a re-evaluation of the forensic evidence which reduced the original random match probability of 1 in 700,000 to 1 in 33. The court's decision was also based on the fact that the expert at the court of first instance had committed the prosecutor's fallacy by presenting the random match probability as the probability of guilt (see further Balding 2005). The overestimation of the strength of forensic (DNA) evidence and the presentation of evidence in the form of posterior probability has also provided the grounds for successful appeals in *R v Doheny and Adams* [1996] and *R v Clark* [2003].

Despite this, the courts have largely rejected the formal application of Bayes' theorem to criminal trials. In 1994, Dennis Adams was found guilty of rape, in part based on a DNA random match probability of 1 in 200,000,000, despite alibi testimony and the victim's evidence which suggested that Adams did not look like the offender (Balding 2005: 151). Based on the defence expert's advice the judge directed the jury to assess this and other evidence in the case in terms of Bayes' theorem based on numerical probabilities. However, the conviction was quashed by the Court of Appeal in *R v Adams* [1996] based on the fact that the prosecution expert had committed the prosecutor's fallacy. Further, the court ruled that "to introduce Bayes' theorem, or any similar method, into a criminal trial plunges the jury into inappropriate and unnecessary realms of theory and complexity deflecting them from their proper task."

The courts reception of the LR has been somewhat mixed. In *George v R* [2007], the re-evaluation of the firearm discharge residue evidence (used as one of primary pillars of the prosecution case in the original trial) in the form of a LR was well received

by the England and Wales Court of Appeal. This, in part, led to the conviction being quashed. However, the LR has been challenged by the Court of Appeal in *R v T* [2010]. The appeal focused on the use of the LR to express the expert's opinion based on an analysis of footwear mark evidence at the original murder trial. The Court of Appeal concluded that "outside the field of DNA (and possibly other areas where there is a firm statistical base) ... Bayes' theorem and (LRs) should not be used." Morrison (2012) argues forcefully that the court displayed fundamental misunderstandings of the LR. Morrison claims that the court equated the LR directly with the use of quantitative data and statistical models, rather than treating the logical framework and the numerical implementation as two distinct elements. That is, the LR can and should be used as a conceptual framework to evaluate evidence without the need for databases to estimate a numerical value to express the strength of evidence. Although the ruling applies to forensic shoeprint evidence, it has been criticised by Berger *et al.* (2011), Aitken *et al.* (2011), Redmayne *et al.* (2011) and others for its potential implications as precedent in other areas of forensic science.

2.2 The LR in FVC

2.2.1 Acceptance of the LR framework for FVC

The first explicit discussion of the application of the LR to speech is in Champod and Meuwly (1998, 2000) (see also Lewis 1984; Broeders 1995; Rose 1998, 1999). Champod and Meuwly outline the value of Bayes' theorem in other branches of forensic science, the logical fallacies associated with posterior probability, and the conceptual correctness of the LR for FVC, but do not discuss the application of these principles to research or casework. Broeders (1999) was the first author to outline specific limitations of classical probability scales (see §2.1.2), but also acknowledges the difficulties of applying the LR to speech data. Champod and Evett (2000) provide a response to Broeders (1999), calling more strongly for the use of the LR in FVC. Discussion on the appropriateness of the LR for FVC is also found in Nolan (2001), Rose (2002) and Morrison (2010).

The last decade and a half has seen a rapid rise in the level of general acceptance of the LR framework, in principle, for FVC. However, in practice, the UKPS (2.1.3) still holds sway for evaluating FVC evidence in casework in many countries (e.g. UK). Therefore, as highlighted by Morrison "the paradigm shift (in FVC) is incomplete and those working in the new paradigm still represent a minority within the (FSS) community" (2009a: 298).

2.2.2 LR-based research

The increasing acceptance of the framework has been reflected in an exponential increase in the amount of LR-based FVC research. The first quantitative LR-based study of linguistic-phonetic variables appears to be Kinoshita (2001), who analysed midpoint F3 values from /o m s/ in spontaneous realisations of *moshimoshi* (*hello* in Japanese), as well as F2 of /i/ and F2 and F3 of /e/ from target words elicited via a map task. Participants were ten male speakers of Standard Japanese aged between 21 and 36. Cross-validated comparisons (see further §3.2.2.3) were performed using the formula in Aitken (1995) and classified using a LR distance approach based on posterior odds. An overall LR (OLR) was generated by taking the product of the LRs from individual variables. Optimally 81 of the 90 (90%) same speaker (SS) comparisons achieved OLRs of greater than one (i.e. support for the prosecution), while 174 of the 180 (97%) different-speaker (DS) comparisons achieved OLRs of less than one (i.e. support for the defence).

Kinoshita (2001) is limited primarily by the procedures available at the time for computing and analysing LRs. As highlighted in Kinoshita (2001: §6.6.1.3), the only available formula (Aitken 1995; based on Lindley 1977) was developed for analysing refractive indices of glass, where $p(\mathrm{E}|\mathrm{H}_p)$ is expected to be one (i.e. if the recovered and comparison samples came from the same source there will be a 1:1 match between the samples). Therefore, Aitken (1995) does not adequately capture the occasion-to-occasion variability in speech produced by the same individual. Aitken (1995) also assumes that variables are normally distributed, which is not necessarily appropriate for linguistic-phonetic variables. Furthermore, individual LRs were combined using naïve Bayes, despite evidence of correlations between variables for certain speakers

(Kinoshita 2001: §6.6.2.2).

Since 2001, considerable methodological advancements have been made in LR-based research, such as the application of LR formulae from other areas of forensic science. These include the Multivariate Normal and Kernel Density (MVKD) approaches (Aitken and Lucy 2004; see §3.2.2.1) developed for glass fragment analysis. The availability of different formulae has also expanded the range of variables which can be analysed using the LR. Much of the research in LR-based linguistic-phonetic FVC has focused on the speaker discriminatory power of vowel acoustics. Studies have considered the performance of formant midpoints (Alderman 2004a; Rose 2010) and trajectories modelled with parametric curves to reduce dimensionality (Kinoshita and Osanai 2006; Morrison 2009b; Enzinger 2010).

The performance of other segmental variables has also received some attention. A number of studies have considered the LR-based speaker discriminatory power of different acoustic properties of nasals (Enzinger and Balazs 2011; Kavanagh 2012; Kasess *et al.* 1993; Yim and Rose 2012; Enzinger and Kasess 2013). LR-based analyses have also been conducted using spectral properties of fricatives (Rose 2011b; Kavanagh 2012). Less attention has been given to suprasegmental variables, although studies have considered the performance of long-term f0 (LTf0) (Kinoshita 2005; Gold 2014), localised intonation contours (Wang and Rose 2012; Pang and Rose 2012) and articulation rate (AR) (Gold 2014; see Chapter 9). Recently, LR-based studies have also been conducted to evaluate the speaker discriminatory performance of the acoustic properties of VQ (Enzinger *et al.* 2012) and the glottal waveform (Vandyke 2014) which captures the vibratory motion of the vocal folds.

The development of techniques for the application of the LR to ASR has a somewhat longer history. Notably, the Gaussian Mixture Model - Universal Background Model (GMM-UBM) approach was developed for ASR and outlined in Reynolds *et al.* (2000), although the use of GMMs for modelling ASR data is reported earlier (Reynolds 1995). The development of the GMM-UBM approach has led to a considerable amount of research into the speaker discriminatory power of cepstral coefficients (CCs) and derivatives (Rose 2011a, 2013a) and semi-automatic variables such as long-term formant distributions (LTFDs: Becker, Jessen, and Grigoras 2008; Becker, Jessen, and Grigo-

ras 2009; Gold, French, and Harrison 2013; Jessen and Enzinger 2014). A number of studies have also considered the comparative performance of GMM-UBM and MVKD for linguistic-phonetic variables (Rose and Winter 2010; Morrison 2011a).

Finally, considerable advancement has been made in the methodologies available for empirically assessing validity (Brümmer and du Preez 2006; van Leeuwen and Brümmer 2007) and reliability (Morrison, Thiruvaran, and Epps 2010; Morrison, Zhang, and Rose 2011), optimising system performance through calibration (Brümmer and du Preez 2006), and combining LRs from correlated pieces of evidence (Pigeon *et al.* 2000; Brümmer 2007). These advancements have been made primarily in the field of ASR and have subsequently been applied to linguistic-phonetic FVC. Given the developments made in modelling and analysing FVC data over the last 20 years, speech is now situated towards the forefront of the claimed *paradigm shift* and informs LR practice in other disciplines (Ramos-Castro 2012; Morrison 2013).

2.2.3 LR-based casework

Despite widespread acceptance of the LR in principle, the mass of LR-based research and the developments in techniques for LR testing, only seven of the 34 (20.5%) experts surveyed by Gold and French (2011) use the LR in FVC casework and, of those, just four (11.8%) use the numerical LR approach. Rose (2013b) is the only published report of the application of the numerical LR to casework in which LR-based evidence was received by the court (reference is also made to the presentation of LR-based FVC evidence in the courts in Australia in Morrison 2009a). The case came to trial in 2008 and involved a fraudulent telephone call (containing 14 seconds of offender speech) made to an Australian bank requesting the transfer of $150 million. The suspect samples were a series of recordings made during police interviews and house searches, as well telephone intercepts of the suspect talking to a friend.

The comparison focused on the word *yes* and the phrase *not too bad*. From *yes*, the onset, midpoint and offset of the first three formants of /je/ were analysed along with a "crude" (Rose 2013b: 304) analysis of the lower cut-off in the spectrum of /s/. From *not too bad*, Rose analysed time-normalised f0 contours sampled across their trajectory

and F1, F2 and F3 midpoints from /o/ in *not*, /uː/ in *too* and /æ/ in *bad* (using Rose's phoneme symbols). In the absence of a specific alternative hypothesis, the relevant population was defined as adult male speakers of General Australian English (AusEng) (based on assumptions about the offender; see §2.3). The reference data consisted of 35 adult males aged between 20 and 70, recorded over the telephone. The analyst prompted responses of *yes* and *not too bad* using questions such as *how's it going?* and attempted to "indirectly prime" the speakers by producing the phrase *not too bad* with the "correct intonation" at the beginning of the conversation (Rose 2013b: 285). Each participant was recorded twice to obtain non-contemporaneous (i.e. random variability introduced by recording speakers on two separate occasions separated by some period of time) assessments of within-speaker variability.

Modelling the reference data both normally and with kernel density (KD), Rose achieved a LR of 70 for the formant analysis of /jɐ/. The low cut-off analysis of /s/ generated a roughly estimated LR of 2.5. The acoustic analysis of the f0 pattern in *not too bad* generated a LR of 20, while a categorical analysis of the tonal structure of the phrase generated a rough LR estimate of marginally greater than one. For the formants extracted from *not too bad*, LRs of 24 (/o/), five (/uː/) and 11 (/æ/) were estimated respectively. Despite calculating an OLR of 11 million using naïve Bayes, a more conservative OLR of 300,000 was arrived at by "simply discard(ing) the putatively correlated LRs (e.g. from individual formants in *not*)" (2013b: 305).

Rose (2013b) also provides a critique of the procedures applied, claiming that system performance should ideally have been presented to the court as a means of interpreting the validity and reliability of the final OLR. The availability of data for pre-testing would also have allowed for the OLR to be calibrated (see §3.2.4), thus potentially improving system validity. Rose's analysis also fails to empirically account for between-variable correlations in determining a conservative OLR. As outlined in §2.2.2, since 2008 techniques for doing this have been developed for FVC. Finally, Rose highlights that the use of relatively small amounts of suspect and offender data means that the LR estimate will be relatively imprecise and that, in such cases, "it is better, if possible, to try to avoid (absolute numerical values for the OLR)" (Rose 2013b: 305).

Beyond Rose's critique, there are other limitations of the analysis. Firstly, the motivation

for the choice of variables is not made explicit. The analysis is based on a limited set of continuous, acoustic variables. However, within *yes* and *not too bad* there are other variables that may have affected evidential support. There may also have been variables of evidential value in the other sections of the offender sample aside from *yes* and *not too bad*. Secondly, the ecological validity of the procedures used to collect reference data is questionable. The context in which the samples for the reference data were made is cognitively different from that of the evidential samples. Further, reference speakers were prompted to produce the target word and phrase and were primed to produce the appropriate intonation contour. The potential effects of such mismatch, or of the non-Bayesian, subjective decisions made by the analyst, on LR output are not explored in Rose (2013b). There are also a number of issues with the definition of the relevant population which are considered in §2.3.1.

There are no published guidelines for the application of the numerical LR to FVC casework. However, given the current state of methodological techniques, a set of procedures can be determined based on the paradigm advocated in Morrison (2014) and its application in Enzinger and Morrison (2014). The procedures for computing a LR for a single variable are:

1. Extraction of acoustic data from the variable of interest from the suspect and offender samples.
2. Decision regarding the relevant population (see §2.3).
3. Multiple recordings, matching the facts of the case at trial, from a sample of the relevant population collected for use as development, test and reference data.
4. Extraction of acoustic data from the variable of interest for the development, test and reference speakers.
5. SS and DS scores (prior to calibration LRs are referred to as *scores*) computed (using an appropriate LR formula) for the development and test data using the reference data to assess typicality (*feature-to-score* stage; see §3.2.2).
6. Calibration coefficients generated by applying logistic regression (see §3.2.4.1) to the scores from the development data (training stage).

7. Calibration coefficients applied to the scores from the test data to convert the scores into calibrated LRs (*score-to-LR* mapping; see §3.2.4).
8. Validity and reliability (see §3.2.3 calculated based on calibrated LRs from the test data (this is presented to the court as a means of interpreting the performance of the system under the conditions of the case at trial).
9. Score computed for the suspect and offender data using the same LR formula as in (5).
10. Calibration coefficients generated from the development data applied to the score for the suspect and offender data to convert the value into a calibrated LR.

If multiple correlated variables are analysed, as is typical in linguistic-phonetic FVC, further stages of analysis are implemented:

1. Stages (1) to (5) repeated for each variable.
2. Logistic regression fusion coefficients (Brümmer *et al.* 2007) derived from the scores for the development set.
3. Fusion coefficients applied to the scores for the test set to convert the scores for individual variables into a calibrated OLR (which incorporates the correlation between the variables).
4. System validity and reliability metrics calculated based on the OLRs for the test data.
5. As in stage (9), scores computed for each variable using the suspect and offender data.
6. Scores for the suspect and offender data combined using the fusion coefficients from the development data to generate a calibrated OLR.

2.2.4 Issues with the LR in FVC

The reluctance of experts to use the LR in casework is due, primarily, to the difficulties in implementing a fully numerical LR analysis in FVC (highlighted by the complexity of the procedures in §2.3.2). Indeed, French and Harrison acknowledge "the desirability

of . . . (the LR) . . . framework" (2007: 142), but claim that the difficulties involved in its application make the UKPS a preferable alternative for FVC. Rose and Morrison (2009) highlight that the theoretical definition of the relevant population and practical issue of the collection of reference data are "real problems" for the numerical LR approach, but state that "these problems, however real, do not prevent the use of (LRs)" (p. 156). In response, French *et al.* (2010) claim that "it is unrealistic to see it as merely a matter of time and research before a rigorously and exclusively quantitative LR approach can be regarded as feasible" (pp. 149-150). They claim that this is primarily because there is insufficient available data to estimate the distribution within the relevant population of all of the potential variables analysed in a given case. French *et al.* (2010) also challenge the available methods for collecting such reference data (see further §2.4).

There are two further issues with the practical application of the LR to FVC evidence, which underlie the arguments in French *et al.* (2010). Firstly, by examining the probability of the evidence rather than the probability of the hypotheses, the LR framework introduces a level of analyst objectivity which is not present in posterior probability-based approaches. However, in computing a numerical LR the analyst must make methodological decisions at several steps: the initial sampling of suspect and offender speech, choice of variables for comparison, methods of analysis, definition of the relevant population, collection of representative reference data, sampling the reference data, selecting the formula for LR computation and calibration procedure, the means of combining LRs from separate variables, and so on. The extent to which these decisions affect LR output is rarely considered. Secondly, as highlighted by Nolan (2001), the intrinsic link between the method of analysis and the conclusion framework in FVC means that there is inevitably an extent to which the framework dictates how the analysis is performed (e.g. what can be analysed).

Such issues and concerns relating to the application of the LR to FVC stem primarily from the inherent, and arguably unique, complexity and multidimensionality of speech as evidence. The following section explores the complexity of speech evidence and the issues that this causes for LR-based FVC.

2.2.5 Complexity of speech evidence

Firstly, LR-based analyses typically focus on a small number of acoustic variables which yield continuous, quantitative data compared across the same word or phrase. The benefit of using the same word or phrase is that the material is directly comparable across the evidential samples and within-speaker variability is minimised by normalising for phonological context and syntactic function. Underlying this approach is also an assumption that a small subset of variables is able to accurately represent the properties of the suspect and offender voices. This *sampling* approach is claimed to be akin to the proportion of the genome analysed in forensic DNA analysis (Morrison p.c.). However, this assumption is potentially insufficient given the number of variables available to the analyst in componential linguistic-phonetic FVC (§1.1.3) which may have a substantial effect on the resulting strength of evidence.

Gold and Hughes (2014) state that the primary reason for the focus on continuous, acoustic data is that this is the only type of data which can be handled by current LR formulae. Speech is complex for LR modelling since it consists of numerous variables which may be continuous or discrete, normally or non-normally distributed, display different distributions within and between speakers, and contain multiple, highly correlated features. At present, there are no means of empirically computing LRs for discrete data such as allophonic consonantal variation (although see Schwartz *et al.* 2011), frequencies of lexical items and the analysis of VQ and vocal settings. As outlined in Gold and Hughes (2014), for the numerical LR to become a more realistic proposition in FVC, it is necessary for new LR models to be developed. Only a small amount of work has considered these issues (Aitken and Gold 2013; Foulkes *et al.* 2013-2015; Nair *et al.* 2014; Neocleous *et al.* 2014).

Secondly, speech variables display complex patterns of structured between-speaker variation (Rose 2002; Foulkes and Docherty 2006; French *et al.* 2010). Such variation is found as a function of factors such as regional background, socio-economic class, age, and ethnicity, as well as the social networks and communities of practice in which a speaker participates. Within a single regional variety different variables are often stratified in different ways. For example, /uː/-fronting is a widespread change in progress in English, and is generally correlated with age. By contrast, another on-going

change, /əʊ/-fronting, correlates with both age and sex, being led by young females (Haddican *et al.* 2013; Williams and Kerwill 1999). Across regional varieties, the social stratification of variables may also differ. For instance, /əʊ/ and /eɪ/ carry considerable socially conditioning in the north east of England, but much less in the south east (Watt 2000, 2002).

In collecting development, test and reference data, it is important that samples match the (relevant) facts of the case at trial otherwise the resulting strength of evidence may be misrepresentative. This is because a speaker will never produce the same word or sentence in exactly the same way even consecutively, meaning that $p(\mathrm{E}|\mathrm{H}_p)$ for speech evidence will never be one (unlike forensic DNA analysis). In current LR-based research and casework, within-speaker variability is captured using non-contemporaneous samples from each speaker separated by some undefined period of time. Non-contemporaneity encompasses multiple sources of structured and random within-speaker variability across samples. Results from Enzinger and Morrison (2012) show that system validity and reliability are overoptimistic when using contemporaneous, compared with non-contemporaneous, samples. However, very little research has empirically tested these issues with different variables commonly analysed in FVC (with the exception of Coe 2012).

Further, evidence from sociolinguistics and sociophonetics indicates that there are numerous, complex sources of within-speaker variability which affect speech production. These include interlocutor, conversational topic and function, level of formality, self-consciousness, physical setting, time of day, illness, fatigue and intoxication. In a given FVC case, a large number of these factors are likely to be relevant: suspect and offender samples are typically recorded with different interlocutors who have a different level of familiarity with the speaker, talking about different topics in different degrees of formality at different times of day. At the present time, no empirical research has investigated the extent to which such factors affect LR output, or whether such factors have a much bigger effect on the resulting LRs than the use of non-contemporaneous samples.

Finally, speech variables form highly correlated sub-systems due to a range of factors. The biological structure of the vocal apparatus means that variables such as vowel

formants are inherently interrelated within and between phonemes. Correlations are also determined by the linguistic system, such that in cases of (vowel) change there are push-pull effects which are thought to ensure that sounds remain acoustically distinct. There is also evidence of linguistically arbitrary correlations due to social factors, for example speakers in Derby with TH-fronting (/θ ð/ → [f v]) also typically produce labial-r (/ɹ/ → [ʊ w]) (Milroy 1996). Although methods for empirically combining LRs have been developed in ASR (logistic regression fusion), Gold and Hughes (2014) argue that it is an empirical question as to whether such methods capture the linguistic-phonetic complexity of the correlations in the raw data.

This thesis considers the implications of the complexity of speech evidence for two specific issues in LR-based FVC: the definition of the relevant population and the collection of development, test and reference data.

2.3 Definition of the relevant population

One of the primary benefits of the LR framework is that the evidence is evaluated under the competing propositions of both prosecution and defence. In most FVC cases, H_p will be the straightforward proposition that the offender sample contains the voice of the known suspect. The definition of H_d is more problematic for the evaluation of comparison evidence. This is because in order to calculate $p(E|H_d)$ it is necessary to assess the probability of the evidence relative to a model of the relevant population, which is defined by the defence (or alternative) proposition. For example, if the defence were to claim that the suspect did not commit the crime but that one of his brothers did, the relevant population would necessarily consist solely of the suspect's brothers.

However, often the defence offer a non-specific alternative proposition such as *it was not the defendant who committed the crime, it was someone else*, where the relevant population technically consists of any member of the population excluding the suspect. In many cases the defence offer no alternative proposition at all. As highlighted by Robertson and Vignaux, "it is … difficult if not impossible to determine the probability of the evidence with a vague and ill-defined (alternative) hypothesis" (1995b: 31). This is because with a non-specific alternative proposition the prior odds will be extremely

small. Assuming, for the sake of exposition, that the relevant population consists of all other people in the world and that there is no other information available, the prior odds are:

$$\frac{p(\mathrm{H}_p)}{p(\mathrm{H}_d)} = \frac{\left(\frac{1}{N}\right)}{\left(\frac{N-1}{N}\right)} = \frac{\left(\frac{1}{7.243\times10^9}\right)}{\left(\frac{7.243\times10^9-1}{7.243\times10^9}\right)} = 1.38 \times 10^{-10} \tag{2.5}$$

where N is the size of the relevant population, i.e. the estimated world population of around 7.243 billion.[4] The LR based on such a broadly defined relevant population will also be small (Aitken and Taroni 2004: 206) and the evidence will offer little probative value to the court. Therefore, it is necessary to reduce the relevant population "to more manageable proportions" (Aitken and Taroni 2004: 206), unless "there is no evidence to separate the perpetrator from the ... population (at large)" or where the evidence is independent of variation within sub-populations (Robertson and Vignaux 1995b: 36).

In most cases certain pragmatic assumptions about the defence proposition must be made. The concept of the relevant population was first defined by Coleman and Walls (1974: 276) as:

> those persons who could have been involved (in the crime); sometimes it can be established that the crime must have been committed by a particular class of persons on the basis of age, sex, occupation or other sub-grouping, and it is then not necessary to consider the remainder of, say the United Kingdom.

This definition has subsequently been developed by Smith and Charrow (1975), who use the term *suspect population* to refer to "the smallest population known to possess the culprit as a member" (p. 556). Similarly, *suspect population* is used by Lempert (1977) to refer to the population of potential offenders. Following such definitions, the relevant population must be based on what is known (or can be assumed) about the offender, rather than the suspect (Robertson and Vignaux 1995a). Furthermore, since the relevant population is defined by the defence proposition, it must, logically, remain constant across all forensic evidence presented to the court.

[4] `http://www.worldometers.info/world-population/` (accessed: 30th June 2014).

2.3.1 Logical relevance

As in Coleman and Walls (1974), assumptions about the alternative proposition may be made based on factors which define sub-groups within the population at large, such as regional background, age and sex. This approach is referred to as *logical relevance* (Kaye 2004, 2008) and factors may be considered logically relevant if they affect the distribution of a variable in the wider population. This approach has been used extensively in forensic DNA analysis. Since allele frequencies differ between racial groups (Gill and Clayton 2009), the logically relevant population is typically defined by race. In the UK, three databases are used to evaluate DNA evidence based on broad racial groups: white Caucasian, Afro-Caribbean and Asian. As it is not possible to infer racial background from the offender sample, multiple LRs are often presented based on different assumptions about the relevant population.

Variation in allele frequencies between sub-populations within racial groups has generally been shown to be relatively minor (Gill and Clayton 2009; Balding *et al.* 1996; Budlowe *et al.* 1999). Gill *et al.* (2000) assessed the level of regional variation in DNA profiles across 24 European populations making up the ENFSI DNA Short Tandem Repeat (STR) Population Database.[5] They concluded that for white Caucasians a single pan-European database is sufficient for generating stable LR output. Where such variation (e.g. regional variation due to high coancestry - regional groups displaying genetic similarity based on interrelatedness) is considered important, it may be accounted for by incorporating a *coancestry coefficient* (F_{ST}) into the LR calculation (Balding and Nichols 1994). Beyond race and regional background, the National Research Council (NRC) states that, in some cases, it may also be necessary to consider other potentially logically relevant factors such as age and sex in forensic DNA analysis (1996: 30).

Applying the principles of logical relevance to FVC, Rose (2004: 4) claims that, in the absence of a specific alternative proposition, the underlying assumption should be that *the voice in the offender sample does not belong to the suspect, but to "another same-sex speaker of the language."* Following this approach, the relevant population is defined by the sex and language of the offender. This definition has been used in almost all LR-based FVC research (Kinoshita 2002; Alderman 2004a; Kinoshita 2005; Rose 2006;

[5] `http://strbase.org` (accessed: 30th June 2014).

Rose, Kinoshita, and Alderman 2006; Rose 2007a; Morrison and Kinoshita 2008; Morrison 2009b) and casework (Rose 2013b), and in the collection of FVC databases (Rose 2007-2010; Morrison *et al.* 2010-2013; Zhang and Morrison 2011).

2.3.1.1 Limitations

The Rose (2004) application of logical relevance makes two potentially problematic assumptions about FVC cases. Firstly, this approach assumes that language and sex information are readily extractable from the offender sample. However, many cases present themselves where even these matters are not trivial (French et al. 2010: 145). For example, Foulkes and French (2012) describe a case in which the unknown speaker on a telephone recording was assumed to be an adult female drug addict, but was in reality a child. The issue of language is also complex due to issues of multilingualism, mobility and identity. Further issues are encountered which defining language more narrowly in terms of regional dialect, since dialect does not equate directly to geographical background. This is due to linguistic differences associated with the physical and psychological spaces (Britain 2013), meaning that certain regional varieties are linguistically well-defined whilst for other dialects regional patterns may be much more heterogeneous. Such incompatibility between social groupings and linguistic differences is reflective of the broader difficulties in defining what is meant by the term *speech community* (see Patrick 2008).

Secondly, the Rose (2004) approach assumes that sex and language are the most important sources of between-speaker variation, at least for those variables which are typically analysed in LR-based FVC (e.g. vowel formants). However, this reflects a naïve view of the complexity of between-speaker variation in speech. Unlike in forensic DNA analysis, it is in principle possible for the sociolinguistically-informed expert to determine considerably more demographic information about the offender, beyond sex and language (French and Harrison 2006). Furthermore, for many of the available variables in auditory-acoustic FVC, sociolinguistic sources of variation other than language and sex may be far more relevant. For instance, there is no expectation for marked differences between males and females in terms of VOT in British English (BrEng), but there may well be differences between ethnic groups (Heselwood and

McChrystal 2000).

However, only a limited number of studies have acknowledged the complexity of between-speaker variation and the associated issues for defining the relevant population. Alderman (2004b) compared the Bernard (1970) and Cox (1999) databases of AusEng as reference data using F1, F2 and F3 midpoints from the tense monophthongs /i a o ʉ ɜ/. LR testing was conducted using non-contemporaneous recordings of 11 speakers aged between 18 and 26, and OLRs were calculated using naïve Bayes. Output was similar across the two sets although Cox (1999) (72.7%) marginally outperformed (1970) (63.6%) by 9.1% in SS discrimination. Alderman concludes that both are useful for FVC, although "as more time passes and further change occurs (Bernard's) usefulness as a reference distribution will diminish" (2004b: 182). However, other sources of between-speaker variation, such as regional background and age, were not assessed. Further, it is considered problematic to judge the usefulness of reference data purely on the output of speaker discrimination tests, rather than on whether it represents an appropriate definition of the relevant population which answers the question asked by the court.

Rose *et al.* (2006) examined the speaker discriminatory value of AusEng /aɪ/ based on a dual-target analysis (see §3.3.1) of the first three formants. As in Alderman (2004a), typicality was assessed using Bernard (1970) as reference data, and the issue of change over time is again acknowledged. Based on a comparison with Cox's (1999) data, Rose *et al.* (2006) claim that the first target of F2 is now c. 100 Hz lower and that the second target of F1 is now c. 30 Hz higher. Despite acknowledging that such change is "important" (p. 330), the potential effect on LRs was not investigated. Similarly, Morrison's (2008) study of AusEng /aɪ/ acknowledges the use of heterogeneous reference data with regard to regional variation and age (19 to 64 years). However, the logical relevance of these factors and their effect on the resulting LRs were overlooked. Only Zhang *et al.*'s (2008) study of midpoint F1, F2 and F3 values for /i y/ in Standard Chinese extends Rose's (2004) definition in controlling for age, sex and regional dialect.

The most extensive discussion of the complexity of the logically relevant population is Loakes (2006), who investigated the performance of a test set of four pairs of male twins from Melbourne aged 18 to 20. Input consisted of F1, F2 and F3 midpoint

values from the eleven monophthongs of AusEng extracted from non-contemporaneous samples. The twin data were initially compared with the reference data from Sydney from Bernard (1970). However, based on this pre-testing only a subset of the available formants from each phoneme were excluded from LR-based testing due to the levels of divergence between the test and reference data. Loakes (2006: 214) offers a number of potential reasons for the divergence, including regional variation (test data = Melbourne, reference data = Sydney), variation in the tasks performed by the test and reference speakers and the level of sociolinguistic heterogeneity (with regard to age, class etc.) in the reference data. As suggested in Alderman (2004a) and Rose *et al.* (2006), processes of sound change in the time separating the test and reference data may also account for the differences in formant frequencies. These factors lead Loakes (2006) to conclude that in defining the relevant population "tighter controls on (other) social variables might also be applied" (p. 198) such as age, communities of practise, education, occupation and friendship groups.

However, there are also a number of problems with narrowly defining the relevant population according to sociolinguistic factors. The first relates to the appropriateness of the expert defining the relevant population. Given that the relevant population is defined by the defence proposition, it is not, strictly speaking, the role of the expert to make assumptions about it. However, there are good reasons to prefer decisions relating to FVC evidence to be made by the linguistics expert rather than by the court, legal professionals or lay people (although this view is not universal; see §2.3.2). Secondly, an issue for the definition of logical relevance more generally is the paradox that without knowing the identity of the offender, it is not possible to know for certain the logically relevant population of which he/she is a member. This applies equally to the general Rose (2004) default as well as more specific propositions.

2.3.2 Lay listener-judged similarity

The importance of considering the appropriate definition of the relevant population is also addressed by Morrison *et al.* (2012) who present an alternative to logical relevance based on *lay listener-judged similarity*. Morrison *et al.* (2012: 64) argue that the default defence proposition in FVC should be that:

> the suspect is not the speaker on the offender recording but is someone who sounds sufficiently similar to the voice on the offender recording that a police office (or other appropriate individual) would submit the offender and suspect recordings for forensic comparison.

That is, the relevant population consists of speakers who sound similar to the offender based on the judgements of a panel of lay (i.e. linguistically naïve) listeners. This definition is based on the fact that suspect and offender samples are submitted to an expert based on a judgement made by a lay listener, usually a police officer, that the voices sound sufficiently similar to warrant expert analysis.

Morrison *et al.* (2012) also offer suggestions as to how this should be implemented to generate a representative set of reference data. The panel of lay listeners responsible for assessing similarity should match the profile of the listener who made the original decision to submit the recordings for analysis. The degree of match between the lay listeners and the original listener extends to occupation (i.e. police officer), regional background and level of FVC experience or linguistic training. The listeners should be presented with recordings which match with the relevant facts of the suspect sample. Therefore, if the suspect sample contains telephone transmitted speech, then this should be reflected in the samples played to the lay listeners. The samples should also match in terms of the ambient conditions in which the suspect sample was made. Finally, since the original decision was a subjective one, the reference data generated by lay listeners may include speakers of different sociolinguistic backgrounds (e.g. males and females, different regional varieties, different ages).

Morrison *et al.* (2012) also provide three examples of how this approach could have been applied in previous cases. The casework examples outline procedures for dealing with mismatched suspect and offender samples, judgements about speaker sex and accent and identifying the appropriate properties of the lay listeners. Finally, Morrison *et al.* (2012) provide an empirical demonstration of the speaker similarity approach using a generic ASR system to identify speakers who are *closest* to the offender based on distances within the multidimensional Mel-frequency cepstral coefficient (MFCC) space (as a proxy for lay-listener judged *similarity*; although the equivalence between MFCC- and lay listener-based assessments of speaker similarity is questionable). The

performance of this system was then compared with the performance of a system based on data extracted at random from a larger database. The results suggest that systems based on data selected using the similarity approach outperform the use of randomly selected datasets in terms of validity.

2.3.2.1 Limitations

Lay listener-judged similarity overcomes the limitations of logical relevance since the expert is not responsible for making potentially incorrect decisions about the sociolinguistic background of the offender. Further, the relevant population is defined by a single grouping factor, namely speaker similarity, rather than the potentially numerous demographic factors involved in logical relevance. However, there are also significant limitations of this approach which make it problematic for FVC. As with logical relevance (§2.3.1), these limitations stem from a naïve view of the linguistic-phonetic complexity of variation in both speech production and perception.

First, evidence from the literature on ear-witness reliability (Bull and Clifford 1999), perceptual dialectology (Montgomery 2007) and voice parades (Atkinson in progress) shows that listeners' linguistic backgrounds can have a substantial effect on their decisions about speaker identity and similarity. Therefore, the background of the listeners may need to be controlled far more narrowly than suggested in Morrison *et al.* (2012). Specifically, the controls over the listeners should focus on those factors which affect the perception of similarity and ignore other factors. For example, ensuring that the panel of lay listeners consists of individuals of the same age may be more important than ensuring that the listeners are police officers. Further research is required to identify the potential sources of variability in lay listener judgements of similarity. The question of which of these factors to control is therefore an empirical one. However, even with knowledge of which factors to control, it is not clear how narrowly to control them (e.g. do listeners need to be exactly the same age, or within a certain age range?).

Second, even for lay listeners of the same background, any set of data judged according to similarity is expected to display a high degree of within-group variation since individual listeners attend to different elements of the speech signal (McDougall 2013). Therefore, it is possible, if listeners are presented with samples from a truly sociolin-

guistically heterogeneous database, that the resulting dataset would contain males and females, as well as speakers of different regional and social backgrounds. Morrison *et al.* (2012) claim that this is acceptable since the decisions were made using logical assumptions about the relevant population, irrespective of the sociolinguistic make-up of the resulting dataset. However, given that decisions may not be linguistically principled, the resulting dataset will not necessarily consist of "those persons who could have been involved (in the crime)" (Coleman and Walls 1974: 276). This also has potential implications for the courts. The decisions about which speakers are included in as population data may not be transparent. The LR-based results are also not likely to be replicable because different panels of lay listeners will make potentially wildly different judgements about similarity.

Thirdly, there are a number of practical issues with the way in which decisions are made by the police officer in the first instance and subsequently by the lay listeners. In most cases, it is questionable whether the original decision to submit recordings for expert analysis will be based purely on an objective judgement about the similarity of the voices in the suspect and offender samples. Rather, it is likely that this decision is made based, at least to some extent, on other information in the case. Therefore, it is not clear whether the panel of lay listeners is really making the same type of decision as that made in the first instance. Further, many FVC cases involve initial analysis by the expert to assess the viability of the samples prior to a full analysis for the courts (in terms of whether there is sufficient similarity between the samples to generate a SS proposition, as well as technical factors). Thus, the decision to analyse the recordings is often made by the expert, rather than a police officer.

There are necessarily differences in the conditions under which the original decision was made and the conditions under which the lay listeners make their judgements of similarity (e.g. time of day, motivation in performing the task, structure of the task). It is essential to understand how these differences affect perceptions of speaker similarity and their effect on the resulting LR output. There are also issues with the recordings presented to the panel of listeners for comparison with the offender sample. Although in theory listeners would be presented with samples from a sociolinguistically diverse range of speakers, in reality, as much for practical reasons, it is necessary for

the analyst to make prior decisions as to which speakers the listeners hear (Enzinger and Morrison 2014). Such decisions will almost certainly relate to the database from which potential speakers for LR testing are identified. For the majority of databases which may be used for this task (see §2.4.2.2), speakers are controlled only for sex and language (reflecting Rose 2004). This is not equivalent to the panel making decisions from a heterogeneous database representative of the entire population, and introduces a degree of analyst subjectivity which this approach is designed to avoid.

Furthermore, the issue of within-speaker variability is not resolved in Morrison *et al.*'s (2012) claim that samples should match the "speaking style" of the suspect sample. The issue of style is an extremely complex one which has received considerable attention in the sociolinguistic literature (e.g. Coupland 2007). As highlighted in §2.2.5, there are potentially numerous sources of within-speaker variability. It may be possible for the analyst to infer these from the suspect sample (e.g. intoxication), but in many cases it may not (e.g. time of day). Such complexity is oversimplified in Morrison *et al.* (2012). Yet, the range of sources of within-speaker variability are expected to have a significant effect on how listeners make judgements relating to similarity. Therefore, it is necessary to know empirically how controls over the samples presented affect the population data identified by lay listeners.

2.4 Collection of development, test and reference data

Another significant issue for the application of the LR framework is how the relevant population should be sampled once it has been appropriately defined. In forensic DNA analysis, databases are collected using convenience sampling from blood banks and disease screenings. This is possible for DNA since allele profiles are "uncorrelated with the means by which samples are chosen" (National Research Council 1996). However, as highlighted by the multitude of sources of within-speaker variability in §2.2.5, speech variables are intrinsically affected by the situation in which they were elicited. This means that it is extremely difficult to collect a sample of the relevant population which sufficiently matches the facts of the case at trial.

It is also important to emphasise that any set of data used for LR-based testing will

necessarily display some degree of mismatch with the facts of the case at trial, due to the wide range of factors affecting within- and between-speaker variation. The extent to which such mismatch affects the resulting LR estimates is an empirical question and has received little attention in the literature. However, it is essential that the influence of any mismatch is acknowledged and understood by practitioners in casework. There are currently three alternatives for assembling quantitative data for LR testing: case specific data, existing non-forensic corpora and existing forensic databases. The benefits and limitations of these approaches are considered below. A further approach, of course, which is not considered in detail here, is that the analyst estimates patterns in the relevant population based on experience and previous research.

2.4.1 *Going and getting it* (Rose 2007b)

Rose (2007b) argues that "we (forensic speech scientists) have ... to be prepared to go and get a suitable reference sample for each case." The use of case-specific data provides the expert with much greater scope to control relevant elements of the facts of the case at trial. An example of the collection of case-specific reference data in FVC is given in Rose (2013b). The limitations of the specific procedures in Rose (2013b) are outlined in §2.2.3. There are also more general limitations of the *going and getting it* approach.

Firstly, there will still inevitably be some degree of mismatch with the facts of the case at trial, given that the expert is responsible for making subjective decisions over which factors to control and which to ignore. In Rose (2013b), the limitations of the case-specific reference data were not considered in terms of their potential effect on LR output. Secondly, there are considerable financial and time constraints imposed when conducting casework. Given these constraints, there is a danger that case-specific reference data will have more shortcomings, in terms of the facts of the case, than *off-the-shelf* data (§2.4.2). Thirdly, such constraints mean that it is only possible to collect reference data for analysing very short amounts of speech or a limited set of variables. For example in Rose (2013b) analysis is limited to the word *yes* and the phrase *not too bad*. It is therefore considered prohibitively difficult to collect case-specific data for a componential analysis of a range of linguistic-phonetic variables (§1.1.3), especially

where variables occur in different utterances, words and phonological contexts.

2.4.2 *Off-the-shelf* data

2.4.2.1 General corpora

It may be possible to use general corpora which were not originally collected for forensic purposes in LR testing. A significant benefit of this approach is that data need not be collected for each case. This improves the time and cost efficiency of the analysis and potentially extends the range of variables which may be analysed. The most suitable corpora are probably those collected as part of sociolinguistic research. Sociolinguistic corpora often contain speakers controlled for numerous sociolinguistic factors, allowing for a definition of the logically relevant population with varying narrowness. The breadth of sociolinguistic research means that data are available for a range of different regional and social groups. Such corpora also contain relatively long samples (c. 40-60 mins) and, in some cases, multiple samples of speakers in different speaking styles (e.g. spontaneous speech, ethnographic interview, read text). Examples of such corpora include ONZE (Gordon *et al.* 2007), the Big Australian Speech Corpus (Wagner *et al.* 2010) and the Northern (British) Englishes corpus (Haddican 2008-2013).

However, the lack of forensic realism in such corpora is potentially problematic, since they are likely to display considerable divergence from the facts of any case at trial. In particular, samples recorded for general corpora do not generally involve transmission mismatch (Künzel 2001; Byrne and Foulkes 2004), mismatch in background noise and signal-to-noise ratio (SNR), or stylistic variability relevant for forensic purposes such as speech under stress or different emotional states. Sociolinguistic corpora can also be very small, with few containing more than 30 speakers from the same sociolinguistic community. Further, corpora containing multiple recordings from each speaker are generally made in a single session, rather than non-contemporaneously.

2.4.2.2 Forensic databases

Finally, there are a small number of databases available which were collected specifically for forensic purposes. The only existing forensic corpus for any variety of BrEng is the Dynamic Variability in Speech corpus (DyViS) (Nolan *et al.* 2005-2009; see §3.1.1). Forensic databases also exist for AusEng (Rose 2007-2010; Morrison *et al.* 2010-2013) and Standard Chinese (Zhang and Morrison 2011). In the field of ASR, considerably more forensic databases are available (see Campbell and Reynolds 1999). Forensic databases have the benefit of having been controlled for the typical facts of casework, such as transmission mismatch, non-contemporaneity samples and mismatch in speaking style. Such databases are also commonly much larger than general corpora, allowing for testing using different subsets of the available data.

However, there are also limitations with forensic databases. There is limited availability of forensic databases. In the case of BrEng, even DyViS is limited since it contains only speakers of Standard Southern British English (SSBE). This is inadequate for narrower definitions of the logically relevant population, even with regard to regional background. Conversely, other forensic databases contain speakers from very wide, sociolinguistically heterogeneous populations, reflecting the Rose (2004) default relevant population based on sex and language. For example, Morrison *et al.* (2010-2013) contains male speakers of AusEng, with no control over other potentially sociolinguistically relevant factors. The relative lack of usable forensic databases is highlighted by French and Harrison as a primary reason "for precluding the quantitative application of (the LR) approach" (2007: 142) in casework. Further, forensic databases, in particular those used for ASR, often contain relatively small short samples for each speaker and little spontaneous material.

2.5 Amount of development, test and reference data

A final issue for the application of the LR to FVC is the amount of development, test and reference data needed for robust system testing. The limited amount of previous research in this area has focused on the effects of the number of reference speakers on LR output. Ishihara and Kinoshita (2008) analysed LTf0 from non-contemporaneous

samples of spontaneous speech produced by 241 male speakers of Japanese. The samples were extracted from the larger non-forensic Corpus of Spontaneous Japanese (CSJ) (Maekawa *et al.* 2000). LTf0 was parameterised using the long-term mean and standard deviation (SD) as well as skew, kurtosis, mode and modal density. The speakers were divided into two groups and within each group 12 differently sized population samples were created containing between ten to 120 speakers. This allowed for the computation of two LRs for each comparison for the same population size. Cross-validated (§3.2.2.3) MVKD (§3.2.2.1) LRs were computed using all 241 speakers as test data and typicality assessed against the differently sized reference sets.

Ishihara and Kinoshita (2008) found the median SS $\log_{10}$ LR (LLR; §3.2.2.4) to be up to three orders of magnitude greater when using ten reference speakers compared with using all 120. The overall range of LR scores also decreased as the amount of reference data increased. DS pairs were found to be more sensitive to the size of the reference sample. With ten speakers the median DS LLR value was around -30, although for certain pairs values extend far beyond -30. In the 120 speakers condition, the DS median was located between -2 and -3. As with the SS results, the overall range of scores decreased as the size of the sample increased. Ishihara and Kinoshita (2008) also found that equal error rate (EER; §3.2.3.1) generally improved as the number of speakers increased, although "improvement seems more rapid up to the population size 30" (p. 1943). As a crude form of calibration, the study also included an analysis of the EER threshold relative to the LLR zero threshold (i.e. neutral evidence; see further §3.2.2.4). When using ten reference speakers the EER threshold was found to be furthest away from zero, with increasing convergence as the number of speakers increased.

Ishihara and Kinoshita (2008) conclude that "we do need a large population data in order to produce reliable (LRs)" and that "(LRs) produced using anything smaller than 30 (reference speakers) (are) highly unreliable" (p. 1944). Although their results provide evidence against the use of small amounts of reference data, there is no explicit discussion as to why small samples should produce such imprecise LRs. Further, given the intrinsic properties of how MVKD LRs are computed (particularly for variables with different numbers of dimensions) (§3.2.2.1), there is reason to predict that sample size

should affect different variables from different regional and social groups in different ways. Further, although the issue of calibration was considered with regard to accept-reject thresholds, Ishihara and Kinoshita (2008) did not assess the effect of sample size on calibrated LRs (§3.2.4) or log LR cost function (C_{llr}; validity metric which is logically consistent with the Bayesian approach, see §3.2.3.1).

Whilst Ishihara and Kinoshita (2008) focus on the effects of small numbers of reference speakers, Rose (2012) investigated an upper limit for reference sample size at which point LR performance becomes asymptotic. Rose (2012) used Monte Carlo simulations (MCS; see Chapters 9 and 10 in this thesis) to synthesise F1, F2 and F3 midpoint values for AusEng /aː/ for up to 10,000 speakers based on values in Bernard (1970). Using both the multivariate normal and KD approaches, LRs were computed for real suspect and offender data which were known to have been produced by the same speaker. Typicality was assessed as a function of the number of reference speakers between five and 60. Output was compared against the *true* LR, which was defined as the LR computed using the maximum amount of reference data (in this case 10,000 speakers).

The results of Rose (2012) are comparable with those of Ishihara and Kinoshita (2008). Based on univariate LR analyses of F1, F2 and F3, SS scores were generally higher in magnitude than the *true* value when using small amounts of reference data (fewer than ten speakers). The overall range of LRs was also considerably greater when using small numbers of reference speakers. However, relatively stable scores were achieved (within two SDs of the *true* scores) by the inclusion of 30+ reference speakers. This was the case even for F2, which displayed the greatest sensitivity to sample size. A similar pattern was found in the multivariate analysis, with the distributions of values skewed towards stronger scores when using small samples. Compared with the univariate analysis, however, the range of scores was far more sensitive to sample size using MVKD.

However, Rose's (2012) preliminary study has a number of limitations. The test data are based on a single suspect and offender comparison. It would be preferable to assess the performance of a large set of test data, where it is known *a priori* whether samples came from SS or DS pairs, as a function of the number of reference speakers. In the absence of such data, Rose (2012) was unable to assess how system validity metrics

such as EER and C_{llr} are affected by sample size. Further, the scores were not calibrated based on coefficients generated from an appropriate set of development data. Therefore, it was not possible to assess the role of calibration in determining the overall sensitivity of LR output to reference sample size.

A limited amount of work has also considered the issue of sample size for ASR. Van der Vloed *et al.* (2011) investigated the inbuilt reference population optimisation algorithm in Batvox[6] which identifies the N *closest* speakers, based on Kullback-Leibler distances calculated from the MFCC vectors, to the suspect (note that Batvox bases population selection on the suspect rather than the offender) from a larger database of speakers. LRs were computed for a test set (i.e. mock suspects and offenders) of 16 male speakers of Swiss-French in Batvox using three population data conditions. The first contained 35 speakers extracted from a 45-speaker subset of the 1995 speakers in the Swiss-French PolyPhone database (Chollet *et al.* 1996). The second condition contained 35 reference speakers extracted from the whole database of 1995 speakers, and the third condition contained 1400 speakers extracted from the 1995 speakers. Tests were conducted using samples of speech transmitted via the Global System for Mobile Communications (GSM) and the Public Switched Phone Network (PSTN) (see Bigelow 1997; Kondoz 2004).

For both transmission types, condition two (35 speakers out of 1995) produced the weakest LRs but, for the GSM condition, the lowest C_{llr}. Conditions one (35/45 speakers) and three (1400/1995 speakers) performed equally well in terms of C_{llr}. Van der Vloed *et al.* (2011) explain this result in terms of the ratio of the size of the subset to the total size of the database rather than the absolute size of the reference data (i.e. the two systems generate similar C_{llr} values because the ratio of speakers used as population data extracted from the larger database is roughly the same). This is because in condition two the 35 reference speakers will be more like the suspect and more homogeneous, since they were identified from a much larger sample. However, the choice of absolute sample sizes appears arbitrary and the results do not provide useful information in addressing how LR output is affected by monotonic increases in sample size. Further, given that these results were computed using Batvox based on CC

[6] `http://www.agnitio-corp.com/products/government/batvox` (accessed: 9th July 2014).

input and inbuilt algorithms, their transferability to other variables and LR formulae is not clear.

The only study to have investigated sample size beyond the number of reference speakers is Ishihara and Kinoshita (2012). They assessed the effect of the number of tokens per test speaker on the two components of C_{llr} (C_{llr_min} and C_{llr_cal}). C_{llr_min} is the lowest C_{llr} value achievable when the system is optimally calibrated, while C_{llr_cal} is system calibration loss (i.e. the difference between the C_{llr} and the C_{llr_min}). Input data consisted of ten tokens of the Japanese filler expression *e-* (/eː/) produced by 118 male speakers of Japanese from the CSJ. 16 MFCCs were extracted from a 20ms hamming window at the temporal midpoint of each token. MVKD LRs based on non-contemporaneous samples were computed using two, four, six, eight and ten tokens per test speaker. To assess how the inclusion of different tokens affected LR output the experiment was conducted using consecutive tokens from each sample, and by reversing the order of the tokens.

Ishihara and Kinoshita (2012) found different patterns for the two elements of C_{llr}. C_{llr_cal} increased considerably as the number of tokens per speaker increased. This had the overall effect of worsening C_{llr} as sample size increased (from around 0.5 with two tokens to 2.5 with ten tokens). This pattern was found in both forms of the experiment. However, C_{llr_min} decreased as the number of tokens increased. The magnitude of this decrease was around 0.2 (from 0.4 to 0.2). The different patterns found for C_{llr_min} and C_{llr_cal} lead Ishihara and Kinoshita to conclude that "additional data can improve the quality of LRs, as long as we calibrate the obtained LRs" and that "uncalibrated LRs can be extremely misleading" (2012: 3). Unfortunately, however, the distributions of calibrated LLRs as a function of the amount of data per test speaker were not provided.

It seems that no empirical work has yet analysed how much data per reference speaker is required to generate stable estimations of within-speaker variation for LR testing. Furthermore, no empirical work has considered the effects of the size of the development and test sets in LR-based testing. Therefore, the experiments in this thesis address these issues.

2.6 Research questions

This thesis considers the definition of the relevant population and the collection of sufficient data in LR-based testing, by addressing the following research questions:

Definition of the relevant population

1. To what extent is LR output affected by different definitions of the logically relevant population with regard to regional background?
   (a) Are different formants more robust to the effects of regional variation?
   (b) Are ASR variables more robust to the effects of regional variation than linguistic-phonetic variables?
2. To what extent is LR output affected by different definitions of the logically relevant population with regard to sources of between-speaker variation other than language and sex, specifically socio-economic class and age?
   (a) Are certain sources of between-speaker variation more important than others? How does the sensitivity of LRs based on definitions of class and age compare with the sensitivity of LRs based on regional background?
3. Are there alternative approaches to defining the relevant population other than logical relevance (§2.3.1) and lay listener-judged similarity (§2.3.2) which are more appropriate for the inherent complexity of speech data?

Collection of development, test and reference data

4. To what extent is LR output affected by the number of reference speakers used in system testing?
   (a) Are different variables affected in different ways by the number of reference speakers?
   (b) How does calibration affect the sensitivity of LR output to variation in the number of reference speakers?

5. To what extent is LR output affected by the number of tokens (i.e. amount of data) per reference speaker in system testing?
   (a) Are different variables affected in different ways by the number of tokens per reference speaker?
   (b) How does calibration affect the sensitivity of LR output to variation in the number of tokens per reference speaker?
6. To what extent is numerical LR output affected by the number of development and test speakers used in system testing?
   (a) Is LR output most sensitive to the size of the development, test or reference data?
   (b) Are there trade-offs between the number of speakers used for development, test and reference data?

As highlighted earlier, it is clear that numerical LR output is dependent on the decisions made by the analyst at all stages of a case (e.g. adding a single speaker to the reference data will necessarily change the absolute numerical value of the LR for a pair of suspect and offender samples). Therefore, the primary concern of the experiments in this thesis is the magnitude of the effects on LRs of key analytic decisions made by the analyst, and the extent to which such variation is systematic. In terms of the practical implications of the results of this thesis, by far the best outcome would be to find very little difference in LR output when varying the definition of the relevant population or sample size. However, the extent to which such factors are important is an empirical question.

In Chapter 3, the methods used throughout the experiments in this thesis are presented. The subsequent chapters present the individual experiments which test the issues of the definition of the relevant population (research questions 1-3; Chapters 4-7) and the collection of data for LR-based testing (research questions 4-6; Chapters 8-10).

Chapter 3

General Methodology

This chapter provides an overview of the methodologies used throughout this thesis. It discusses the corpora used, the principles of testing LR-based FVC systems and evaluating their performance, the general structure of the experiments, the choice of input variables and data extraction. Separate methods sections are also included in each data chapter to explain experiment-specific procedures.

3.1 Corpora

Multiple corpora were used in this thesis, each chosen to meet the specific needs of individual experiments. The justifications for the choice of datasets are outlined in the specific chapters in which they are used. This section provides an overview of their structure.

3.1.1 Dynamic Variability in Speech (DyViS)

DyViS (Nolan *et al.* 2005-2009) is a corpus designed for forensic research containing 100 male speakers of SSBE, aged between 18 and 25. All participants were students at the University of Cambridge. SSBE is described as a prestige variety spanning across the south of England (Hughes *et al.* 2005; Wells 1982).SSBE is defined linguistically by the FOOT-STRUT split and BATH-broadening (Hawkins and Midgley 2005), as

well as innovations such as GOOSE fronting (Harrington *et al.* 2008; Chládkova and Hamann 2011) and /t/ glottaling (Fabricius 2000). For an overview of SSBE see Kerswill (2006). DyViS participants were not controlled for geographical background, but were included based on self-assessment and the judgement of a DyViS researcher. Therefore, the extent to which the speakers make up a sociolinguistically "homogeneous group" (Nolan *et al.* 2009: 37) is potentially questionable.

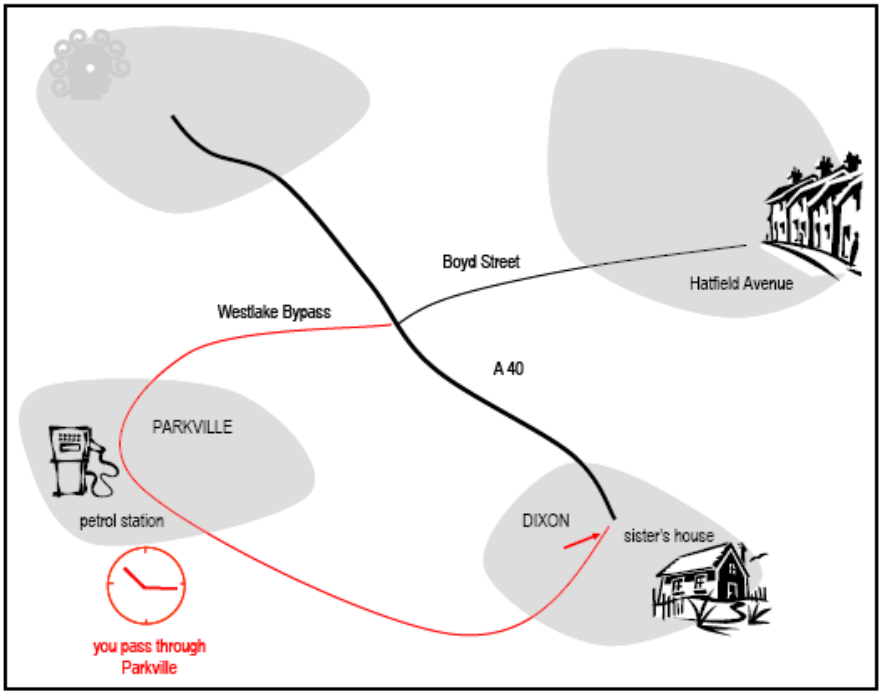

Figure 3.1: Example of a slide from DyViS Task 1 containing information about the mock suspect's story (Nolan *et al.* 2009: 42)

DyViS Task 1 involves a mock police interview, which elicited "spontaneous speech in a situation of 'cognitive conflict,' where speakers (were) made to lie" (Nolan *et al.* 2009: 41). Participants were presented with slides containing prompts (Figure 3.1) and asked to describe the information in black type, avoiding incriminating information in red. The slides displayed target words (Nolan *et al.* 2009: 53) containing segmental variables of interest. Task 1 recordings were digitised at a sampling rate of 44.1 kHz and a 16-bit depth, and are between 20 and 30 minutes in duration. An issue with Task 1 for FVC is the extent to which the speech elicited is entirely spontaneous, since target items were read from a screen. Interviewers were also careful to ensure that participants produced as many of the target items as possible. Where target words were missed

the interviewers questioned the participants to ensure elicitation or directly asked the participants to say the word.

In Task 2, participants conducted a landline telephone conversation with a mock accomplice. The accomplice was also a male SSBE speaker who had attended the University of Cambridge. The mock accomplice was chosen to elicit "a reasonably relaxed speaking style ... such as they might use when talking to a friend" (Nolan *et al.* 2009). As in Task 1, the participant recalled information from slides presented on a computer. Recordings were made directly and sampled at a rate of 44.1 kHz and 16-bit depth. Task 2 was also recorded at the opposite end of the telephone line, following typical landline band-pass filtering of between 300 and 3400 Hz (Byrne and Foulkes 2004).

3.1.2 Origins of New Zealand English (ONZE)

ONZE consists of three corpora containing speakers born between 1850 and 1987. This thesis utilises the Canterbury Corpus (CanCor) (Maclagan and Gordon 1999; Gordon *et al.* 2007) which constitutes the most up to date database of New Zealand English (NZE). CanCor contains 418 speakers born between 1935 and 1985 with almost equal numbers of younger (20-30) and older (45-60) speakers, males and females, and professionals and non-professionals. All participants were born in NZ, with the majority from Canterbury. CanCor has been collected since 1994 as part of an on-going undergraduate module at the University of Canterbury. Each student records two participants in spontaneous, sociolinguistic interviews of around 30 minutes. Given that recordings were made by different students, there is some variability in interview style and recording quality.

Phoneme-level forced-alignment (see Sjölander 2003) was performed as part of ONZE using the Hidden Markov Model Toolkit (HTK; Young *et al.* 2006), based on orthographic transcriptions for each sample. For more information on the specific details of how forced-alignment of the ONZE data was conducted see Fromont and Hay (2012). The ONZE sound files along with orthographic transcriptions, coding at different levels of representation and Meta-data about speakers are embedded within the LaBB-CAT software (Fromont and Hay 2008). LaBB-CAT is an online platform for storing and sharing large corpora. It is optimised to search for specific linguistic variables according

to different speaker groups and can be used to automatically extract large amounts of acoustic data. Due to ethics restrictions, it is not possible to access CanCor outside of the University of Canterbury. For the purposes of Chapters 4 and 8, formant data were generated automatically without access to the sound files. Fortunately, it was possible to visit the University of Canterbury to extract the data in Chapter 7.

3.1.3 Northern Englishes (NE)

The NE dataset was collected by Haddican (2008-2013). NE consists of sociolinguistic corpora from Manchester, Newcastle, Derby and York. For the purposes of the experiments in this thesis, only the Manchester and York corpora are used. The Manchester corpus contains 47 speakers from a wide age range (17-82), with roughly equal numbers of males and females and working and middle class speakers. All participants were recorded in spontaneous sociolinguistic interviews of roughly 45 minutes in duration. In the same session, speakers were recorded in spontaneous ethnographic interviews. This involved questions relating to speakers' attitudes towards their hometown, identity and accent. All of the NE recordings were digitised at a sampling rate of 44.1 kHz and a 16-bit depth

The York corpus consists of eight males and ten females aged between 18 and 22. Participants were recorded performing the same tasks as the Manchester corpus. In Haddican (2008-2013), the 2008 York recordings were combined with an older corpus recorded in 1998 for the Roots of Identity (RoI) project (1996-1998), to generate a real-time dataset of York English. RoI consists of 32 speakers divided equally between males and females and older (59-78) and younger (17-31) speakers. Although no explicit control was made over speakers' socio-economic class, Haddican *et al.* state that "speakers were all judged to be from the upper working or lower middle class" (2013: 376). RoI consists of spontaneous sociolinguistic interviews of c. 45 minutes. RoI recordings were digitised at a rate of 22.05 kHz.

3.1.4 Phonological Variation and Change (PVC)

The PVC corpus consists of recordings from Newcastle and Derby (Milroy, Milroy, and Docherty 1994-1997; for an overview see Milroy *et al.* 1999). The dataset contains 64 speakers, divided equally between Newcastle and Derby, older (45-65) and younger (16-25), and working and middle class speakers. Each participant was recorded in casual conversation with a peer group member and sessions lasted between 48 and 64 minutes. Recordings were digitised at a sampling rate of 16 kHz and a 16-bit depth.

3.1.5 TIMIT

The TIMIT Acoustic-Phonetic Continuous Speech Corpus, released by NIST in 1990, was designed by SRI International, Texas Instruments (TI) and the Massachusetts Institute of Technology (MIT). TIMIT contains 630 speakers (438 males) aged between 21 and 65 (mean = 31) from seven major dialect regions (DRs) (Figure 3.2) within the United States and an eighth group named *Army Brats* who moved geographical location during childhood. According to Garofolo *et al.* (1993), a speaker's DR was defined as "the geographical area in the US where he or she had lived during their childhood years (age 2 to 10)" (1993: 15). The number of speakers in each DR is shown in Table 3.1.

Speakers were recorded in a noise-isolated sound booth reading a set of ten target sentences. Target sentences were defined as: (i) dialect sentences (SA) designed to "expose ... dialectal variants" (TIMIT 1990: 2), (ii) phonetically compact sentences (SX) to provide a "good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest" (TIMIT 1990: 2), and (iii) phonetically-diverse sentences (SI) taken from the Brown Corpus (Kuchera and Francis 1967) or the Playwrights Dialog (Hultzen *et al.* 1964). Each speaker read both SA sentences, five random SX sentences and three random SI sentences. Two-channel recordings were made using a head-mounted microphone and a far field pressure microphone, although only the recordings made using the head-mounted microphone were included in the released version of TIMIT. Samples were initially digitised at a sampling rate of 20 kHz and then downsampled to 16 kHz (Fisher *et al.* 1986).

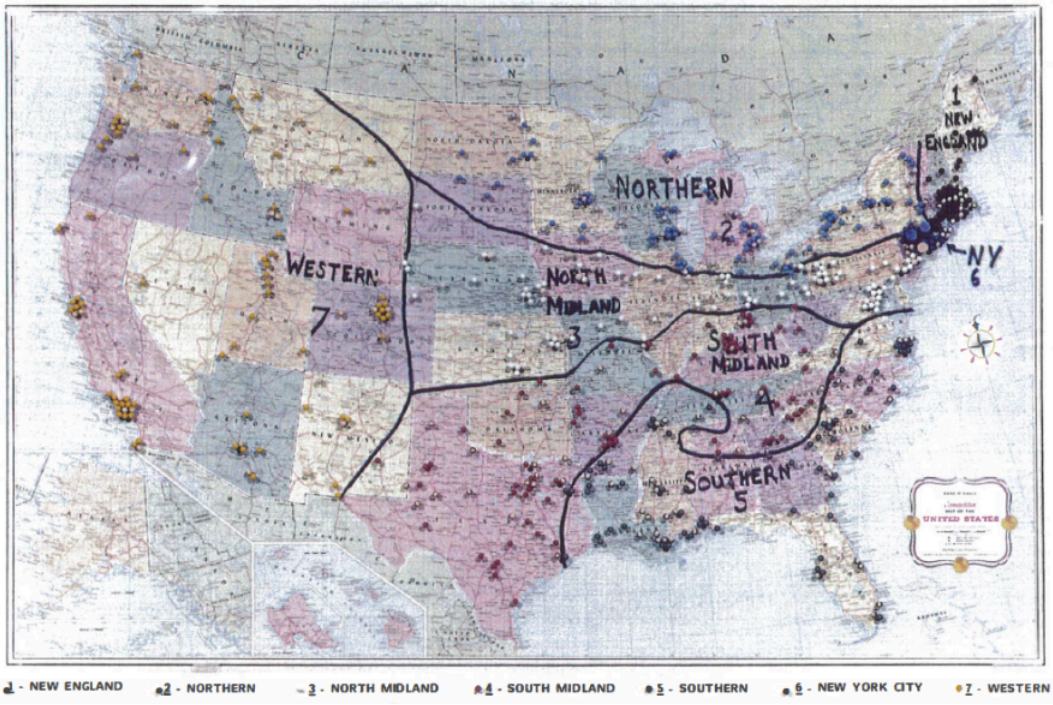

Figure 3.2: Map of USA with TIMIT dialect regions marked (from Garofolo *et al.* 1993: 17)

TIMIT has been used widely in linguistic research to investigate acoustic-phonetic variation between speakers (Byrd 1992; Sun and Deng 1995), the performance of phoneme (and speech) recognition systems (Kapadia *et al.* 1993), the performance of speech segmentation and labelling systems (Ljolje and Riley 1991), formant measurement errors (Harrison 2013), and the performance of speaker recognition systems (Reynolds 1995; Reynolds *et al.* 1995).

3.2 LR testing

This section provides a detailed overview of the principles involved in LR-based system testing and the specific procedures used in this thesis. The term *system* is used in its broadest sense to refer to "a set of procedures and databases that are used to compare two samples, one of known origin and one of questioned origin, and produce a (LR)" (Morrison 2013: 174). This use of the term system is not to be confused with the much narrower use of the term in ASR referring to a stand-alone piece of commercial software.

Table 3.1: Number of male speakers in each of the dialect regions (DRs) in the TIMIT corpus (from Garofolo *et al.* 1993: 16)

Dialect Region (Number)	N **Males**	**Total** N **Speakers**
New England (1)	31 (63%)	49 (8%)
Northern (2)	71 (70%)	102 (16%)
North Midland (3)	79 (67%)	102 (16%)
South Midland (4)	69 (69%)	100 (16%)
Southern (5)	62 (63%)	98 (16%)
New York City (6)	30 (65%)	46 (7%)
Western (7)	74 (74%)	100 (16%)
Army Brat (8)	22 (67%)	33 (5%)
TOTAL	438 (70%)	630

3.2.1 Development, test and reference data

The first stage of LR testing is the feature-to-score stage. This involves the computation of LRs (called *scores*; Morrison 2013) from multiple pairs of samples where it is known, *a priori*, whether they were produced by the same (SS) or different (DS) speaker(s). The feature-to-score stage is initially implemented to compute scores for a set of SS and DS pairs called the test data, which function as mock suspect and offender samples matching the facts of the case at trial. The typicality element of LR computation is based on reference data containing representative speakers from the relevant population. The scores for the test data can be used to assess validity and reliability of the system (§3.2.3).

Preferably, the second stage of testing involves score-to-LR mapping (referred to as calibration; Morrison 2013). Calibration is a means of optimising system performance (see §3.2.4). To calibrate, it is necessary to have a set of development (or training) data, consisting of speakers who are representative of the relevant population and samples which reflect the facts of the case at trial. The feature-to-score stage is implemented to compute scores for the development data and these scores are used to train a calibration model. The model is then applied to the scores for the test data to generate calibrated LRs from which system validity and reliability can be calculated.

3.2.2 Feature-to-score stage

3.2.2.1 Modelling

Both the Multivariate Kernel Density (MVKD; Aitken and Lucy 2004) and Gaussian Mixture Model - Universal Background Model (GMM-UBM; Reynolds *et al.* 2000) approaches were used during the feature-to-score stages of the experiments in this thesis.

MVKD

The numerator of the MVKD LR is given as:

$$f_0(\bar{y}_1, \bar{y}_2|U, C) = \tag{3.1}$$

$$(2\pi)^{-p}|D_1|^{-\frac{1}{2}}|D_2|^{-\frac{1}{2}}|C|^{-\frac{1}{2}}$$

$$(mh^p)^{-1}|D_1^{-1} + D_2^{-1} + (h^2C)^{-1}|^{-\frac{1}{2}}$$

$$exp\left\{-\frac{1}{2}(\bar{y}_1 - \bar{y}_2)^T(D_1 + D_2)^{-1}(\bar{y}_1 - \bar{y}_2)\right\}$$

$$\sum_{i=1}^{m} exp\left\{-\frac{1}{2}(y^* - \bar{x}_i)^T((D_1^{-1} + D_2^{-1})^{-1} + (h^2C))^{-1}(y^* - \bar{x}_i)\right\}$$

and the denominator is given as:

$$f_1(\bar{y}_1, \bar{y}_2|U, C) = \tag{3.2}$$

$$(2\pi)^{-p}|C|^{-1}(mh^p)^{-2}$$

$$\prod_{i=1}^{2}[|D_l|^{-\frac{1}{2}}|D_l^{-1} + (h^2C)^{-1}|^{-\frac{1}{2}}$$

$$\sum_{i=1}^{m} exp\left\{-\frac{1}{2}(\bar{y}_l - \bar{x}_i)^T(D_l + h^2C)^{-1}(\bar{y}_l - \bar{x}_i)\right\}]$$

from Aitken and Lucy (2004: 116-117)

where:

U, C = within-, between-speaker variance/covariance matrices

n_1, n_2 = number of replicates per speaker

m = number of speakers in reference data

p = number of assumed correlated variables per speaker

D_l = D_1, D_2 = offender, suspect variance/covariance matrices = $n_1^{-1}U, n_s^{-1}U$

h = optimal smoothing parameter for KD = $(4/2p+1))^{1/(p+4)}m^{-1/(p+4)}$

$\bar{y}_l$ = $\bar{y}_1, \bar{y}_2$ = offender, suspect mean vector

from Rose (2013a: 94)

The MVKD LR is the ratio of Equations 3.1 and 3.2. Using this approach, the suspect data are modelled using a Gaussian distribution while the background data are modelled using speaker-dependent Gaussian KD estimation. The reference model is speaker-dependent meaning that it is generated using equally-weighted Gaussians for each reference speaker based on the mean and variance of their values. Equation 3.1 is equivalent to the probability of the offender value at the intersection of the suspect model and Equation 3.2 is the probability of the offender value at the intersection of the reference model. Where there are multiple offender values, the mean of the LRs is taken (Morrison 2010) to give a single LR for the offender data.

MVKD is generally preferred for analysing linguistic-phonetic variables, with Morrison claiming that it is "considered the standard procedure ... in acoustic-phonetic (FVC)" (2011a: 244). This is because it is suited to small amounts of data per speaker where the distributions of speakers' values are normally distributed (Jessen and Enzinger 2014). MVKD is also preferable for multivariate data which consist of a relatively small number of correlated features (Nair *et al.* 2014). As stated by Rose (2013a: 95), correlations are handled through the variance-covariance matrices between-speakers (C) and the inversion of the within-speaker variance-covariance matrices for the suspect and offender data (D_1,D_2), which "contribute towards the decorrelation of the individual features ... and the equalisation of their contribution" (Khodai-Joopari 2006: 145). Figure 3.3 displays a visualisation of MVKD based on bivariate suspect and reference distributions of F1 and F2 values for /uː/ from the same data as in Figure 2.2.

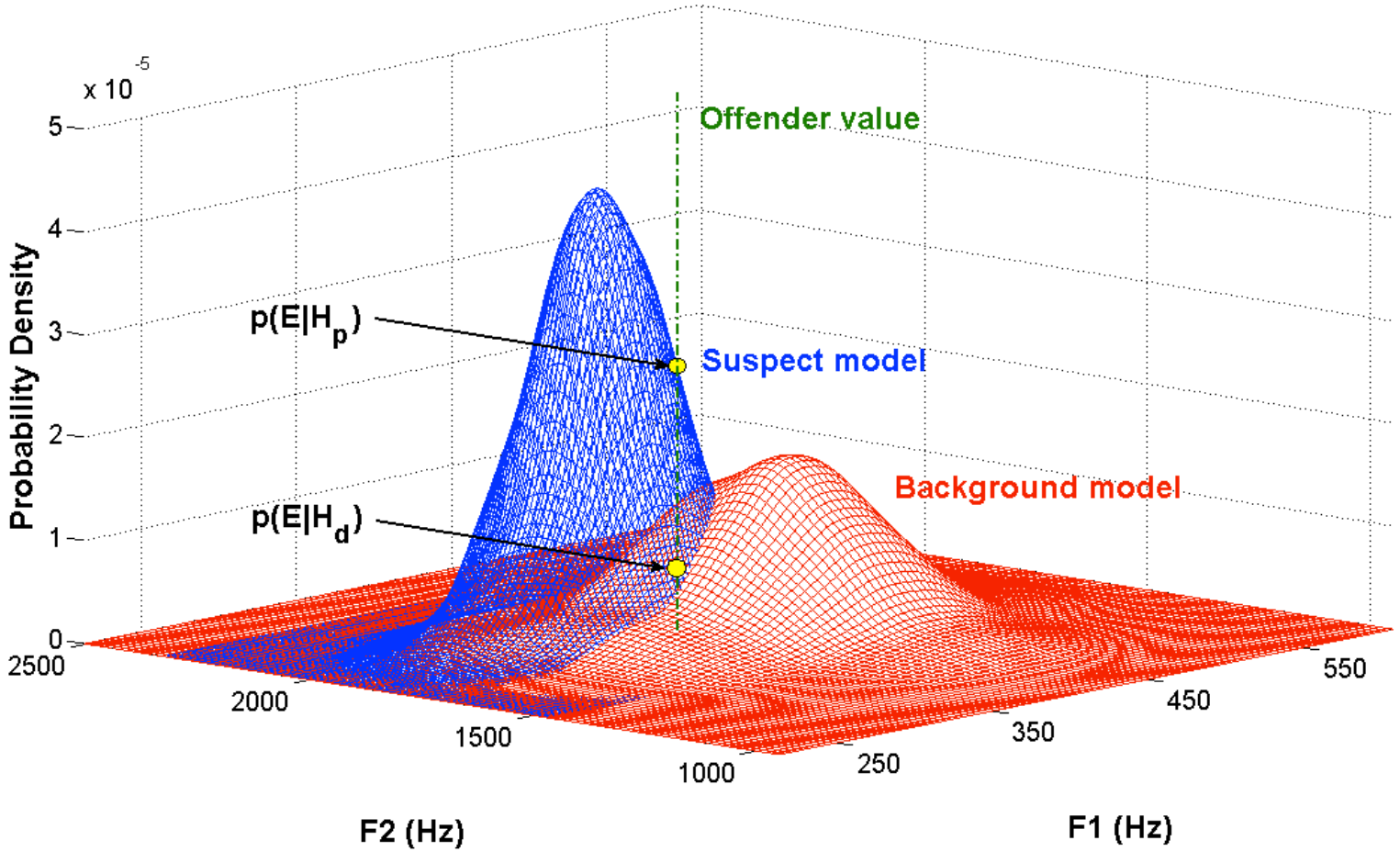

Figure 3.3: Bivariate example of a MVKD LR computed for a same speaker comparison based on midpoint F1 and F2 (Hz) values using data from ONZE

With the exception of Chapter 7, MVKD LRs were computed using a MATLAB implementation[7] of Aitken and Lucy's (2004) formula. Multiple SS and DS MVKD LRs were computed using a loop script.[8] The script performs LR comparisons based on contemporaneous samples by dividing the data for each speaker in half to create suspect (1st half) and offender (2nd half) data. This allows for a single SS comparison per speaker and two DS comparisons per speaker pair (1_{sus} vs. 2_{off} and 2_{sus} vs. 1_{off}). In Chapter 7, MVKD LRs were computed in R using the *Comparison* package.[9] The same format for testing was implemented in Chapter 7 as in the other chapters, in which contemporaneous samples were divided in half to allow for SS comparisons. The limitations of using contemporaneous samples for computing LRs are outlined in §3.4.

GMM-UBM

An alternative to MVKD for multivariate data is the GMM-UBM approach. The formula for computing a GMM-UBM natural log LR (see §3.2.2.4 score is:

$$s = \frac{1}{T}\sum_{i=1}^{T} log\left(p(x_i|\lambda_{sus})\right) - log\left(p(x_i|\lambda_{bkd})\right) \tag{3.3a}$$

$$p(x_i|\lambda) = \sum_{i=1}^{M} \frac{w_i}{(2\pi)^{\frac{D}{2}}|\Sigma_i|^{\frac{1}{2}}} \times exp\left(-\frac{1}{2}(x_t-\mu_i)'(\Sigma_i)^{-1}(x_t-\mu_i)\right) \tag{3.3b}$$

$$\sum_{i=1}^{M} w_i = 1 \tag{3.3c}$$

where:

s = Score

D = Number of variables measured for each token (i.e. dimensions)

x_t = $D \times 1$ vectors of measurements offender (unknown) data, where T is the number of tokens

$\lambda_{sus}, \lambda_{bkg}$ = Suspect, background models

[7] Morrison, G. S. (2007). MATLAB implementation of Aitken and Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation (2007). `http://geoff-morrison.net/#MVKD` (accessed: 31st May 2011).

[8] 'ss_ds_lr_loop.m' written by Philip Harrison (2011).

[9] Lucy, D. (2013). Comparison (version 1.0-4) (R package). `http://cran.r-project.org/web/packages/comparison/index.html` (accessed: 8th August 2014).

M = Number of Gaussians per GMM parameterised by a $D \times 1$ mean vector μ_i, a $D \times D$ covariance matrix Σ_i, and a scalar weight w_i

from Morrison (2011a: 244-245)

A benefit of GMM-UBM is that it is capable of modelling non-normally distributed suspect and reference data. Unlike MVKD, the GMM background model (UBM) is speaker independent in that it consists of pooled data from all reference speakers. The UBM is a GMM trained using the expectation maximisation algorithm (Duda *et al.* 2000). In this thesis, GMM suspect models are constructed using raw suspect data (rather than MAP adaptation; see Morrison 2011a), following the same procedure as the UBM (examples of this approach are in Becker *et al.* 2008 and Becker *et al.* 2009). The number of Gaussians in a GMM is dependent on the amount of data and its multidimensionality. A visualisation of suspect and background GMMs using four Gaussians per model for hypothetical f0 data is in Figure 3.4 (Morrison 2010: 28).

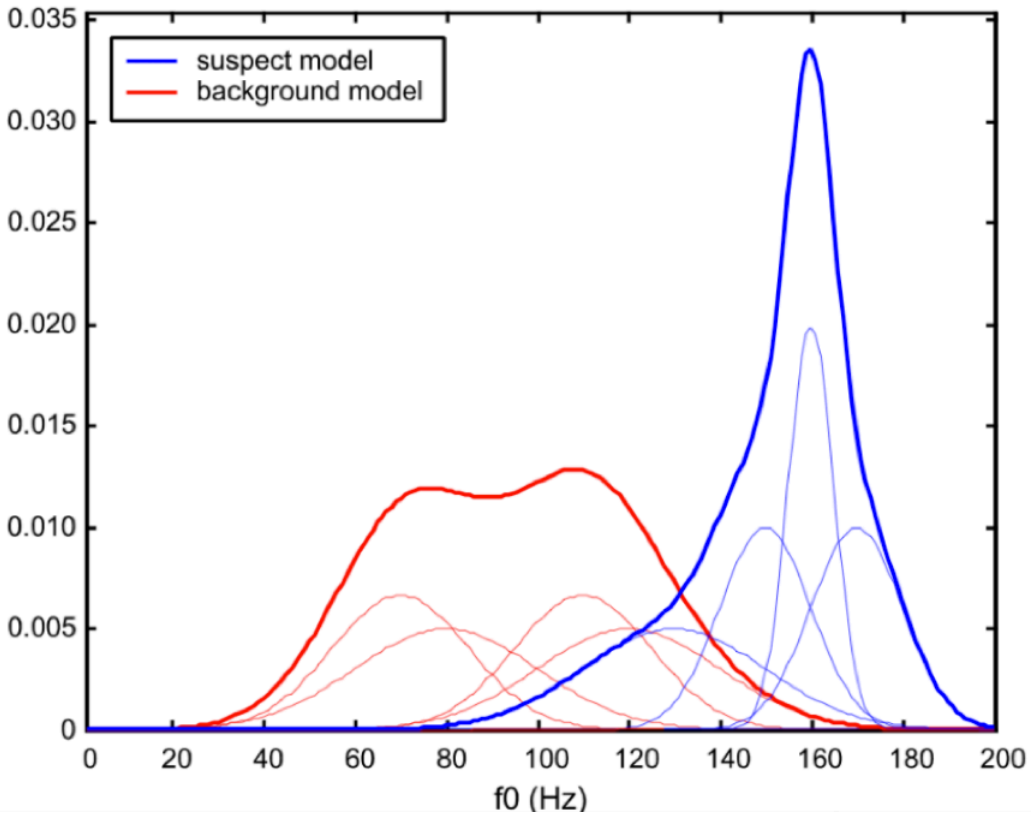

Figure 3.4: GMMs of hypothetical suspect and reference data for f0 constructed using four Gaussians per model (from Morrison 2010: 28)

For each offender value within each feature (x_i), a LR score (s) is computed as the probability density of the value at the intersection of the suspect model $p(x_i|\lambda_{sus})$ divided by probability density of the value at the intersection of the background model $p(x_i|\lambda_{bkg})$. A single score per feature is calculated by taking the mean of the scores from each offender value for that feature (Equation 3.3). The OLR score is the mean of the scores across all features of the variable. GMM-UBM is typically used in ASR research (Alexander and Drygajlo 2004b; González-Rodríguez *et al.* 2006; Ramos-Castro 2007) since the speaker independent UBM is suited to large amounts of data (Jessen and Enzinger 2014). In this thesis, GMM-UBM scores were computed using MATLAB functions from the Speaker Recognition Project (see Alexander and Drygajlo 2004a). Since contemporaneous samples were used when computing GMM-UBM LRs, comparisons were conducted by dividing each speaker's data in half (as above).

3.2.2.2 Intrinsic vs. extrinsic testing

Intrinsic LR testing involves a single database from which all speakers are extracted, whereas extrinsic testing uses separate databases for the development, test and reference sets. Intrinsic testing is more common, due to the paucity of usable datasets which are sufficiently well matched for extrinsic testing (an exception is Rose *et al.* 2006). System performance using intrinsic testing is generally expected to be better than that using extrinsic testing. This is because the use of a single corpus increases the homogeneity of the datasets in terms of the speakers used and the recording conditions. Extrinsic testing is also considered more forensically realistic because in casework evidential samples would not come from the database used to build the FVC system. A combination of intrinsic and extrinsic testing is conducted in this thesis (see §3.4).

3.2.2.3 Independent sets vs. cross-validation

To compute meaningful LRs, it is important that the development, test and reference sets contain different speakers. This is because $p(\mathrm{E}|\mathrm{H}_d)$ refers to the probability of the evidence assuming it was produced by a random member of the relevant population other than the suspect. To include the suspect in the reference data would therefore be

logically incorrect. The easiest way to deal with this issue is to use separate datasets. However, this requires a large database of speakers. Cross-validation (Hastie *et al.* 2009: 241-249) allows speakers to function simultaneously as comparison and reference data in the feature-to-score stage such that for each speaker pair the reference data consists of all speakers excluding the suspect and offender. Therefore, using cross-validation, the reference data changes for each comparison. Applying cross-validation in this thesis, for DS pairs both speakers were excluded from the reference data. For SS pairs, the target speaker and another random speaker were excluded from the reference data. This ensures that the number of reference speakers remains constant across comparisons.

Cross-validation can also be used to calibrate (§3.2.4) a set of SS and DS scores. This is done by generating calibration coefficients for each comparison pair based on the scores from all pairs except those involving the speakers being compared. In this way, the calibration coefficients change for each comparison. While not ideal, this approach is useful where the number of comparisons is small. This approach was used in §10.2.1.

3.2.2.4 Log likelihood ratios (LLRs)

The distributions of raw LRs from FVC systems are often skewed. To account for this LRs are converted to log LRs using a base x logarithm such that:

$$LLR = log_x(LR) \tag{3.4}$$

$$LR = x^{LLR}$$

The value for x is typically either 10 ($\log_{10}$ or base 10) or e ($\approx$2.71828 natural log or base e). The relationship between raw LRs, $\log_{10}$ LRs and natural log LRs is shown in Table 3.2. Where raw LRs take values between zero and ∞ with threshold at one, log LRs take values between $-\infty$ and ∞ with zero as threshold. This means the log scales are symmetrical either side of zero, improving the interpretability of the relative weight of the evidence (Lempert 1977; Edwards 1986).

In this thesis, $\log_{10}$ LR values were used to interpret strength of evidence. The abbreviation LLR is therefore used to refer to $\log_{10}$ LRs (unless otherwise stated). Given that the distributions of LLRs may also be skewed their central tendency is described using the median.

Table 3.2: Raw values with base 10 and base e logarithm values

Raw value	**Base 10 log value**	**Base e log value**
1000	3	6.9078
100	2	4.6052
10	1	2.3026
1	0	0
0.1	-1	-2.3026
0.01	-2	-4.6052
0.001	-3	-6.9078

3.2.2.5 Tippett plots

Throughout this thesis the distributions of LRs are presented using Tippett plots (Meuwly 2001; see Morrison 2011a: appendix 2). Figure 3.5 displays a Tippett plot based on hypothetical LRs produced by a FVC system. The solid line represents SS LRs and the dashed line represents DS LRs. The x-axis displays the $\log_{10}$ LR value with zero marking the threshold between support for the prosecution (positive values) and support for the defence (negative values). The y-axis displays the cumulative proportion (or percentage) of comparisons that achieve a value less than (for SS)/ greater than (for DS) or equal to the value on the x-axis. For example, based on Figure 3.5 approximately 60% of SS pairs achieve a LLR of less than +2. By extension, around 40% of SS pairs achieve a LLR of more than +2. The further the SS line to the right and the further the DS line to the left the stronger LRs.

The Tippett plot also provides information about the validity of the system. The point on the y-axis at which the lines cross is the equal error rate (EER) (§3.2.3.1). The proportion of misses (i.e. SS pairs offering support for the defence) occurs at the point on the y-axis where the SS line crosses the zero threshold into negative values. Conversely the proportion of false hits (i.e. DS pairs offering support for the prosecution) occurs at the point on the y-axis where the DS line crosses zero into positive values. The magnitude of the contrary-to-fact LRs is determined by how far the SS line extends into negative values and how far the DS line extends into positive values.

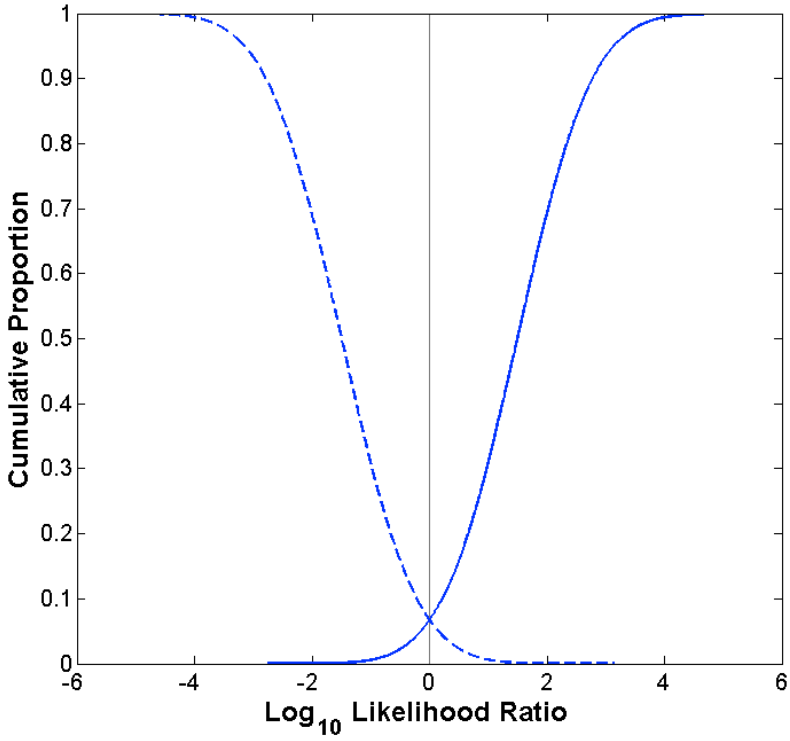

Figure 3.5: Tippett plot based on hypothetical SS and DS LRs produced by a FVC system

3.2.2.6 Verbal LRs

Whilst Lucy claims that "a (LR) is an ... easily interpretable quantity which expresses the persuasive power of evidence" (2005: 133), the extent to which triers-of-fact are able to comprehend numerical estimates of strength of evidence is a significant concern for the courts. Therefore, numerical LRs may be converted into a verbal expression. Table 3.3 shows the scale proposed in Champod and Evett (2000: 240). However, since the meaning of such scales "can vary both between and within the several interested groups" (Rose 2002: 62), the verbal outcome is claimed to be arbitrary (Buckleton *et al.* 2005). Further, categorical distinctions impose *cliff-edge* effects whereby the difference between two LLRs of 9.9 and 10 is equivalent to the difference between *limited* and *moderate* support for the prosecution (Table 3.3). The difficulties in the interpretation of verbal scales for the trier-of-fact are highlighted in Mullen *et al.* (2013) and Martire *et al.* (2014).

Table 3.3: Verbal expressions of raw and LLRs according to Champod and Evett's (2000: 240) scale

LLR	Verbal expression
$\pm4 : \pm5$	*Very strong* support
$\pm3 : \pm4$	*Strong* support
$\pm2 : \pm3$	*Moderately strong* support
$\pm1 : \pm2$	*Moderate* support
$0 : \pm1$	*Limited* support

Nonetheless, in this thesis, Table 3.3 was used as a means of contextualising numerical values, and for broad cross-comparison of the LR output from different systems.

3.2.3 System performance: validity and reliability

The performance of a forensic comparison system can be analysed in terms of validity (or accuracy) and reliability (or precision). Validity refers to how well a system performs the task it is claimed to do. In the case of FVC systems, this is to discriminate between SS and DS pairs and generate low magnitude contrary-to-fact values (i.e. SS LLRs < 0 and DS LLRs > 0). Reliability refers to the degree of variability (or imprecision) in LR estimates from the same comparisons (i.e. how close the observed LR is to the mean). Figure 3.6 displays a visual representation of validity and reliability from Morrison (2011b: 92) based on the proximity of a distribution of measurements of a given object to the true value of that object.

3.2.3.1 Validity

Equal error rate (EER)

EER is a metric for assessing categorical system validity using the LR as a discriminant function. The categorical decisions made by a LR-based system are defined in Table 3.4. In calculating EER, the percentages of false hits and misses are assessed using a series of thresholds between the minimum and maximum LLRs. EER is the percentage at

which the proportion of false hits and misses is equal. In this thesis, EER was calculated using a MATLAB function[10] which tests for false hits and misses with 2000 thresholds.

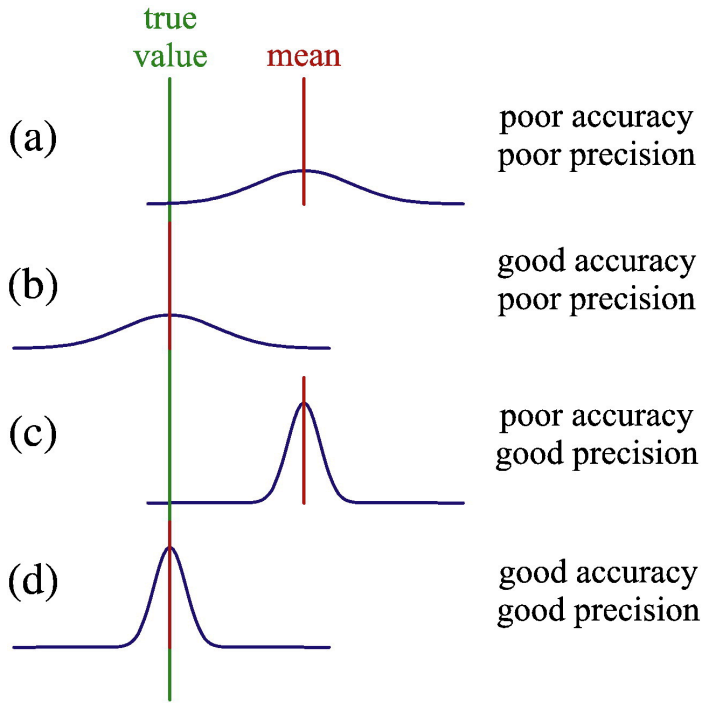

Figure 3.6: Visual representation of validity (accuracy) and reliability (precision) (from Morrison 2011b: 92)

Table 3.4: Categorical *correct* (consistent-with-fact) and *incorrect* (contrary-to-fact) decisions made by a LR-based biometric system (equivalent to that in Morrison 2011b: 93)

	SS comparison	DS comparison
LLR > 0	✓	*false hit*
LLR < 0	*miss*	✓

There are two primary limitations of EER. Firstly, the threshold for EER is often not zero. Therefore, sets of non-overlapping SS and DS LLRs with an EER of 0% may still have a high proportion of one type of *error* when using zero as threshold. Secondly, EER considers each contrary-to-fact LR as an *error* irrespective of its magnitude. Therefore, EER fails to capture the fact that a system which produces high magnitude contrary-to-fact LRs is worse than a system which produces low magnitude contrary-to-fact LRs, even if the absolute percentage of *errors* is the same.

[10]Ketabdar, H. (2004). 'jEER_DET.m' (version 1.2 with amendments by Anil Alexander).

Log LR cost function (C_{llr})

An alternative to categorical validity is C_{llr}. C_{llr} was developed for ASR (Brümmer and du Preez 2006; van Leeuwen and Brümmer 2007) but has since been used extensively in linguistic-phonetic FVC (Morrison 2009b; Morrison 2011b; Rose 2010; Morrison and Kinoshita 2008). C_{llr} penalises the system based on the magnitude, rather than the proportion, of contrary-to-fact LLRs (Rose and Winter 2010). The assessment of the "gradient goodness of a set of LRs" (Morrison 2009c: 6) provides a logically appropriate means of analysing system validity.

It is defined as:

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{ss}}\sum_{i=1}^{N_{ss}} log_2\left(1+\frac{1}{LR_{ss_i}}\right) + \frac{1}{N_{ds}}\sum_{i=1}^{N_{ds}} log_2\left(1+\frac{1}{LR_{ds_i}}\right)\right) \quad (3.5)$$

where:

N_{ss} = Number of SS comparisons
N_{ds} = Number of DS comparisons
LR_{ss} = SS LR
LR_{ds} = DS LR

from González-Rodríguez *et al.* (2007)

By assessing the magnitude of contrary-to-fact LRs, there is an assumption that not all *errors* are equally problematic for system performance. That is, contrary-to-fact LRs closer to zero are preferred to contrary-to-facts LRs of a higher magnitude. The closer C_{llr} is to zero the better the performance. Values approaching one (*unity*) reflect bad system validity, whilst values of above one indicate very bad performance (van Leeuwen and Brümmer 2007: 343-344). A system which produces LRs of one (LLR = zero), irrespective of the input, will produce a C_{llr} of one, and so offers no useful information for the purposes of FVC. However, the interpretation of the C_{llr} is difficult, since its value does not correspond directly to system decisions (unlike EER, where interpretation is clear). Therefore, the power of C_{llr} lies in comparing the validity of multiple systems. In this thesis, C_{llr} was calculated using MATLAB functions from Brümmer's FoCal toolkit.[11]

[11]FoCal toolkit: `https://sites.google.com/site/nikobrummer/focal` (accessed:

3.2.3.2 Reliability

Credible intervals (CIs)

The most common approach for assessing the reliability of FVC systems is to use 95% Credible intervals (CIs). The CI is the Bayesian equivalent of the frequentist confidence interval, which is "philosophically consistent with the (LR) framework" (Morrison 2011b: 95). CIs are typically used to capture the imprecision across multiple non-contemporaneous comparisons of the same comparison pairs using the same system. However, in this thesis CIs are used to estimate the imprecision of LRs from the same test comparisons across systems based on different definitions of the relevant population (Chapters 6 and 7). They are then compared across experiments to assess which input variables and logically relevant factors generate the greatest imprecision.

In this thesis, CIs were calculated using the non-parametric procedure in Morrison, Thiruvaran and Epps (2010), which assumes unequal variance across LRs from the same comparisons (for an alternative approach see Morrison *et al.* 2011). Using this approach, the mean LLR value ($\bar{x}_i$) is calculated for each SS and DS comparison across conditions by:

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \tag{3.6}$$

where:

i = Specific SS or DS pair

n_i = Number of LLRs per pair (up to eight in Chapter 6)

x_{ij} = j^{th} LLR for comparison pair i

3rd June 2011).

The deviation of each LLR (y_{ij}) from the mean for a given comparison is then:

$$y_{ij} = x_{ij} - \bar{x}_i \tag{3.7}$$

from Morrison, Thiruvaran and Epps (2010: 65)

The CI for each comparison is calculated using local linear regression with a nearest neighbour kernel, following the eight-stage procedure in Morrison, Thiruvaran and Epps (2010: §2.2). The 95% CI is a probabilistic region of a posterior distribution within which one can be 95% certain the true value is found, where the wider the CI the greater the imprecision in the LLR estimate. The overall reliability of a FVC system is analysed using the mean of the 95% CIs. The mean 95% CI is the average $\pm$ difference between the upper and lower bounds of the CI and the mean value for a given comparison. Its value can be interpreted in terms of $\log_{10}$ magnitude.

3.2.4 Score-to-LR mapping

As outlined by Morrison (2013), scores are interpretable as comparative estimates of strength of evidence (i.e. a larger score provides stronger evidence), but "the absolute values of scores are, in general, not interpretable as … (LRs)" (p. 174). This is because "all models are wrong and should be recalibrated empirically" (Neumann *et al.* 2012: 410). Calibration uses knowledge of how the system performs using development data to update, and ultimately improve, performance in testing. Grigoras *et al.* (2013) claim that "calibration can ameliorate what would otherwise be very misleading results, and … it is essential if one wishes to interpret system output as (LRs)" (p. 620).

3.2.4.1 Logistic regression calibration

Logistic regression (Brümmer and du Preez 2006) is an approach for calibrating FVC systems, which minimises the C_{llr} and also typically minimises the magnitude of contrary-to-fact log scores as a result. A visual representation of the procedure is displayed in Figure 3.7.

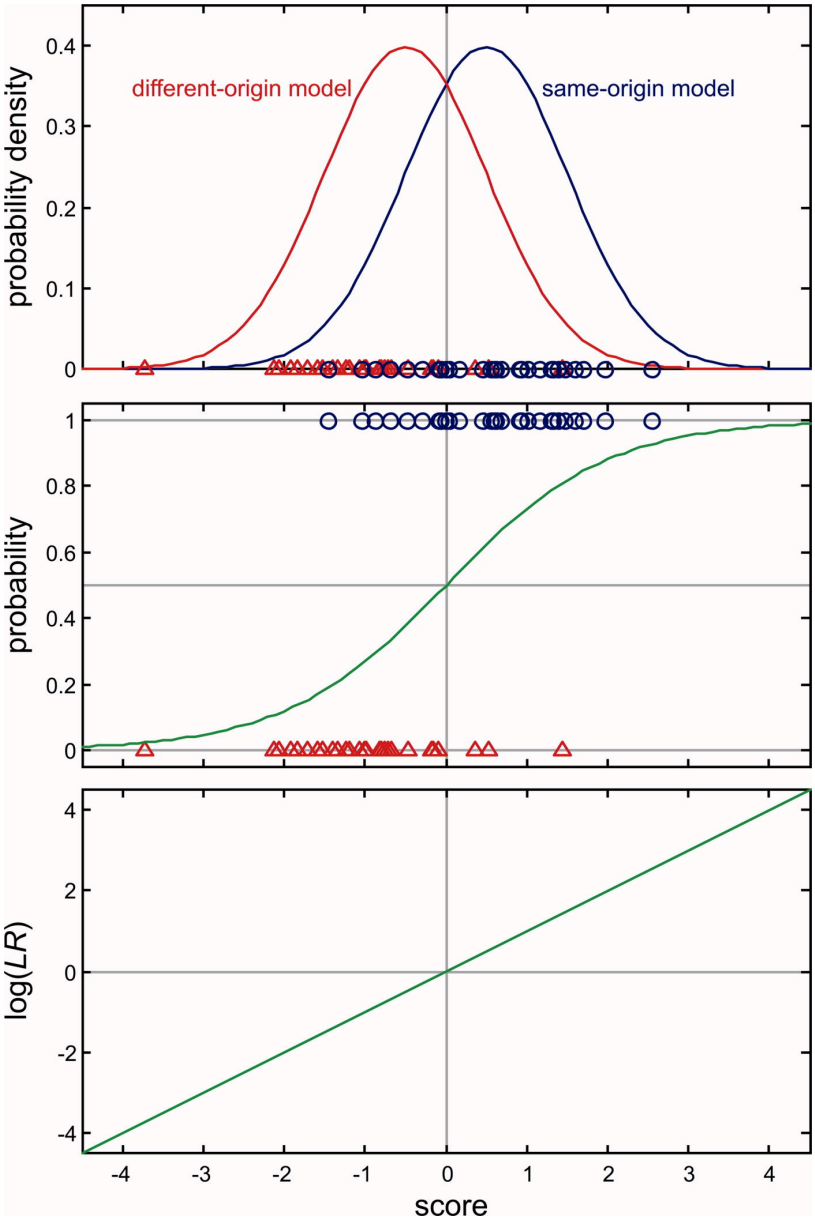

Figure 3.7: Visual representation of logistic regression calibration involving modelling of SS (red) and DS (blue) scores for a set of development data with Gaussian curves (panel 1), with a probability curve (panel 2) and the linear relationship between the score and the LLR in the log-odds space (panel 3) (from Morrison 2013: 182)

As explained in Morrison (2013), the distributions of SS and DS log scores (panel 1) from a set of development data are transformed based on the modelled probability of the samples coming from a SS pair given the score $p(\mathrm{H}_p|s)$ (panel 2). This involves coding scores from SS comparisons as one and scores from DS comparisons as zero. The probability space (panel 2) is then mapped to the logged odds space (panel 3) using

the equation:

$$log\left(\frac{p(H_p|s)}{p(H_d|s)}\right) = log\left(\frac{p(H_p|s)}{1-p(H_p|s)}\right) \tag{3.8}$$

Morrison (2013: 181)

where:

s = Score

H_p = Prosecution proposition (same-speaker)

H_d = Defence proposition (different-speakers)

The logistic regression model is defined by the linear relationship between the score and the LLR:

$$LLR = as + b \tag{3.9}$$

where:

LLR = Calibrated LLR

s = Score

Once trained, the logistic regression model in the log-odds space is applied to the scores for the test data to convert them into calibrated LLRs using the linear equation in 3.9. The linear term is the calibration scale value (a) which is multiplied to the score (s) and the intercept (b) is the calibration shift value which is added to the product of the score and the scale value.

In this thesis, calibration coefficients were calculated using a robust implementation[12] of the logistic regression procedure from Brümmer's (2007) FoCal toolkit.[11] The results in §4.3.1 are analysed using uncalibrated scores, since the number of test speakers (eight per set) is considered insufficient for meaningful calibration.

3.3 Input variables

This section describes the input variables used in this thesis: namely formant dynamics and cepstral coefficients (CCs) and derivatives. Although AR data were also used,

[12]Morrison, G. S. (2009). 'train_llr_fusion_robust.m' `http://geoff-morrison.net/#TrainFus` (accessed: 14th December 2011).

the justification for using AR and the procedures for data extraction are explained in Chapter 9.

3.3.1 Formant dynamics

Formants are resonant frequencies which characterise any sounds with a phonatory source filtered by the vocal tract, but are most commonly analysed in the context of sonorants such as vowels, liquids and nasals. Formants are defined as high amplitude harmonic peaks in the spectrum and can be seen in the spectrogram as a series of bands spread across the frequency range. Source-filter theory (Fant 1960) assumes that formant frequencies are a consequence of the interaction between the configuration of the vocal tract from the larynx to the lips (determined by dynamic articulators such as the tongue) and the overall physiology of the tract itself (e.g. vocal tract length). This accounts for the intrinsic relationship between articulatory configuration and acoustic output. For vowels, F1 is correlated with the open-close dimension, while F2 is related to the front-back dimension, although the configuration of other articulators also affects vowel acoustics (e.g. lowering of F2 and F3 during lip rounding; Stevens 2000).

Nolan (1983) describes speech as a series of phonetic targets which are the result of an interaction between communicative intent, phonological representation and physical implementation. Phonetic targets are perceived by listeners in decoding a speaker's communicative intent. In the analysis of vowel targets in mainstream phonetics and sociophonetics, the acoustic structure of monophthongs has traditionally been defined by a single formant measurement at the steady-state, approximately at the midpoint, to minimise the coarticulatory effects of adjacent sounds (Daniloff and Hammarberg 1973: 239). As dynamic events with movement between two phonetic targets, the phonetic quality of diphthongs is often captured using the dual-target model (Morrison and Assmann 2013) involving two measurements from the steady-states of the onset and offset targets.

However, an alternative to the analysis of vowels based on phonetic targets is to extract multiple measurements across formant trajectories. This is often referred to as the dynamic approach. The dynamic approach captures considerably more phonetic detail

in vowel production and a number of studies have shown that it outperforms steady-state analyses based on midpoints of monophthongs or the dual-target characterisation of diphthongs in speaker discrimination (Greisbach *et al.* 1995; Ingram *et al.* 1996; Rodman *et al.* 2002; Eriksson *et al.* 2004a, 2004b; McDougall 2004, 2005, 2006). Nolan claims that the dynamics of speech are useful for FVC because they capture information which is acquired individually through "trial and error" (1997: 749), whereas phonetic targets are learned as part of shared knowledge of sociolinguistically homogeneous speakers. However, Koops (2010) and Hughes *et al.* (2011) offer evidence to suggest that vowel dynamics may also encode socio-indexical information such as regional background, age and sex.

3.3.1.1 Speaker discrimination and individual formants

Individual formants are also expected to display different patterns in terms of the information which they encode. The *speech-speaker* dichotomy (Mokhtari 1998) refers to the two broad types of information encoded within the speech signal: information relating to linguistic content (*speech*) and information relating to the individual (*speaker*). Speaker information can broadly be thought of as the source of between-speaker differences in the speech signal. According to Garvin and Ladefoged (1963: 194), speaker information can be categorised as *organic*, relating to the anatomy of the speech apparatus, or *learned*, defined by "behavioural differences in (the) usage of the moveable articulators during speech production" (Mokhtari 1998: 4). However, as highlighted by Nolan and Oh (1996), "it is normally impossible ... to assign observable differences to one source or the other" (p. 39). Further, Nolan (1983) claims that the *organic-learned* dichotomy is a "gross oversimplification (which) conceals the complexity of the bases of speaker-specific information in speech" (p.27).

Garvin and Ladefoged (1963) offer a second distinction for categorising *speaker* information based on the *group* and the *individual*. *Group* information relates to a speaker's regional and social background, while *individual* information relates to speaker-specificity. Mokhtari (1998) argues that the *group-individual* distinction can be thought of in terms of *homogeneity* and *heterogeneity*. That is, variables which encode considerable regional and social *speaker* information generally display a high degree

of *homogeneity* across speakers of the same linguistic background. Conversely, variables which are *heterogeneous* carry less regional and social information and therefore potentially offer greater discriminatory power. However, as with the *organic-learned* distinction, it is questionable whether the distinction between *group* and *individual* variation is dichotomous since all linguistic-phonetic variables respond, at least to some extent, to individual, regional, social and contextual factors.

It is traditionally argued that linguistic information relating to phonetic contrast is encoded in F1 and F2 (Ladefoged and Johnson 2010; Clermont and Mokhtari 1998) since these lower formants relate to broad articulatory differences in vocal tract configuration. In this way, F1 and F2 encode a considerable amount *speech* related information. The two lowest formants are also responsible for carrying considerable *speaker* information relating to regional and social background (i.e. *group* information). Since higher formants in English are not responsible for phonetic contrast, there is reason to predict that they are not regionally and socially stratified to the same extent as lower formants. Furthermore, higher formants have been identified as carriers of speaker-specific information since "they are less susceptible to ... behavioural and anatomical variation (within) speakers" (McDougall 2004: 123). It is argued that this is because they are more closely related to resonances in smaller cavities within the vocal tract (Peterson 1959; Rose 2002). From this observation it is reasonable to hypothesise that higher formants furnish greater inter-speaker variability and intra-speaker stability.

The results of a number of studies offer considerable support for the claim that F3 and higher formants are strong carriers of speaker-specificity. Based on formant contours of German long vowels, Greisbach *et al.* (1995) found greater between-speaker variation in F3 compared with F2 and particularly F1. Using 20 speakers from DyViS Task 1, Simpson (2008) analysed the comparative performance of F1 to F4 across five short vowel phonemes of SSBE based on F-ratios (the ratio of between- and within-speaker variation) generated using ANOVAs and comparisons of SS and DS pairs. F3 and F4 consistently outperformed F1 and F2 across all metrics, with the highest proportion of variance in F3 and F4 associated with *speaker* as compared with *lexical set* for F1 and F2. These results are also consistent with the traditional view that F1 and F2 primarily encode information for phonetic contrast.

McDougall (2004) investigated the speaker discriminatory potential of formant trajectories of /aɪ/ preceding /k/ under different speaking rates and levels of prosodic stress. Linear discriminant analysis (DA) was performed on five speakers of General AusEng. DA is a closed-set form of Bayesian posterior analysis which generates a classification rate based on the proportion of cases (tokens) correctly assigned to a given group on the basis of a series of input predictors (see Morrison 2008: 261-264). F3 classification rates were generally found to be higher than those for F1 and F2, particularly with small numbers of predictors. Further, analyses of combined formant performance were consistently better with the inclusion of F3. Based on the magnitude of LRs, Kinoshita (2001) found that F3 provided the best speaker discrimination for Japanese mid-point vowel data, followed by F2, F1 and finally F4, although performance varied across phonemes.

Despite the comparative performance of F3 relative to lower formants in previous studies, there are a number of factors which may introduce systematic regional and social variation, potentially diminishing its speaker discriminatory potential at least for certain regional and social groups. F3 has been shown to have direct articulatory correlates, with lip rounding and protrusion in particular causing a decrease in F3 (2001). Stevens (2000) states that such lowering of F3 (and indeed of all resonant frequencies) during rounding is an acoustic consequence of a decrease in the "cross-sectional area of the anterior end of the vocal tract, and lengthening of the front part of the tract" (p. 291) Consistent with this claim, Maeda (1990) maintains that the reduction in the proximity of F3 and F2 is responsible for the auditory-phonetic distinction between [i] and [y].

Many lingual settings responsible for habitual differences in vocal tract configuration have also been shown to cause variability in resonant frequencies (Laver 1994: §13.5.2.3). Settings may also be variety-specific, causing systematic variation in F3 (e.g. velarised vocal setting in Liverpool and Birmingham; Esling and Dickson 1985). Such systematic articulatory and phonatory factors are capable of encoding socio-demographic information, which may compromise the speaker discriminatory power of F3 within dialects. Conversely, such systematicity may improve the discriminatory power of F3 across dialects.

3.3.1.2 Data extraction

On the basis of previous research, the dynamic approach was adopted in this thesis for the analysis of vowel formant trajectories. With the exception of Chapter 4, vowels were analysed using F1, F2 and F3 (F1∼F3). For the majority of the data, the onset and offset of vowel tokens were manually defined using interval tiers on PRAAT (Boersma and Weenink 2011) TextGrids according to the criteria in §3.3.1.3. Where forced-aligned TextGrids were available (i.e. ONZE), the onset and offset of tokens determined by automatic segmentation were used with some manual correction based on auditory analysis and visual inspection of the spectrogram where possible. Following the procedures in McDougall (2004), time-normalised formant measurements were taken at +10% steps across the duration of each vowel token (exemplified in Figure 3.8).

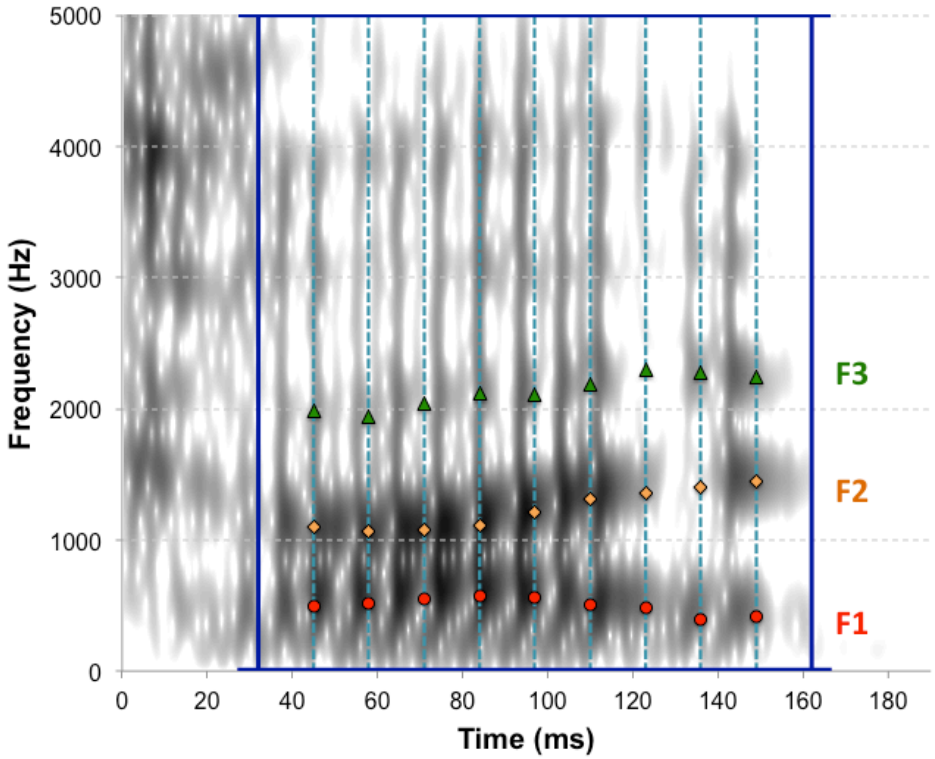

Figure 3.8: Example of points for time-normalised dynamic formant analysis with measurements taken at +10% steps (McDougall 2004) for a token of /aɪ/ from the word *skype* from DyViS sample 027-1-060425.wav

A number of PRAAT scripts[13] were used to extract dynamic data. The scripts create a Formant object for the entire sound file using the *To formant (burg)...* function which performs short-term spectral analysis of windowed frames of 2.5ms (Gaussian window = 5ms) shifted at 2.5ms steps. For each window, the function estimates formant frequencies based on linear predictive coding coefficients using the burg algorithm (see Harrison 2013). Pre-emphasis was also applied, which was set to amplify frequency components above 50 Hz to account for the spectral tilt. The script then uses the regions from the associated PRAAT TextGrids to extract the dynamic data at the appropriate points.

Given that the *To formant (burg)...* function in PRAAT performs spectral analysis on frames across the whole speech sample, prior to data extraction for specific time stamps, memory issues were encountered in PRAAT were processing large sound files. This issue was overcome in a number of ways. For certain datasets in Chapters 4, 5 and 7, recordings were resampled to reduce file size or tokens were manually extracted to separate sound files to preserve sampling rate. A similar approach was also used in Chapter 10 whereby individual tokens were automatically extracted to separate sounds files using a PRAAT script.[14] For ONZE (Chapters 4, 7 and 8), the original sampling rate was preserved by extracting data automatically using LaBB-CAT.

3.3.1.3 Defining the onset and offset of vowel tokens

When manually segmenting vowel tokens, the onset and offset were defined according to a series of criteria depending on adjacent phonological context. These criteria were implemented to ensure that vowel segmentation was accurate and consistent across experiments. Table 3.5 displays the criteria for defining the onset of tokens based on the preceding sound, and Table 3.6 displays the criteria for defining the offset based on the

[13]Chapters 4, 5, 7 and 8: Hudson, T. and Williams, C. 'IntervalFormants_use_me3.praat' and 'common.praat'/ Chapter 10: adapted version of Lennes, M. (2003) 'Collect_formant_data_from_files.praat'. `http://www.helsinki.fi/~lennes/praat-scripts/public/collect_formant_data_from_files.praat` (accessed: 15th May 2013)

[14]Lennes, M. (2003). 'Save_intervals_to_wav_sound_files.praat' `http://www.helsinki.fi/~lennes/praat-scripts/public/save_intervals_to_wav_sound_files.praat` (accessed: 29th July 2013)

following sound (Turk *et al.* 2006). Tokens adjacent to /r/ and /l/ were largely avoided based on their expected long-term resonance effects (West 1999, 2000), although some of these tokens were included where data were limited. In all cases, boundaries were moved to the nearest zero crossing.

Table 3.5: Criteria used to define the onset of vowel tokens based on the preceding sound

Preceding	**The vowel onset is defined by…**
pause	… full periodicity (excluding any period of creak) in the waveform, coinciding with the presence of vertical striations and formant structure in the spectrogram.
nasal	… the end of the simple waveform structure and low overall amplitude (relative to the vowel) with particularly weak higher formants due to the absorption of energy "from the main nasal-pharyngeal tube" (Harrington 1997: 114) in nasal production, the absence of evidence of anti-formants in the spectrogram (Stevens 2000), and the onset of a complex waveform structure and clear, high amplitude formants.
plosive	… the offset of aperiodicity in the plosive burst and the onset of full periodicity (excluding any period of voiced friction).
fricative	… the offset of aperiodicity and onset of full periodicity (excluding any period of voiced friction).
lateral	… a marked change in the spectrogram at which point the amplitude of higher formants is considerably greater than that in the lateral, consistent with the "abrupt change in articulation" (Ladefoged and Johnson 2010: 52) whereby the tongue tip is released from the closure at the alveolar ridge, and possibly an increase in F2 (which is typically quite low in /l/ realisation; F2 values for the lateral vary on a continuum based on the darkness of the realisation (amount of velarisation) with the darkest /l/ having the lowest F2).

glide	… the point at which F2 stabilises, following a decrease in F2 during the transition from the palatal glide into the vowel and following an increase in F2 (partly due to lip unrounding) during the transition from the labial-velar glide into the vowel.

Table 3.6: Criteria used to define the offset of vowel tokens based on the following sound

Following	**The vowel offset is defined by…**
pause	… the absence of acoustic energy in the signal.
nasal	… the change in waveform from a complex vowel structure to simplistic nasal structure and a marked decrease in overall amplitude (particularly in the higher formants).
plosive	… the offset (or weakening) of energy in F2 indicating the presence of a closure in the oral tract (Foulkes *et al.* 2010: 67).
fricative	… the offset of full periodicity in the vowel and the onset of any aperiodicity characteristic of the fricative, or the offset of F2 to indicate the presence of an oral closure for affricates.
glide	… the point at which F2 begins to increase into the palatal glide.

3.3.1.4 Parametric representations of formant trajectories

More recently, attention in FVC has focused on curve fitting techniques which are able to capture the dynamic properties of formant trajectories using a smaller number of predictors than raw frequency input. The use of such parametric representations of the data improves the statistical efficiency (and precision) involved in LR computation by reducing the number of potentially correlated dimensions. One of the most common curve fitting techniques applied to FVC is polynomial regression.

For formant trajectories, polynomial regression provides an approximation of the non-linear relationship between time and frequency. This relationship can be described as an equation, of increasing complexity:

$$\tilde{y}(Hz) = f(x) = a_1 + a_2x + a_3x^2 + a_4x^3 \dots a_nx^i \tag{3.10}$$

where $\tilde{y}$ is the frequency value on the curve of best fit (y-fit) and x is the +10% step. By fitting i^{th} order polynomials, the raw data are reduced to a series of coefficients (Seber and Wild 1989) which describe properties of the trajectory. Coefficients are calculated using the least squares method, which minimises the sum of the squared residuals (ϵ) (Whittle 1983), where residuals are the difference at each point of x between the raw data and the fitted data. The goodness of the fit is determined by the R^2 value, which increases towards one as a function of polynomial complexity:

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{N} \epsilon_i^2}{\sum_{i=1}^{N} (y_i - \tilde{y})^2} \right) \tag{3.11}$$

The first three coefficients can be interpreted as linguistically meaningful. The intercept (a_1) represents the y value at the point where x = zero (i.e. where the line crosses the y-axis). The linear term (a_2x) captures the slope of the trajectory defined by the amount of movement between the onset and offset. The squared term (a_3x^2) captures the magnitude of the parabola (i.e. deviation from a straight line). Beyond this, coefficients become increasingly abstract in terms of their linguistic correlates. It is also important to emphasise that polynomial coefficients are typically highly correlated with each other, reflecting the trade-offs in least squares regression (Whittle 1983).

An issue for the experiments in this thesis is the choice of polynomial order applied. Although increasing polynomial order improves the goodness of the fit, the principle of parsimony in least squares regression demands the use of "models and procedures that contain all that is necessary for modelling but nothing more" (Hawkins 2004: 1). Where parsimony is violated, "overfitting" (Morrison 2008: 253) may occur, since an overly complicated regression model exaggerates noise caused by factors such as measurement errors, potentially resulting in worse speaker discriminatory performance.

McDougall (2006) compared the performance of quadratic (2nd order) and cubic (3rd order) polynomials extracted from formant trajectories of /aɪ/ with raw values at nine time-normalised sampling points. Based on DA the cubic system plus duration achieved the highest classification rate (optimally 96%). Despite containing four fewer predictors, the performance of the quadratic system was not markedly lower (91%). Similarly, based on an analysis of the F1 and F2 contours of /uː/, McDougall and Nolan conclude that "although the cubic polynomials provide a better fit . . . it appears that a worthwhile amount of speaker-distinguishing information can be captured with the quadratic approximations" (2007: 1828). Morrison (2009b) performed LR-based comparisons using polynomial representations of formant trajectories from five diphthongs of AusEng. The highest order representation (cubic) was found to achieve the lowest C_{llr}. However, Morrison emphasises that "the parametric curve with the best performance for each vowel phoneme (may need to) be determined on a case-by-case basis" (2009b: 2395).

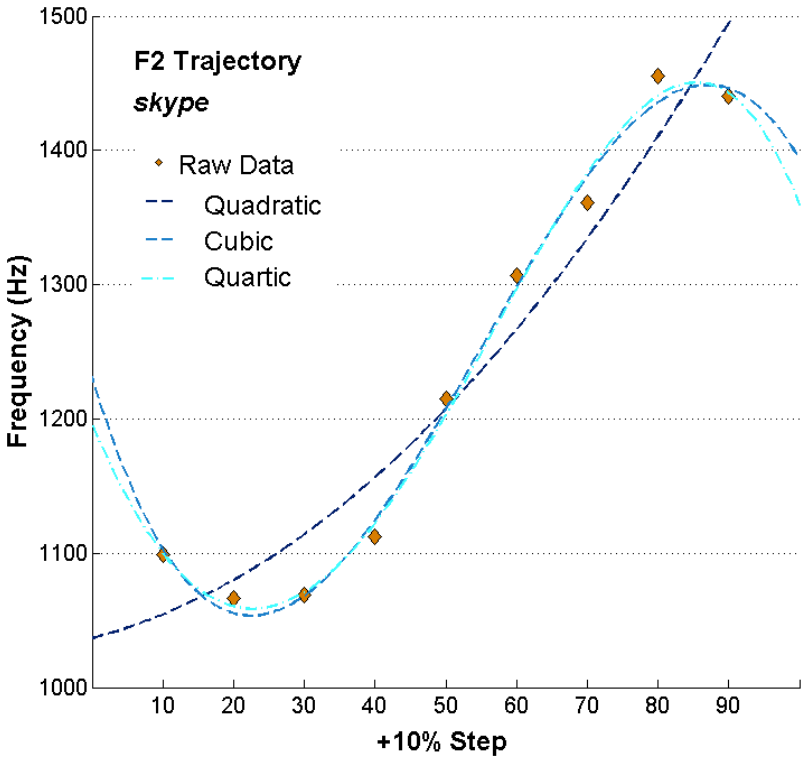

Figure 3.9: Raw F2 (Hz) trajectory for a token of /aɪ/ from the word *skype* from DyViS sample 027-1-060425.wav (as in Figure 3.8) fitted with quadratic, cubic and quartic polynomial curves

This section presents the results of pre-testing of different polynomial orders to establish which should be used throughout this thesis. The comparative performance of quadratic, cubic and quartic polynomial representations of F1∼F3 trajectories of /aɪ/ are analysed with regard to the distributions of calibrated LLRs, C_{llr} and EER.

Method

Task 1 recordings for all 100 DyViS speakers (§3.1.1) were used. Dynamic F1∼F3 data for /aɪ/ for the first 20 speakers were available from Hughes (2009). Tokens of /aɪ/ from the remaining 80 speakers were manually segmented following the criteria in Tables 3.5 and 3.6. Only /aɪ/ tokens occurring in DyViS target words (Table 3.7) were analysed. The relative lack of available data for each speaker meant that it was not possible to ensure the same number of tokens in equivalent phonological contexts for each speaker.

Table 3.7: Target items containing /aɪ/ elicited by the interviewer for DyViS

/aɪp/	**/aɪt/**	**/aɪk/**
type	heights	bike
pipeworks	kite	pike
typesetter	tightrope	hike
hypermarket	pighty /ˈpaɪtiː/	sky-coloured
skype		tyke

The recordings were resampled at a rate of 11.025 kHz. Dynamic formant data were then extracted (§3.3.1.2) searching maximally for between five and six formants (i.e. an LPC order of 10 or 12) over a 0 to 5 kHz range and errors were hand-corrected. Of the 100 speakers, three were removed due to small numbers of available tokens. The resulting dataset contained 97 speakers with between 11 and 19 tokens per speaker. Formant trajectories from all tokens were fitted with quadratic, cubic and quartic polynomials (Figure 3.9) using an implementation of the MATLAB *polyfit* function.[15] This reduced the nine raw frequency values to between three and five coefficients per formant.

[15] 'polyfitcaller.m' written by Ashley Brereton (2011). This function was used to fit polynomial curves to the formant trajectory data throughout this thesis.

To ensure comparable models of within-speaker variability, only the first ten tokens per speaker were included. Of the 97 available speakers, 20 were identified at random to function as development data and a further 20 were used as test data. Typicality was assessed using a background model based on the remaining 57 speakers. MVKD (§3.2.2.1) LR scores (20 SS/ 380 DS) were computed for the development and test data. The test scores were then converted to calibrated LLRs based on calibration coefficients generated from the development scores (§3.2.4.1).

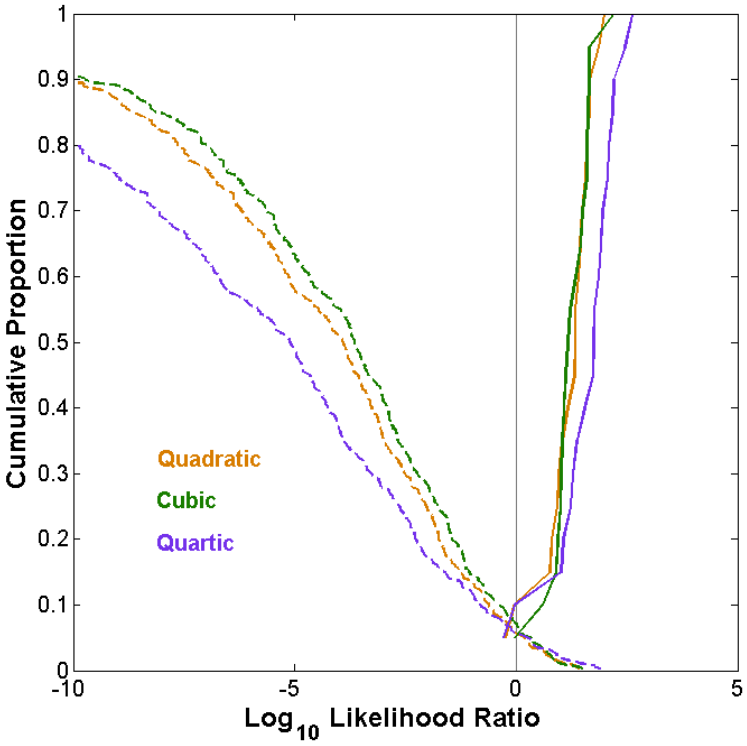

Figure 3.10: Tippett plot of SS (solid) and DS (dashed) LLRs using quadratic (orange), cubic (green) and quartic (purple) representations of the F1∼F3 trajectories of /aɪ/

Results

Figure 3.10 is a Tippett plot of the calibrated LLRs based on different polynomial representations of the raw data. The general strength of LLRs was greatest using the quartic model. The median SS LLR using quartic coefficients was 0.46 greater than that using quadratic input and 0.56 greater than that using the cubic input, although in verbal terms the medians were all equivalent to *limited* support for the prosecution. The distributions of SS LLRs were similar across the quadratic and cubic systems. The

lowest proportion of misses (5%) was recorded using cubic input, where only one of the 20 comparisons achieved a negative value. Using the quadratic and quartic input the miss rate was marginally higher (10%).

The differences between polynomial orders were greater for DS LLRs. The median DS LLR using the quartic data (-5.12) was two orders of magnitude greater than with either quadratic (-3.92) or cubic (-3.66) input. Verbally, this is equivalent to the difference between *very strong* (quartic) and *strong* (quadratic, cubic) support for the defence. As with SS LLRs, the distributions of quadratic and cubic DS LLRs are similar to each other, although the LLRs were marginally weaker using cubic representations. The proportion of misses was highest using the cubic data (6.84%), with both quadratic and quartic input achieving a miss rate of 5.53%.

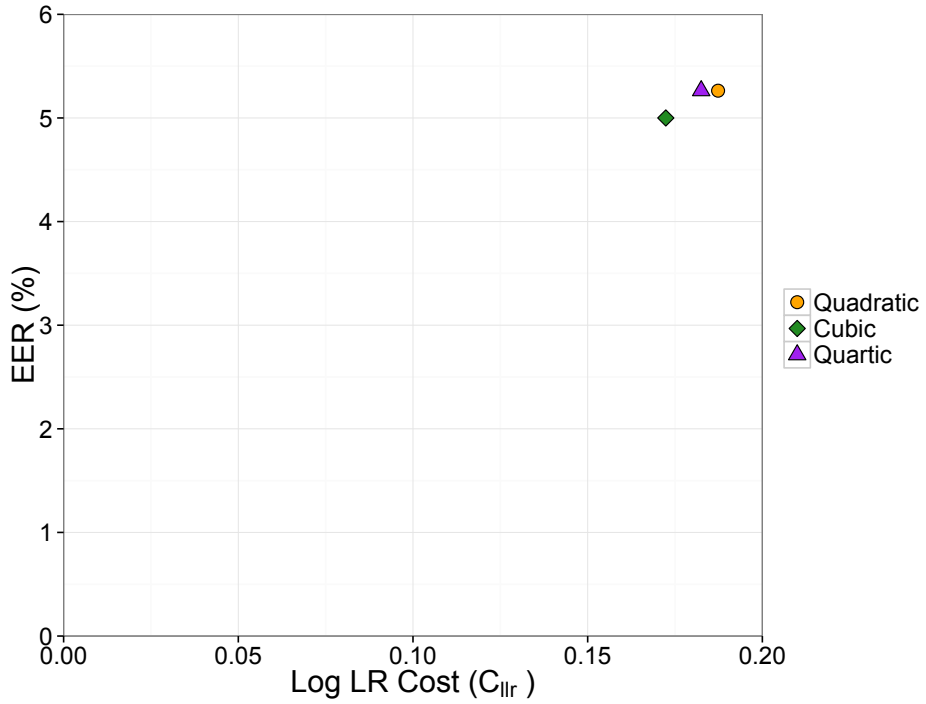

Figure 3.11: C_{llr} plotted against EER using quadratic (orange), cubic (green) and quartic (purple) input

Figure 3.11 shows that there was very little difference between the systems in terms of validity. Optimum EER or C_{llr} values were achieved using cubic input, reflecting the smallest proportion misses and the lowest magnitude contrary-to-fact LLRs. The quadratic and quartic systems achieved the same, marginally worse, EER (5.26%), although the quadratic system produced a higher C_{llr} (0.48).

Conclusions

Across LR output, the three systems performed similarly well. The EER and C_{llr} values also suggest that all three representations captured a substantial amount of speaker-specific information. Further, there was no evidence of overfitting using higher order polynomials with the quartic system achieving lower C_{llr} than the quadratic system as well as the highest magnitude LLRs. Equally, there was also no improvement in performance through increasing polynomial order. For the purposes of the experiments in this thesis, it is considered preferable to use the polynomial representation with the best validity rather than the strongest LLRs. Therefore, cubic coefficients were chosen for modelling dynamic formant data.

3.3.2 Cepstral coefficients and derivatives

ASR typically involves the analysis of coefficients and derivatives from the power cepstrum. The cepstrum is used in ASR primarily because data extraction can be automated across an entire speech recording and because a considerable amount of useful, speaker discriminatory, information can be extracted efficiently. The power cepstrum is the inverse Fourier transform of the logarithm of the short-term power spectrum of a signal, defined in Bogert *et al.* (1963) as:

$$\text{power cepstrum} = |F^{-1}\{log(|F\{f(t)\}|^2)\}|^2 \tag{3.12}$$

where:

F = Fourier transform

t = Signal

In this thesis, two forms of the cepstrum were used: Mel-frequency cepstrum (MFC) and linear prediction cepstrum (LPC). The primary difference between these is the frequency scale onto which the powers of the spectrum are mapped during processing. In the case of the MFC, frequency bands are spaced according to the non-linear Mel scale, while the linear frequency scale is used for the LPC. Both forms of the cepstrum are used extensively in ASR research.

The primary data extracted from the cepstrum are CCs. CCs contain considerable information about the supralaryngeal vocal tract with Rose (2002) claiming that the cepstrum "effectively decouples the part(s) of the speech wave that were due to the glottal excitation from those that were due to the supralaryngeal response" (p. 262). However, Rose also claims that the cepstral-spectral envelope does reflect "aspects of the phonatory activity of the source" (2011a: 1718). ASR performance has also been shown to improve with the addition of derivatives (Campbell 1997). The derivatives of CCs used in ASR are delta, or differential, coefficients (Ds)[16] and delta-delta, or acceleration, coefficients (As). Ds are based on the spectral change between CC vectors from preceding and following frames. Since it is not possible to calculate Ds without CCs, the number of Ds extracted must be equal to the number of CCs. The calculation of As is based on the same principles, using change in Ds rather than CCs as input.

This section outlines the general procedures for extracting MFC and LPC coefficients and derivatives used in Chapter 6. Specific choices relating to input data and settings are explained in Chapter 6 itself.

3.3.2.1 Extracting cepstral coefficients and derivatives

The extraction of cepstral information consists of seven steps. These steps, with the exception of step four, are identical for MFC and LPC analysis. This process is visualised, for MFC analysis, in Figure 3.12.

The entire speech sample is initially divided into frames based on a window size of x(ms), shifted across the sample at intervals of y(ms). Typically, either a rectangular or hamming window is used for this. In this thesis a hamming window was used. A pre-emphasis filter is then applied to the signal using the first order difference equation:

$$s'_n = s_n - ks_{n-1} \tag{3.13}$$

from Young *et al.* (2006: 62)

[16]The abbreviations used here are the same as those used in the HTK toolkit documentation (Young *et al.* 2006).

where s_n ($n = 1, 2, 3 \ldots n$) is the signal from a given frame and k is the pre-emphasis coefficient. The pre-emphasis filter accounts for the spectral tilt by increasing the amplitude of lower intensity, high frequency components relative to the amplitude of higher intensity, low frequency components.

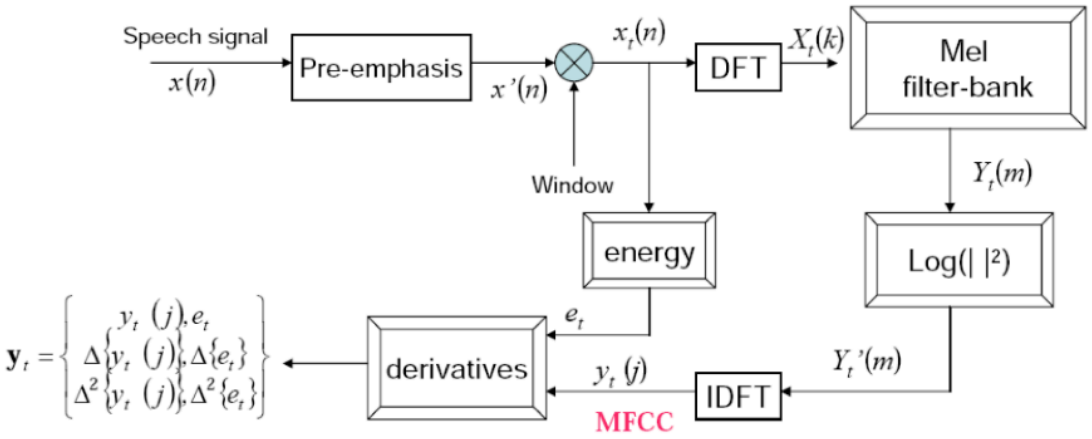

Figure 3.12: Visual representation of extraction of cepstral (in this case MFC) information from a speech signal (Jurafsky 2007)

At step three, the signal from each frame is converted to a power spectrum by applying a Discrete Fourier Transform (DFT) and a filterbank is then applied. The filterbank consists of a number of triangular filters applied across the entire frequency range. At this stage, the processes of analysing the MFC and LPC differ slightly. For the MFC, the filterbank is based on the Mel-frequency scale; a perceptual scale which captures the non-linearity of the human auditory system (Johnson 2008). The relationship between linear frequency (f) and Mel-frequency (m) can be expressed as (Figure 3.13):

$$m = 2595\, log_{10}\left(1 + \frac{f}{700}\right) \tag{3.14}$$

from O'Shaughnessy (1987)

In the case of the MFC, the filterbank consists of filters whose width and absolute degree of overlap increases with frequency (Figure 3.13). For the LPC, the filterbank is applied to the linear frequency scale (i.e. with no transformation of the power spectrum), involving filters of equal width and absolute and proportional overlap. The energy in each filter is then summed and, in step five, the values logged. The penultimate step, involves fitting a discrete cosine transform (DCT) to the logged filterbank energies. The coefficients associated with the DCT are the CCs.

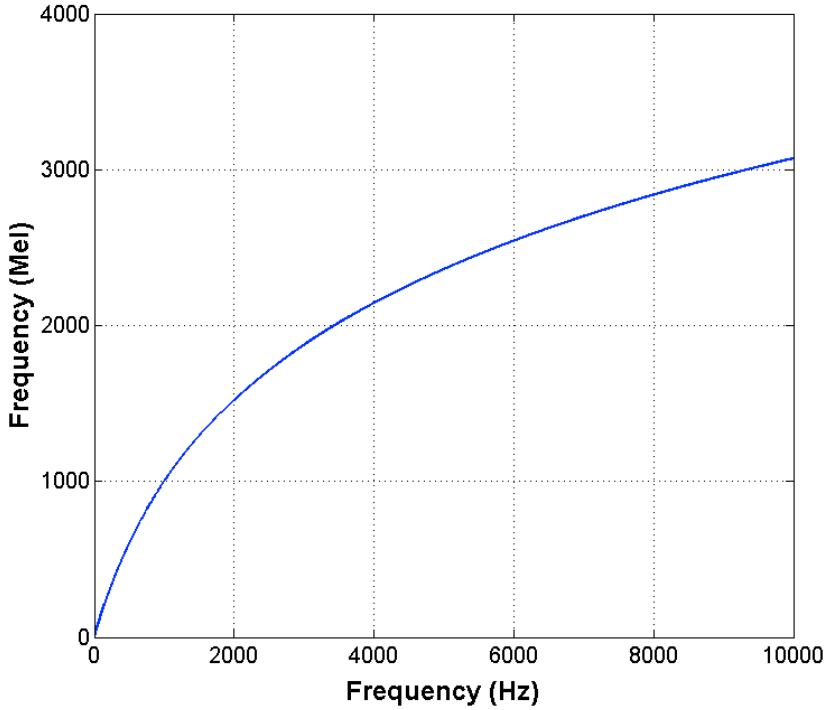

Figure 3.13: Relationship between linear and Mel frequency scales

The final stage involves extracting derivatives based initially on the vectors of CCs for adjacent frames. Deltas (Ds) are calculated by:

$$D_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{s \sum_{n=1}^{N} n^2} \tag{3.15}$$

where:

D_t = Delta coefficient

t = Frame

c_{t+n}, c_{t-n} = Static CCs from adjacent frames

from Young *et al.* (2006: 62)

Delta-deltas (As) are calculated by applying Equation 3.15 to the Ds rather than the CCs.

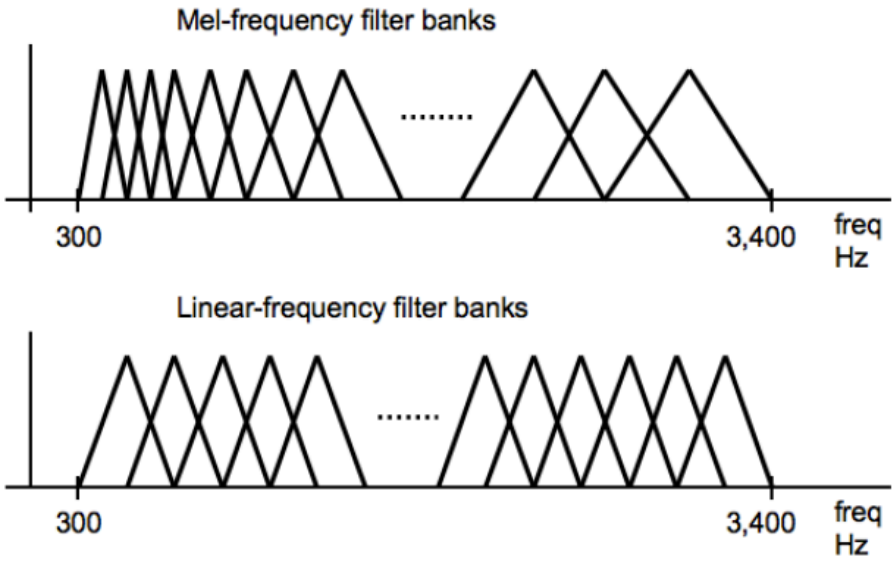

Figure 3.14: Graphical representation of the Mel (above) and linear frequency (below) filterbank applied to the power spectrum from a given window, with 50% overlap between filters (from Lei and Lopez-Gonzalo 2009: 2324)

3.4 Limitations

There are a number of general limitations with the experiments in this thesis. Firstly, across all experiments LRs are computed using contemporaneous data from single sample per speaker, i.e. divided in half to compute SS comparisons. This is due to the fact that databases with non-contemporaneous samples (i.e. two recordings per speaker separated by some period of time) generally do not contain sufficient numbers of speakers from the narrowly defined sociolinguistically groups relevant to the experiments in this thesis. Secondly, the ONZE, NE and PVC datasets were collected primarily for sociolinguistic research and are therefore liable to the limitations outlined in §2.4.2.1. For TIMIT, the level of forensic relevance is further limited by the use of read speech (the preference for TIMIT over other ASR databases is explained in Chapter 6). All of the samples used were also recorded directly, and most in high quality and digitised with optimum sampling rates.

It is predicted that the use of contemporaneous, high quality samples will lead to overly optimistic performance compared with real forensic conditions. The potential implications of using optimal data are discussed in individual data chapters. However,

as highlighted in §2.2.5, relatively little work has empirically tested the impact of non-contemporaneity on the outcome of numerical LRs (with the exception of Enzinger and Morrison 2012 and Coe 2012). Furthermore, given that little work has considered the research questions of this thesis, it is considered preferable to test these questions initially using optimal data. This will help to reveal the specific effects of variability in the definition of the relevant population and sample size in LR-testing, without the confounding issues of various sources of mismatch between suspect and offender samples encountered in forensic casework.

In Chapters 4 and 5, a combination of intrinsic and extrinsic testing (§3.2.2.2) was used. This may exaggerate the differences in LR output between the regionally Matched and Mixed/Mismatched sets, since intrinsic testing predicts greater similarity between datasets extracted from the same database. The use of auto-generated data in Chapters 4 and 8 is also a substantial limitation, since the accuracy of segmental boundaries is reduced for forced-aligned TextGrids. Further, the procedures implemented to correct and remove errors serve to identify clear outliers, rather than more subtle measurement errors (although the use of parametric representations of the trajectories does help to reduce the noise in the raw data) or errors due to incorrect segmental boundaries. Further experiment-specific limitations are also discussed in the relevant data chapters.

Chapter 4

Regional Background: /uː/

This chapter explores the extent to which LRs are affected by different definitions of the relevant population with regard to regional background, using the formant trajectories of /uː/ as input. Firstly, LRs were computed using multiple sets of regionally defined test data and a single set of reference data, where one test set matches the reference set for regional background. Secondly, calibrated LLRs for a single test set were computed using multiple systems containing: (a) regionally **Matched** development and reference data and (b) regionally **Mixed** development and reference data.

4.1 Introduction

As outlined in §2.3.1, logical relevance based on offender language and sex has been the preferred approach for defining the relevant population in the vast majority of LR-based research and casework. However, a substantial issue for the application of logical relevance to FVC is the extent to which analysts' decisions relating to these sources of between-speaker variation affect LR output. This chapter presents the results of two experiments which address this issue by considering different definitions of the relevant population with regard to regional background using the formant trajectories of /uː/ (GOOSE; Wells 1982) as input.

In Experiment (1), LRs were computed using a single set of regionally homogeneous reference data and multiple sets of regionally defined test data, where one matches the

reference data for dialect. This experiment reflects the practical issue in LR-based FVC of the limited availability of databases for assessing typicality. Therefore, in the vast majority of cases the analyst would currently need to use reference data (forensic or non-forensic; §2.4.2) which displays some degree of mismatch with the offender in terms of the regional background of the reference speakers. Experiment (1) compares the effects of such mismatch on LR output relative to a set of appropriate reference data.

Experiment (2) relates more directly to analyst decisions regarding the relevant population. LR scores were computed for a set of regionally homogeneous test data using systems which represent different controls over regional background. Since the definition of the relevant population in casework informs the choice of speakers used as development and reference data, the effects on LR output are considered across both the feature-to-score (§3.2.2) and score-to-LR (§3.2.4) stages. The systems are defined as (a) **Matched**: using development and reference speakers who match the test data for regional background, reflecting a situation where the analyst defines the regional variety of the relevant population narrowly and correctly relative to the offender, and (b) **Mixed**: using speakers from different regional varieties as development and reference data, reflecting limited control over regional background. The Mixed condition is, to some extent, consistent with Rose (2004) where regional background is defined broadly as *language*.

4.2 Method

4.2.1 Data

A total of 134 speakers were used. The speakers were divided into five groups: a reference set of 102 speakers, and four test sets, each comprising eight speakers. The reference data consisted of NZE speakers, drawn from CanCor (§3.1.2). The four test sets differed in terms of the regional background of the speakers. Three sets contained speakers of British English (BrEng) varieties from the north of England: Manchester (§3.1.3), Newcastle (§3.1.4), and York (§3.1.3). The York data consisted

of five speakers from Tagliamonte (1996-1998) and three speakers from Haddican (2008-2013). The three sets of BrEng data are classed as Mismatched conditions. The fourth test set contained NZE speakers and is classed as the Matched condition. Aside from differences in regional background, the test speakers are considered well matched for sociolinguistic factors such as age, social class, and speaking style.

4.2.2 Variation and change in /uː/

There are a number of reasons why /uː/ was considered a good choice for these experiments. Firstly, /uː/ is not a vowel with a high degree of social or regional variation (until recently it has not been the subject of extensive attention by linguists working in the UK or NZ, for example). It is most appropriately categorised as a sociolinguistic *indicator* (Labov 1971) in all four dialects. *Indicators* are features that display systematic variation but which generally remain below the level of speaker consciousness. They contrast with *markers*, which display stylistic variation, and *stereotypes*, which may be the subject of overt commentary.

A further advantage of using /uː/ is that patterns of variation and on-going change are predicted to be consistent across varieties. The apparent-time fronting of /uː/ has been attested in NZE (Easton and Bauer 2000) and in Manchester and York (Hughes *et al.* 2011; Haddican *et al.* 2013), although evidence from Watt (2000) suggests that /uː/ may be more retracted in Newcastle English. /uː/-fronting has also been found to be correlated with age, such that F2 values are generally higher for younger speakers. Thus only younger speakers were included in the four test sets. Further, consistent patterns of variation according to adjacent phonological context have been found across regional varieties. Ash (1996) and Hall-Lew (2005) establish maximally fronted realisations following /j/ and maximally retracted realisations preceding /l/, especially in varieties where coda /l/ is velarised. Such internal phonological factors are important in ensuring that within-speaker variability is controlled across test sets.

The experiments in this chapter therefore test the effects on LR output of different definitions of the relevant population using a variable which is not expected to display marked differences between regional varieties. In this case the use of a general set of

English data may, *a priori*, be considered adequate for LR testing.

4.2.3 Dynamic formant extraction

Time-normalised dynamic measurements (§3.3.1) of F1 and F2 were auto-generated for 169 male speakers from CanCor using LaBB-CAT. The formant extraction script was set to identify five formants within a range of 0 to 5 kHz, based on an expectation for roughly one formant per kHz for adult males (Keller 2005). This approach was used to generate a large amount of data in a short space of time, since manual formant extraction is labour intensive. As highlighted in Zhang *et al.* (2012), the reliability of auto-generated formant data is expected to be worse than human-supervised formant extraction even with high quality recordings. However, Zhang *et al.* (2013) claim that human supervised formant extraction is not necessary for FVC casework "given the high-cost … and the relatively small levels of meaningful improvement it provides" (p. 808) relative to the performance of a much cheaper, generic MFCC-based system.

To remove measurement errors, heuristic thresholds were set to constrain acceptable measurements. F1 measurements outside the range of 250-600 Hz were considered errors and the entire token was removed. This allows for considerable F1 variation, as NZE has a variant with a central offset /uː/ → [ʉə] (Hay *et al.* 2008: 24). An upper limit of 600 Hz was considered sufficient to capture variation in vocal tract length, without accepting erroneous values. Tokens with F2 values outside 800-2400 Hz were also removed. The wide threshold for F2 values was implemented to account for the expected range of phonological variation. Univariate outliers were identified by calculating between-speaker z-scores, such that values greater than ± 3.29 standard deviations (SDs) from the mean were removed (Tabachnick and Fiddell 2007: 73)

The NZ test set was reduced from the full cohort of 169, in order to exert more control over speaker age. Only those speakers born in 1970 or later, who would have been between 20 and 30 years old at the time of recording, were eligible for inclusion in the test set. From this group, speakers with fewer than 20 tokens were removed. Based on between-speaker z-scores, the eight speakers closest to the group mean were identified as test speakers. To ensure a fair estimate of within-speaker variability across test sets,

within-speaker z-scores were calculated for each speaker and tokens ranked within phonological grouping for each speaker. The 16 tokens with the lowest z-scores in the four phonological conditions in Table 4.1 were used as input data.

Table 4.1: Phonological categorisation of /uː/ tokens and the maximum number of tokens in such contexts shared by every test speaker

Phonological Context	**N Tokens per Speaker**
j_____	6
j_____#	4
non-j_____	4
non-j_____#	2

With the removal of the eight NZ test speakers, 161 males born between 1932 and 1987 were eligible for inclusion in the reference data. Beyond the removal of pre-/l/ tokens, it was not possible to control fully for phonological conditioning while simultaneously ensuring that reference speakers had the same number of tokens overall. Instead, combined z-scores were used to rank tokens according to speaker such that ten tokens per speaker were identified on the basis of minimal between-speaker variation. Therefore, there is some divergence between the test and reference data in the proportion of tokens in each context (Table 4.2), but for both sets 40-50% of tokens were post-/j/. The resultant reference data consisted of 102 speakers with 13 tokens per speaker.

Table 4.2: Percentage of tokens in each of the four phonological contexts for the NZ test and reference sets

Phonological Context	**% Tokens (Test)**	**% Tokens (Reference)**
j_____	37.5	23.7
j_____#	12.5	18.0
non-j_____	25.0	31.5
non-j_____#	25.0	26.8

The dynamic formant data for Manchester, Newcastle and York were extracted manually. Tokens were segmented using the criteria in §3.3.1.3. The Manchester audio files were resampled at a rate of 11.025 kHz. Spectrograms from the original and resampled audio

files were inspected visually to ensure that resampling had not significantly affected acoustic output. Formant measurements for one token were also extracted from the original and resampled files. Comparison reveals a mean difference of 7 Hz between the 11.025 kHz and 44.1 kHz samples at each +10% step. This difference was considered negligible in terms of the resulting LRs.

To avoid resampling, tokens from the York and Newcastle data were extracted to a separate sound file. Dynamic formant data were extracted by identifying between five and six formants within a range of 0 to 5 kHz, determined on a token-by-token basis according to visual inspection of the spectrogram. Obvious measurement errors were hand-corrected. From the Manchester, Newcastle and York data, 16 tokens per speaker were identified following the same methods as the NZ test data (i.e. z-scores and phonological context). The raw F1 and F2 trajectories were then fitted with cubic polynomial curves (§3.3.1.4) and the coefficients used for computing LRs.

4.2.4 Variability in the data

The raw data were inspected to assess the degree of regional variability across test sets. /uː/ was not expected to display a marked degree of variation across the four varieties. This was confirmed by the overlap in post-/j/ mid-point measurements within the F1∼F2 plane (Figure 4.1), although some of the oldest York recordings (from Tagliamonte's data) produced notably back realisations. Mean values for F1 and F2 at each +10% step for all tokens from each set were also calculated. The range of between-set variation was low, with mean F1 spread over 100 Hz and mean F2 spread maximally over 300 Hz. Visual comparison of the raw formant trajectories also suggests considerable overlap between tokens from all speakers in the acoustic space both in terms of absolute frequency and dynamic implementation[17].

Given the degree of overlap in the F1∼F2 plane (Figure 4.1) and the patterns of variation predicted by the sociophonetic literature, there is no particular evidence to suggest that these data come from four distinct regional speech communities. Of course, in a

[17]However, using the data in Figure 4.1, pairwise t-tests did reveal significant differences F1 differences between the NZ set and the Manchester and Newcastle sets ($p =< 0.01$) and F2 differences between the York and Manchester sets ($p =< 0.01$)

FVC case an analyst would ideally not want to evaluate BrEng suspect and offender samples relative to population data from NZE. However, the choice of vowel here is illustrative of possible analytic procedures where the regional background of the offender is unknown and where potential databases for LR evaluations are limited. Further, no set of reference data is perfect and will necessarily display some degree of mismatch with logically relevant characteristics of the offender, of which the analyst may or may not be aware. It is, therefore, essential to assess the effect of such mismatch on the resulting strength of evidence.

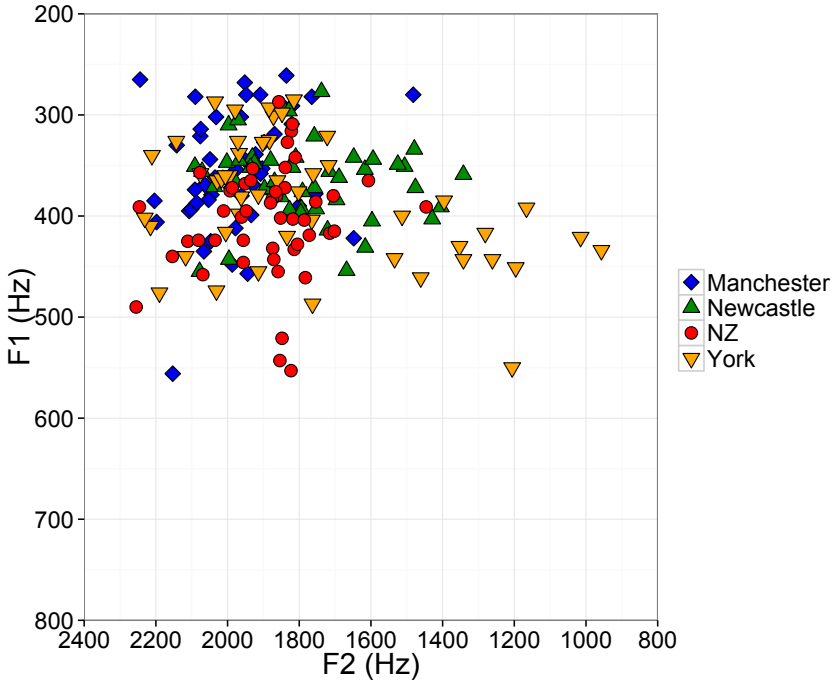

Figure 4.1: F1∼F2 plots of individual tokens of /uː/ (post-/j/ and in open syllables) at the +50% step (mid-point) of formant trajectories for each of the test speakers

4.2.5 Experiments

In Experiment (1), MVKD (§3.2.2.1) scores (8 SS/ 56 DS) were computed using F1 and F2 combined and F2-only input for each of the test sets independently. Typicality was assessed relative to the NZ reference set consisting of 102 speakers. As is common in LR-based research, the F2-only condition was intended to recreate forensic conditions in which F1 may be compromised due to bandwidth restrictions imposed by telephone transmission (Künzel 2001; Byrne and Foulkes 2004). A limitation of Experiment (1)

is the use of multiple test sets and a single reference set, meaning that the estimation of similarity differs across test sets. This compromises the extent to which the differences between the sets of scores can be attributed exclusively to typicality based on regional variation.

This limitation is, however, resolved in Experiment (2) through the use of a single set of test data where the estimation of similarity is consistent across systems. The four test sets from Experiment (1) were initially combined to create a single set of **Mixed** English system data, and the number of tokens per speaker reduced to ten. The 32 youngest speakers from the 102 NZE speakers used as reference data in Experiment (1) were identified to act as a set of regionally **Matched** system data. The youngest speakers were used to ensure that differences in LR output were not due to age, given the expected processes of change over time for /ʉː/ in these varieties. The Matched dataset also consisted of ten tokens per speaker. From the remaining 70 NZE speakers, 40 were identified at random to function as test data. No controls over age were implemented over this dataset as it remained constant across the Matched and Mixed systems. F1 and F2 were used as input.

Initially, cross-validated (§3.2.2.3) MVKD scores were computed for the 32 speakers in each of the Matched and Mixed sets (32 SS/ 992 DS). Using these scores, logistic regression calibration (§3.2.4.1) coefficients were calculated for each system (Matched and Mixed). Parallel sets of SS (40) and DS (1560) scores were then generated for the 40 NZ test speakers using the Matched and Mixed sets as reference data. The use of different numbers of speakers at different stages is not considered problematic since the size of each set was constant across systems. The test scores were converted to calibrated LLRs using the coefficients derived from the Matched and Mixed development scores.

The results of both experiments are evaluated in terms of the distributions of LRs (Experiment (1): scores and Experiment (2): LLRs) and system validity (EER and C_{llr}).

4.3 Results

4.3.1 Experiment (1): Multiple test sets

F1 and F2

Figure 4.2 displays the Tippet plot of scores for the Matched and three Mismatched test sets computed using NZ reference data. There were marked differences in the magnitudes of scores across test sets. The median SS scores were one order of magnitude greater for the three BrEng sets, compared to the NZ set. This is equivalent to the difference between *limited* and *moderate* support for the prosecution. The Manchester and Newcastle sets also generated 0% misses, compared with 12.5% for the Matched set. Further, for Manchester and Newcastle, the ranges of SS scores were narrower than for the NZ set with almost all pairs achieving values between +1 and +2. The largest range of SS scores, however, was found for York, minimally achieving values below +1 (*limited* support) and maximally values of over +3 (*moderately strong* support). This reflects the high within-group variability between the York speakers due to the use of two York corpora, separated by ten years (§3.1.3).

More complicated patterns of variation were displayed across DS results. The highest median DS score was achieved using the York test data (-6.4; off the scale on Figure 4.2). Given the high between-speaker variation in the York data, this finding is unsurprising and not directly attributed to regional variation (i.e. greater dissimilarity between DS pairs generates stronger negative scores). A concerning outcome of the DS results was the high proportion of contrary-to-fact scores when using the Manchester and Newcastle sets. For the Manchester set 57% of DS pairs generated positive scores, while the proportion of false hits for the Newcastle set was 71%. Therefore, the median strength of evidence for Manchester and Newcastle DS pairs was positive, equivalent to *limited* support for the prosecution, compared with the NZ median which was equivalent to *limited* support for the defence.

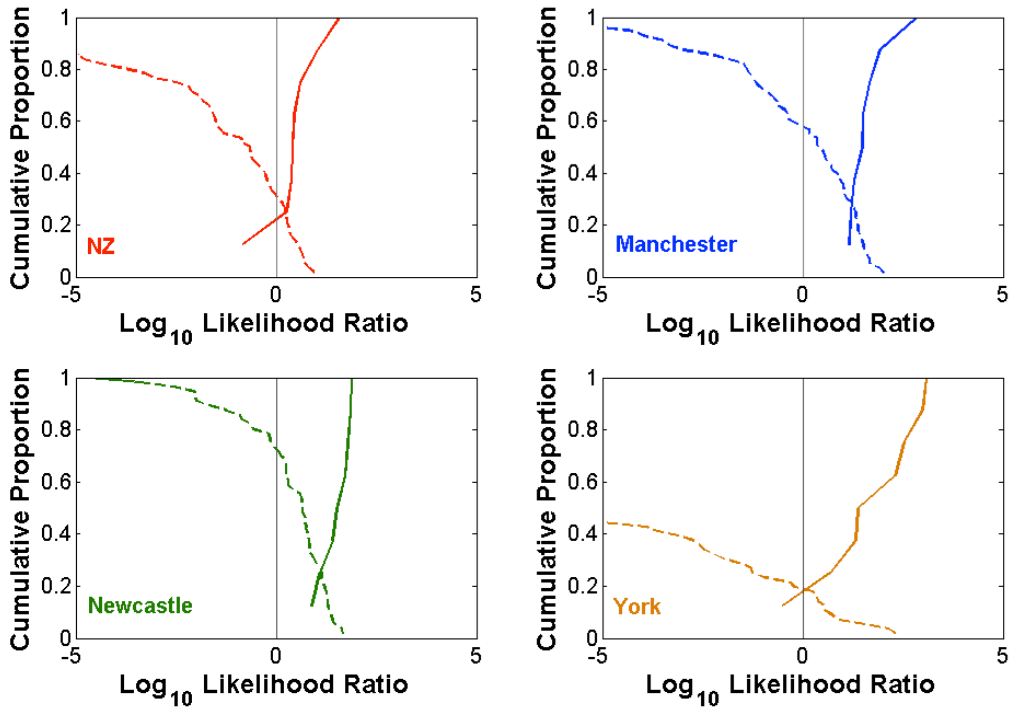

Figure 4.2: Tippett plots of SS and DS scores based on F1 and F2 trajectories from /uː/ for the NZ (top left) (Matched), Manchester (top right), Newcastle (bottom left) and York (bottom right) (Mismatched) test sets

The effects of regional mismatch were also reflected in system validity (Figure 4.3). Consistent with the relatively good separation of SS and DS pairs in Figure 4.2, the best EER and C_{llr} values were achieved using the York data. Again, this is attributed to the high degree of between-speaker variation in this set. Validity was somewhat worse for the NZ test set (EER > 20%, $C_{\text{llr}} > 0.7$) reflecting a large proportion of false hits and the relatively high magnitude of one contrary-to-fact SS score. EER and C_{llr} values were considerably greater for the Manchester and Newcastle sets, reflecting the high proportion (considerably worse than chance) of DS pairs offering support for the prosecution.

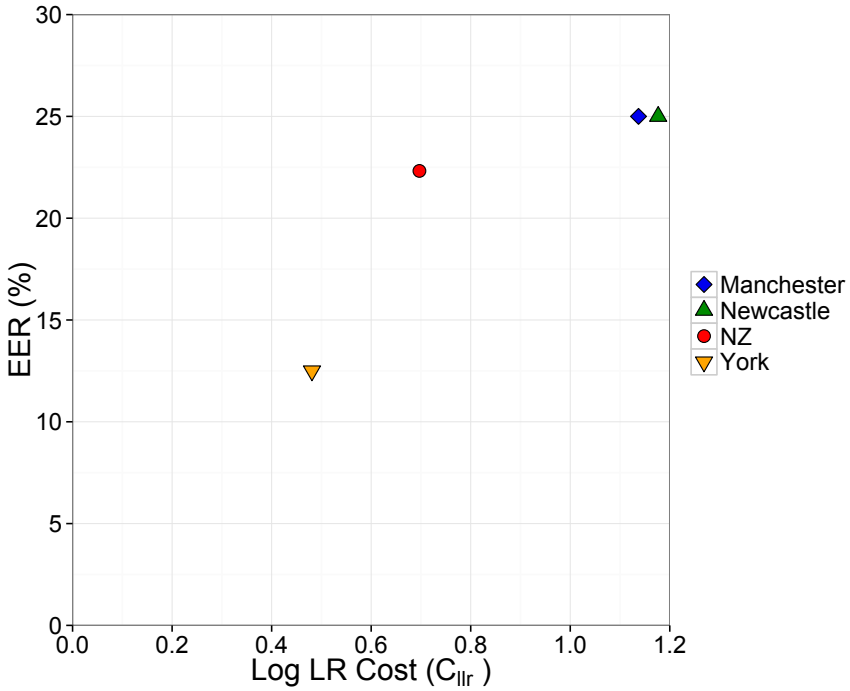

Figure 4.3: C_{llr} plotted against EER (%) for each of the test sets based on F1 and F2 from /uː/

F2-only

Figure 4.4 displays the Tippett plot of SS and DS scores based on F2-only input. Across all sets, SS scores were weaker with the removal of F1. As with F1 and F2 input, the lowest median SS score was achieved with the Matched data. However, for the NZ test data, two SS pairs generated contrary-to-fact support for the defence, equivalent to *limited* strength of evidence, compared with one pair using F1 and F2. The effect of the mismatch between test and reference data was considerably reduced with F2-only information. The median SS scores for the NZ, Manchester and Newcastle sets were all equivalent to *limited* support for the prosecution. As with F1 and F2, the median score for the York test set was one order of magnitude greater than for the Matched set. The York set also produced the widest range of scores.

The removal of F1 also reduced the overall magnitude of DS scores across all sets. As with the SS scores, this also reduced the differences between the Matched and Mismatched sets in terms of LR output. Although differing in terms of verbal equivalent, the numerical differences between DS medians based on F2-only for the NZ

(-0.04), Manchester (+0.07) and Newcastle (+0.16) sets were extremely small, with all values located close to zero. The removal of F1 also increased the proportion of false hits in the Matched condition (44%), and reduced the proportion of false hits in the Manchester (55%) and Newcastle (64%) sets. The median DS score for the York set was again considerably higher than for the Matched and other Mismatched sets, while the proportion of contrary-to-fact scores was lower.

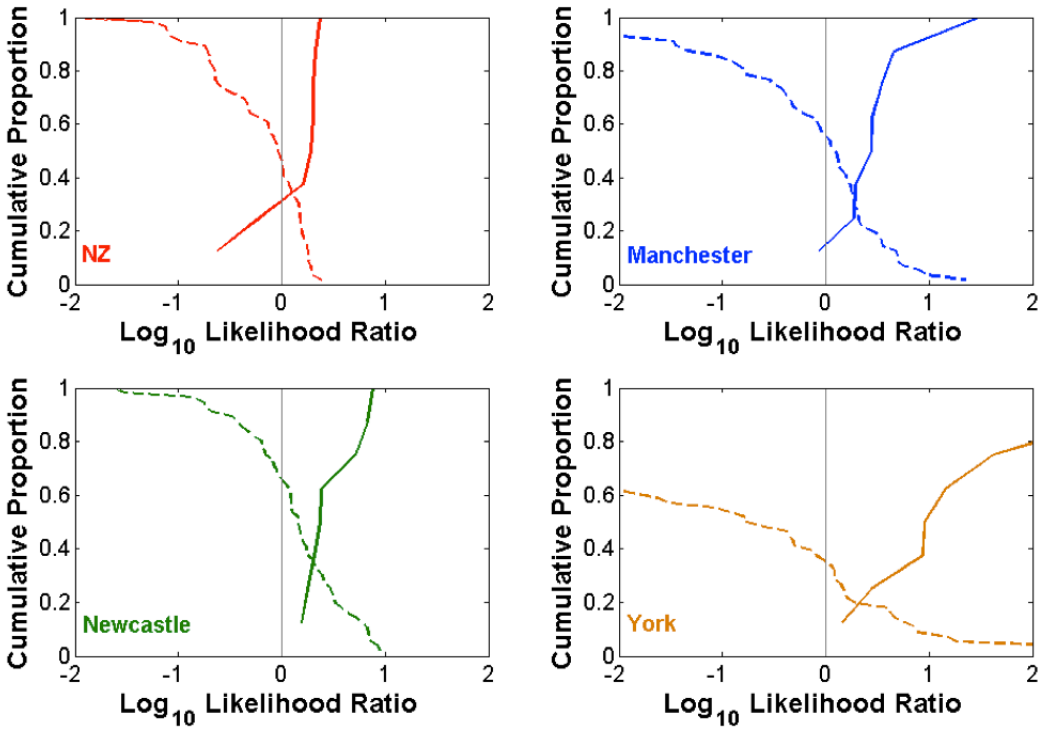

Figure 4.4: Tippett plots of SS and DS scores based on F2-only trajectories from /uː/ for the NZ (top left) (Matched), Manchester (top right), Newcastle (bottom left) and York (bottom right) (Mismatched) test sets

Predictably the best validity, in terms of both EER and C_{llr}, was again achieved using the York data (Figure 4.5). The Matched set produced relatively high EER (25%) and C_{llr} (0.88) values. This reflects a high proportion of high magnitude contrary-to-fact SS scores and a very high proportion of low magnitude contrary-to-fact DS scores. Validity for the Manchester data was marginally better than the Matched set in terms of C_{llr} and marginally worse in terms of EER, reflecting a higher proportion of lower magnitude contrary-to-fact scores. The worst validity was produced by the Newcastle test set. This shows that mismatch between test and reference data can have different effects on validity relative to the performance based on an appropriate reference set.

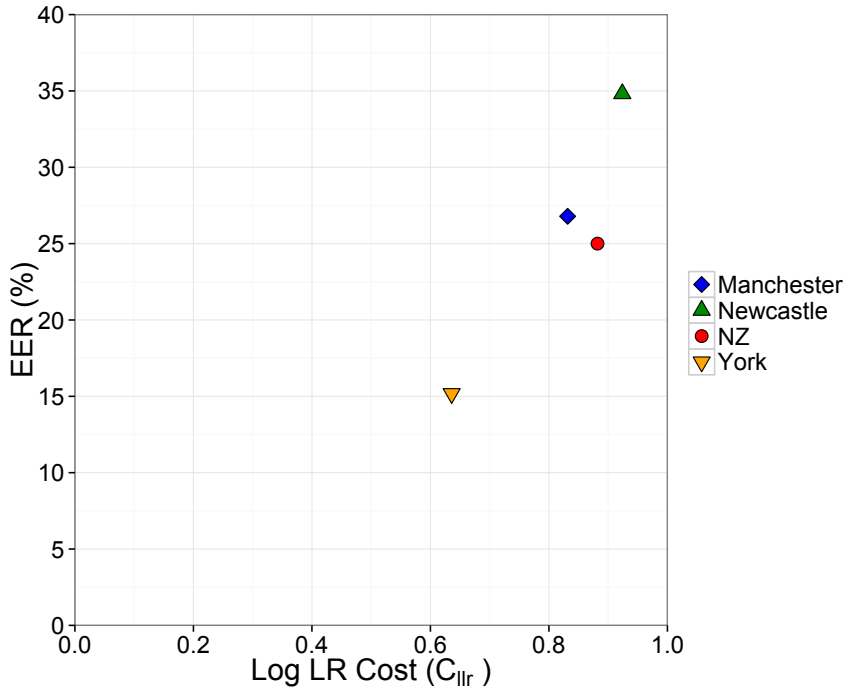

Figure 4.5: C_{llr} plotted against EER (%) for each of the test sets based on F2-only from /uː/

4.3.2 Experiment (2): Multiple systems

Figure 4.6 displays the Tippett plot of calibrated LLRs based on F1 and F2 input using Matched and Mixed data at both feature-to-score and score-to-LR stages. For both systems the magnitudes of the LLRs generated were very low, with the majority located between -1 and +1. There were also high proportions of false hits and misses. There are two likely reasons for this. Firstly, the test set consisted of speakers of all ages, while the reference data consisted of only younger speakers. Therefore, it is conceivable that values from certain DS pairs lie so far onto the tails of the reference distribution that they generate positive scores. Secondly, data for each of the test speakers were not as tightly controlled as in Experiment (1), in terms of the number of tokens in different phonological environments. This may account for very wide ranges of within- relative to between-speaker variation.

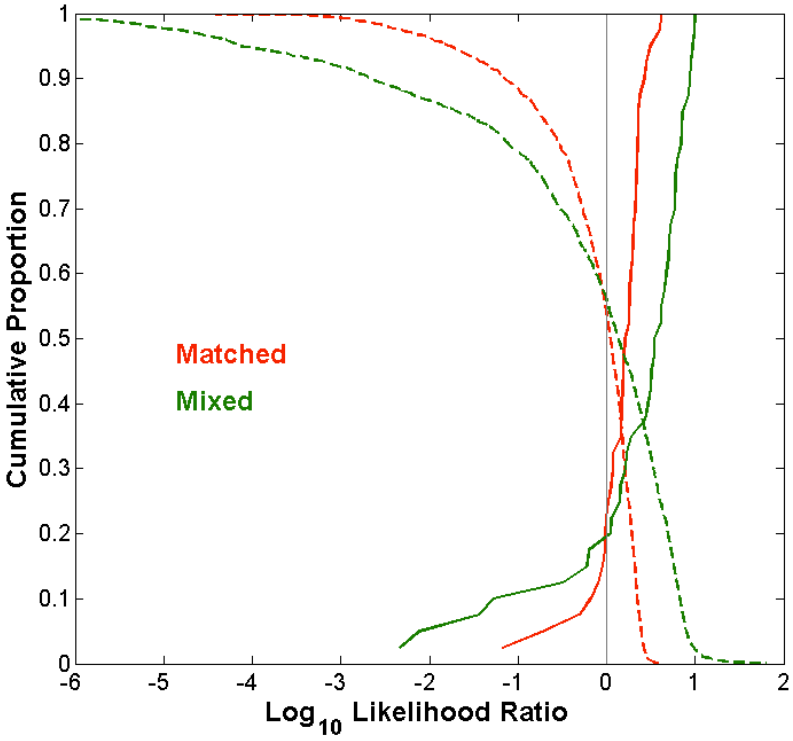

Figure 4.6: Tippett plot of SS and DS LLRs based on F1 and F2 from /u:/ using Matched (red) and Mixed (green) data in both the feature-to-score and score-to-LR stages

Figure 4.6 does, however, show differences in the distributions of LLRs across systems. The median SS LLR was marginally higher using the Mixed system (+0.58) than the Matched system (+0.23), although in verbal terms both are equivalent to *limited* support for the prosecution. More marked differences were displayed in the overall range of SS LLRs, particularly for pairs which generated contrary-to-fact support. LLRs from the Mixed system extended maximally from -2.33 (*moderately strong* support) to +0.99, compared with a range of -0.62 (*moderate* support) to +1.18 using the Matched system. Similarly, for DS LLRs, the median was marginally higher in the Mixed system (+0.13) than in the Matched system (+0.04), although both were very close to zero. As with SS LLRs, the overall range of DS LLRs was considerably greater using the Mixed system.

As shown in Figure 4.6, the proportions of false hits (c. 57%) and misses (c. 20%) made by both systems were almost identical. This is reflected in almost identical EERs in Table 4.3. Table 4.3 also displays C_{llr} values for the two systems. Consistent with the differences in the ranges of LLRs (particularly into contrary-to-fact support) in Figure 4.6, C_{llr} was considerably better using the Matched data than when using the

Mixed data. This is due to the higher magnitude contrary-to-fact LLRs (in particular for SS comparisons) produced by the Mixed system. However, EER and C_{llr} values for both systems reflect bad system performance, due to both the high proportion and high magnitude of LLRs offering contrary-to-fact support.

Table 4.3: EER and C_{llr} using Matched and Mixed data in both the feature-to-score and score-to-LR stages

	Matched	**Mixed**
EER	35.51%	35.22%
C_{llr}	0.92	1.19

4.4 Discussion

Experiment (1) reveals a number of systematic effects of regional mismatch between test and reference data. Given the markedly greater between-speaker variation in the York dataset (due in part to the use of two corpora separated by ten years), patterns for the Mismatched data are only considered in terms of the Manchester and Newcastle data. Firstly, when using F1 and F2, SS scores for the Mismatched sets were generally stronger by one order of magnitude compared with the Matched scores. Secondly, the strength of DS evidence was weaker in the Mismatched conditions compared with the DS scores for the Matched set. Thirdly, a considerably higher proportion of DS scores achieved contrary-to-fact support (i.e. false hits) in the Mismatched conditions. Therefore, the validity of the systems based on Manchester and Newcastle Mismatched data was substantially worse than that for the Matched NZ condition, especially when using F1 and F2. Importantly, the removal of F1 reduced the effect of regional mismatch between test and reference data, making the distributions of scores in the Manchester and Newcastle sets more like those from the NZ set. This suggests that for this particular variable, F1 contains sufficient region-specific information to substantially affect LR output. However, while LR output was more similar across conditions using F2-only, the removal of F1 also generally produced lower magnitude scores (i.e. weaker evidence) across all sets.

An explanation for the differences in LR output between the Matched and Mismatched test sets relates to the location of the suspect and offender data relative to the reference distribution. For SS pairs in the Mismatched conditions, the offender data are likely to be situated on the tails of the distribution of the reference data, meaning that $p(\mathrm{E}|\mathrm{H}_d)$ is lower than it would be in the Matched data where such values are much more typical. This has the effect of generating higher SS scores in the Mismatched conditions (assuming the similarity between suspects and offender is broadly the same across sets). A second implication of this is that $p(\mathrm{E}|\mathrm{H}_d)$ will be lower for DS pairs, leading to weaker DS scores than in the Matched condition. In some cases, $p(\mathrm{E}|\mathrm{H}_d)$ may be so low that the score offers contrary-to-fact support for the prosecution, which would account for the higher EER and C_{llr} values generated for the Manchester and Newcastle sets. Further discussion of these issues based on a comparison of the results across experiments is provided at §11.1.

The calibrated results in Experiment (2) also revealed differences between the Matched and Mixed systems in terms of LLRs. The median SS and DS LLRs across the two systems were within the same order of magnitude, although in numerical terms SS LLRs were generally stronger using the Mixed system. More importantly the magnitudes of contrary-to-fact LLRs were considerably greater using the Mixed system. This accounts for the relatively large difference (0.28) in C_{llr} between the Mixed and the better performing Matched system. However, the distributions of LLRs and validity metrics, even using the Matched system, suggest that /uː/ is not a very good speaker discriminatory variable for FVC. This is likely due to the high level of potential within-speaker variation (particularly in F2) as a function of phonological context. Nonetheless, the results of both experiments highlight that LR output may be substantially affected by different definitions of the relevant population according to regional background even where the variable is not expected to display marked patterns of regional or social variation.

4.5 Chapter summary

Experiment (1): Multiple test sets

F1 and F2

- Mismatched SS scores stronger than Matched SS scores (by the equivalent of one order of magnitude).
- Manchester and Newcastle Mismatched DS scores weaker than Matched DS scores.
- Higher EER and C_{llr} values for Manchester and Newcastle Mismatched sets compared with Matched set.

F2-only

- Weaker LRs and worse system validity compared with F1 and F2 input across all sets.
- Manchester and Newcastle much closer to NZ in terms of the magnitude of scores and system validity.

Experiment (2): Multiple systems

- SS LLRs marginally stronger using Mixed system at both feature-to-score and score-to-LR stages.
- Magnitude of contrary-to-fact LLRs much higher for the Mixed system.
- C_{llr} worse using the Mixed system (1.19) compared with the Matched system (0.92).

Chapter 5

Regional Background: /aɪ/

The experiments in this chapter also explore the effects on LR output of different definitions of the relevant population with regard to regional background. The experiment in §4.3.2 was replicated using the formant trajectories of F1, F2 and F3 of /aɪ/ as input. The results of further experiments are also presented which consider how regional (*group*) and speaker-specific (*individual*) information are encoded in the individual formants of /aɪ/.

5.1 Introduction

The results of Chapter 4 highlighted that for a variable which is not expected to display marked patterns of sociolinguistic variation, LR output may be substantially affected by different definitions of the relevant population according to regional background. The experiments in this chapter explore the same issues as in Chapter 4 using the formant trajectories of /aɪ/ (PRICE; Wells 1982) as input, but improve on the methods in a number of ways. First, all of the data were extracted from varieties of BrEng. The results therefore have more direct implications for the collection of data for LR-based FVC in a BrEng context.

Second, the formant dynamics of /aɪ/ have been the subject of considerable study in FVC (McDougall 2004, 2006; Rose *et al.* 2006; Morrison 2008). This is because the frequency of /aɪ/ words is very high, while lexical variety of /aɪ/ words is relatively low

(Cruttenden 2001). Therefore, /aɪ/ tokens in comparable phonological environments are likely to be available in most FVC cases (e.g. *hi* and *bye*). /aɪ/ also has "reasonably easily measurably formants" (Rose *et al.* 2006: 330) which, for most varieties of English, display considerable movement within the acoustic space (in particular on the F1∼F2 plane) between onset and offset. Therefore, /aɪ/ offers considerable scope for individual differences in dynamic implementation.

Third, as well as potentially offering useful speaker discriminatory power, /aɪ/ is also expected to display more marked patterns of regional variation compared with /uː/ in Chapter 4. Therefore, relative to the results for /uː/ it is predicted that LR differences between Matched and Mismatched sets using F1 and F2 of /aɪ/ will be considerably greater. Comparison of the results of these experiments with those in Chapter 4 will provide insight into the potential extent to which LR output is affected by the definition of the relevant population for different linguistic-phonetic variables.

Finally, the acoustic information analysed in this chapter is expanded to include F3. As outlined in §3.3.1.1, there is considerable theoretical and empirical evidence to suggest that F3 (and higher formants more generally) offers greater speaker-specific information since it is not responsible for phonetic contrast (i.e. *speech* information) in the same way as F1 and F2. Further, unlike F1 and F2, F3 is not expected to display marked patterns of regional and social variation (although there are potentially confounding factors which may introduce systematic differences in F3 across sociolinguistic groups; e.g. VQ and vocal setting). Given that the differences in LR output from regionally Matched and Mixed systems were minimised with the removal of F1 in Chapter 4, it may be that LRs based on F3-only are relatively robust to differences in the definition of the relevant population. This would offer the potential for using general BrEng data (or even inappropriate data from another regional variety) in LR-testing using F3.

In this chapter three experiments are presented. Experiment (1) replicates §4.3.2 using all three formants (F1∼F3), F2 and F3 and F3-only of /aɪ/ as input. Results are compared across different input to assess the sensitivity of LR output (LLRs and system validity) to (a) **Matched** and (b) **Mixed** definitions of the relevant population with regard to regional background. Two further experiments are also presented which test claims about the regional/social (*group*) and speaker-specific (*individual*) information

encoded in the individual formants (in particular F3) of /aɪ/. The degree of regional information encoded in individual formants is explored in Experiment (2) (§5.3.2) using discriminant analysis (DA; see §3.3.1.1). Experiment (3) investigates the speaker discriminatory power of F3 relative to F1, F2 and a combination of all three formants in a homogeneous population of speakers. The results of Experiments (2) and (3) are then compared relative to the *group-individual* dichotomy (Garvin and Ladefoged 1963) which predicts that variables that encode speaker-specific information will be less susceptible to regional and social variation.

5.2 Method

5.2.1 Data

A total of 121 speakers from four varieties of BrEng were used. The data included the 97 DyViS Task 1 speakers from §3.3.1.4, and three datasets each containing eight speakers from Manchester (§3.1.3), Derby and Newcastle (§3.1.4). Eight DyViS speakers were initially extracted at random, to create a DyViS subset. These speakers were later combined with the Manchester, Derby and Newcastle sets to form a balanced, Mixed BrEng dataset. The speakers used for testing are considered relatively well matched in terms of age, sex and style, although the social class of the speakers in each set is potentially problematic. All of the DyViS speakers were students at the University of Cambridge and can broadly be defined as middle class. The Newcastle data consists exclusively of working class speakers, who display extensive use of localised variants (e.g. glottal reinforcement of medial plosives and centering diphthongs for /eɪ/ and /əʊ/: see Foulkes and Docherty 1999; Hughes *et al.* 2005). According to Haddican *et al.* 2013, the Manchester speakers can be classified as upper working or lower middle class. The Derby data are divided equally between working and middle class speakers, although Foulkes (p.c.) claims that there were not large linguistic differences between the class groups. Therefore, acoustic differences between the sets may also reflect class variation rather than purely regional variation.

5.2.2 Variation and change in /aɪ/

BrEng /aɪ/ is considered to be much more regionally variable than /uː/. Indeed, for many regional varieties of BrEng, /aɪ/ can be considered a linguistic *stereotype* (Labov 1971) owing to the degree to which speakers are aware of regional patterns. The extent of regional variation in BrEng /aɪ/ is highlighted by the isoglosses in Figure 5.1.

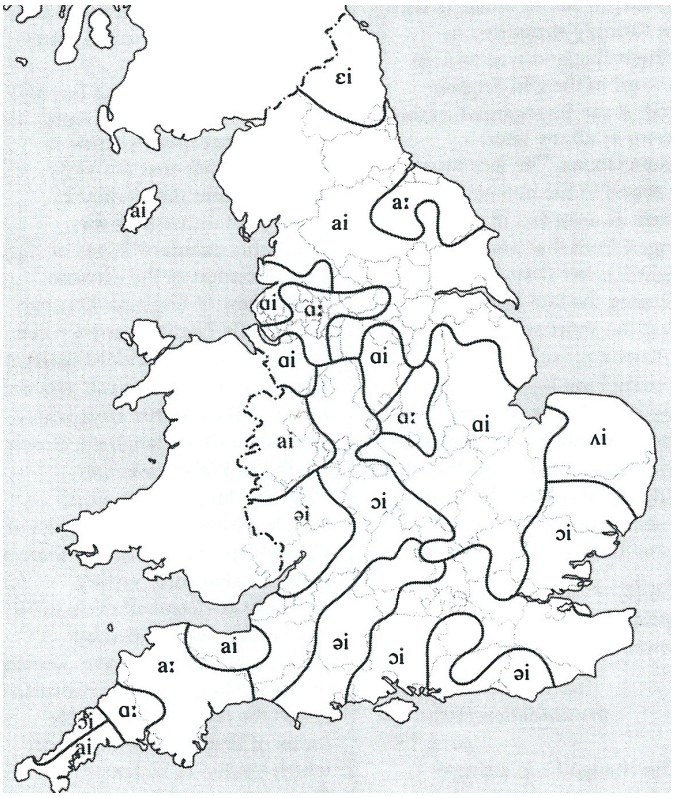

Figure 5.1: Dialect map of the British Isles with isoglosses marking regional variants of /aɪ/ (from Upton and Widdowson 2006: 32)

There are also specific predictions about patterns of variation across the four regional sets. Cruttenden describes modern RP (SSBE) /aɪ/ as having an open and centralised onset [ä] followed by an extensive glide towards [ɪ]. The realisation of SSBE /aɪ/ is also claimed to involve some degree of change in lip configuration from "a neutral to a loosely spread position" (Cruttenden 2001: 131-132) which may introduce further variation in F2 and F3 across the vowel (Stevens 2000). Considerably more variation is expected for Derby. Foulkes and Docherty (1999) describe three variants of /aɪ/ → [aɪ ∼ ɑ ∼ ɒɪ] in the speech of young males in Derby which are stratified by class. Working class males generally display a retracted onset position which may involve some lip

rounding. Foulkes and Docherty (1999: 50) report that for young middle class males around 60% of tokens in their data (the same as used here) were categorised auditorily as [aɪ].

Wells (1982) suggests that, consistent with patterns across northern England, the onset element in Manchester is typically a front open vowel [a]. Further, Wells (1982) claims that "diphthongs with a weakened second element ... occur widely as optional variants" (p. 150) in Manchester, with the most typical realisation of /aɪ/ being [aɛ] (p. 358). Hughes *et al.* (2005) state that /aɪ/ is predominantly realised as [ɛɪ] in Newcastle. However, Tyneside also has a stereotypical [iː] variant occurring in lexically restricted, high frequency words such as *alright* and *tonight*. According to Beal (2004), these variants are primarily associated with working class speakers. Finally, a process of allophonic variation in /aɪ/ has been reported for Newcastle English which is similar to the Scottish Vowel Length Rule (SVLR), but which affects vowel quality rather than length (Milroy 1995; Scobbie *et al.* 1999; Watt and Milroy 1999). Before voiceless consonants, /aɪ/ may be realised as [ɛɪ], while [aɪ] occurs before voiced consonants. However, Foulkes (p.c.) claims that there was little to no evidence of SVLR variation in the Newcastle data analysed in this chapter.

5.2.3 Dynamic formant extraction

The process of extracting formant data from the 97 DyViS speakers is described in §3.3.1.4. The Derby set consisted of existing dynamic data for the first three formants of /aɪ/ (18-43 tokens per speaker) from Rhodes (2009; extracted using the same scripts as in this thesis). Tokens with adjacent /r/ and /w/ were removed from this dataset (West 1999). Due to the high frequency of *like* in the recordings some tokens with onset-/l/ were included in the analysis, however coda-/l/ tokens were not included. The data were visually inspected and tokens with obvious measurement errors removed. Within-speaker z-scores were calculated at each +10% step and tokens with outlying values greater than ± 3.29 SDs from the mean were removed. The ten tokens per speaker with the lowest combined z-score across the three formant trajectories were used as input data.

/aɪ/ tokens were extracted from the resampled (11.025 kHz) Manchester (10-16 tokens per speaker) audio files, while tokens from the Newcastle (15-19 tokens per speaker) recordings were extracted to a new audio file to preserve sampling rate. Beyond the removal of tokens with adjacent /l r w/, the availability of only a small of numbers of tokens per speaker meant that there was no control over potential SVLR contexts. Tokens of /aɪ/ realised as [iː] were not included in the analysis. Given the potential sensitivity of /aɪ/ to phonological environment, it is expected that the lack of narrow controls will introduce extraneous within-speaker variation, which may generate weaker LRs and higher proportions of errors compared with more controlled data. Tokens were manually segmented following the criteria in §3.3.1.3. Dynamic formant data were then extracted with the script set to identify maximally between five and six formants (determined by-token) between 0 and 5 kHz. Ten tokens per speaker for the Manchester and Newcastle sets were identified based on the sum of their within-speaker z-scores.

All of the data from all sets were fitted with cubic polynomial curves and the coefficients used as input for LR computation.

5.2.4 Variability in the data

Figure 5.2 displays mean /aɪ/ trajectories within the F1∼F2 plane for the four regional sets (eight speakers), with mean mid-point values for reference vowels (FLEECE /iː/, GOOSE /uː/, NORTH /ɔː/ and TRAP /a/) based on measurements from the first 20 DyViS speakers (Simpson 2008; Atkinson 2009). The raw data revealed marked between-variety differences as predicted in §5.2.2. The DyViS data displayed considerable F1 and F2 movement within the acoustic plane, with the onset situated at a position more open and retracted from schwa [ɑ ∼ ä]. The offset was located towards a close-front position, although less peripheral than /i/.

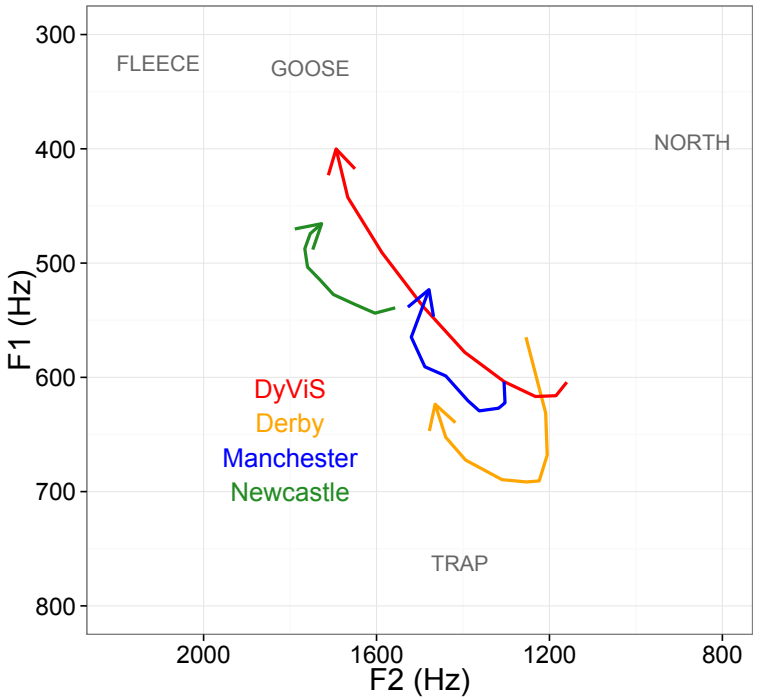

Figure 5.2: F1∼F2 plot of mean /aɪ/ trajectories for DyViS (red), Derby (orange), Manchester (blue) and Newcastle (green) (eight speakers per set, ten tokens per speaker) with mean mid-point values for FLEECE /iː/, GOOSE /uː/, NORTH /ɔː/ and TRAP /a/ (based on the first 20 DyViS speakers)

The mean trajectory for Derby revealed less F1 and F2 movement between onset and offset compared with DyViS, with mean F1 values across the trajectory consistently greater than 500 Hz. The mean trajectory for Manchester was consistent with Wells' (1982) description of an open [a]-like onset and an open-mid or centralised offset. The acoustic consequence of this is less F1 movement between the two targets compared with the DyViS data. Finally, the Newcastle data conformed to expectations in §5.2.2. The Newcastle data displayed the narrowest diphthong trajectory. The onset was phonetically much closer (i.e. higher F1) than for the other regional sets with a mean F1 of around 550 Hz. The offset was front (i.e. high F2), but with a relatively high F1 (F1 movement < 100 Hz between onset and offset), compared with the close-front offset of the DyViS set.

There was also a high degree of between-speaker variation within each group. Figure 5.3 displays mean F1∼F3 trajectories for /aɪ/ for each speaker grouped by regional set. The greatest homogeneity between speakers in mean trajectories was found for

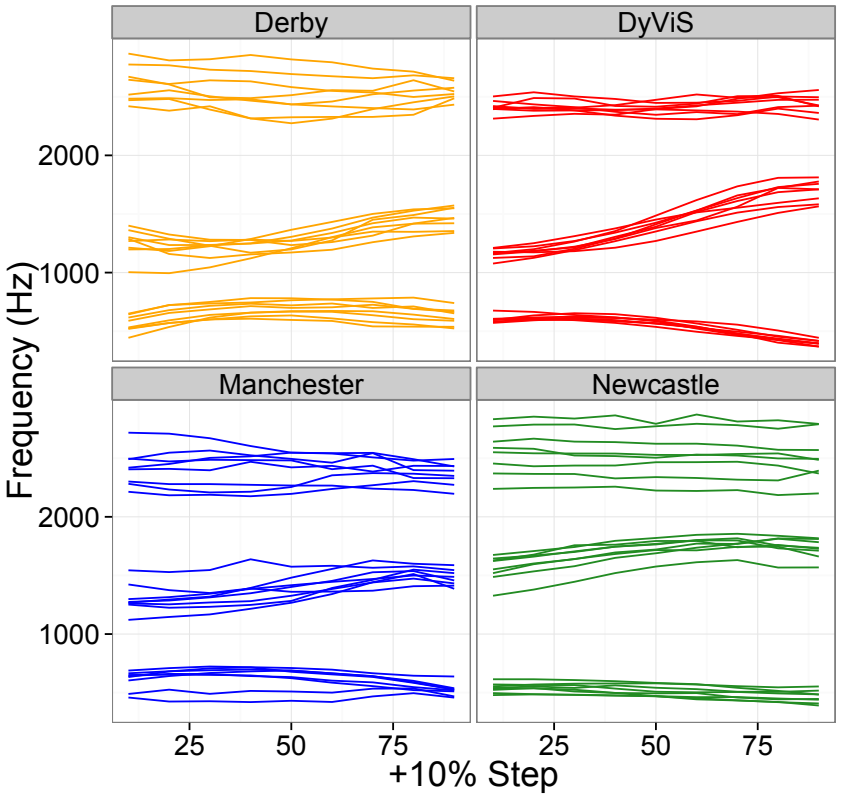

Figure 5.3: Mean F1∼F3 trajectories for each speaker by regional set (based on ten tokens per speaker)

DyViS, with values at any +10% step maximally spread over a range of around 300 Hz. Within the Derby, Manchester and Newcastle sets, greater between-speaker differences were found. The magnitude of the variation differed according to formant and regional group, although the widest range of between-speaker within-group variation was found in F3. However, even for F1 and F2, individual speakers in the Derby, Manchester and Newcastle sets displayed considerable divergence from the group pattern.

5.2.5 Experiments

For Experiment (1), 40 speakers were initially extracted at random from the remaining 89 DyViS speakers to function as test data. From the remaining 49 speakers, a further 32 were identified at random to function as **Matched** system data (development/reference speakers). The eight DyViS speakers extracted initially (§5.2.1) were combined with the eight speakers from each of the northern BrEng varieties to create a set of **Mixed** system

data containing 32 speakers. The same procedures as in §4.3.2 were then followed to evaluate the two systems at both the feature-to-score and score-to-LR stages. Cross-validated (§3.2.2.3) MVKD (§3.2.2.1) scores (32 SS/ 992 DS) were initially computed for the Mixed and Matched sets. Based on these scores, logistic regression calibration coefficients (§3.2.4.1) were calculated. MVKD scores were then computed for the test data (40 SS/ 1560 DS) using the Mixed and Matched sets as reference data, and calibrated using the coefficients generated from the appropriate development data. The experiment was run using F1∼F3, F2 and F3, and F3-only input.

Experiment (2) examines the extent to which regional information is encoded in the individual formants of /aɪ/, with a focus on F3. Linear DA (§3.3.1.1) was performed using the four sets of eights speakers from DyViS, Manchester, Derby and Newcastle. DA was used to assign cubic polynomial coefficients of individual formants from individual tokens to one of the four regional groups. The procedure uses leave-one-out cross-validation whereby the questioned token is not used to generate the four regional models against which it is compared. DA requires the number of input elements (i.e. features of a variable) to be less than the number of tokens per group (Tabachnick and Fiddell 2007: 23-24). Given that tokens were pooled by regional group (80 tokens per set; ten tokens per speaker), it was possible to include all cubic coefficients from F1∼F3 (4 features per formant) as input.

Finally, Experiment (3) considers the LR-based speaker discriminatory performance of individual formant trajectories of /aɪ/ using only the DyViS data. A homogeneous set was used to highlight speaker-specific rather than regional patterns across formants. The same 40 DyViS test speakers as in Experiment (1) were divided equally into development and test sets (20 per set). The remaining 57 speakers were used as reference data. Test scores were computed using MVKD and calibrated based on coefficients from scores for the development set.

5.3 Results

5.3.1 Experiment (1): Multiple systems

The distributions of LLRs are firstly considered for each combination of formants separately. The comparative performance of the Matched and Mixed systems with regard to validity (EER and C_{llr}) for the different input data is then considered.

F1, F2 and F3

Figure 5.4 displays the Tippett plot of LLRs from the Matched and Mixed systems using F1∼F3 input. There was considerable similarity in the distributions of SS LLRs across systems. The median SS LLR was +1.68 in the Matched condition and +1.49 in the Mixed condition, in both cases equivalent to *moderate* support for the prosecution. The two systems were also comparable in terms of the overall ranges of SS LLRs, with values extending maximally to greater than +3. However, the Matched condition (15%) recorded a higher proportion of misses than the Mixed condition (5%). Further, the magnitude of the errors using the Matched system was marginally higher with values approaching -1, although for both systems contrary-to-fact SS LLRs did not extend beyond *limited* support for the defence.

Marked differences across the systems were found in terms of the DS LLRs. The median DS LLR for the Matched system was -5.38 (*very strong* support), which was four orders of magnitude greater than the median for the Mixed system (-1.44; *moderate* support). There were also substantial differences in the maximum strength of support for the defence with values for the Matched system extending beyond -44 compared with -16 for the Mixed system. The Mixed DS LLRs also performed considerably worse in terms of categorical validity, with 20.7% of comparisons achieving contrary-to-fact support. For the Matched system, the false hit rate was 6.3%. Finally, the magnitude of the errors using the Mixed system was marginally higher with values extending to +2.88.

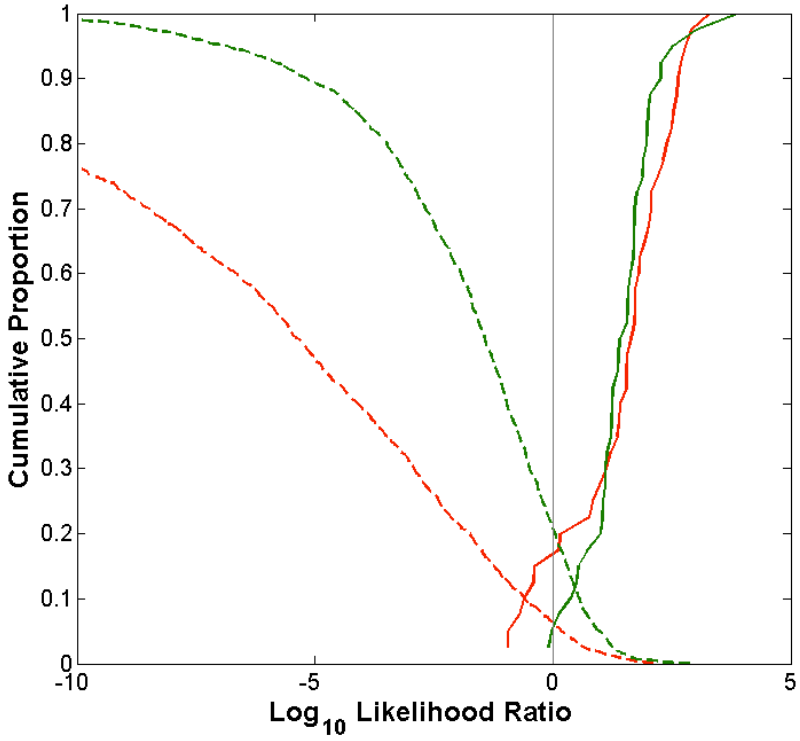

Figure 5.4: Tippett plot of SS and DS LLRs based on F1∼F3 trajectories from /aɪ/ using Matched (red) and Mixed (green) system data

F2 and F3

The omission of F1 reduced some of the differences between the two systems (Figure 5.5). As with F1∼F3, the distributions of SS LLRs from the Matched and Mixed systems based on F2 and F3 overlapped considerably. The median SS LLR in the Matched condition was +1.38, compared with +1.36 in the Mixed condition. Similarly the highest magnitude LLRs across both systems offered *moderately strong* support for the prosecution. Comparison with Figure 5.4 also shows that the miss rates were more similar with the omission of F1, with 12.5% of pairs in the Matched condition and 10% of pairs in the Mixed condition offering contrary-to-fact support. For both systems the magnitudes of the contrary-to-fact LLRs were greater without F1, although across systems their magnitudes were broadly similar.

Similar patterns in the distributions of DS LLRs are displayed in Figure 5.5 as those in Figure 5.4. DS LLRs were generally weaker when using the Mixed system, such that the median DS LLR was -1.28 (*moderate* support) compared with -4.24 (*very strong*

support) using the Matched system. Further, the strongest LLRs were found using the Matched system with values extending to -40 compared with -17 for the Mixed system. The proportion of false hits was again highest using the Mixed system (24.5%). Further, the magnitude of contrary-to-fact DS LLRs was greatest using the Mixed data with values extending to almost +4, compared with +2.33 for the Matched system.

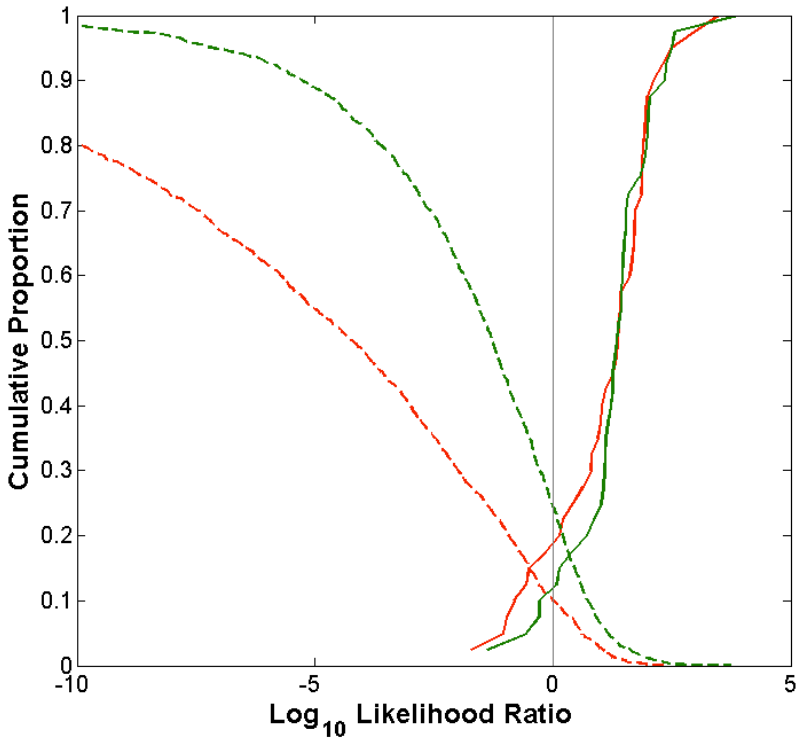

Figure 5.5: Tippett plot of SS and DS LLRs based on F2 and F3 trajectories from /aɪ/ using Matched (red) and Mixed (green) system data

F3-only

Figure 5.6 displays the Tippett plot based on F3-only input. The removal of F2 reduced the strength of the LLRs, offering further evidence to suggest that F1 and F2 are carriers of speaker-specific information for this vowel in these varieties. The removal of F2 also further minimised the effects of using Mixed system data compared with the Matched system. The median SS LLRs based on F3-only were very similar across systems, although in verbal terms they were equivalent to one order of magnitude weaker than with the inclusion of F1 and F2 (difference between *limited* and *moderate* support). The

overall ranges of SS LLRs were also broadly comparable, although the maximum LLR for the Mixed system (+3.01) was greater than that for the Matched system (+2.05). Contrary-to-fact SS LLRs were of a similar magnitude with no SS pairs achieving LLRs of less than -1. Unlike with the inclusion of F1 and F2, the miss rate was lower for the Matched system (10%) than for the Mixed (15%) system.

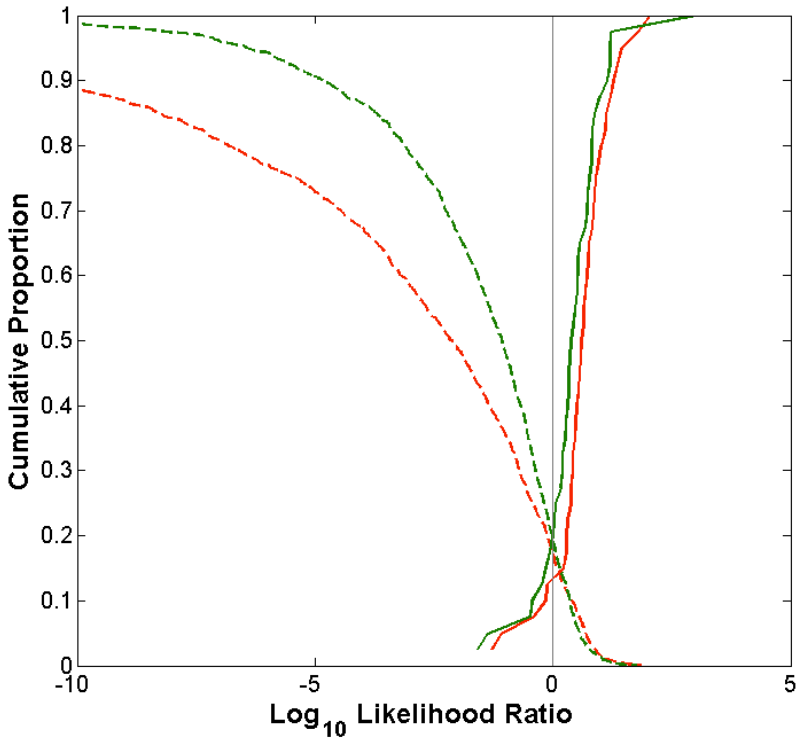

Figure 5.6: Tippett plot of SS and DS LLRs based on F3-only trajectories from /aɪ/ using Matched (red) and Mixed (green) system data

The differences between the Matched and Mixed systems were also less marked in terms of DS LLRs when using F3-only. DS LLRs for the Mixed system were generally weaker by only one order of magnitude compared with the Matched system (difference between *moderately strong* and *moderate* support). In terms of the maximal support for the defence, however, the large differences between the systems found using F1 and F2 were, to some extent, preserved using F3-only. The proportion of false hits was again lower using the Matched system (17.2%) compared with the Mixed system (19.2%), although performance was more similar than with the inclusion of F1 and/or F2.

Overall performance

Overall system performance was assessed using EER (Figure 5.7) and C_{llr} (Figure 5.8). Across all three sets of input data, EER was worse for Mixed system than for the Matched system. Differences between the systems based on F1∼F3 were relatively small (ca. 1%), reflecting the fact that the Matched system produced more misses and the Mixed system produced more false hits. The biggest EER difference between systems, however, was found when using F3-only input. This is partly due to the improvement in EER for the Matched system with the omission of F2 information.

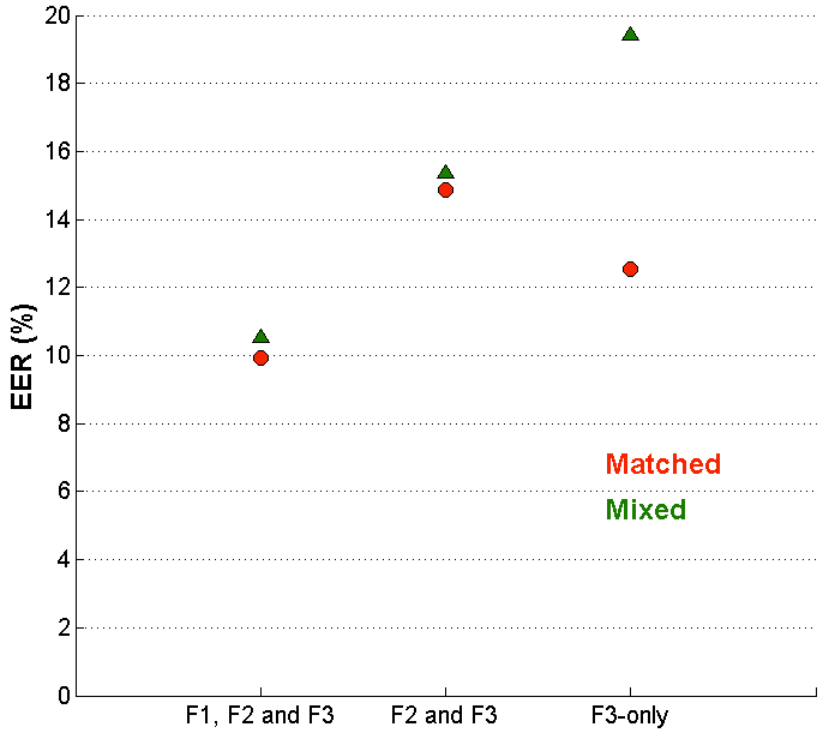

Figure 5.7: EER (%) based on F1∼F3, F2 and F3, and F3-only input from /aɪ/ using Matched (red) and Mixed (green) system data

The results based on C_{llr} were more systematic. Across both the Matched and Mixed systems, C_{llr} increased as the amount of acoustic input data was reduced. As with EER, C_{llr} was also consistently higher using the Mixed system than the Matched system. Interestingly, the smallest C_{llr} difference between the systems was found using all three formants as input. The difference between the systems increased as F1 was removed, and increased again with the removal of F2. This finding is contrary to earlier

predictions about the potential lack of regional stratification in F3, since the effect of regional differences on C_{llr} was greatest when using F3-only input.

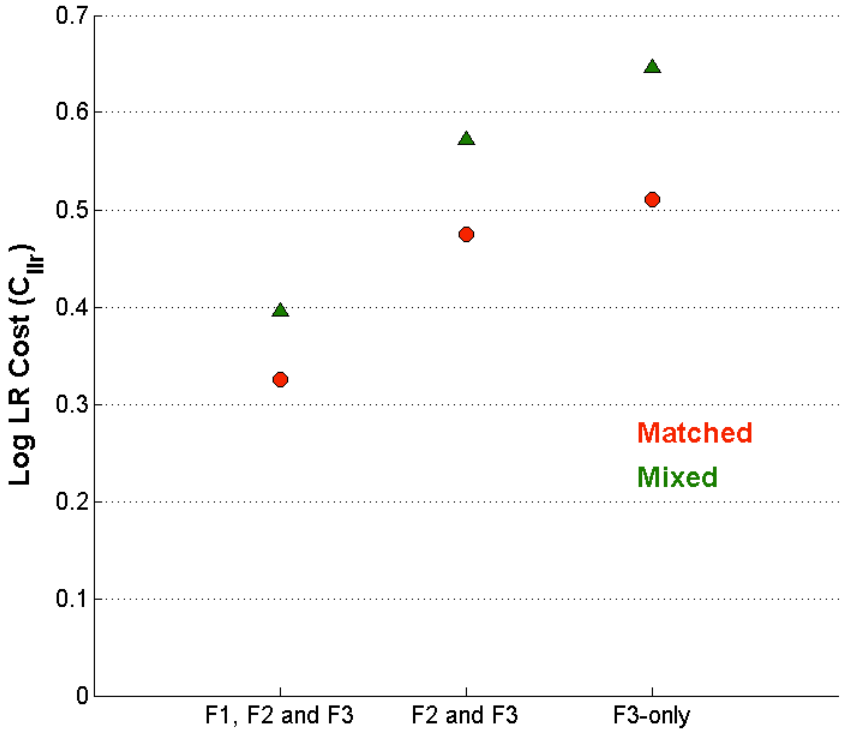

Figure 5.8: Log LR Cost (C_{llr}) based on F1∼F3, F2 and F3, and F3-only input from /aɪ/ using Matched (red) and Mixed (green) system data

5.3.2 Experiment (2): Regional patterns

In Experiment (1), the removal of F1 and F2 reduced the sensitivity of LR output to regional differences in the definition of the relevant population. Specifically, the distribution of SS LLRs produced by the Matched and Mixed systems converged as F1 and F2 were removed, while DS LLRs were found to become most similar across system using F3-only. Despite this, the largest differences between the systems in terms of validity were found using F3-only input. This finding suggest that the regional differences between the Matched and Mixed sets are captured, at least to some extent, using F3 and that such differences may substantially affect LR output. This section explores the issue of the regional stratification of the individual formants of /aɪ/ using the same SSBE, Derby, Manchester and Newcastle data.

Table 5.1 displays cross-validated classification rates based on DA for each formant. The classification rate is the percentage of the 320 tokens (10 tokens per speaker $\times$ 8 speakers $\times$ 4 sets) assigned correctly to the regional group of the speaker that produced the token. Since there are four possible groups, chance is 25%. The highest classification rate was achieved using F2 (64.7%), followed closely by F1 (63.8%). This means, as expected, that both F1 and F2 encode a high degree of region discriminatory information. The classification rate for F3 (40.6%) was considerably lower than for F1 and F2, but was better than chance. Therefore, as suggested by the results in Experiment (1), there is evidence that F3 contains information which is able to discriminate between regions. DA was re-run using individual coefficients to assess which elements of the formant trajectories carry the most region-specific information.

Table 5.1: Cross-validated classification rates of tokens correctly assigned to regional set based on DA using F1, F2 and F3

Formant	**Classification Rate**
F1	63.8%
F2	64.7%
F3	40.6%

Table 5.2 displays cross-validated classification rates for individual cubic coefficients from each formant of /aɪ/. In all cases, no one coefficient outperformed the combination of coefficients. This suggests that all of the coefficients provided some region-specific information. Interestingly, for all three formants the same general ordering of the coefficients was found. The cubic (a_4x^3) and quadratic (a_3x^2) terms, relating to the more phonetically fine-grained shape of the trajectory (see §3.3.1.4), generated the lowest classification rates. For all three formants the slope (a_2x) and intercept (a_1) terms generated the highest classification rates. Therefore, predictably, for all three formants, information relating to absolute frequency and the magnitude of onset to offset movement were the best predictors of regional background.

Table 5.2: Cross-validated classification rates of tokens correctly assigned to regional set based on DA using individual cubic coefficients from F1, F2 and F3

Formant	Coefficient	Classification Rate
F1	a_4x^3	29.1%
	a_3x^2	32.8%
	a_2x	38.4%
	a_1	38.4%
F2	a_4x^3	30.9%
	a_3x^2	29.7%
	a_2x	40.6%
	a_1	41.9%
F3	a_4x^3	25.6%
	a_3x^2	25.4%
	a_2x	27.5%
	a_1	34.7%

5.3.3 Experiment (3): Speaker-specific patterns

Figure 5.9 displays LLRs based on input from the first three formants of /aɪ/ analysed individually and in combination using DyViS speakers only (20 development/ 20 test/ 57 reference). The SS median LLRs based on F1-only and F2-only were within the same order of magnitude (*limited* support), although numerically strength of evidence was generally better using F2-only. The ranges of SS LLRs for F1-only and F2-only were also broadly equivalent, with values spread from marginally less than zero to around +1. Although the median SS LLR for F3-only was also located within the zero to +1 range, the absolute numerical value was much closer to +1. Further, the maximum strength of SS evidence for F3-only was +2.72 (*moderately strong* support) indicating that F3 in some cases outperformed F1 and F2 by up to two orders of magnitude. The strength of SS evidence was, however, greatest when using a combination of all three formants, with LLRs generally one order of magnitude higher compared with any formant individually (*moderate* support). The proportion of misses also decreased from maximally 15% using F1-only to 5% using all three formants.

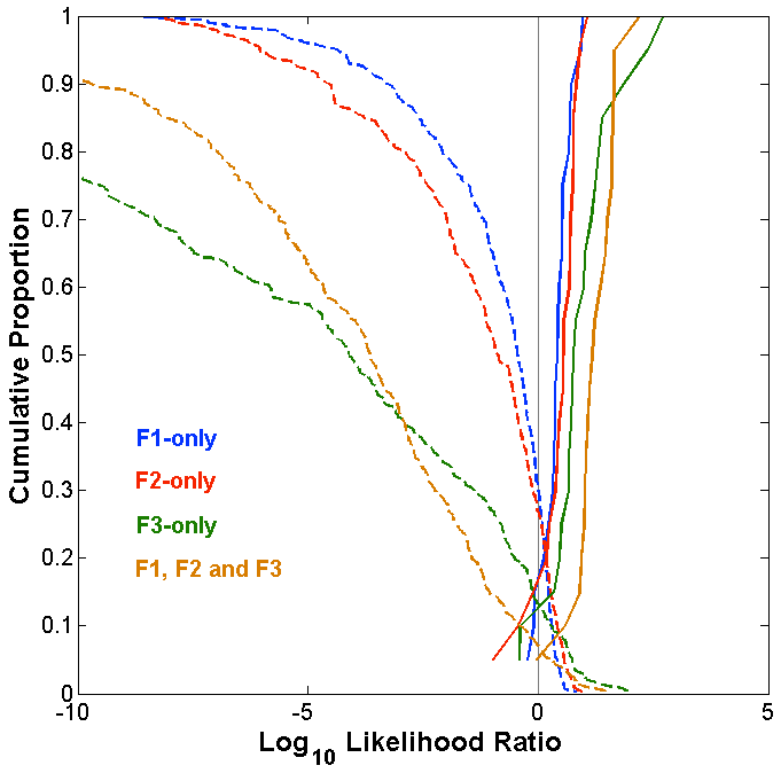

Figure 5.9: Tippett plot of SS and DS LLRs using F1-only (blue), F2-only (red), F3-only (green) and a combination of the three formants (orange) of /aɪ/ from DyViS

Similar results are revealed in the distributions of DS LLRs. Numerically, the weakest DS LLRs were achieved using F1-only, followed by F2-only. The difference in median values was equivalent to one order of magnitude from *limited* (F1-only) to *moderate* (F2-only) support for the defence. However, unlike the SS comparisons, F3-only input generated generally stronger LLRs than the combination of the three formants. The median DS LLR based on F3-only was -4.11 (*very strong* support), compared with -3.66 (*strong* support) using F1∼F3. Further, the range of DS LLRs for F3-only extended to -35.4, compared with -19.5 for F1∼F3. However, F3-only input also generated a higher false hit rate, as well as higher magnitude contrary-to-fact DS LLRs compared with the combination of formants.

Figure 5.10 displays EER and C_{llr} values for each of the four sets of formant data. Despite achieving somewhat weaker DS LLRs compared with F3-only, the combination of formants produced the best performing system in terms of both EER and C_{llr}. F1∼F3 outperformed F3-only by 5% in terms of EER and 0.2 in terms of C_{llr}. The worst performance was found using F1-only and F2-only, which performed similarly,

achieving EER values of around 20% and C_{llr} values of around 0.6. Consistent with patterns in Experiments (1) and (2), the improved performance of the combination of formants over F3-only in terms of the strength of SS LLRs and system validity provides evidence that F1 and F2 do carry speaker-specific information. However, given that F1 and F2 encode so much *speech* information (i.e. they are carriers of contrast), their value as individual discriminants is relatively minimal. Clearly in terms of individual formants, F3 dominates with regard to speaker discrimination.

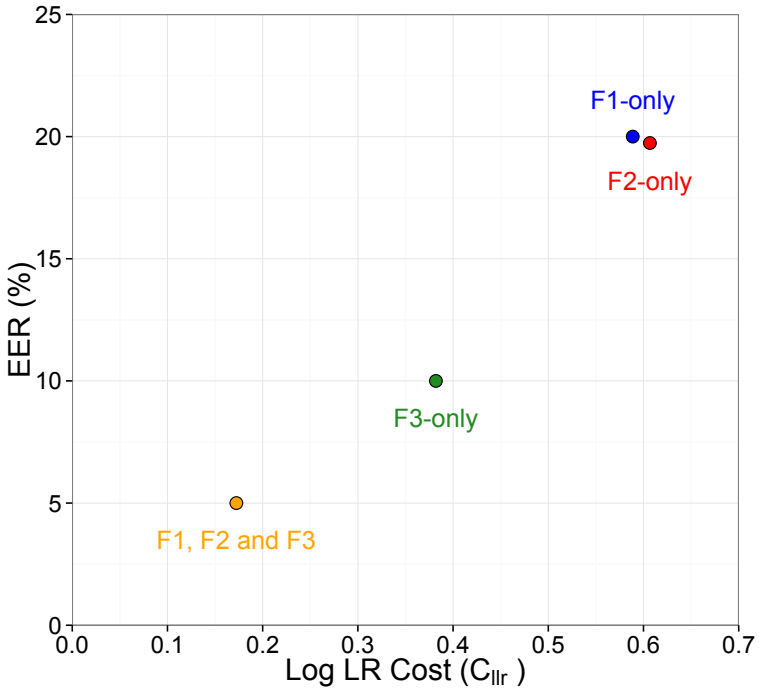

Figure 5.10: Log LR Cost (C_{llr}) plotted against EER (%) for different DyViS formant input for /aɪ/

5.4 Discussion

The results of Experiment (1) revealed a number of effects of using regionally Matched and Mixed BrEng data at both the feature-to-score and score-to-LR stages of system testing using /aɪ/. Consistent with predictions in §5.1, the effects of using regionally Mixed system data were considerably more severe for /aɪ/ than for /uː/, owing primarily to the regional variation encoded in /aɪ/ in BrEng. The distributions of SS LLRs were generally comparable in terms of the median LLR and overall range when using the

Matched and Mixed system. However, DS LLRs were weaker by up to four orders of magnitude in the Mixed condition (using F1∼F3). Further, consistent with the results in §4.3.2, validity was consistently worse (by up to 7% EER and 0.15 C_{llr}) when using the Mixed system compared with the Matched system.

The removal of F1 and then F2 in Experiment (1) generated lower magnitude LLRs and generally worse system validity across both systems. This, along with the results of Experiment (3), suggests that F1 and F2, which are primarily thought to encode phonetic contrast and systematic regional and social variation, are capable of carrying considerable speaker discriminatory information. Further, the removal of F1 and F2 in Experiment (1) reduced the divergence between the Matched and Mixed systems in terms the distributions of LLRs, such that LLRs were most similar across systems when using F3-only input. These results suggest that there may be a trade-off between the speaker discriminatory potential that lower formants (F1 and F2) provide and the regional sensitivity they introduce into the LR-based analysis. That is, with the removal of F1 and F2, the strength of evidence and overall system performance may be lower, but the effects of regional variation, at least in terms of the magnitudes of the LLRs themselves, may be minimised.

Somewhat different patterns were revealed in terms of the Matched and Mixed validity across the three sets of /aɪ/ input. The EER for the Mixed system was only marginally higher than that of the Matched system when using all three formants and with the removal of F1. However, the largest difference between the systems in terms of EER was found when using F3-only (c. 7%). Similarly, the smallest difference between the systems in terms of C_{llr} was found using F1∼F3, followed by F2 and F3. As with EER, the largest C_{llr} difference between systems was found using F3-only (c. 0.15). This finding runs contrary to the earlier prediction that LR output based on F3 may be most robust to different definitions of the relevant population based on the hypothesis that it encodes more information relating to the *individual* rather than regional and social information relating to the *group* (Garvin and Ladefoged 1963).

In Experiment (2), the cubic coefficients of F1 and F2 were both able to correctly assign around 64% of the 320 tokens to the regional group (four regional groups) of the speaker, and both outperformed F3. This suggests, predictably, that F1 and F2 (and

in particular the intercept (absolute frequency) and slope elements of the trajectory) are primarily responsible for the differences between the four sets (as shown in Figure 5.2). F3 generated a classification rate of 40.6% which, although worse than F1 and F2, was better than chance (25%). Further, when analysing the individual elements of the trajectory using DA, the intercept generated the highest classification rate compared with coefficients relating to the dynamics of the trajectory. This suggests that F3 does encode some region-specific information primarily in the absolute frequency element of the trajectory. This may be due to intrinsic factors (i.e. an inherent property of F3 itself) such as VQ and vocal setting (see Stevens and French 2012), as well as extrinsic factors (i.e. extraneous) such as correlation with F2 (although no consistent correlations between elements of F2 and F3 were found when this was tested using these data). Formal analysis of these factors was not possible, however, due to the small number of speakers and regional sets available.

Despite evidence of region-specific patterns of F3 variation, consistent with previous studies, in Experiment (3) F3 outperformed F1 and F2 in terms of the magnitude of LLRs and system validity. There was also evidence of speaker-specificity in the lower formants, with F1∼F3 generating higher magnitude SS LLRs and better overall system performance than any individual formants. However, the addition of F1 and F2 to F3 did generate lower magnitude DS LLRs. The combined results of Experiments (2) and (3) suggest that for F3, Garvin and Ladefoged's (1963) *group-individual* distinction is a continuum rather than a dichotomy, since F3 was found to encode at least some regional information along with considerable speaker discriminatory power. More importantly when considered in terms of the results of Experiment (1), it is clear that the inevitable regional and social information to which linguistic-phonetic variables respond may affect different elements of LR output (e.g. magnitude of LLRs, validity) in potentially unpredictable ways and to unpredictable extents. Potential explanations for the results in §5.3.1 are offered in §11.1.

5.5 Chapter summary

Experiment (1): Multiple systems

- Distributions of SS LLRs broadly similar across Matched and Mixed systems for F1∼F3, F2 and F3 and F3-only input.
- DS LLRs weaker using Mixed system compared with Matched.
- Convergence in the distributions of LLRs as F1 and then F2 were removed.
- Largest validity differences between Matched and Mixed systems using F3-only.

Experiment (2): Regional (group) patterns

- Absolute frequency and slope elements of F1 and F2 best predictors of regional background.
- F3 classification rate (40.6%) better than chance (25%) with intercept the strongest carrier of regional information.

Experiment (3): Speaker (individual) patterns

- F3 outperforms F1 and F2 in LR-based speaker discrimination using a sociolinguistically homogeneous set of speakers.
- Evidence of *group-individual* continuum rather than dichotomy.

Chapter 6

Regional Background: Cepstral Coefficients and Derivatives

This chapter explores the sensitivity of LR output to regional variation using ASR variables as input. Calibrated GMM-UBM (§3.2.2.1) LLRs were computed for a set of regionally homogeneous test data using multiple systems defined as: (a) **Matched**: development and reference data of the same regional background as the test set, (b) **Mixed**: regionally mixed development and reference data, and (c) **Mismatched**: development and reference data containing speakers of a different regional background from the test set. Testing was performed using cepstral coefficients (CCs) and derivatives (delta (D) coefficients and delta-delta (A) coefficients) extracted from the Mel-frequency cepstrum (MFC) and the linear prediction cepstrum (LPC) (see §3.3.2).

6.1 Introduction

This chapter develops on the results of Chapters 4 and 5 to explore the sensitivity of LR output to regional variation in the definition of the relevant population using ASR variables: namely CCs and derivatives (§3.3.2) extracted holistically from across speech samples. Cepstral input was used since it was not expected to display the same sensitivity to sources of social and stylistic variation as linguistic-phonetic variables. There are two reasons for this claim, relating to the internal structure of the cepstrum

and of CCs, and the way in which such data are extracted and typically analysed in ASR systems.

Firstly, as shown in Equation 3.12, the cepstrum is a representation of the power spectrum of a signal based on an inverse discrete Fourier transform (DFT) from which CCs are extracted using discrete cosine transforms (DCTs). Therefore, the CCs themselves are linguistically abstract in the sense that they do not have direct articulatory correlates in the way that formants do (although see Clermont and Itahashi 1999, 2000; Clermont 2013). Further, as outlined in Rose (2013a), CCs capture information about spectral shape rather than just spectral peaks (as in formants), and therefore have the potential to encode considerably more information about the *individual* useful for speaker discrimination. Finally, as cited in Rose (2013a: 81), the cepstral smoothing involved in extracting CCs displays "strong immunity to non-information variabilities in the speech spectrum" (Rabiner and Juang 1993). Thus, smoothing is able to better preserve spectral differences which can be attributed to *speech* and *speaker* (both *group* and *individual*).

Secondly, as outlined in §1.1.4, ASR systems typically analyse CCs holistically from frames across an entire speech sample. This introduces further abstraction into the analysis, since the data are not extracted from linguistically meaningful units of speech (although segmental cepstral analysis is possible; Rose 2011a). When analysed holistically, a multivariate model of CCs from across a sample captures overall physical properties of the supralaryngeal vocal tract as well as the long-term configuration of articulators. Thus, it is predicted that CCs analysed holistically will not be as sensitive to *group* variation in the realisation of individual phonemes (such as those in §4.2.2 and §5.2.2), although CCs are likely to be sensitive to regional differences in long-term vocal setting (e.g. velarised setting in Liverpool English). Such assumptions have led to claims that ASR systems based on cepstral input are robust to the sources of structured variation outlined in §2.2.5, such as regional background. BATVOX, specifically, is claimed to be "language and speech independent and thus deliver(s) results irrespective of the language or accent used by the speaker."[18] If this is the case then it may be possible to use a dataset of sociolinguistically heterogeneous speakers when performing

[18]Agnitio (2013). Solution Brief: Criminal ID. `http://www.agnitio-corp.com/sites/default/files/SOL_BRIEF_Criminal_ID.pdf` (accessed: 27th January 2014).

LR-based testing using holistic cepstral input.

However, such claims are dependent on two issues. First, regional (and social) differences in the distributions of cepstral coefficients in the population must be shown empirically to be small. One field in which regional variation in cepstral variables is of central concern is the development of speech and accent recognition systems. Huang *et al.* (2004) investigated the potential of using accent-dependent speech recognition systems to improve speech recognition performance using Mandarin Chinese. They developed a GMM-based (32 Gaussians) accent recognition system based on 12 MFCCs and derivatives extracted holistically, as a means of classifying speakers into one of four accent groups, before then applying the appropriate accent-dependent system for speech recognition. Based on training models containing 300 speakers (males and females), between 77.5% and 98.5% of the 60 test speakers were assigned to the correct regional group. Similar findings are presented in Yan *et al.* (2012), although improved classification rates are reported in Huckvale (2004, 2007) using a metric based on text dependent segmental cepstra (see also Brown and Wormald 2014). These results confirm Salvi's (2003) claim that accent variations "have proved to be important variables in the statistical distribution of the acoustic features usually employed in ASR" (p. 1149). Clearly, cepstral coefficients when analysed both globally, and in particular segmentally, can encode sufficient information to achieve relatively high closed-set classification of speakers according to regional background.

This leads to the second more specific issue: the extent to which such language and accent information affects FVC systems. Van Leeuwen and Bouten (2004) present the results of the NFI-TNO forensic speaker recognition evaluation based on 12 ASR systems (similar to the NIST evaluations, but based on Dutch wire-tap recordings from real cases). Two conditions in the evaluations considered cross-language suspect and offender samples. EERs were up to 9% greater for cross-language conditions (EERs = 20-44%) compared with same-language comparisons (Dutch-Dutch) (EERs = 12-35%). Similarly Przybocki *et al.* (2007) maintain that "performance (of ASR systems) is clearly superior for ... matched trials (same-language) than for the unmatched (cross-language)" (p. 1957). However, Künzel (2013) found that by applying a normalisation procedure (Lu *et al.* 2009) using 75 bilingual speakers across seven languages, EERs

for cross-language comparisons may be equal to, or in some case better than, those based on same-language comparisons. These studies indicate a degree of sensitivity to language variation in cepstrum-based ASR, but do not directly test the question of the definition of the relevant population where suspect and offender language (or more specifically regional background) are matched.

This issue has received only a limited amount of attention. Moreno *et al.* (2006) considered the sensitivity of LR output using BATVOX to differences in the regional make-up of the reference data. Using a test set of 43 Spanish speakers from Andalusia, LRs were computed using three sets of test data (50 speakers): one matched (Andalusian) and two mismatched (Castilian and Galician Spanish). Based on 19 MFCCs and Ds, relatively small EER differences were found between sets, although the matched data generally produced the best performance. When using the optimised reference data option in BATVOX, EER was found to improve by a further 1.5%. These findings lead Moreno *et al.* (2006) to conclude that "it looks like dialect influence is not a relevant variable for (A)SR systems." However, the study did not consider effects of mismatch on the magnitude of LRs or on C_{llr}. Further, the degree of sociolinguistic homogeneity in the data, other than regional background, was not made explicit.

Harrison and French (2012) present an exploration into the specific issue of cepstral regional variation in British English (BrEng). They calculated Kullback-Leibler distances (from BATVOX) based on MFCCs between a set of 97 DyViS (§3.1.1) speakers and a set of 118 speakers from multiple accent backgrounds (containing Manchester, Northern Irish, south east and west Yorkshire speakers). For the set of mixed accent data, recordings were taken from real police interviews. They found differences between the DyViS and the mixed set in terms of the overall distributions of distances, as well as differences between groups when the 118 mixed speakers were categorised separately by accent. The differences lead Harrison and French to conclude, contra to the claims of Moreno *et al.* (2006), that Batvox is "accent sensitive", possibly as a result of the MFCCs capturing variability in "vocal tract settings ... across accents." However, the extent to which such variability affects LR output was not tested.

The studies of Moreno *et al.* (2006) and Harrison and French (2012) are developed in this chapter by replicating the experiments in §4.3.2 and §5.3.1 using Mel-frequency

(MFC) and linear prediction (LPC) cepstral input (for an overview of the differences between MFC and LPC input see §3.3.2.1). CCs, Ds and As were extracted holistically from the MFC and LPC and analysed in various combinations. GMM-UBM (§3.2.2.1) LLRs were computed for a homogeneous set of test data using systems (development and reference data) based on three regional definitions of the relevant population. As in Chapters 4 and 5, (a) **Matched** and (b) **Mixed** systems were used. The Matched system represents a narrow and appropriate definition of the relevant population with regard to regional background according to that of the offender. The Mixed system represents the current approach to defining the relevant population based on language and sex. The test data in this chapter were also evaluated using multiple (c) **Mismatched** systems. These systems represent narrowly but inappropriately defined relevant populations relative to the regional background of the offender. The use of Mismatched systems is intended to account for the paradox in FVC that without knowing who the offender is, the population of which he is a member cannot be known for certain. Therefore, it is possible that the analyst would define the regional background of the offender narrowly but incorrectly.

6.2 Method

6.2.1 Data

The data in these experiments were extracted from the TIMIT (§3.1.5) database of North American English (AmEng). TIMIT was chosen primarily because it contains a large number of speakers (438 male speakers) from a large number of different regional backgrounds (eight regional groups). This allows for large-scale testing with multiple sets representing different definitions of the relevant population. However, as highlighted by Campbell and Reynolds (1999), TIMIT is limited for the purposes of evaluating the performance of speaker recognition systems. It contains only highly controlled read speech in the form of ten (randomised) sentences per speaker, meaning that there is relatively little available data for each speaker. Further, the samples are high-quality wideband recordings made in a studio in a single session. In this way, they do not reflect typical mismatch conditions in forensic casework.

A number of other databases were considered, including Switchboard I-II and POLYCOST (see Campbell and Reynolds 1999). Although in many respects these databases are more forensically realistic, none of the available alternatives fulfilled the essential requirements of a large number of speakers controlled for regional background within a single language. Therefore, despite the limitations of TIMIT, it does allow for the research questions in this chapter to be tested, initially under optimal experimental conditions. The results are interpreted in light of these limitations (see §6.4).

From the entire TIMIT database, the 22 speakers from dialect region (DR) 8 (Army Brat) (see §3.1.5) were firstly removed from the analysis. This decision was based on the small number of available speakers, given that so many speakers were available for the other regions. DR 3 (North Midland) was identified as the Matched set since it contained the largest number of speakers (79). 25 speakers were first identified at random from DR 3 to function as test data. The same test data were used throughout this chapter. From the remaining 54 speakers, 28 were identified at random to act as a **Matched** system which functioned as both development and reference sets. For each of the other six DRs, 28 speakers were extracted at random to form six **Mismatched** systems (development and reference data). Six speakers were then chosen at random from the Matched set and each of the Mismatched sets to create a **Mixed** system (development and reference data) containing 28 speakers. In this way the Mixed set was regionally balanced. A total of 221 TIMIT speakers were used in this chapter (one test set = 25 speakers; eight development/reference sets = 28 speakers each).

6.2.2 Preparation of samples

The TIMIT database contains individual sounds files for each sentence produced by each speaker. Therefore, for each speaker the sound files for the ten sentences were compiled using PRAAT to create a single sample. To ensure that cepstral data were extracted only from the speech-active portions of the sound files, silences were removed using Morrison's Sound File Cutter Upper software[19] in MATLAB. The software performs a Root Mean Square (RMS) (Johnson 2008: 31-33) amplitude analysis across the speech

[19]Morrison, G. S. (2010). Sound file cutter upper. `http://geoff-morrison.net/#CutUp` (accessed: 22nd January 2014).

signal. The default setting in the software defines the threshold between silence and non-silence as:

$$\left((max_{RMS_{amp}} - min_{RMS_{amp}}) \times \frac{1}{3}\right) + min_{RMS_{amp}} \tag{6.1}$$

such that portions of the signal with amplitude greater than threshold are classed as speech and portions with amplitude lower than threshold classed as silence. Following Künzel (1997), a pause was defined as a period of silence greater than 100ms in duration. This removed larger periods of silence between sentences but preserved speech-related silences such as the hold phases of stops (Figure 6.1).

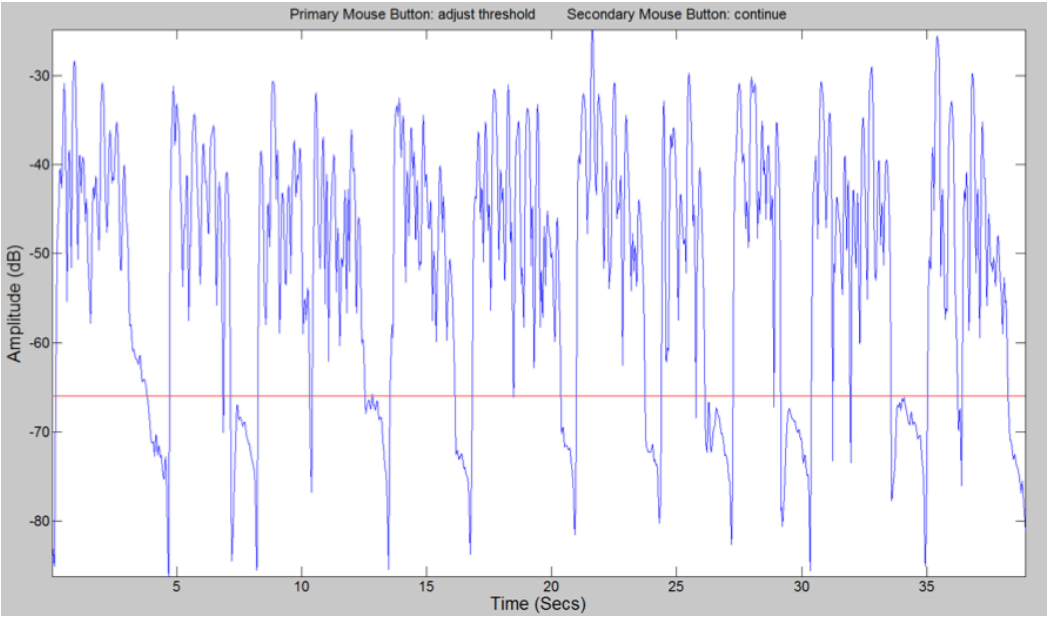

Figure 6.1: RMS amplitude analysis using the Sound File Cutter Upper software for speaker MCEW0 007 from DR 2 with the default threshold between silence and non-silence marked by a red line

To check the appropriateness of the default threshold in Equation 6.1, two tests were performed. A small number of sound files were manually edited in PRAAT and the output compared with that of the automatic approach. The software with the default threshold setting performed very well, removing the same silences as the manual segmentation without removing any low amplitude speech. The success of the software in this instance is primarily due to the relatively high quality of the recordings. Different settings for defining silence and non-silence were also tested in the software itself. A higher amplitude threshold generally resulted in low intensity speech being classed as

silence (such as in voiceless fricatives), while a lower threshold underestimated the true duration of the silent portions.

The software saves the speech active portions between silences as individual sound files. Following the same procedure as above, single sound files were created for each speaker by compiling the individual silence-free samples using PRAAT. The resulting samples were each around 30s in duration (c. 15s per suspect and offender sample when divided in half). This was considered adequate given that between 13 and 15s of net speech was used to compute LRs using the generic MFCC system in Lindh and Morrison (2011).

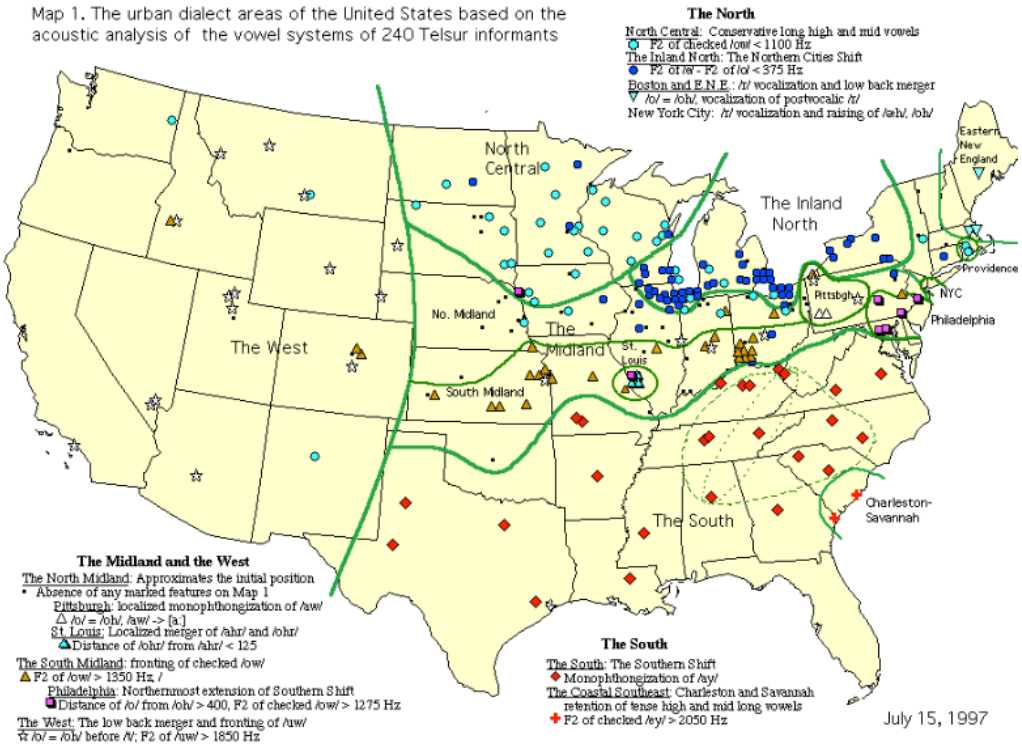

Figure 6.2: Map of the major urban dialect areas of North American English as identified through analysis of 240 participants (marked as points on the map) as part of the Telsur project (Labov *et al.* 1997)

6.2.3 Linguistic differences between dialect regions

A potential issue with TIMIT for investigating issues of regional variation is the extent to which the seven DRs represent linguistically distinct regional varieties. This is

especially relevant given that there is expected to be less marked patterns of regional variation for AmEng, compared with BrEng. Figure 6.2 displays the major urban dialect areas of AmEng as identified in the Telsur project (Labov *et al.* 1997); data which were subsequently used to develop the Atlas of North American English (ANAE) (Labov *et al.* 2006). The data were generated through auditory analysis of the vowel systems of 240 speakers from urban areas (two speakers per area) and larger metropolises (between four and six per area) within North America. Figure 6.2 includes the 'three dialects' of AmEng (Inland North, South and West) identified in Labov (1991), the Midland area, identified by Kurath and McDavid (1961), as well as the North Central, Philadelphia, New York and New England regions.

Table 6.1: Number of development, test and reference speakers used in each system within each experiment

TIMIT DR Number	**TIMIT DR**	**Telsur (Labov *et al.* 1997)**
1	New England	New England
2	Northern	North Central + Inland North
3	North Midland	Midland (inl. North + South Midland)
4	South Midland	The South
5	Southern	The South
6	New York City	New York City
7	Western	The West

Comparison of Figures 3.2 and 6.2 shows that there is general geographical agreement between Telsur and TIMIT, with minor differences primarily in naming conventions (see Table 6.1). The boundary defining the West is almost exactly the same across both TIMIT and Telsur, consisting of an area which includes California, Arizona, New Mexico, Nevada, Utah, Colorado, Oregon, Idaho, Wyoming, Washington and Montana. Consistent with Figure 6.2, New York and New England are also defined as individual DRs. The large area in Figure 6.2 defined as the South almost precisely contains within it the South Midland (4) and Southern (5) DRs of TIMIT. This suggests that it may be preferable to consider these sets as linguistically homogeneous. Further, in Figure 6.2 the North Central and Inland North areas are classified as separate linguistic varieties, where this is a single DR in the TIMIT data (Northern). However, evidence from ANAE

(2006: 204) indicates that the Northern Cities Shift is also present across the North Central area in Figure 6.2. Therefore, the TIMIT definition of a single DR encompassing the North Central and Inland North areas may be linguistically preferable.

Having identified the North Midland (DR 3) region as the test data, it is also necessary to assess the extent to which this constitutes a single variety of AmEng. Figure 6.3 displays this area using the same data and boundaries as in Figure 6.2. The definition of the North Midland area in TIMIT is clearly different to the North Midland area in Figure 6.3, but does overlap almost completely with the Midland area. Therefore, Figure 6.3 raises possible issues with DR 3 as a single linguistic entity. There are potential differences between the North and South Midland areas (Figure 6.3) although such variation is expected to be relatively subtle (Labov *et al.* 1997). Further, Philadelphia is identified as a separate linguistic area within the Midland DR in Figure 6.2 and so speakers from Philadelphia included in TIMIT DR 3 may introduce greater between-speaker variation into this set.

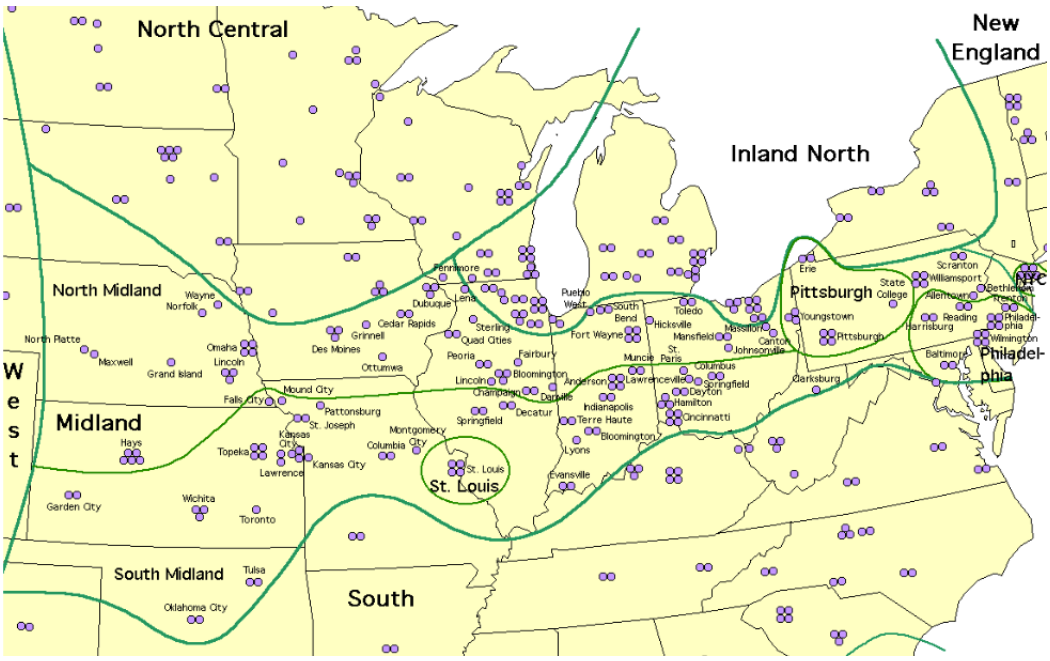

Figure 6.3: Map of the North Central, Inland North, New England, New York, Midland and South dialect areas as identified by Labov *et al.* (1997)

Evidence of linguistic differences between AmEng DRs make it possible to predict which Mismatched systems should generate the most divergent LR output relative to the Matched system, assuming the cepstral input captures linguistically salient regional

information. Figure 6.4 displays what Labov *et al.* (1997) refer to as a *phonological taxonomy* containing the primary vocalic differences between the major AmEng dialects. As a means of interpreting Figure 6.4, Table 6.2 displays the symbols used by Labov *et al.* (1997) (and used extensively in sociolinguistics in North America) relative to the equivalent lexical set (Wells 1982) and IPA phonemic transcription. The IPA phoneme symbols are used throughout this chapter for consistency.

Table 6.2: Transcription conventions used by Labov *et al.* (1997) with the equivalent lexical set (Wells 1982) and IPA phonemic transcription

Labov *et al.* (1997)	**Lexical Set**	**Phonemic Transcription**
/iy/	FLEECE	/iː/
/i/	KIT	/ɪ/
/ey/	FACE	/eɪ/
/e/	DRESS	/e/
/æ/	TRAP	/æ/
/ɜ/	SCHWA	/ə/
/æh/	START	/ɑː/
/o/	LOT	/ɒ/
/aw/	MOUTH	/aʊ/
/oh/	THOUGHT	/ɔː/
/ʌ/	STRUT	/ʌ/
/ow/	GOAT	/əʊ/
/u/	FOOT	/ʊ/
/uw/	GOOSE	/uː/

The TIMIT West (DR 7) set should display the greatest (linguistic) similarity to the Matched Midland (DR 3) set, since it is a related strand of the general Midland region identified by Labov *et al.* (1997). There should also be linguistic similarity between DR 3 and DRs 4 (South Midland) and 5 (Southern) since these three regions share laxing of long high and mid vowels (although differ on a number of other dimensions). As outlined earlier, Figure 6.2 suggests that, on the basis of linguistic evidence, DRs 4 and 5 should behave in similar ways. The most divergent results should be found for the Northern (DR 2) set, due to the Northern Cities Shift and New York (DR 6), due to /ɪ/

vocalisation and /ɔː/ lowering.

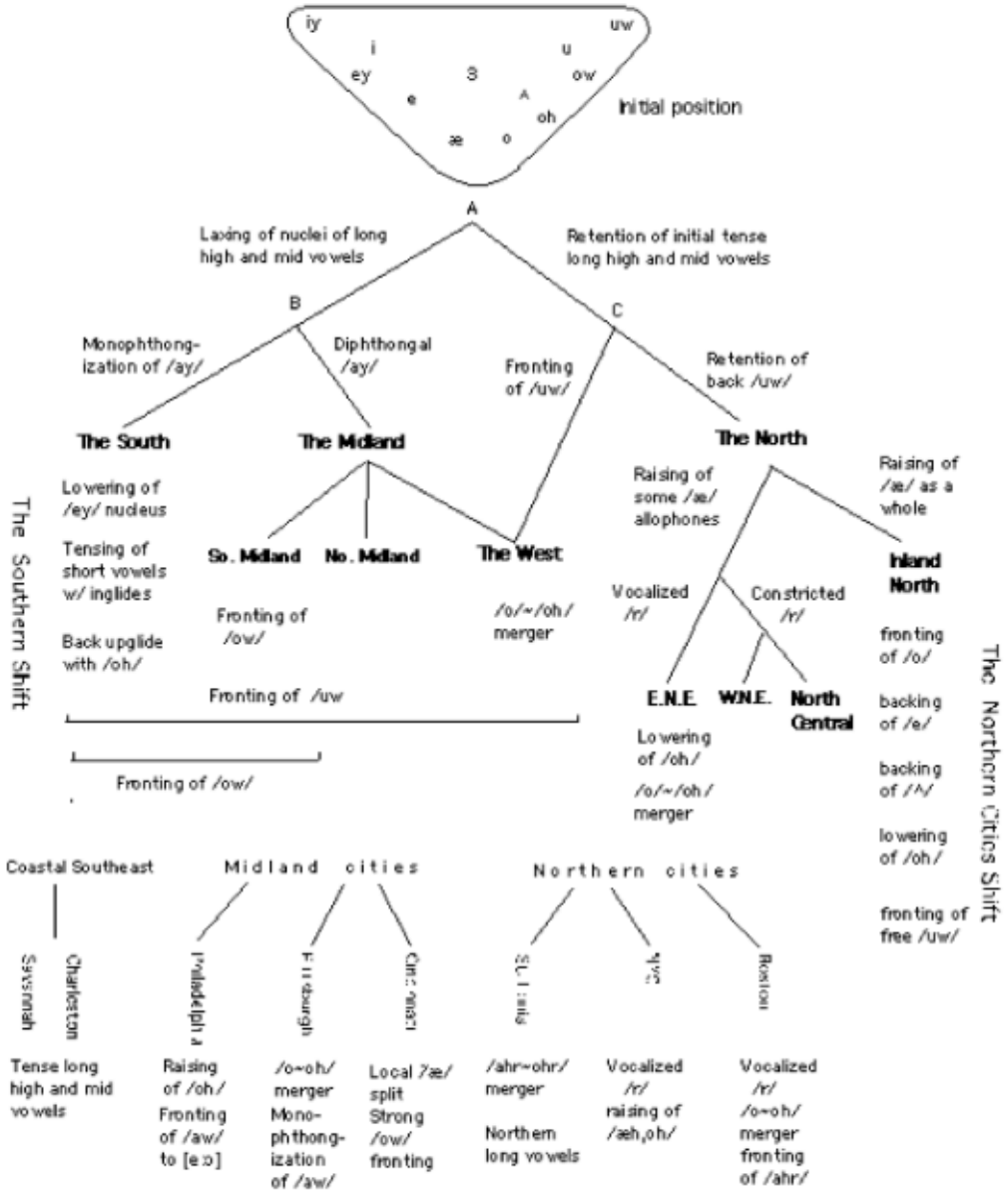

Figure 6.4: *Phonological taxonomy* of vocalic differences between the major dialect regions of North American English (Labov *et al.* 1997)

6.2.4 Feature extraction

In this chapter, elements of both the MFC and LPC were used as input (see §3.3.2 for an overview). MFC variables were chosen primarily due to the extensive use of MFCCs in ASR systems and ASR-based FVC research. Although less common, LPC input, analysed both holistically and segmentally, has been used considerably in ASR-based

FVC (Rose 2011a, 2013a; Nair *et al.* 2012). Further, the differences in the frequency scales used to extract MFC and LPC data may mean that there are differences in the sensitivity of LR output to regional variation depending on the form of the cepstrum used.

MFC and LPC data were extracted from the silence-removed sound files for each speaker using HTK.[20] Initially, a pre-emphasis filter with a coefficient value of 0.97 was applied to the signal (Equation 3.13). The signal was then divided in a series of frames using a 20ms Hamming window shifted at 10ms steps, resulting in 50% overlap between adjacent frames. The power spectrum of each frame was processed by applying Mel (MFC) and linear (LPC) filter banks consisting of 26 filters to the entire frequency range (0-8 kHz). A DCT was fitted to the log of the filter outputs and 12 coefficients (CCs) extracted. Ds were calculated from the CCs using Equation 3.15. The same equation was used to calculate As using the Ds as input.

6.2.5 Experiment

In this chapter, the sensitivity of LR output to different degrees of regional dialect match in both the feature-to-score and score-to-LR stages is assessed using different input data. Throughout testing, the same set of 25 test speakers (from DR 3) was used. GMM-UBM (§3.2.2.1) scores were initially computed for SS (28) and DS (756) pairs within each of the Matched (DR 3 - North Midland), Mismatched (DR 1 - New England; DR 2 - Northern; DR 4 - South Midland; DR 5 - Southern; DR 6 - New York City; DR 7 - Western) and Mixed sets (comprising 28 speakers). The first half of the data for each speaker was used to build a suspect GMM. The offender values from the second half of the data for each speaker were then used to compute $p(x_i|\lambda_{sus})$ (see §3.2.2.1). Due to the computational load involved in performing cross-validation using the GMM-UBM approach (i.e. creating a new UBM for each comparison), a single GMM background model was built for each system using all of the data for all 28 speakers within each of the eight sets. In this way, the background model also contained the data from each of the suspect and offender samples being compared. While this is not ideal, it is expected

[20] Hidden Markov Model Toolkit (HTK) `http://htk.eng.cam.ac.uk` (accessed: 11th September 2013) (Young *et al.* 2006).

to have limited effect on the overall results. The offender values were then evaluated against the background model to compute $p(x_i|\lambda_{bkg})$. The scores for each set were used to train a logistic-calibration model for each system (§3.2.4.1).

SS (25) and DS (600) scores were then computed for the test data. GMM suspect models were built for the suspect samples from the first half of the data for each of the test speakers. The numerator of the score was calculated by evaluating the offender values (i.e. the second half of the data for each speaker) relative to the suspect models. In this way, $p(x_i|\lambda_{sus})$ for the test comparisons remained constant across all eight systems. $p(x_i|\lambda_{bkg})$ was computed using the background models for each of the eight 28-speaker reference sets created when generating development scores. The calibration coefficients from each of the eight sets of development scores were then applied to the scores for the test data based on the dataset used to create the background model. The experiment was run using CCs and derivatives and CCs-only from both the MFC and the LPC.

Although ASR systems, such as BATVOX, typically model data with up to 1024 Gaussians, 32 Gaussians per variable were used here to build the GMM suspect and reference models. This is because the suspect samples were short and so contained a relatively small number of frames from which to extract data. Given the relative lack of data there is a risk of overfitting by using too many Gaussians. Pre-testing also revealed no marked differences in LR output from models generated using 32, 64 or 128 Gaussians. Therefore, the model with the lowest computational load (i.e. 32 Gaussians) was chosen to maximise efficiency. Further, 32 Gaussians have previously been used to model the TIMIT data in Reynolds (1995).

LR output is compared across systems in terms of the distributions of calibrated LLRs and validity (EER and C_{llr}). The imprecision of LLRs from individual comparisons across the Matched, Mismatched and Mixed systems was quantified by calculating the mean 95% CIs (§3.2.3.2) for each form of input data.

6.3 Results

This section presents LR output from the Matched, Mismatched and Mixed systems. Mismatched systems are referred to using their TIMIT DR number. The results based on MFC input are considered first, followed by the results based on LPC input. The comparative performance of MFC and LPC input is assessed in §6.4.

6.3.1 Mel Frequency cepstrum

CCs and derivatives

Figure 6.5 displays the distributions of calibrated LLRs based on CCs and derivatives from the MFC. The median SS LLR for the Matched system was +3.25, equivalent to *strong support* for the prosecution. Marginally weaker median SS LLRs were generated using Mismatched systems 6 and 7, while marginally stronger medians were achieved using the Mixed, Mismatched 4 and 5 sets, although in all cases these values were within the same order of magnitude as the Matched median. However, the SS medians using Mismatched sets 1 and 2 were stronger by one order of magnitude. There were also considerable differences in the overall ranges of LLRs. Relative to the Matched condition, the minimum LLR was up to two orders of magnitude stronger using Mismatched system 2, while the maximum was up to four orders of magnitude stronger using Mismatched system 1.

The DS median for the Matched condition was -5.57, equivalent to *very strong* support for the defence. With the exception of Mismatched 2, all systems produced DS medians within the same order of magnitude as the Matched median (between -5 and -6). Mismatched 2, however, generated a median value one order of magnitude weaker (-4.47) than the Matched median, although verbally both values are equivalent to *very strong* support for the defence. As with SS LLRs, there were some differences in the overall distributions of DS LLRs. The minimum DS LLR was up to four orders of magnitude weaker using the Mixed system than when using the Matched system, while the strongest contrary-to-fact LLRs were three orders of magnitude greater using Mixed data. This pattern was also found using Mismatched sets 1, 2, 4 and 5.

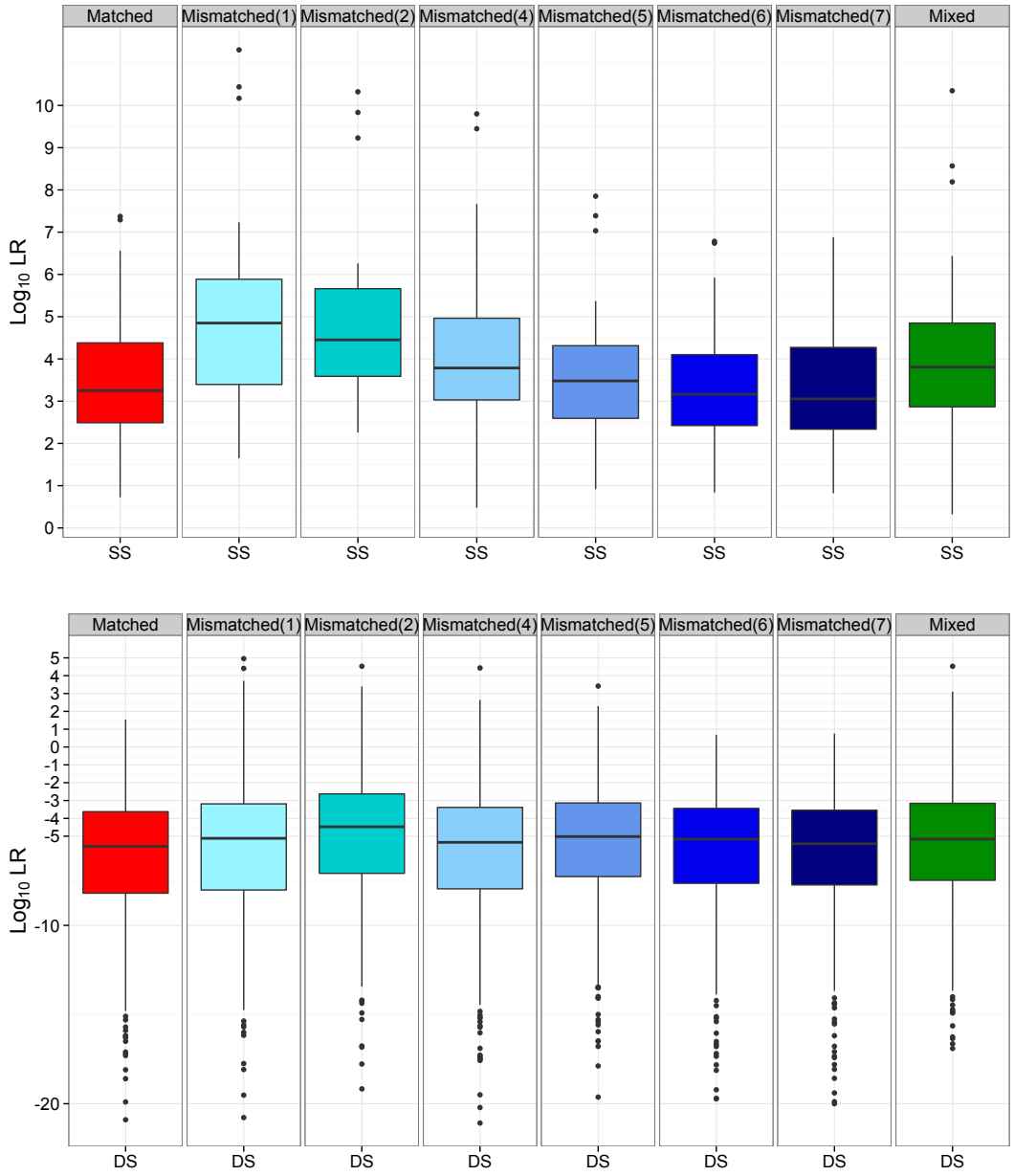

Figure 6.5: Boxplots (mid line = median, filled box = interquartile range (containing middle 50% of the data), whiskers = scores outside the middle 50%, dots = outliers) of SS (above) and DS (below) LLRs for each system using CCs and derivatives from the MFC

Figure 6.6 displays EER and C_{llr} values based on CCs and derivatives from the MFC. Despite evidence of differences in the distributions of LLRs, Figure 6.6 suggests that there were only limited differences across systems in terms of validity. Overall performance was generally extremely good, with values for both EER and C_{llr} very close to zero across all eight systems. This primarily reflects the use of optimal, forensically

unrealistic data. The implications of this in terms of the variation across systems according to regional background are discussed in §6.4. Maximally, EER values were spread across a range of 0.75%, with Mismatched systems 6 and 7 achieving the lowest value (0%) and the Mixed system achieving the highest (0.75%). C_{llr} values were also spread over an extremely narrow range (0.063). The lowest C_{llr} was again achieved using the Mismatched 6 and 7 data, although the Matched C_{llr} was just 0.003 greater. The highest C_{llr} was produced using Mismatched 1 data (0.082), although the absolute differences between sets were extremely small.

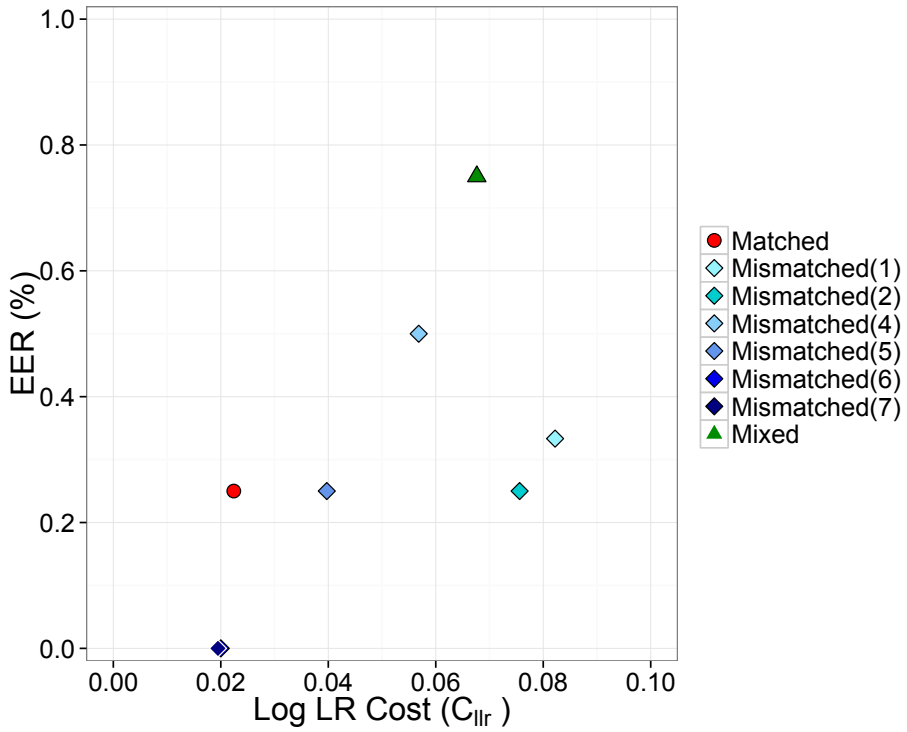

Figure 6.6: C_{llr} plotted against EER (%) for each system using CCs and derivatives from the MFC

CCs-only

Figure 6.7 reveals marginally less variation in the distributions of LLRs across systems using CCs-only compared with Figure 6.5. With the exception of Mismatched 2, median SS values for all of the Mismatched and Mixed systems were within the same order of magnitude as that for the Matched system (+3.58). The median SS LLR for Mismatched system 2 was one order of magnitude stronger (+4.05), although in numerical terms it

was much closer to the Matched median than with the inclusion of derivatives. Further, there was considerably more overlap in the interquartile ranges of SS LLRs for the Mismatched/Mixed systems and the interquartile range for the Matched system. The strongest SS LLRs were generated using the Mixed and Mismatched 1, 2 and 4 sets, and were maximally two orders of magnitude stronger than when using the Matched data.

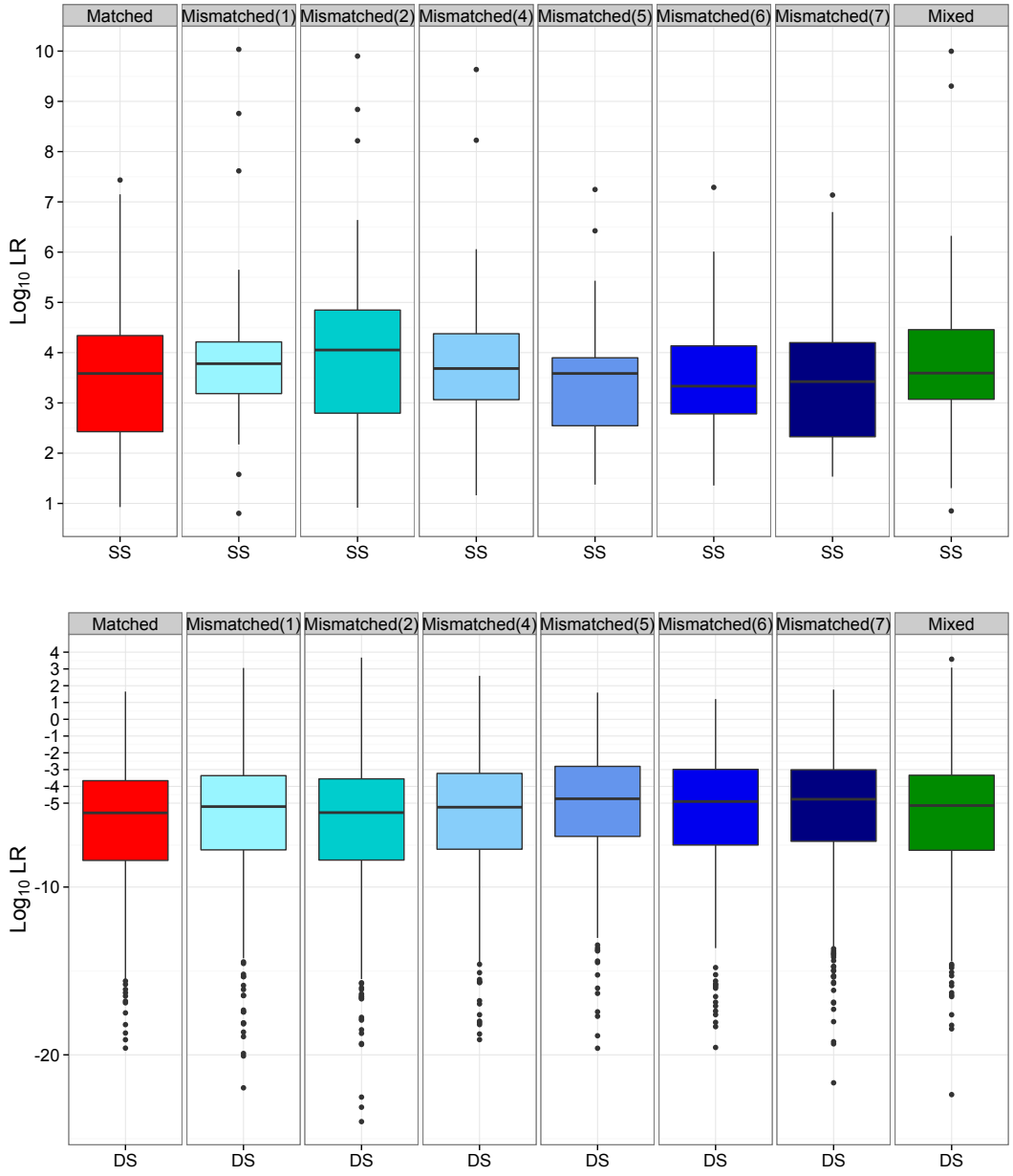

Figure 6.7: Box plots of SS (above) and DS (below) LLRs for each system using CCs-only from the MFC

The median DS LLR for the Matched system was -5.58, equivalent to *very strong* support for the defence. Median values for the Mixed and Mismatched 1, 2 and 4 systems were within the same order of magnitude as the Matched median. However, DS medians were weaker by one order of magnitude for Mismatched sets 5, 6 and 7. As with the inclusion of derivatives, the overall ranges of DS LLRs were considerably greater for the Mixed and Mismatched 1 and 2 systems, compared with the range for the Matched system. This was caused by stronger outlying negative values, by as much as four orders of magnitude, and stronger contrary-to-fact support for the prosecution, by maximally two orders of magnitude. Despite differences in the magnitudes of the DS medians, the most similar distributions of LLRs to those produced by the Matched system were found using Mismatched systems 6 and 7.

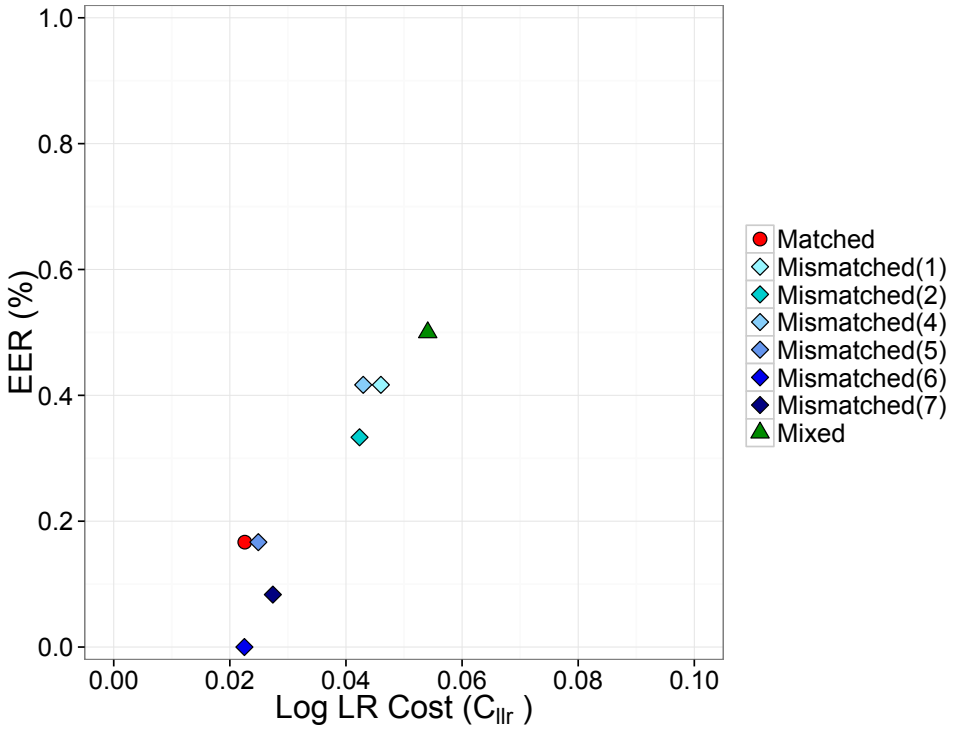

Figure 6.8: C_{llr} plotted against EER (%) for each system using CCs-only from the MFC

Figure 6.8 reveals even less variability in validity using CCs-only compared with Figure 6.6. EER values were spread over a range of 0.5%, with optimum performance achieved using Mismatched system 6 (0%) and the worst performance produced by the Mixed system. EER for the Matched system was 0.17%, compared with which, the Mixed and Mismatched 1, 2 and 4 EERs were marginally higher, and the Mismatched 6 and 7 EERs were marginally lower. The range of C_{llr} variability was also very narrow,

with values spread maximally over a range of 0.03. The best performing systems were based on the Matched and Mismatched 6 sets. For the remaining Mismatched systems, C_{llr} was marginally higher than the Matched value, while the highest C_{llr} value was produced by the Mixed system.

6.3.1.1 Reliability

Figure 6.9 displays Tippett plots of mean LLRs with 95% CIs across the eight systems for each set of MFC input data (CCs and derivatives and CCs-only). For both forms of input data, imprecision (i.e. the width of the 95% CIs) increased as the magnitude of the LLRs increased. This applied to both SS and DS LLRs, although the widest 95% CIs occurred for the highest magnitude DS LLRs (possibly due to the fact that the strongest DS LLRs are of considerably higher magnitude than the strongest SS LLRs). The largest mean 95% CI was generated using CCs-only ($\pm$1.88), although the mean CI with the inclusion of derivatives was only marginally smaller ($\pm$1.86). This indicates that the addition of derivatives had relatively little effect on the overall sensitivity of individual LLRs to regional variation.

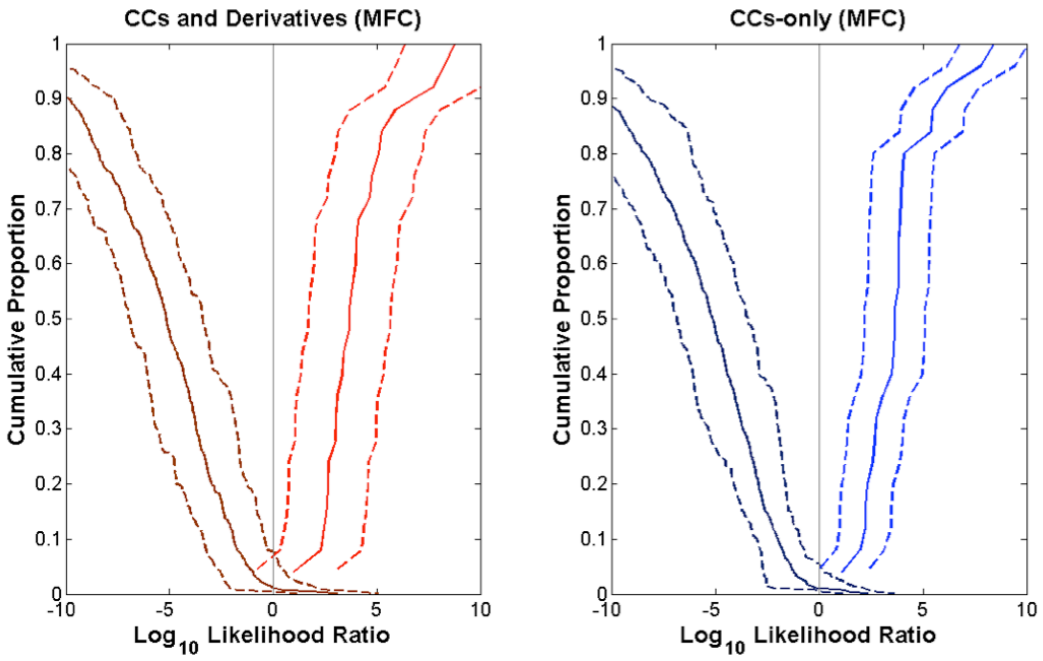

Figure 6.9: Tippett plots of mean SS (light) and DS (dark) LLRs and 95% CIs across systems using CCs and derivatives (left; red) and CCs-only (right; blue) from the MFC

6.3.2 Linear prediction cepstrum

CCs and derivatives

Figure 6.10 reveals marked differences in the distributions of SS LLRs across systems based on CCs and derivatives from the LPC. The median SS LLR for the Matched system was +4.12, equivalent to *very strong* support for the prosecution. Although also categorised as *very strong* evidence, the SS median for Mismatched 1 was one order of magnitude higher than the Matched value. With the exception of Mismatched system 2, the medians for the remaining Mismatched systems were one order of magnitude weaker. Despite this, as with MFC input, in terms of overall SS distributions the Mismatched 6 and 7 sets were most similar to the Matched output. For Mismatched systems 1 and 2, the minimum and maximum LLRs were up to two orders of magnitude greater than those generated by the Matched system, and up to four orders of magnitude greater using the Mixed system.

Similar differences between systems were found in the distributions of DS LLRs. The median DS LLR was one order of magnitude weaker for Mismatched sets 1 and 6 compared with the Matched set. Median values for the Mixed and other Mismatched systems were within the same order of magnitude as the Matched median, equivalent to *very strong* support for the defence. Although there was considerable overlap of interquartile ranges across all systems, there were marked differences in the overall ranges of DS values. The strongest outlying negative LLR was four orders of magnitude weaker using Mismatched system 5 compared with the Matched system. For the majority of systems the strongest contrary-to-fact LLRs were around two orders of magnitude greater than for the Matched system.

As with MFC input, the differences in the distributions of LLRs were not reflected in differences in system validity (Figure 6.11). EERs were spread over a range of 0.67% with Mismatched systems 2, 5 and 7 achieving marginally better EER than the Matched system, and the Mixed and Mismatched 6 systems performing marginally worse. Mismatched 1 and 4 achieved EERs equal to that of the Matched system (0.5%). C_{llr} values were also spread over a narrow range (0.08), with systems achieving both marginally better and marginally worse performance relative to the Matched system.

The best performing system overall was Mismatched 7, although the absolute difference between systems were extremely small.

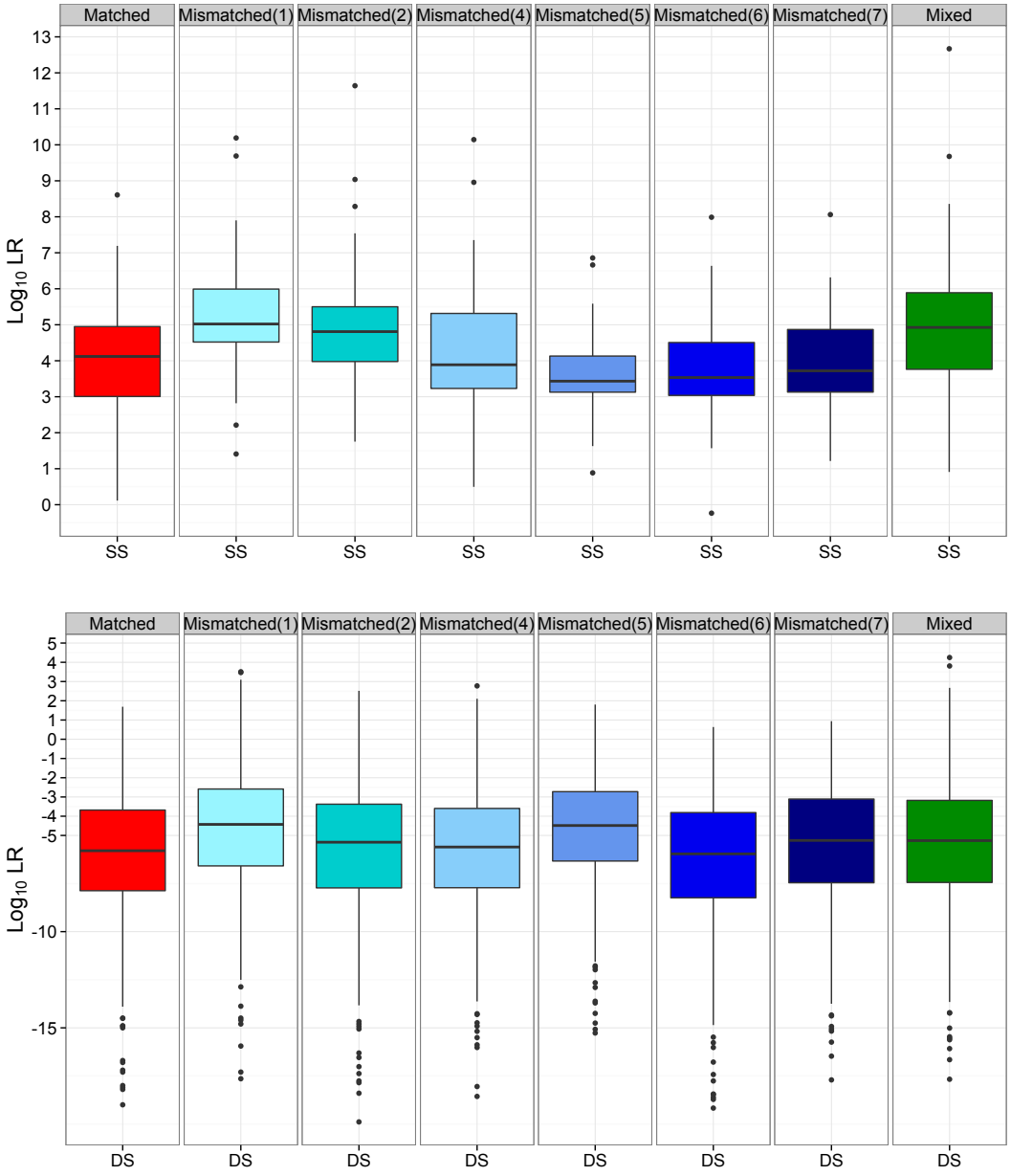

Figure 6.10: Boxplots of SS (above) and DS (below) LLRs for each system using CCs and derivatives from the LPC

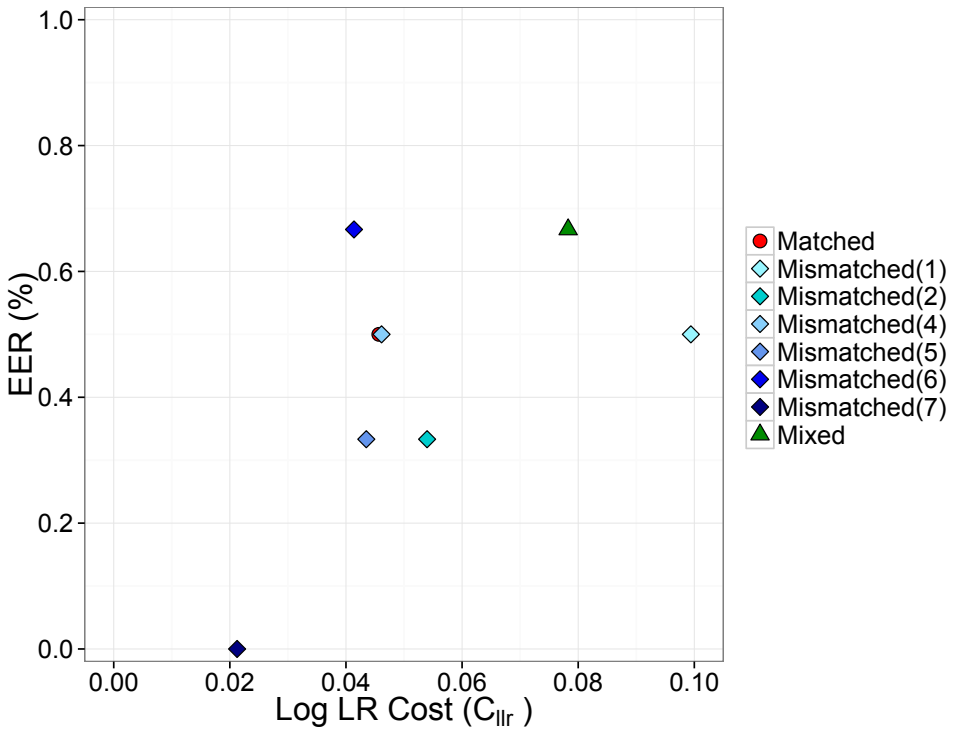

Figure 6.11: C_{llr} plotted against EER (%) for each system using CCs and derivatives from the LPC

CCs-only

Figure 6.12 displays the distributions of LLRs using CCs-only from the LPC. There was greater stability across systems in the distributions of SS LLRs compared with the inclusion of derivatives. The median was consistently within the range of +3 to +4, equivalent to *strong* support for the prosecution, other than for Mismatched 1 where the median was one order of magnitude stronger than the Matched median. There was also considerable overlap of the interquartile ranges of LLRs produced by all eight systems. There were, however, some differences across systems in terms of overall range, with the minimum LLR weaker by one order of magnitude using the Mixed and Mismatched 1, 5 and 7 systems, compared with the Matched system.

Slightly more variation was displayed in terms of DS LLRs. Relative to the Matched system (-5.55), median DS LLRs were weaker by one order of magnitude using Mismatched systems 1 and 5. The Mixed and remaining Mismatched systems produced DS medians equivalent to that of the Matched system. As with SS comparisons, the

interquartile ranges of DS LLRs displayed considerable overlap, with some variation in overall ranges. Specifically, contrary-to-fact DS LLRs offered stronger support for the prosecution by up to three orders of magnitude for the Mixed and Mismatched 1, 2, 4 and 5 systems compared with the Matched system.

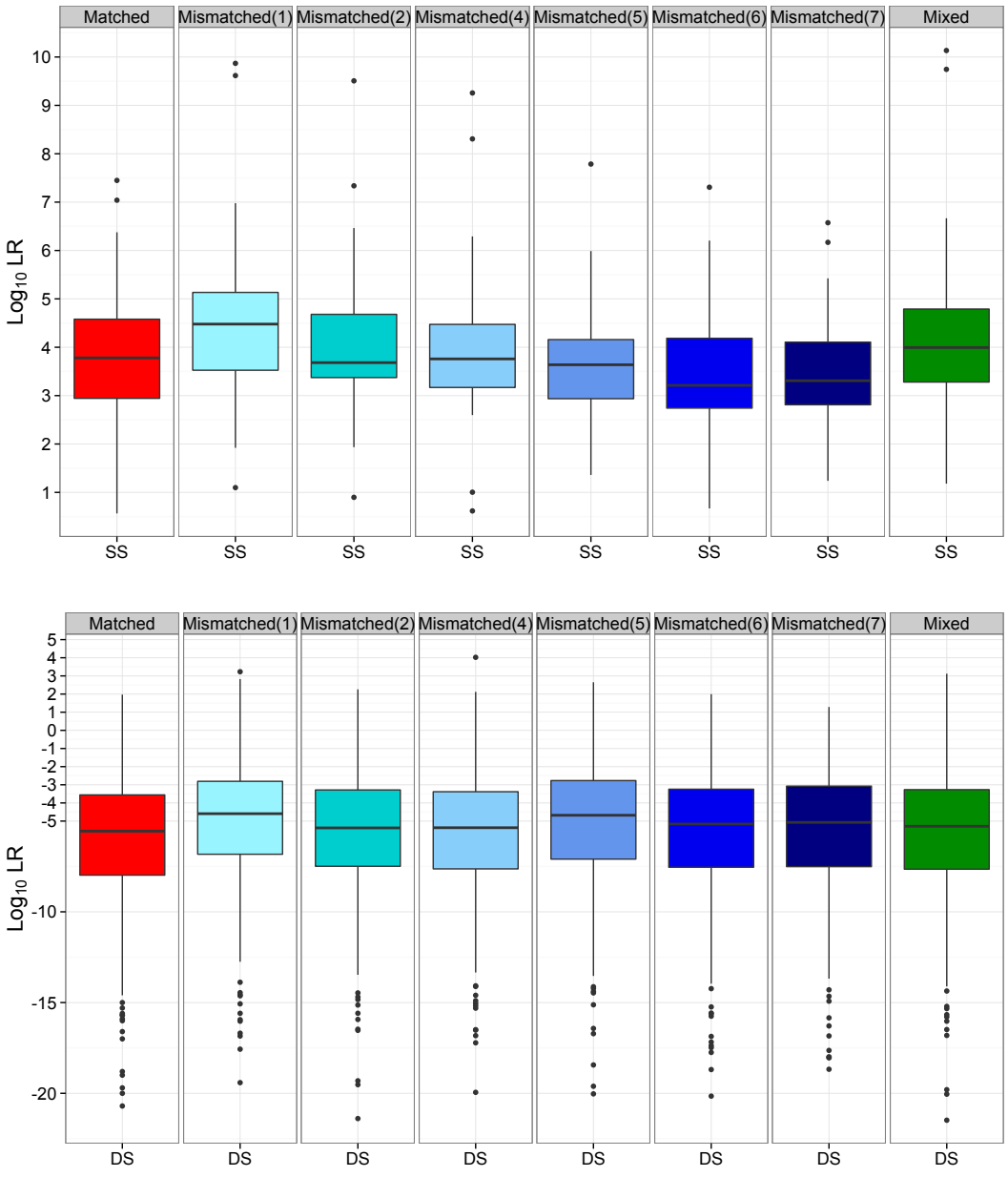

Figure 6.12: Boxplots of SS (above) and DS (below) LLRs for each system using CCs-only from the LPC

Figure 6.13 displays validity metrics for the eight systems based on CCs-only. EER values were spread over a range of 0.42%, smaller than the EER range in Figure 6.11. Mismatched EERs were both under and overestimated relative to the EER for the

Matched system, with the best performance achieved using Mismatched systems 5 and 6 (0.083%), and the worst performance produced using Mismatched systems 1 and 2 (0.5%). The Mixed system achieved the same EER as the Matched system (0.417%). Similar patterns were found for C_{llr}. Values were spread over 0.039, a narrower range than that with the inclusion of derivatives. Mismatched sets 6 and 7 produced the best C_{llr} values. Relative to the Matched C_{llr}, values based on the other Mismatched and Mixed systems were higher. However, importantly, the absolute differences in EER and C_{llr} values across systems were again extremely small.

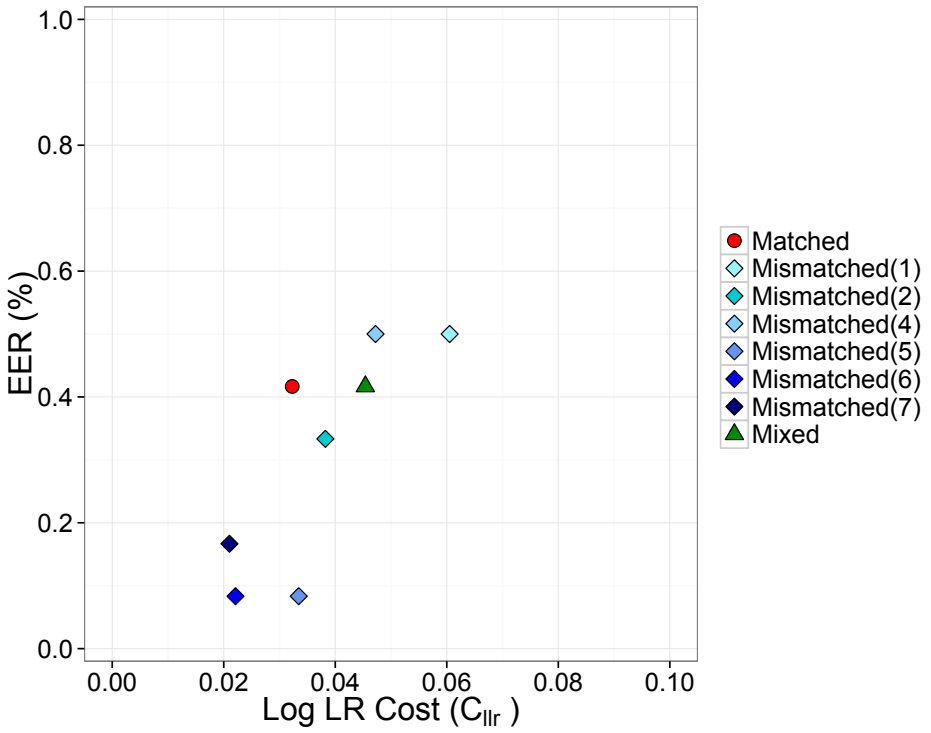

Figure 6.13: C_{llr} plotted against EER (%) for each system using CCs-only from the LPC

6.3.2.1 Reliability

Figure 6.14 displays mean LLRs and 95% CIs across the eight systems using LPC-based CCs and derivatives and CCs-only. As in Figure 6.9, across both forms of input data, there was an increase in the width of the 95% CIs as the magnitude of the mean LLRs increased. Again, the widest CIs were found for the highest magnitude DS LLRs. The greatest imprecision in individual LLRs across systems was found with the inclusion of

the derivatives, which generated a mean 95% CI of ±1.84. A marginally narrower mean 95% CI was found when using CCs-only (±1.80), although the absolute difference was relatively small. As in §6.3.1.1, this indicates that the inclusion of derivatives had essentially no effect on the imprecision in individual LLRs across systems.

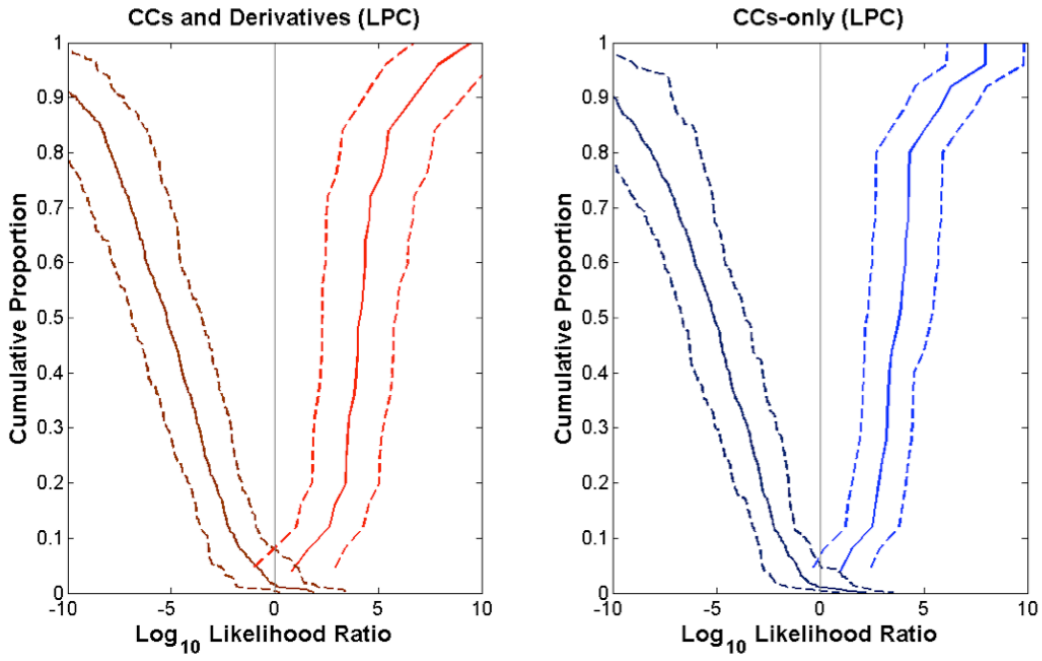

Figure 6.14: Tippett plots of mean SS (light) and DS (dark) LLRs and 95% CIs across systems using CCs and derivatives (left; red) and CCs-only (right; blue) from the LPC

6.4 Discussion

The results in §6.3 are to some extent consistent with Harrison and French (2012) in that the cepstrum-based ASR systems tested in this chapter displayed some sensitivity to regional variation. Across all forms of input data, this sensitivity was manifested in the distributions of LLRs and the imprecision in LLRs from individual comparisons across the eight systems. The patterns of variation were also broadly consistent with the results in §4.3.2 and §5.3.1. Marginally stronger SS LLRs were produced using the Mixed and certain Mismatched sets across all forms of input data, compared with the Matched set. The overinflation of SS LLRs was evident in the magnitude of the largest positive SS LLRs, which were up to four orders of magnitude stronger using Mismatched and Mixed systems, compared with output from the Matched systems. DS

LLRs were generally weaker for the Mismatched and Mixed sets. As in §4.3.2 and §5.3.1, these patterns are attributed to the shifting of the reference distribution relative to stable suspect models and offender values in feature-to-score conversion. A more detailed discussion of this is provided in §11.1.

Across all forms of input data the distributions of LLRs produced by the Mixed systems were generally closer to the distributions of the Matched LLRs than those produced by the Mismatched systems. However, not all Mismatched systems were found to perform in the same way. The greatest divergence from the Matched results in terms of the distributions of LLRs was found for Mismatched sets 1 (New England) and 2 (Northern). These patterns are consistent with the expected linguistic differences between the Mismatched New England and Northern DRs and the Matched Midland DR (outlined in §6.2.3). According to Labov *et al.*'s (1997) *phonological taxonomy* of AmEng (see Figure 6.4), the New England and Northern DRs differ from the Midland DR in that they display retention of backed /uː/ and retention of initial tense long high and mid vowels (e.g. /iː/). New England is also claimed to display /ɔː/ lowering and /r/ vocalisation, while the Northern area displays the Northern Cities Shift patterns.

The distributions of LLRs from Mismatched set 6 (New York City) and 7 (Western) were found to be most similar to the distributions of LLRs from the Matched systems. This finding is predictable for the Western DR based on the linguistic patterns in Figure 6.4. Both the Midland (3) and Western (7) DRs are claimed to display a merger in the quality of /ɒ/ and /ɔː/ (referred to as the COT∼CAUGHT merger) and fronting of /əʌ/. The convergence between the Matched and Mismatched 6 (New York City) sets in terms of the distributions of LLRs is, however, not predicted based on linguistic similarity. In fact, Figure 6.4 suggests that there should be considerable linguistic divergence between these DRs, with New York City sharing some of the patterns of the New England and Northern DRs (outlined above) as well as /r/ vocalisation and raising of /ɑː/ and /ɔː/. The patterning of the Mismatched systems suggest that the cepstral input used in this chapter did capture some of the linguistically meaningful variation between the DRs. However, the linguistic definition of the DRs in §6.2.3 only considered variation in segmental (and primarily vocalic) variables. There may, of course, be regional differences in long-term vocal setting which also explain the patterns

in the Mismatched sets, although very little work in sociolinguistics has considered systematic regional differences in vocal settings for varieties of AmEng.

Consistent with the variation in terms of the distributions of LLRs, relatively large mean 95% CIs were generated across all forms of input data (Table 6.3), indicating considerable imprecision in the LLRs from individual comparisons across systems. The largest mean 95% CI was produced using CCs and derivatives from the MFC, although the absolute differences across all forms of input data (MFC and LPC, CCs and derivatives and CCs-only) were extremely small. The fact that the mean CIs were broadly similar across input suggests that the frequency scale used to represent the cepstrum has little effect on the sensitivity of LR output to regional variation. Further, the addition of derivatives had essentially no effect on the imprecision in individual LLRs. Therefore, it can be inferred that the regional variation in these data was primarily encoded in the CCs while derivatives were relatively robust to regional variation.

Table 6.3: Mean 95% CIs (±LLR) for CCs and derivatives and CCs-only from the MFC and the LPC

	Mean 95% CI	
	MFC	**LPC**
CCs and derivatives	1.867	1.836
CCs-only	1.880	1.799

Despite the sensitivity of individual LLRs to the different regionally defined systems, almost no variation was found in terms of validity. This is consistent with Moreno *et al.* (2006) who found limited difference between Matched and Mismatched systems in terms of EER when using BATVOX. Contrary to §4.3.2 and §5.3.1, no systematic ordering patterns were found in terms of EER or C_{llr} across systems (i.e. the Matched system did not always produce the best validity). Table 6.4 displays the ranges of EER and C_{llr} values across the eight systems for each form of input data. The ranges of variation were extremely small for both EER (maximally within 0.75%) and C_{llr} (maximally within 0.08). Further, the ranges of variation were comparable across the MFC and LPC results for CCs and derivatives and CCs-only. As with the 95% CIs in Table 6.2, the results in Table 6.2 suggest that the different frequency scales (MFC and

LPC) did not affect the overall sensitivity of the systems to different definitions of the relevant population. Further, no greater sensitivity to regional variation was introduced with the addition of derivatives.

Table 6.4: Ranges of EER (%) and C_{llr} values across all systems for CCs and derivatives and CCs-only from the MFC and the LPC

	EER (%)		C_{llr}	
	MFC	**LPC**	**MFC**	**LPC**
CCs and derivatives	0.75	0.67	0.06	0.08
CCs-only	0.42	0.42	0.03	0.04

There are two reasons for the narrow ranges of validity variability displayed in Table 6.4. Firstly, all of the systems appeared to reach a ceiling in terms of performance. In all cases, both EER and C_{llr} values were extremely close to zero. Indeed, contrary to Campbell (1997) the addition of derivatives did not improve the overall validity of the systems in terms of either EER or C_{llr}. Similarly, no systematic validity differences were found between MFC and LPC input. This is primarily due to the use of optimal data, in the form of short contemporaneous samples of highly controlled read speech, with no transmission, quality or style mismatch across suspect and offender samples. Therefore, performance is necessarily overoptimistic relative to that based on more forensically realistic conditions.

Secondly, the use of optimal data is confounded by the fact that the speaker discriminatory power of cepstral input is generally very good (and typically much better than individual linguistic-phonetic variables; Rose 2002, 2013a). This is reflected in the fact that the magnitudes of the LLRs in this chapter were very high (SS LLRs generally > +3 and DS LLRs generally < -5, across all systems). Therefore, the variability in individual LLRs as a function of regional variation, reflected in the mean 95% CIs, occurred so far away from the zero threshold that it had essentially no effect on the resulting validity of the systems. For example, based on CCs-only from the MFC, one SS speaker comparison achieved a LLR of +7.43 using the Matched system. Using the Mixed system, the LLR was over two orders of magnitude stronger (+9.99) while using Mismatched system 1 this value was stronger by three orders of magnitude. Such

variability would not contribute towards differences in either EER or C_{llr} across these systems. However, given the evidence of sensitivity to regional variation in the 95% CIs there is potential for considerably greater validity variability under more forensically realistic conditions (where LLRs are closer to zero).

6.5 Chapter summary

- Evidence of stronger (by up to four orders of magnitude) SS LLRs and weaker DS LLRs using Mismatched and Mixed systems compared with Matched systems.
  - Wide 95% CIs across all input data reflecting considerable imprecision in LLRs from individual comparisons across systems.
  - No differences in mean 95% CIs for MFC vs. LPC input, or CCs and derivatives vs. CCs-only.
- Patterns of divergence from the Matched results to some extent consistent with expected linguistic differences based on segmental (vocalic) variables.
- Essentially no differences across systems in terms of EER and C_{llr}.
  - Ceiling effect for system validity due to the use of forensically unrealistic data and the speaker discriminatory power of cepstral input.
  - Imprecision in individual LLRs (95% CIs) not reflected in validity since the variability occurred in such high magnitudes.

Chapter 7

Socio-Economic Class and Age: /eɪ/

This chapter expands on previous applications of logical relevance to FVC by exploring the definition of the relevant population in terms of socio-economic class and age. Using cubic polynomial estimations of the F1∼F3 trajectories of New Zealand English (NZE) /eɪ/, calibrated LLRs were computed for a sociolinguistically homogeneous set of test data using (a) **Matched**, (b) **Mixed** and (c) **Mismatched** systems. The distributions of calibrated LLRs and system validity are compared across systems. The imprecision in LLRs for individual pairs based on class and age variation is compared using 95% CIs.

7.1 Introduction

As highlighted in §2.2.5, it is well known in phonetics and linguistics, particularly sociolinguistics and sociophonetics, that speech is affected by a wide range of factors that generate both within- and between-speaker variation (Rose 2002; French *et al.* 2010). Despite such inherent complexity, the potential logical relevance of socio-indexical factors beyond sex and language is rarely considered in LR-based testing (with the exception of Loakes 2006; Zhang *et al.* 2011). This chapter therefore explores the extent to which different controls over socio-economic class and age in the definition of the relevant population affect LR output using the F1∼F3 trajectories of /eɪ/ (FACE; Wells 1982) in NZE. Class and age were chosen as illustrative of the socio-indexical factors which may affect LR output, but which are typically overlooked.

This chapter replicates the structure of experiments in previous chapters to evaluate the LR output from a sociolinguistically homogeneous set of test data using three systems, which represent different definitions of the relevant population: (a) **Matched**, (b) **Mixed** and (c) **Mismatched**. Each system contains development and reference data to test the effects of class and age variation at both the feature-to-score and score-to-LR stages. In each experiment, the distributions of calibrated LLRs and system validity are compared across systems. As in Chapter 6, the results for both class and age variation are compared using 95% CIs to assess the imprecision of LLR estimates across systems. In §7.4, the relative importance of class and age variation is compared with the patterns for regional variation from Chapter 6.

7.2 Method

7.2.1 /eɪ/ in New Zealand English

The choice of /eɪ/ is motivated by known patterns of variation and change in NZE and the availability of a large amount of acoustic data. There has been considerable change in the quality of NZE diphthongs over the last century. There is clear evidence of *diphthong shift*, attested as far back as 1887, in which the onset element has lowered and backed from /e/ towards [a ∼ ɐ] (Ellis 1889; Adams 1904; Maclagan 1982; Gordon *et al.* 2004; Sõskuthy *et al.* in press). Gordon *et al.* (2004: 149) claim that a second phase of change involved *glide weakening*, reducing the amount of articulatory and acoustic movement between onset and offset. Despite broad processes of change over time there remains considerable variation in the phonetic realisation of /eɪ/ in NZE (Hay *et al.* 2008).

Hay *et al.* (2008) distinguish between *cultivated* and *broad* varieties of NZE. Variation in the quality of the closing diphthongs aligns with this distinction. For /eɪ/, phonetic variation relates primarily to the onset, which is more open and back in *broad* NZE than in *cultivated* NZE. In terms of the acoustic output, Figure 7.1 predicts that *broad* NZE speakers will display higher F1 and lower F2 values at the initial phonetic target than *cultivated* NZE speakers. Variation in the position of the onset element also causes

differences in the amount of articulatory and acoustic movement across the duration of the vowel within the vowel plane, since the offset position for both groups is predicted to be in a similar position (close-mid [e]).

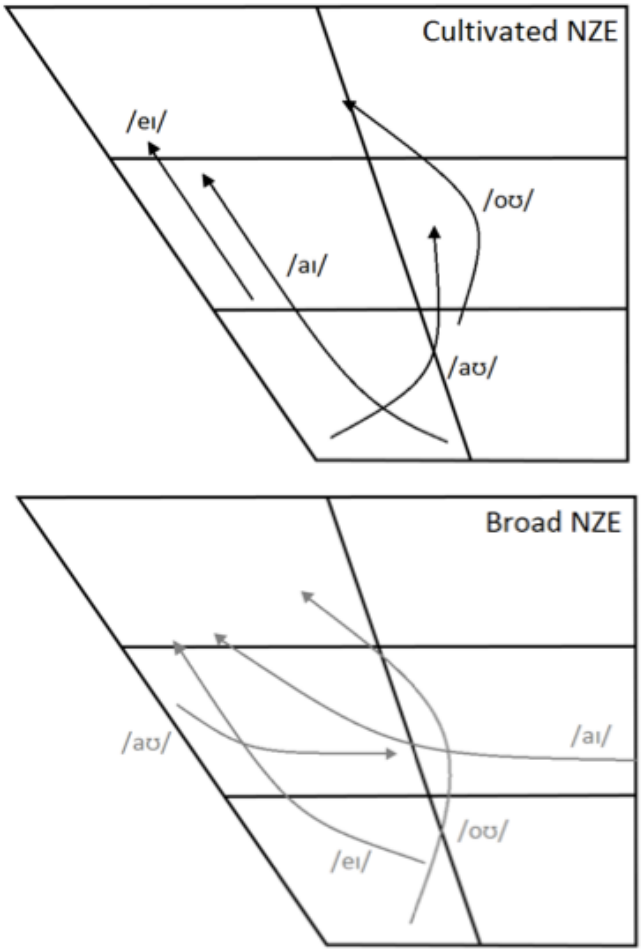

Figure 7.1: Schematic representation of variation in the closing diphthongs of *cultivated* (above) and *broad* (below) NZE (adapted from Hay *et al.* 2008: 97)

7.2.2 Data

Data were extracted from the male speakers in CanCor (§3.1.2). The forced-aligned phoneme-level TextGrids embedded within LaBB-CAT were inspected relative to the waveform and spectrogram, and auditory analysis performed. Erroneous segmental boundaries were manually hand-corrected for target /eɪ/ tokens. Boundaries were determined by the criteria in §3.3.1.3. Dynamic data were extracted from the first three formants of /eɪ/ using LaBB-CAT (§3.3.1.2) with the script set to find five formants

within a 0 to 5 kHz range. Along with the raw formant values, the output from LaBB-CAT also included speaker, year of birth, class, sex, phonological conditioning and syntactic category. The dataset initially contained 211 male speakers with between one and 410 tokens per speaker.

A series of heuristic procedures were implemented to correct or remove errors as in Chapter 4. Broad accept-reject thresholds were firstly applied to all of the data to remove obvious measurement errors (such as F2 measured as F1). A wide pass-band for F1 of between 200 and 900 Hz was chosen based on expectations for considerable movement on the open-close dimension between onset and offset. For F2 a range of 1100 to 2200 Hz was implemented, to capture the maximal amount of potential F2 movement assuming the onset of /eɪ/ can be central [ɐ] and the offset can be front [ɪ]. For F3, a range of 2000 to 3000 Hz was used. Tokens with values outside of these ranges were removed.

Given the relatively small number of tokens for most speakers it was not possible to ensure that the same number of tokens in each phonological context were included for each speaker. Rather, all tokens with adjacent /l/ and /r/ were removed. The data also contained multiple tokens of the indirect object *a*, all of which were removed since in spontaneous speech it is predicted that these will be reduced to schwa [ə]. Given the predicted patterns of class and age variation in §7.2.1, it was considered preferable, in terms of ensuring accurate formant measurement, to then divide the data into class-by-age sub-groups. Further heuristic error-removal procedures were then applied to the separate sub-groups.

7.2.3 Dividing the data

Within ONZE, speakers are classified according to social class, and labelled as either *professional* or *non-professional* based on occupation and education level (Gordon *et al.* 2007: 91). A six-point version of the Elley-Irving scale was used as a metric of occupation level (Elley and Irving 1985). A similar six-point scale adapted from Gregersen and Pedersen (1991) was used to code for education level. Scores were added together, with low values representing higher social class. In the ONZE data, those

classed as *professional* scored on average between 4 and 4.5, while *non-professionals* scored between 8.5 and 9.5. The labels assigned to speakers in the ONZE coding were used to divide the current data.

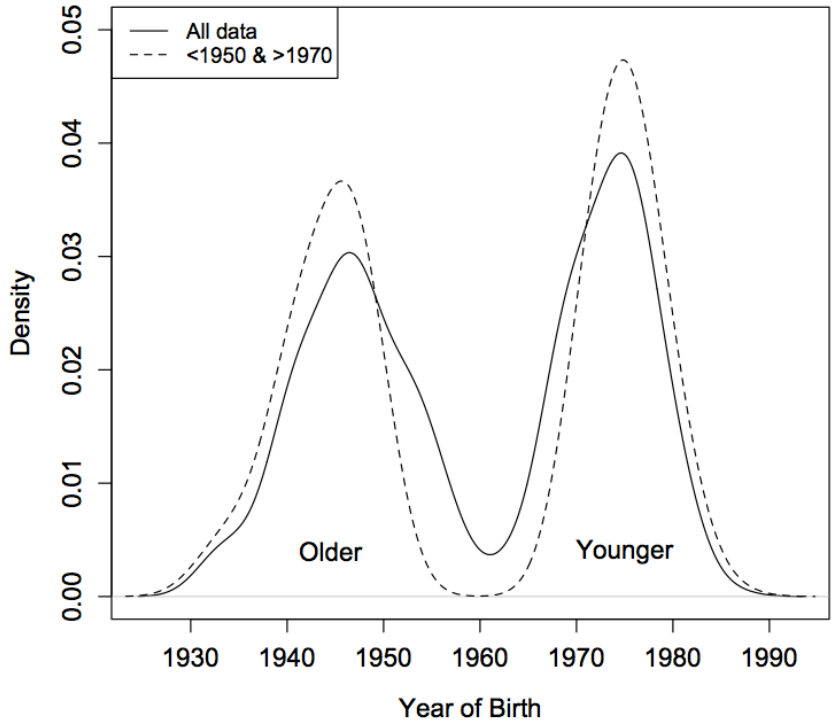

Figure 7.2: Density plot of bimodal distribution of year of birth from the entire dataset (solid) and from the subdivided dataset consisting of speakers born before 1950 and after 1970 (dashed)

Information relating to age in CanCor was limited to year of birth, although it is possible to deduce a range for age at the time of recording based on when the recordings were made. The continuous *year of birth* variable was converted into a discrete variable with two levels, *older* and *younger*. Across the entire dataset there is a wide age range with speakers born between 1932 and 1982. The distribution of year of birth is also bimodal, with a dip around the mean (c. 1960; see Figure 7.2). For the purposes of these experiments, speakers born after 1970 were classed as *younger*, and those born before 1950 were classed as *older*. Speakers born between 1951 and 1969 were removed (c. 50 speakers). The decision to divide the sample in this way ensured that a *cliff-edge* turning point at 1960 was avoided.

The class-by-age classification generated four sub-groups: younger professionals, younger non-professionals, older professionals, and older non-professionals. After speakers had been separated into sub-groups, z-scores for each formant measurement were calculated relative to the pooled mean across all speakers within each sub-group to remove univariate outliers. Tokens containing a value greater than ± 3.29 SDs from the mean were removed. The sub-groups were analysed separately to preserve patterns of sociolinguistic variation across groups whilst also removing measurement errors.

7.2.4 Parametric representations

The procedures outlined above removed the most obvious measurement errors. However, such procedures were reductive in that tokens were removed if any single value did not fit the criteria. Therefore, a final procedure was implemented to correct more localised errors without removing tokens from the analysis. Each formant trajectory from each token was fitted with a cubic polynomial curve (§3.3.1.4). Individual frequency values with residuals of greater than 50 Hz for F1 and F2 or 100 Hz for F3 (relative to the fitted value) were then removed (Figure 7.3). These heuristics were determined based on expectations for the maximal extent of potential movement between adjacent points in the formant trajectory (separated typically by less than 10 ms). A cubic polynomial curve was then re-fitted to the remaining data.

Finally, between-speaker z-scores within each class-by-age group were calculated for each cubic polynomial coefficient from the refitted curve. Tokens with outlying values of greater than ± 3.29 SDs from the group mean were again removed. Speakers with fewer than eight available tokens were also removed. A minimum of eight tokens per speaker was chosen after trial and error procedures comparing the trade-off between number of tokens and maximal number of speakers. The final dataset consisted of 120 speakers with eight tokens per speaker: 33 younger professionals, 31 younger non-professionals, 32 older professionals and 24 older non-professionals.

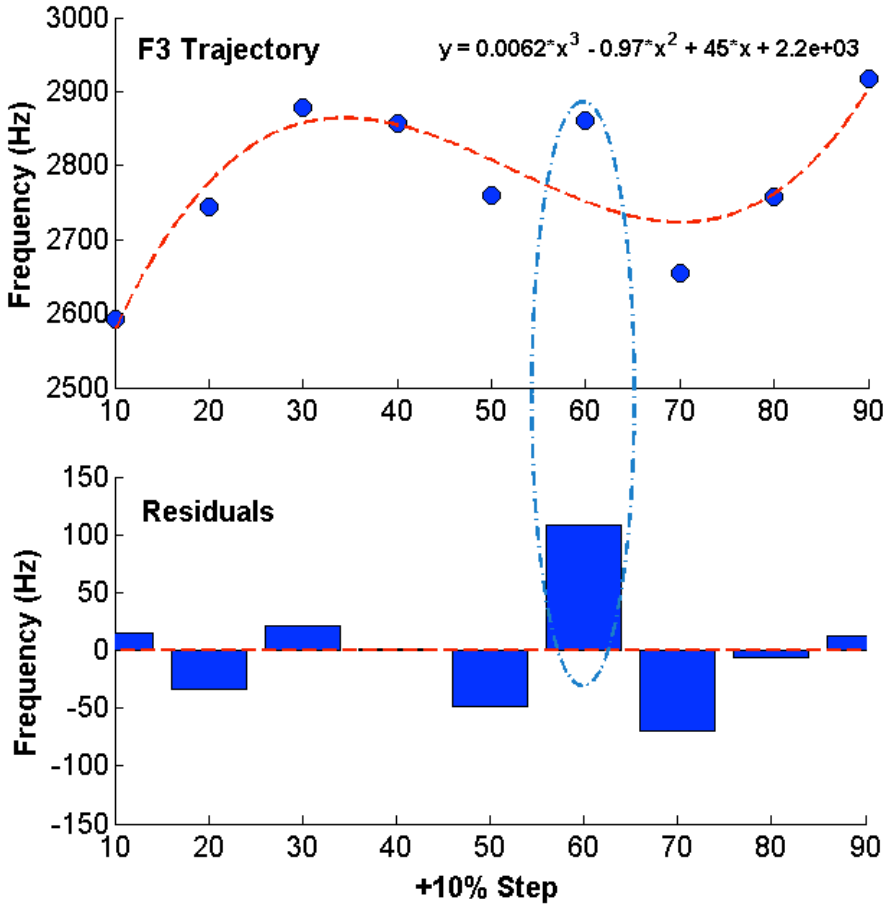

Figure 7.3: Raw F3 values (y) for a single token fitted with a cubic polynomial (y-fit) (red dashed curve) (above) and values with a residual greater than $\pm$100 Hz identified (dashed ellipsis) (below)

7.2.5 Variability in the data

The raw data were analysed to assess the extent to which systematic class and age variation was present. Mean F1, F2 and F3 values were calculated at each +10% step based on the raw data pooled by class and age. At each +10% step for each formant, 95% CIs were calculated. In this case, the 95% CI is a probabilistic region of a posterior distribution, where the probability of the mean being contained within the upper and lower bounds is 0.95. Following Albert (2009: 63), a standard noninformative prior $g(\mu, \sigma^2) \propto 1/\sigma^2$ was used to generate the posterior density:

$$g(\mu, \sigma^2|y) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}+1}} exp\left(-\frac{1}{2\sigma^2}(S + n(\mu - \bar{y})^2)\right) \tag{7.1}$$

where:

n = Sample size

$\bar{y}$ = Sample mean

$S = \sum_{i=1}^{n}(y_i - \bar{y})^2$

A noninformative prior was used due to the lack of prior numerical information about the distribution of values at each +10% step. Upper and lower bounds of 95% CIs at each 10% step were calculated using the *LearnBayes* package in R.[21]

7.2.5.1 Socio-economic class

Figure 7.4 displays the sample mean and 95% CIs for the trajectories of F1, F2 and F3 according to class. There was some consistency between the F1 patterns in Figure 7.4 and those predicted in §7.2.1. Although there was considerable overlap between the CIs at the onset, the upper bound for the non-professional speakers was marginally higher. This suggests that the first target of /eɪ/ (located at around the +20% point of the trajectory) displays a marginally higher mean F1 for the non-professionals indicating a more open, [ɑ]-like onset position. However, at the second phonetic target (located at around the +80% step) the F1 mean for the non-professionals was marginally lower than for the professionals, with almost no overlap in terms of the 95% CIs, indicating a closer offset position. Therefore, for the non-professionals there is no evidence of glide weakening, as predicted by Hay *et al.* (2008).

There was greater separation of the CIs at the onset of F2, with the CI for the professionals covering a higher F2 range than the CI for the non-professionals. The difference between the groups in F2 onset indicates a slightly backer realisation of the nucleus for the non-professionals. However, the magnitude of the mean separation between groups was relatively small. Further, there was considerable overlap between the groups at the offset of F2, meaning that non-professionals generally displayed greater F2 movement within the vowel plane compared with professionals. The differences in absolute frequency at the onset and in overall dynamic implementation are consistent

[21] Albert, J. (2014). LearnBayes: Functions for Learning Bayesian Inference (version 1.0-4) (R package). `http://cran.r-project.org/web/packages/LearnBayes/index.html` (accessed: 5th September 2014).

with Hay *et al.* (2008: 97), suggesting that class differences are manifested not only on the open-close dimension, but also on the front-back dimension.

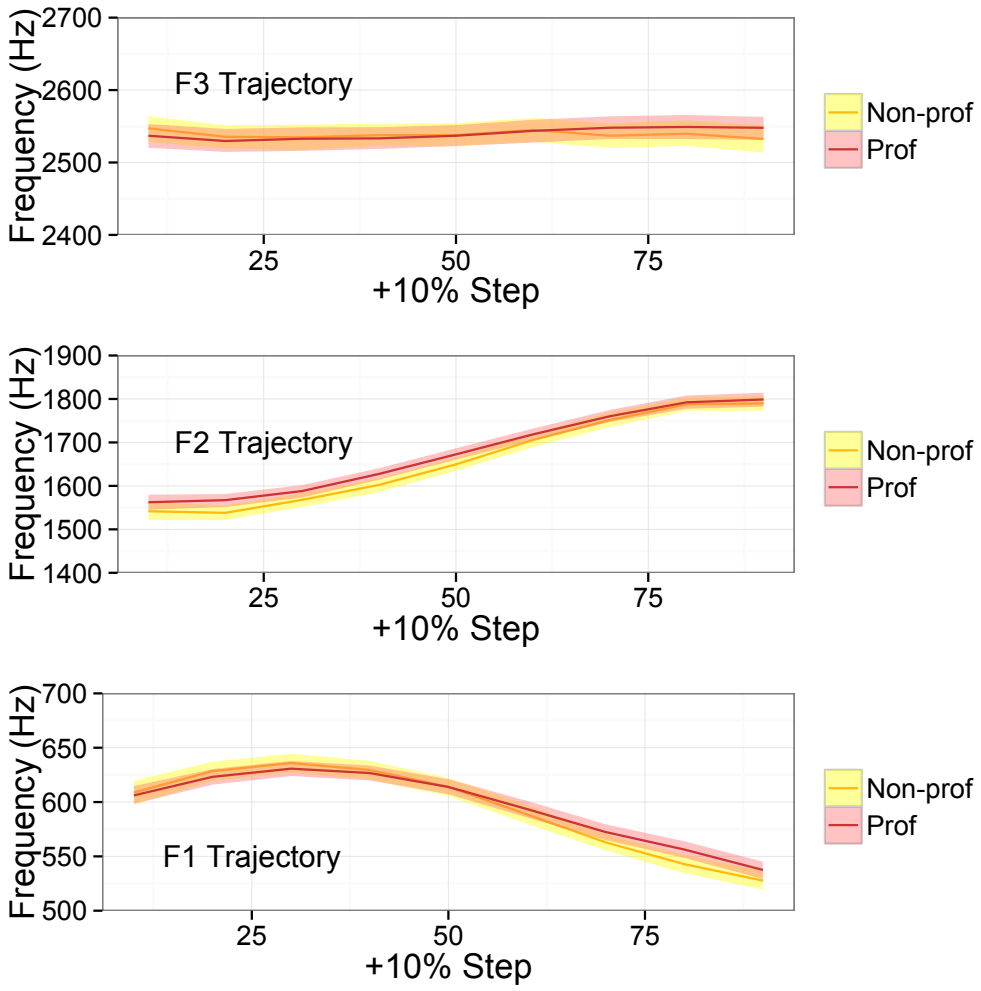

Figure 7.4: Mean F1, F2 and F3 trajectories with 95% CIs plotted by class based on 120 male speakers and eight tokens per speaker

There was considerable overlap in the F3 CIs across both groups. This indicates that the F3 mean for the two groups was located within roughly the same interval. For both groups, there was also very little fluctuation in the CIs across the duration of F3, indicating that the mean F3 trajectory was typically stable between onset and offset.

7.2.5.2 Age

Figure 7.5 displays the sample mean F1, F2 and F3 trajectories with 95% CIs according to age. Compared with Figure 7.4, the age differences were more marked than the

class differences. For F1, there was separation of the CIs for the younger and older groups at different points across the trajectory. The sample mean and the CI for the older speakers covered higher F1 frequencies at the onset of /eɪ/ indicating a typically, more open onset position compared with the younger speakers. However, by the onset the separation of the CIs had reversed. This indicates that there is on average more movement across the F1 trajectory for the older speakers than for younger speakers.

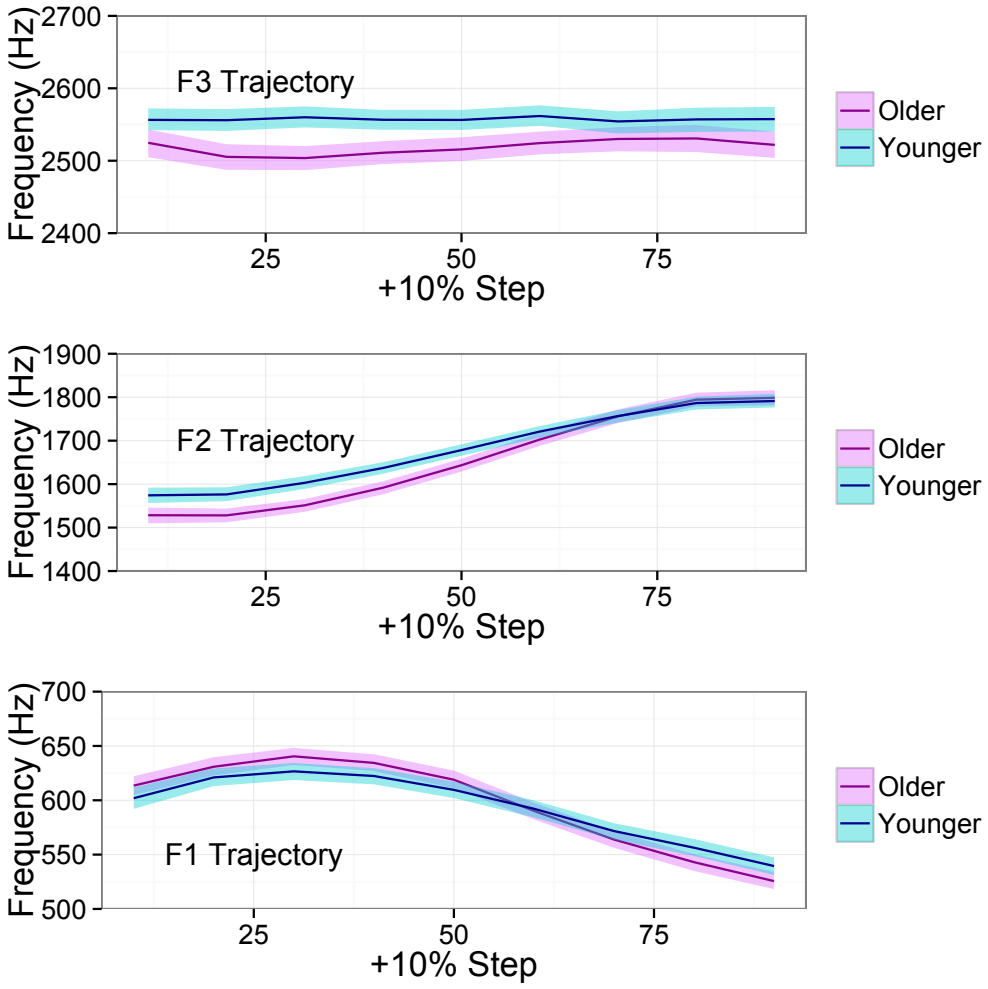

Figure 7.5: Mean F1, F2 and F3 trajectories with 95% CIs plotted by age based on 120 male speakers and eight tokens per speaker

In terms of F2, there was complete separation of the CIs for the younger and older groups between the onset and approximately the temporal midpoint. The CI for the older group covered a lower F2 frequency range than that of the younger group, indicating that the mean onset position for the older speakers had a lower F2, and therefore possibly a more retracted tongue position. The separation between groups was greatest

at the +30% step. Towards the offset, the F2 CIs converge, displaying almost complete overlap between the +60% and +90% steps. As in F1, the F2 trajectories indicate greater movement within the vowel plane for the older speakers. Finally, there was considerable separation of the CIs and sample means across the F3 trajectory. The F3 mean was consistently higher for the younger speakers than for the older speakers. Further, the F3 trajectory for the younger speakers was extremely stable between onset and offset, compared with the older speakers where there appeared to be an increase in the mean between the two phonetic targets at the +20% and +80% steps.

These patterns offer potential evidence of apparent time change, with the formant differences indicating the typical realisation of /eɪ/ as [aɪ] for older speakers and as [eɪ] for younger speakers. However, auditory analysis of these data suggests more subtle differences in realisations across the two groups. The variation in Figure 7.5 is also, to some extent, consistent with the physiological effects of ageing, with higher formant frequencies produced by the younger speaker. However, previous research on real time age differences predicts considerable lowering of F1 and less marked lowering of F2 and F3 as speakers become older (Wilder 1978; Linville and Rens 2001; Reubold *et al.* 2010), which is not the case in these data. Given that neither change nor physiological ageing account fully for the variation it is assumed that there is some interaction between the age-related factors. These data do, however, highlight the complexity and multidimensionality of age as a logically relevant grouping variable for FVC.

7.2.5.3 Interaction between class and age

Figure 7.6 displays mean F1∼F2 trajectories within the vowel plane plotted for each class-by-age group. There is clear evidence of differences between the groups in terms of the onset and offset, as well as the degree of movement within the vowel plane. The older non-professionals displayed the most open first target (at around the +20% step), with the older professionals, younger professionals and younger non-professionals differing primarily on the F2 dimension. Generally, /eɪ/ was more fronted (i.e. higher F2) for the younger professionals, who also displayed a more open offset position, resulting in less F1∼F2 movement across the vowel plane compared with the other

groups. For the other three groups the offset position was very similar.

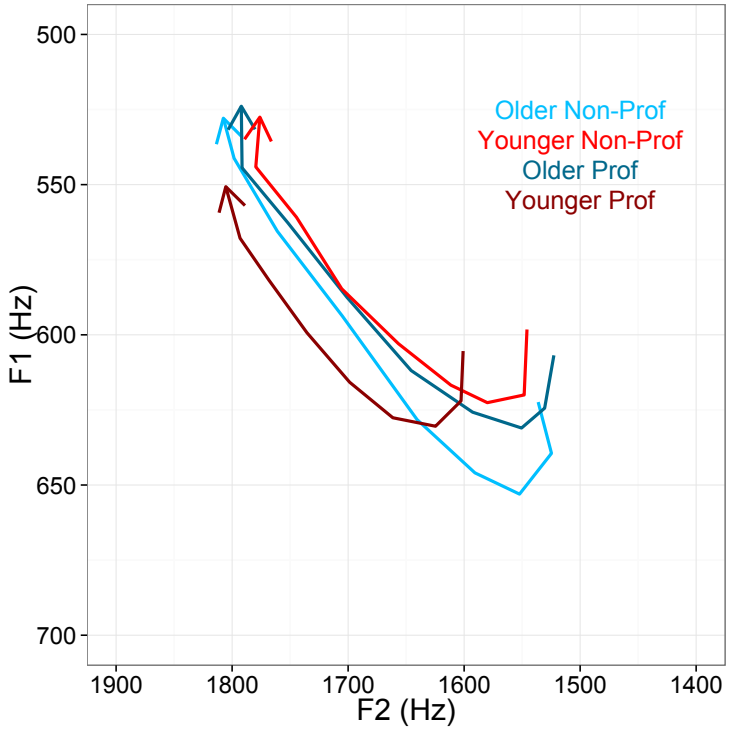

Figure 7.6: F1∼F2 plot of mean /eɪ/ trajectories according to age and class for 120 speakers based on eight tokens per speaker

7.2.6 Experiment

This chapter reports the results of an experiment which considers the effects of the definition of the relevant population according to socio-economic class and age. The structure of this experiment replicates those in §4.3.2, §5.3.1 and §6.3. A single set of homogeneous (with regard to class and age) test data was used across different relevant population systems. A single set of homogeneous test data was used to recreate FVC conditions in which the suspect and offender are typically members of the same sociolinguistic groups. The test set consisted of 20 speakers identified at random from the younger professional group. This group was used because it consisted of the largest number of speakers (33) allowing for separate sets of test and development/reference speakers. For both class and age, calibrated LLRs for the test set were computed using three systems based on different definitions of the relevant population (Table 7.1).

The **Matched** system involved development and reference data consisting of 24 speak-

ers who matched the test data for class (professional) or age (younger). The Matched system reflects an appropriate, narrowly defined relevant population according to the demographic background of the offender. In this case, the defence proposition may be formulated as: *the voice on the offender sample is not that of the suspect, but of another professional/younger male speaker of New Zealand English.*

Table 7.1: Number of development, test and reference speakers used in each system within each experiment

	System	Test	Development/Reference
Class	Matched	20 Younger Profs	24 Profs
	Mixed	20 Younger Profs	12 Profs + 12 Non-profs
	Mismatched	20 Younger Profs	24 Non-profs
Age	Matched	20 Younger Profs	24 Younger
	Mixed	20 Younger Profs	12 Younger + 12 Older
	Mismatched	20 Younger Profs	24 Older

The **Mixed** system contained 24 speakers consisting of equal numbers of professionals and non-professionals, and younger and older speakers. As in previous chapters, the Mixed system represents the current application of logical relevance to FVC casework, whereby neither class nor age are controlled (although the numbers of speakers from each group were balanced). In this case, the defence proposition may be formulated as: *the voice on the offender sample is not that of the suspect, but of another adult male speaker of New Zealand English.*

Finally, the **Mismatched** system used 24 non-professional or older speakers as development and reference data. This represents a narrowly defined but inappropriate relevant population, based on an incorrect judgement about the class or age of the offender. In this case, the defence proposition may be formulated as: *the voice on the offender sample is not that of the suspect, but of another non-professional/older male speaker of New Zealand English.* The use of Matched, Mismatched and Mixed development and reference data ensures that the different definitions of the relevant population were applied during the feature-to-score stage as well as during score-to-LR mapping. When varying the definition of class, development and reference data in all

systems were balanced for age. That is, the Matched and Mismatched sets consisted of equal numbers of younger (12) and older (12) speakers. Similarly, when varying the definition of age, development and reference data were balanced for class, with equal numbers of professionals (12) and non-professionals (12) used in the Matched and Mismatched sets. The same Mixed data, consisting of six speakers from each class-by-age combination, were used for both the class- and age-based experiments. In all cases, the 24 Matched, Mismatched and Mixed speakers were identified at random from the appropriate class-by-age sub-group.

Cross-validated (§3.2.2.3) SS (24) and DS (276) scores were initially computed for the Matched, Mismatched and Mixed development sets based on the suspect and offender data (four tokens each) using MVKD (§3.2.2.1). The input data consisted of four polynomial coefficients per formant, generating a 12 dimensional density function for the suspect and reference data. Due to the relatively small amount of available reference data (in terms of both N speakers and N tokens) and the high dimensionality of the input variable (12 dimensions), the experiments were also repeated using a multivariate normal LR approach (i.e. modelling with reference data with a multivariate normal distribution). This produced the same comparative patterns across conditions as reported in §7.3 but generally much weaker LLRs and worse overall performance.

SS (20) and DS (190) MVKD scores for the 20 test speakers were then computed using the Matched, Mismatched and Mixed reference sets (24 speakers) to generate three sets of parallel scores for both the class and age conditions. The distributions of SS and DS scores for each of the three sets of development data per condition were used to train a logistic regression calibration model for each system (§3.2.4.1). The calibration coefficients for each system were then applied to the appropriate set of test scores (for each system) to convert the scores to LLRs. Results are evaluated in terms of the distributions of LLRs and overall system validity (EER and C_{llr}). As in Chapter 6, the imprecision of LLRs from the same comparisons across different systems is captured using 95% CIs, which are compared across the class- and age-based results in §7.3.3.

7.3 Results

7.3.1 Socio-economic class

Figure 7.7 displays the Tippett plot of LLRs according to class-based definitions of the relevant population. The distributions of SS LLRs were similar across the three systems. In all cases, the median SS LLR was between zero and +1 (*limited* support for the prosecution), although in absolute terms the Matched median (+0.73) was marginally stronger than that for the Mismatched (+0.50) and Mixed systems (+0.58). There were small differences in terms of contrary-to-fact SS LLRs, with the Mixed system producing the strongest support for the defence (as low as -0.73). The Mismatched system produced the weakest contrary-to-fact SS LLRs (up to -0.15), as well as the lowest proportion of contrary-to-fact SS LLRs (5%).

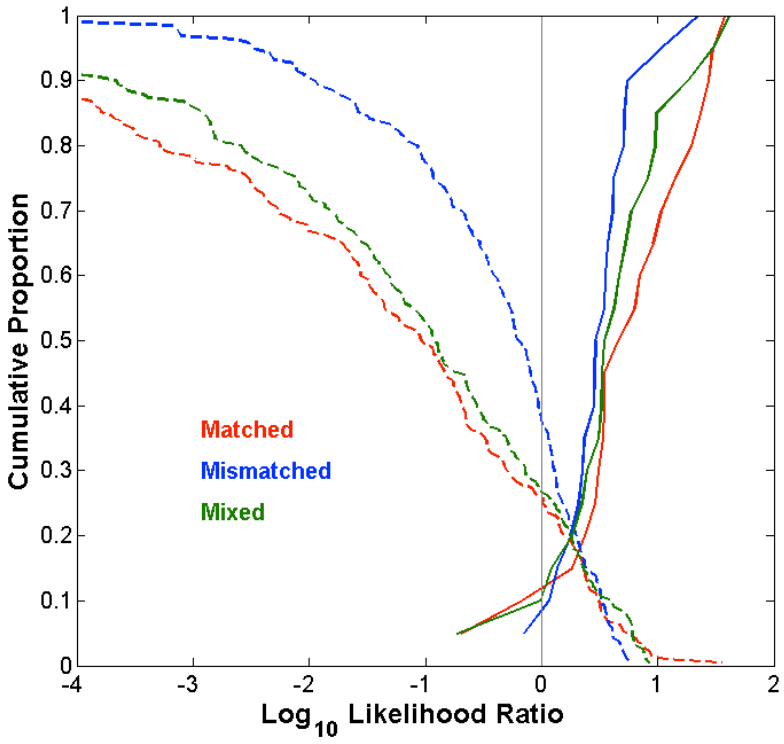

Figure 7.7: Tippett plot of SS and DS LLRs using the three class-based systems

Slightly larger differences between systems were revealed for DS LLRs. The Matched median (-1.04) was one order of magnitude stronger than the Mismatched median (-0.21), indicating that the Matched system generally produced the strongest DS LLRs. The Mixed median was marginally weaker (-0.93) than that of the Matched system. For both the Mismatched and Mixed systems, the difference with the Matched system was equivalent to the difference between *limited* (Mismatched/Mixed) and *moderate* (Matched) support for the defence. However, in absolute terms, the numerical difference between the Mixed and Matched medians was extremely small. The highest magnitude contrary-to-fact values were generated by the Matched system (up to +1.55), while the weakest contrary-to-fact LLRs were produced by the Mismatched system (as low as +0.77). However, the Mismatched system also produced the highest proportion of positive DS LLRs (37.4%).

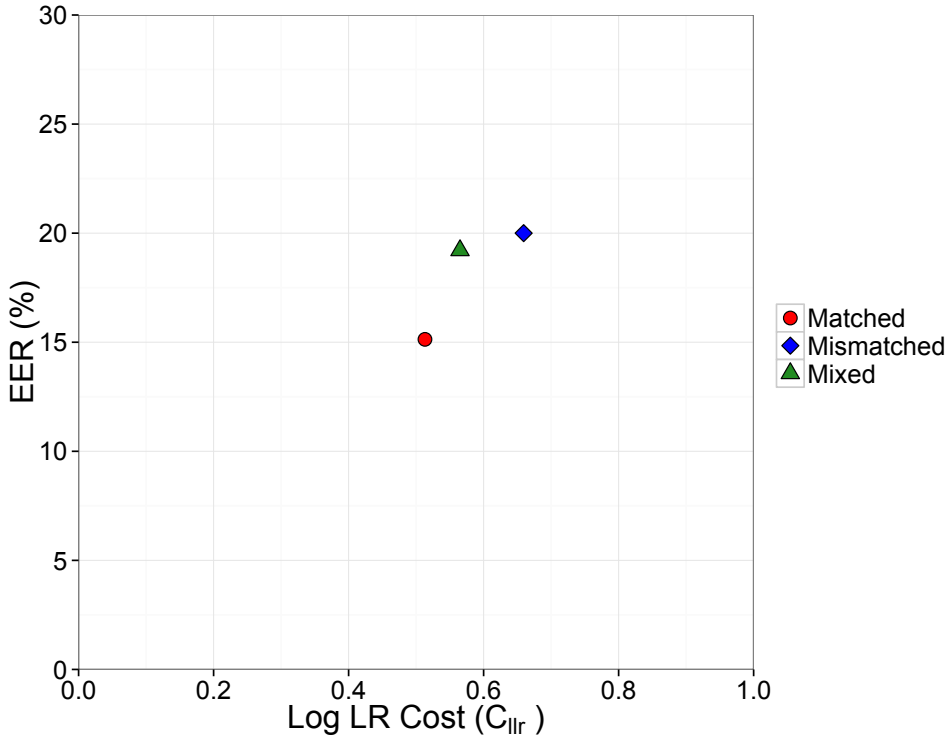

Figure 7.8: C_{llr} plotted against EER (%) for each of the three class-based systems

Figure 7.8 displays EER (%) and C_{llr} values for each of the three class-based systems. The Matched system generated the best EER (15.13%), followed by the Mixed system (19.21%) and finally by the Mismatched system (20%). While the performance of the Matched system was markedly better than the Mismatched or Mixed systems, the absolute EER difference between the Mixed and Mismatched systems was relatively

small. The same ordering of systems was found for C_{llr}. The best performing system was the Matched system with a C_{llr} of 0.51. The most divergent performance from the Matched system was found for the Mismatched system, which produced the highest C_{llr} value (0.66). Consistent with the distributions of LLRs in Figure 7.7, the C_{llr} for the Mixed system (0.57) was much closer to that of the Matched system. Despite this, validity was still worse using the Mixed system.

7.3.2 Age

Figure 7.9 displays the Tippett plot of LLRs for the three age-based systems. The general patterns were similar to those in §7.3.1, although the absolute differences between systems were smaller. SS medians across all three systems were within the same order of magnitude, between zero and +1 (*limited* support for the prosecution: Matched = +0.59; Mismatched = +0.67; Mixed = +0.58). The overall ranges of SS LLRs were also comparable, with values extending maximally to above +1 but below +2 (*moderate* support). The Mismatched system produced no contrary-to-fact SS LLRs. While the magnitude of the contrary-to-fact LLRs in the Matched and Mixed set were all within the range of zero and -1, the strongest contrary-to-fact SS LLRs, in absolute terms, were generated using the Mixed system.

Similar patterns to those in §7.3.1 are revealed in the distributions of DS LLRs. Median DS LLRs were all within the same order of magnitude, between zero and -1 (*limited* support for the defence), although the absolute numerical differences were greater than for the SS LLRs. The median was weakest using the Mismatched data (-0.11), suggesting that LLRs based on the Mismatched system generally offered weaker support for the defence compared with the Matched and Mixed systems. The distribution of Mixed DS LLRs was much closer to that from the Matched system. The magnitudes of contrary-to-fact LLRs were, however, similar across the three systems, producing values consistently lower than +1 (*limited* support for the prosecution). As in §7.3.1, the proportion of contrary-to-fact DS LLRs was greatest using the Mismatched system.

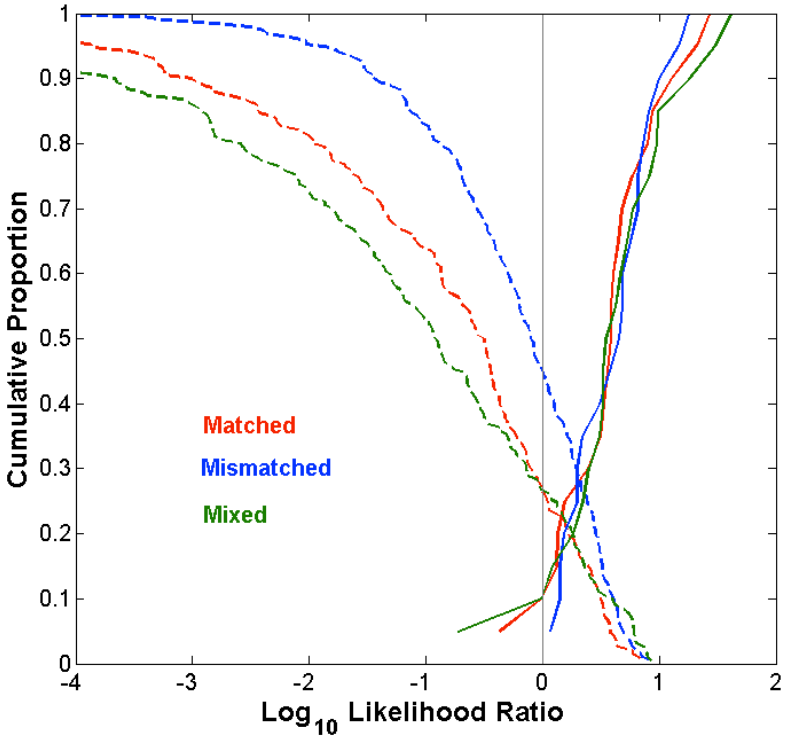

Figure 7.9: Tippett plot of SS and DS LLRs using the three age-based systems

Finally, Figure 7.10 displays system validity for each of the three age-based systems. The EER values pattern slightly differently from those in Figure 7.10. The best EER performance was found for the Mixed system (19.21%), followed by the Matched (20.79%) and Mismatched (29.47%) systems. In absolute terms the difference between the Mixed and Matched EERs was very small, while the differences between the Mixed/Matched and the Mismatched EERs were more considerable. The C_{llr} values were, however, consistent with the patterns in §7.3.1. The best performing system was again the Matched system (0.56), although the Mixed system produced a C_{llr} of almost equal magnitude (0.57). This indicates that the overall performance of the system using Mixed data was considerably closer to that using Matched data, compared with using Mismatched data. The highest C_{llr} was again generated by the Mismatched system (0.71).

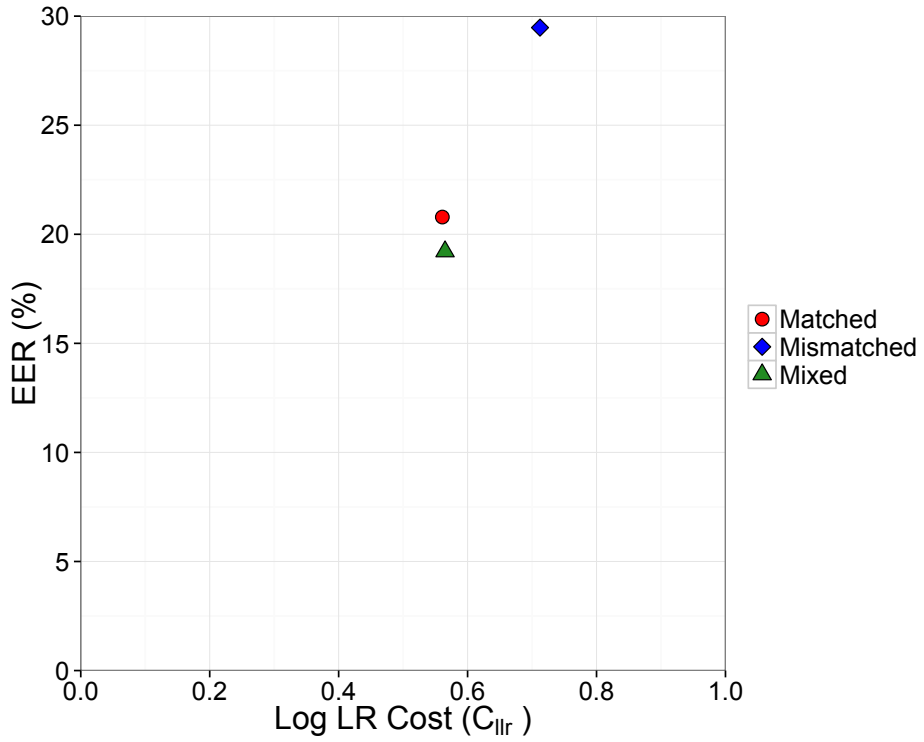

Figure 7.10: C_{llr} plotted against EER (%) for each of the three age-based systems

7.3.3 Reliability

Figure 7.11 displays Tippett plots of mean SS and DS LLRs with 95% CIs based on the output from the Matched, Mismatched and Mixed systems for class (left) and age (right). For both class and age, the width of the 95% CIs increased as the magnitude of the LLRs increased. Since the DS pairs generated larger magnitude LLRs than SS pairs, the CIs for the DS LLRs were generally wider than the CIs for the SS LLRs. Differences between the class and age results were also found. As shown in Figure 7.11, the 95% CIs were marginally narrower (particularly for DS LLRs) for age (mean CI = $\pm$0.95) than for class (mean CI = $\pm$1.12). This is consistent with the slightly smaller differences between systems in the Tippett plots in Figures 7.7 and 7.9. The difference in the mean 95% CIs suggests that individual comparisons were generally more sensitive to variation in class than variation in age, although the absolute difference in the mean CIs was relatively small.

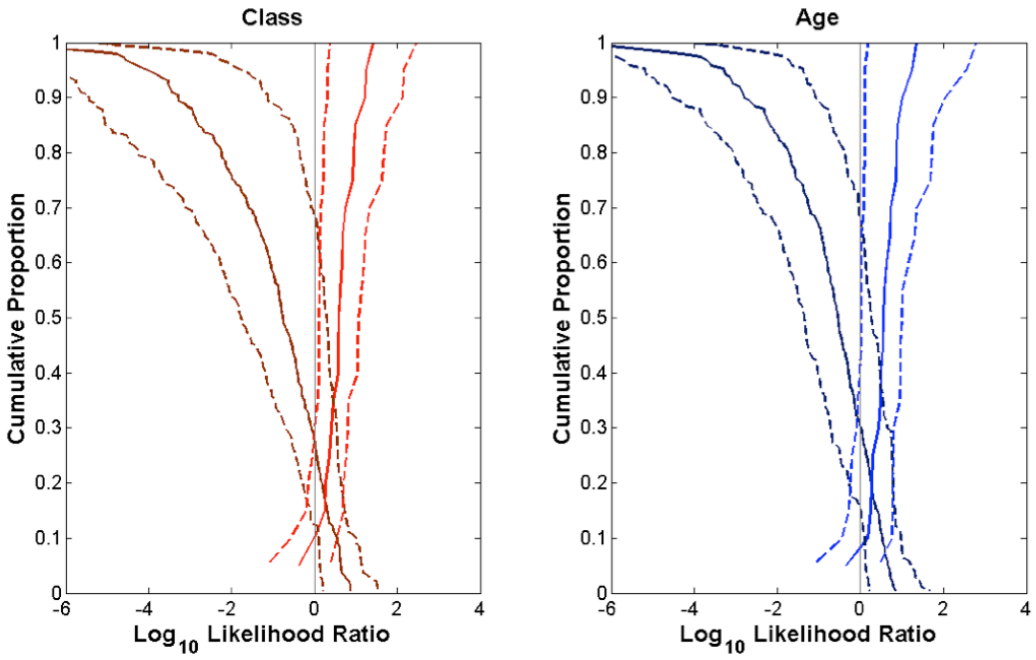

Figure 7.11: Tippett plots of mean SS (light) and DS (dark) LLRs and 95% CIs across the three systems based on class (left; red) and age (right; blue)

7.3.4 Systematic patterns or random variation?

The results in §7.3.1-§7.3.3 have revealed differences in the distributions of LLRs and system validity as a function of the class- or age-based definition of the relevant population. However, it is not clear whether these patterns are an inherent property of using Matched, Mismatched and Mixed data or reflect random variation as a function of sample size. To test this issue, the experiment was re-run 20 times using speakers sampled randomly from the appropriate class-by-age subsets. In each replication, the speakers used as test data and Matched, Mismatched and Mixed system data changed, although sample size remained constant (20 test speakers; 24 development/reference speakers). The test data again contained young professionals and was the same for the class- and age-based conditions in each replication. The results of the replications are evaluated against the main patterns in §7.3.1-§7.3.3.

Socio-economic class

Figure 7.12 displays the distributions of median LLRs for each system based on the 20 replications. As in §7.3.1, the SS median LLRs across all systems were within the range

of zero and +1, although in absolute terms the medians were marginally stronger for the Matched system than for the Mismatched and Mixed systems. The results of §7.3.1 and the replications were also broadly consistent with regard to the patterns for DS LLRs. The weakest median DS LLRs were produced by the Mismatched system. While the Mixed and Matched medians were more similar, the Matched system generally produced stronger DS median values across the replications. However, the differences between the systems in terms of DS medians were smaller than suggested in Figure 7.7.

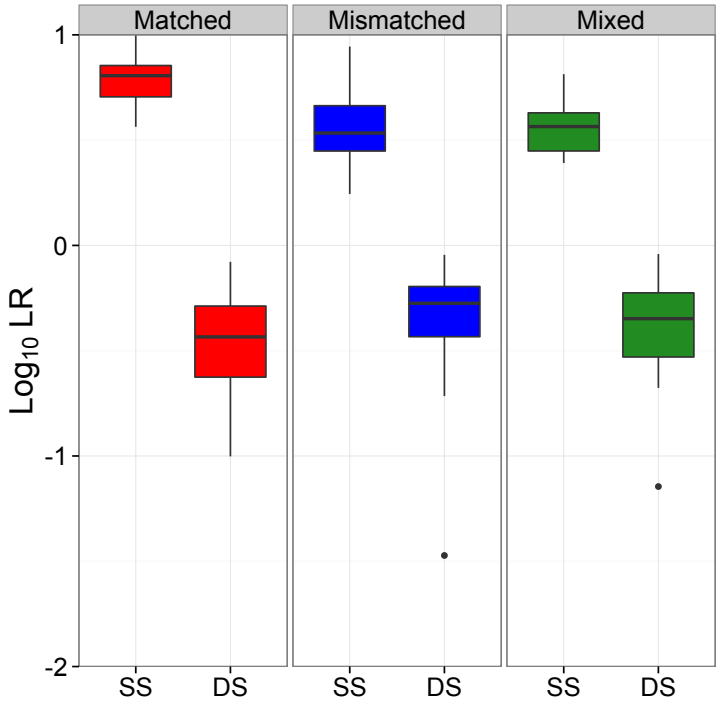

Figure 7.12: Boxplots of median SS and DS LLRs for the three class-based systems across the 20 replications

In §7.3.1, the Matched system produced the best EER and C_{llr} values, followed closely by the Mixed system, while much worse performance was found for the Mismatched system. The same patterns were also found in the replications (Figure 7.13). Both EER and C_{llr} were generally best using the Matched system. Although the Mixed median EER and C_{llr} values across the replications were similar to those from the Matched system, the interquartile range indicates that values were generally marginally higher using the Mixed system. As in §7.3.1, validity was generally worst using the Mismatched data, producing the highest C_{llr} value in 16 of the 20 replications.

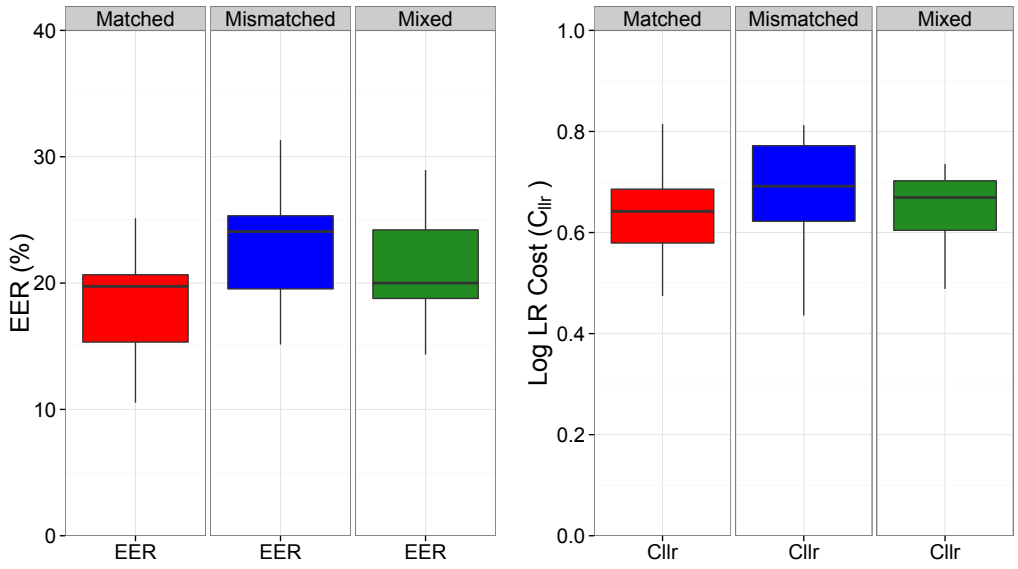

Figure 7.13: Boxplots of the distributions EER (left) and C_{llr} (right) values for the three class-based systems across the 20 replications

Age

In §7.3.2 the distributions of LLRs were found to be more stable across systems than in §7.3.1, although somewhat weaker DS LLRs were again found for the Mismatched system. Figure 7.14 displays the distributions of median LLRs for each age-based system across replications. Consistent with §7.3.2, in the replications SS medians were found to be stable across the three systems with values typically fluctuating between +0.4 and +0.6. The patterns in §7.3.2 were also found for DS LLRs, with the interquartile range of medians for the Mismatched system much closer to zero than those of the Matched and Mixed systems. This suggests that DS evidence was generally weakest using the Mismatched system. The DS medians were marginally stronger for the Matched system than for the Mixed systems, although across all systems the DS medians were within the range of zero and -1 (with the exception of three outlying values).

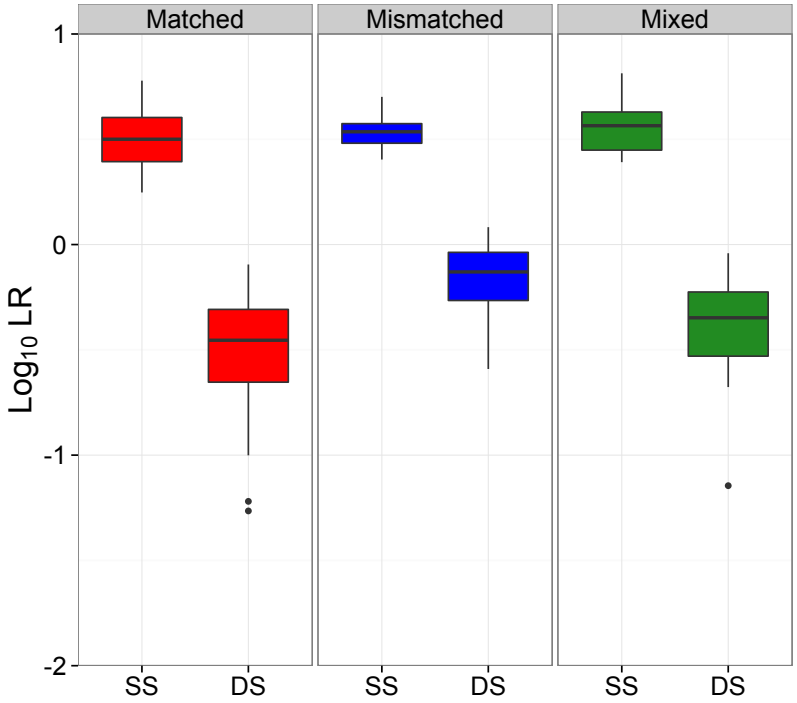

Figure 7.14: Boxplots of median SS and DS LLRs for the three age-based systems across the 20 replications

In §7.3.2, the Mixed system produced the best EER. The Matched system produced a marginally higher EER value (difference = 1.58%), while the reverse ordering was found for C_{llr}. For both EER and C_{llr}, however, the Mismatched system produced the worst validity. These patterns were also found across the replications (Figure 7.15). Although there was considerable overlap in the interquartile ranges of EER values, the median EER was lowest for the Mixed system, followed by the Matched and Mismatched systems. The distributions of C_{llr} values were very similar across the Matched and Mixed systems, although the median was marginally lower for the Matched system. The Mismatched system consistently achieved higher C_{llr} values. However, the absolute differences between the three systems in terms of both EER and C_{llr} were less than those found in §7.3.2.

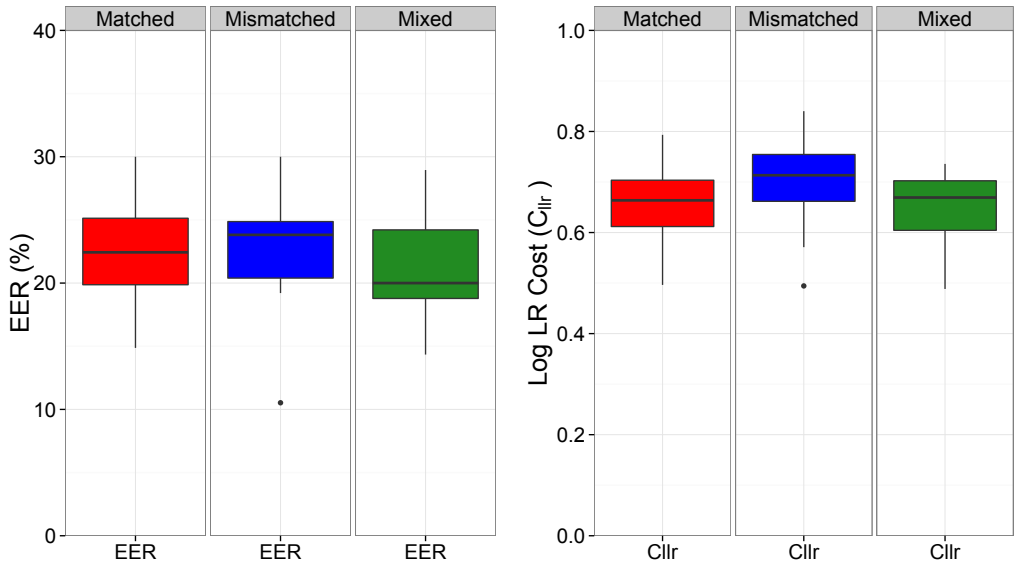

Figure 7.15: Boxplots of the distributions EER (left) and C_{llr} (right) values for the three age-based systems across the 20 replications

7.4 Discussion

The results in §7.3.1 and §7.3.2, supported by the multiple replications in §7.3.4, reveal a number of effects of using different definitions of the class- and age-based relevant population on LR output. For class, the results of §7.3.1 and the replications suggest that the distribution of SS LLRs was relatively stable across different relevant population systems, although in numerical terms the SS LLRs may be marginally stronger using Matched data. §7.3.1 suggests that the distribution of DS LLRs was shifted closer to zero (i.e. weaker evidence) using Mismatched data. As shown in §7.3.1, a consequence of this is that the Mismatched system also produced highest proportion of contrary-to-fact DS LLRs. In the replications, a similar pattern was found whereby DS medians were generally closer to zero using the Mismatched data. However, the magnitudes of the differences were much smaller than in §7.3.1, with all medians within the range of zero and -1.

For age, both §7.3.2 and the replications in §7.3.4 indicate that the distribution of SS LLRs was relatively stable across the three systems, with the median value in the replications fluctuating within a range of just 0.2. In terms of DS LLRs, a similar

pattern was found for age as that found for class. The distribution of DS LLRs was closer to zero (i.e. weaker evidence) using the Mismatched data compared with the Matched and Mixed systems, which generated similarly strong LLRs. However, the differences between the systems in terms of the distributions of DS LLRs were smaller for age than for class. Potential explanations for the patterning of SS and DS LLRs in these experiments, along with comparison of consistent patterns across this chapter and Chapters 4, 5 and 6 are discussed at §11.1.

Despite the relatively small differences across systems in terms of the distributions of LLRs, the 95% CIs suggest that differences in the definition of the relevant population with regard to class or age may have a substantial effect on the magnitude of the resulting LLR. For example, for the strongest mean DS LLR (-7.97) in §7.3.1 (class), the 95% CI was $\pm$2.66, while the CI for strongest mean SS LLR was $\pm$1.04. As this example highlights, imprecision was greater for DS pairs than for SS LLRs. This finding is consistent with the CI results in Chapter 6. The results in §7.3.3 also reveal that individual comparisons were marginally more sensitive to the different relevant population systems for class-based variation (mean 95% CI = $\pm$1.12) than for age-based variation (mean 95% CI = $\pm$0.95). The CIs based on class and age in this chapter were also considerably narrower than the CIs in Chapter 6, indicating that regional variation has a more considerable effect on LR output than class or age variation. However, this pattern is not necessarily generalisable to other FVC variables since the relative importance of regional background, class and age is, of course, determined on a variable-by-variable basis.

Marked differences were found between relevant population systems in terms of validity. For class, both §7.3.1 and the replications suggest that optimal EER performance was achieved using Matched data, followed by the Mixed system, although the absolute difference between these systems was rather small. For age, the Mixed system outperformed the Matched system in terms of EER, although again the differences were extremely small. However, for class and age the Mismatched system consistently produced the highest EER. In terms of C_{llr} the same patterns were found across both experiments in terms of the data in §7.3.1 and §7.3.2, as well as the replications in §7.3.4. The Matched systems consistently produced the best C_{llr}, followed closely by

the Mixed system. As with EER, the worst C_{llr} validity was consistently found using the Mismatched data.

While these results do indicate patterns of divergence across the systems, the effects on LR output were somewhat smaller than expected based on the predictions in §7.2.1. There are a number of potential reasons for this. Firstly, the range of acoustic-phonetic variation in the data was rather less marked that the descriptive literature had suggested. Secondly, the relatively small number of tokens per speaker means that the magnitude of the LRs will necessarily be relatively low, thus offering a narrower range of potential variation across systems. Thirdly, the results suggest that /eɪ/ offers relatively weak strength of evidence meaning that LLRs are inherently closer to zero (neutral evidence). Again, this reduces the range of potential variation across the three systems. However, these factors do not account for the fact that there are bigger differences between systems in terms of the magnitude of LLRs for class than for age, despite more marked age-based variation in the raw data (§7.2.5). This issue is discussed in §11.1.

7.5 Chapter summary

Based on systematic patterns across the results of §7.3.1-§7.3.3 and the replications at §7.3.4:

Socio-economic class

- SS LLRs marginally strongest using the Matched system.
  - Although medians consistently within the range of *limited* support for the prosecution across systems.
- Marginally weaker strength of DS evidence using Mismatched data with LLRs shifted towards zero relative to the Matched and Mixed sets.
  - Although medians consistently within the range of *limited* support for the defence across systems.

- Lowest EER and C_{llr} values found using the Matched system, followed by the Mixed system and then by the Mismatched system.

Age

- Distribution of SS LLRs stable across systems with medians fluctuating within a range of 0.2 (*limited* support for the prosecution).
- DS LLRs weakest using the Mismatched system.
  - Although medians consistently within the range of *limited* support for the defence across systems.
- EER generally better for Mixed system than Matched, although absolute differences extremely small.
- C_{llr} generally better for Matched system than Mixed, although absolute differences extremely small.
- Across both EER and C_{llr} the Mismatched system produced the worst performance.

General conclusions

- LR output from the Mixed systems generally relatively close to that of the Matched systems.
  - Output from the Mismatched systems most divergent from the Matched systems.
- 95% CIs reveal potentially considerable variability in LLRs for individual comparisons across systems.
  - Comparisons which generate high magnitude LLRs most affected.
  - DS comparisons more affected than SS comparisons (i.e. larger mean 95% CI).
- LR output more affected by variation in class than variation in age.

Chapter 8

Reference Sample Size: Raw Data

This chapter explores the effects of reference sample size on LR output. Using the formant trajectory data for /uː/ and /aɪ/ from Chapters 4 and 5, the results of two experiments are presented which address the issues of (1) the number of reference speakers and (2) the number of tokens per reference speaker in LR-based FVC. In both experiments, scores were computed as sample size (N speakers/ tokens) was systematically increased. At each stage, system validity (EER and C_{llr}) was calculated. In §8.4, the results are compared across the two phonemes.

8.1 Introduction

As outlined in §2.5, a substantial practical issue for the application of the numerical LR framework to FVC is the amount of data needed to generate precise estimates of strength of evidence and to adequately test system performance. The limited amount of previous research in this area (e.g. Ishihara and Kinoshita 2008) has focused on the estimation of between-speaker variation in the relevant population, through analyses of the sensitivity of LR output to the number of reference speakers. These studies are generally consistent in their findings, suggesting that the magnitude of the LR is misrepresented and considerably more variable when using small numbers of reference speakers.

The issue of the number of reference speakers used in LR testing is particularly relevant

when using MVKD (§3.2.2.1) to compute scores. This is because, as highlighted in Equation 3.1, the value for h, the smoothing parameter for the between-speaker KD, is "determined by a function of the number of groups (speakers) in the background" data (Morrison 2011a: 243). More generally, as highlighted in Rose (2013a), the degree of precision with which multivariate densities are modelled is proportional to the number of reference speakers and the number of dimensions per variable. That is, the more multidimensional the variable, the more reference speakers are needed for the multivariate model to be adequately precise. Therefore, individual FVC variables are expected to display different levels of sensitivity to the number of reference speakers used. Highly multidimensional variables are predicted to be more sensitive to small reference samples, meaning that stable LR output is achieved with more reference speakers than for variables with fewer dimensions.

A second issue relating to sample size is the number of tokens per reference speaker. This issue is of particular concern when computing LRs using speaker-dependent methods, such as MVKD, in which the distributions of values from each reference speaker are used to generate the model of the reference data. This can be seen in Equation 3.1, in which within-speaker means from the reference data ($\bar{x}_i$) are used to build the between-speaker KD model. Thus, a sufficient number of tokens per speaker are required in order for the estimation of $\bar{x}_i$ to be precise. This is an issue which has received little attention in LR-based research. Predictions of the potential effects of the number of tokens per speaker are the same as those for the number of speakers. LR output is expected to be highly unstable with small numbers of tokens per speaker and become increasingly more precise as the amount of data per speaker increases.

This chapter describes the results of two experiments. Experiment (1) investigates LR output as a function of the number of reference speakers. However, unlike in previous research, EER and C_{llr} were tested as the number of reference speakers increased. Experiment (2) considers the effects of the number of tokens per reference speaker on LR output. These experiments are the first to consider issues of sample size using highly multivariate formant trajectory data: specifically cubic polynomial coefficients from the formant trajectories of /uː/ (F1 and F2; eight dimensions) and /aɪ/ (F1∼F3; 12 dimensions). Further, these experiments develop on previous studies by considering the

comparative performance of two phonemes to test the relationship between sample size sensitivity and dimensionality.

8.2 Method

In this chapter the /uː/ data from Chapter 4 and the /aɪ/ data from Chapter 5 were used as input.

8.2.1 Data

8.2.1.1 /uː/

The /uː/ data consisted of cubic polynomial coefficients from F1 and F2 trajectories (eight dimensions). The eight NZ, eight Manchester, eight Newcastle and eight York used as test data in §4.3.1 were combined to create a single set of test data. This remained constant across all sample size conditions. Input data for each test speaker were 16 tokens in the four phonological conditions in Table 4.1. The available reference data consisted of an expanded version of the reference data used in §4.3.1, containing 120 NZE speakers with 10 tokens per speaker. In Experiment (1) (§8.3.1.1), scores were computed for the 32 test speakers using between ten and 120 reference speakers. In Experiment (2) (§8.3.2.1), scores were computed using the reference set of 102 speakers from §4.3.1 and up to 13 tokens per speaker.

8.2.1.2 /aɪ/

The /aɪ/ data consisted of cubic polynomial coefficients from F1∼F3 trajectories (12 dimensions). The eight speakers from each of the regional varieties (DyViS, Derby, Newcastle and Manchester) used in Chapter 5 were combined into a single set of test data. Data for each test speaker consisted of 10 tokens with broad controls over phonological context (see §5.2.3). The remaining 89 DyViS speakers with 10 tokens per speaker were used as reference data. In Experiment (1) (§8.3.1.2), scores were computed using between ten and 89 reference speakers. In Experiment (2) (§8.3.2.2),

scores were computed using all 89 reference speakers and between two and ten tokens per speaker.

8.2.2 Experiments

In Experiment (1), SS (32) and DS (992) scores were computed firstly using ten random reference speakers. The same comparisons were repeated as the number of reference speakers increased monotonically up to the maximum number of speakers available (/uː/ = 120, /aɪ/ = 89). At each stage the speaker added to the reference data was chosen at random. In Experiment (2), scores were computed initially using the first two tokens per reference speaker. A similar loop to that in Experiment (1) was then run, in which scores were computed after the addition of a single token per speaker up to the maximum number of tokens available (/uː/ = 13, /aɪ/ = 10). Tokens were added according to their position in the original recordings, with those produced earlier added first. This was intended to recreate variable sample length in FVC casework. Therefore, no control was made for the adjacent phonological context of tokens included at each stage.

MVKD (§3.2.2.1) was used to compute scores for both experiments. The effects of sample size were assessed using the distributions of SS and DS scores. System validity, in terms of EER and C_{llr}, was also assessed as a function of sample size. In §8.4, the results of the two experiments are compared across the two phonemes. Given the mix of regional dialects in each of the test sets (i.e. for both /uː/ and /aɪ/) and the lack of calibration, the range of scores is expected to be relatively large even when using the maximum available amount of reference data. Therefore, the distributions of scores and assessments of validity are not expected to be comparable with more forensically realistic systems based on /uː/ and /aɪ/ which use more sociolinguistically homogeneous test data.

8.3 Results

This section presents the results of Experiments (1) (§8.3.1) and (2) (§8.3.2). Following Rose (2012), the most precise system is assumed to be that with the largest amount of data and this is used as a baseline against which the results for other N speakers and N tokens conditions are compared. For both experiments, the results based on /uː/ input are assessed firstly, followed by the results for /aɪ/.

8.3.1 Experiment (1): Number of reference speakers

8.3.1.1 /uː/

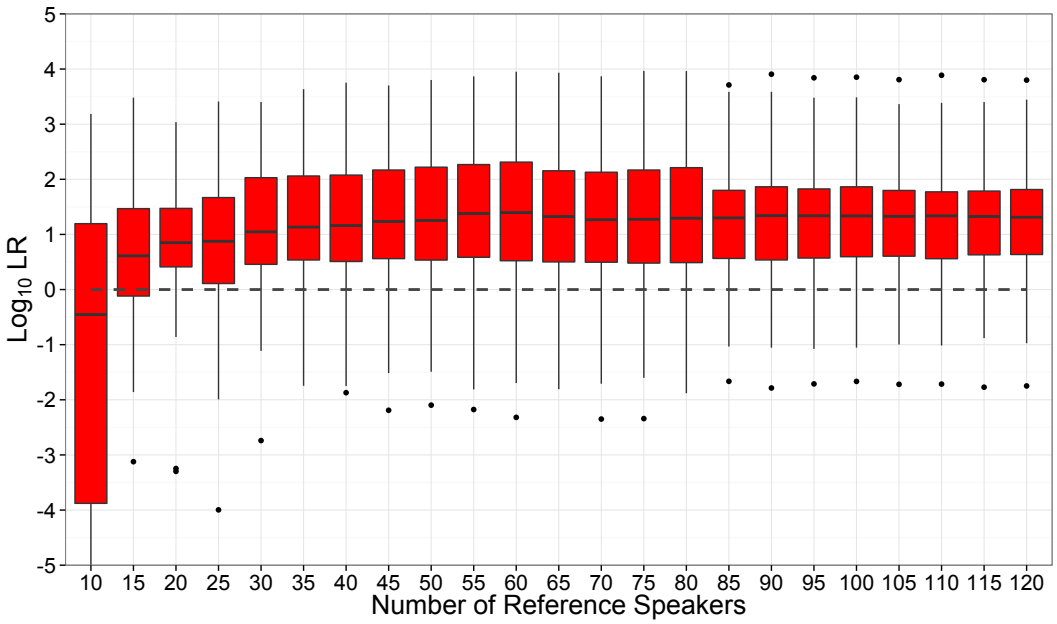

Figure 8.1: Boxplots (mid line = median, filled box = interquartile range (containing middle 50% of the data), whiskers = scores outside the middle 50%, dots = outliers; following Rose 2012) of SS scores based on /uː/ as a function of the number of reference speakers with the y-axis scaled to between +5 and -5 (outliers with ten speakers extent to -16)

Figure 8.1 shows the distributions of SS scores as a function of the number of reference speakers based on /uː/ input. With ten reference speakers the median score was negative, offering *limited* support for the defence. The median SS score became positive with

more than 15 reference speakers and continued to increase as a function of the number of speakers. The overall range of SS scores also narrowed as sample size increased. The proportion of contrary-to-fact (negative) SS scores was considerably higher with smaller numbers of reference speakers. However, with the inclusion of more than 30 speakers, SS scores became more robust to sample size and their distribution was essentially equivalent to that achieved using the maximum amount of reference data.

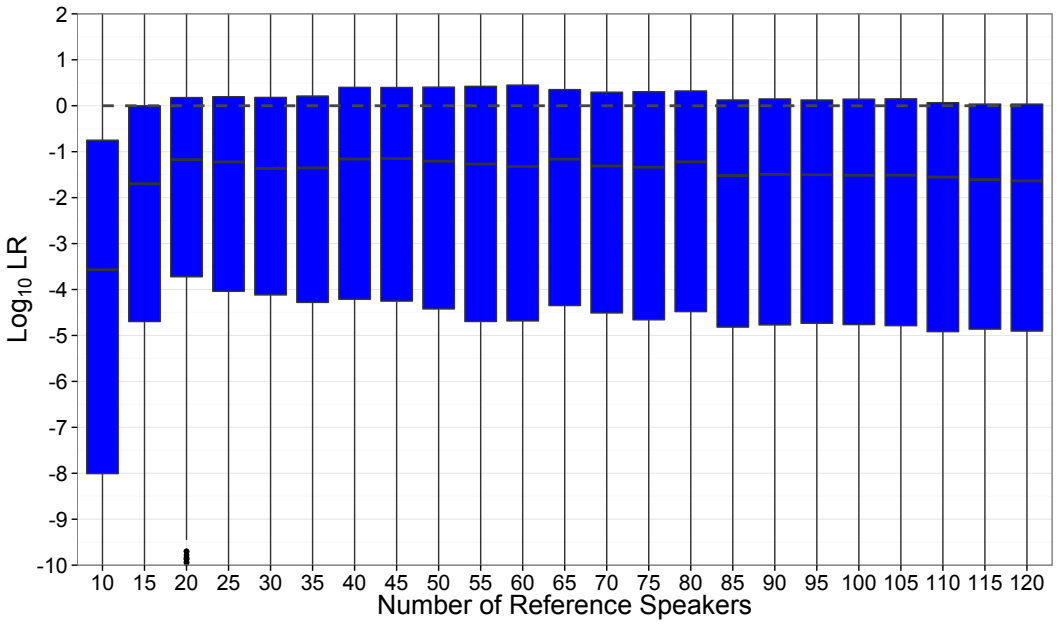

Figure 8.2: Boxplots of DS scores based on /uː/ as a function of the number of reference speakers with the y-axis scaled to between +2 and -10 (for all N speakers outliers extend to -20, with outliers using 10 speakers extending to almost -40)

Figure 8.2 displays the distributions of DS scores according to the number of reference speakers. The DS median was more sensitive to the size of the reference data. The median was two orders of magnitude stronger (i.e. negative) using ten reference speakers compared with the 120-speakers system. Using the smallest amount of reference data, the median score was equivalent to *strong* support of the defence. With the inclusion of between 15 and 120 speakers there was minor fluctuation in the median within the range of -1 to -2 (*moderate* support). As with SS pairs, the widest range and highest proportion of contrary-to-fact DS scores were found using the smallest amount of reference data. With more than 30 speakers the distributions of DS scores appeared to stabilise, although there was still some increase in the interquartile range and decrease in the proportion of errors after this point.

Considering the results on a comparison-by-comparison basis, the addition of certain individuals to the reference data clearly has a greater effect on scores than others. This may be due to the relative lack of phonological conditioning in the reference data. Given that certain reference speakers may have greater numbers of post-/j/ tokens, their addition may have a more substantial effect on the overall reference distribution. This may in turn serve to make pairs of target samples more or less typical relative to the background data.

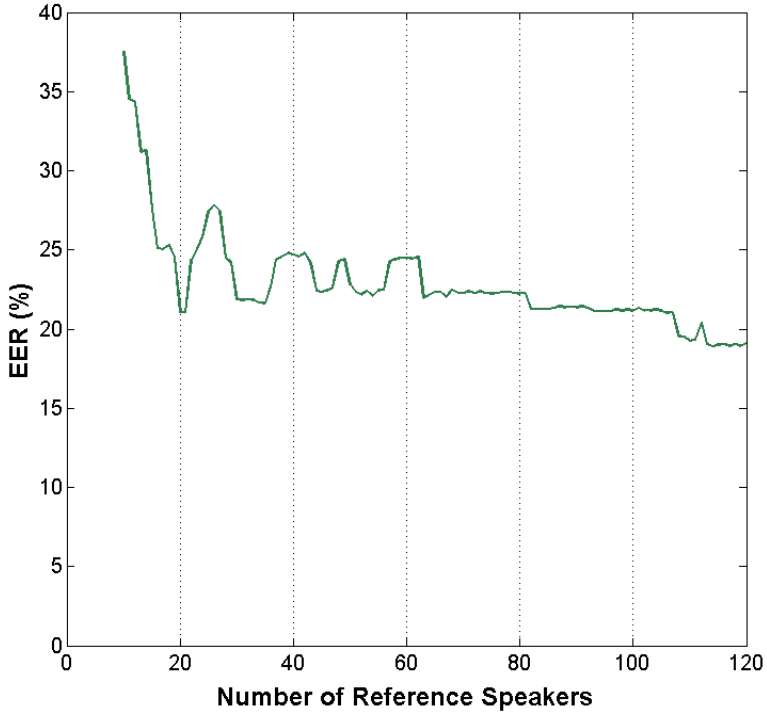

Figure 8.3: EER (%) based on /uː/ as a function of the number of reference speakers

The effect of the number of reference speakers on EER is shown in Figure 8.3. The highest EER (37.6%) was achieved with the smallest number of speakers. This reflects the high proportion of contrary-to-fact SS and DS scores using small amounts of reference data. Between ten and 20 reference speakers, there was marked improvement in EER followed by fluctuation within a range of around 10% (between 22% and 32%) when using between 20 and 60 reference speakers. There was an overall trend for an improvement in EER with the inclusion of more data, highlighted by the fact that the lowest EER was achieved using all available reference speakers (19.1%).

Figure 8.4 shows C_{llr} as a function of the number of reference speakers. As with EER, the worst performing system was that based on the smallest amount of reference data.

With ten speakers, C_{llr} was considerably higher than unity (> 5) indicating very bad system validity. This reflects the high magnitude of contrary-to-fact SS scores, which in one outlying case was less than -16. C_{llr} improved markedly with the inclusion of more speakers, such that validity appeared more stable by the inclusion of 20 reference speakers. However, even with more than 20 reference speakers, validity continually improved with the inclusion of more reference data (Figure 8.4; right).

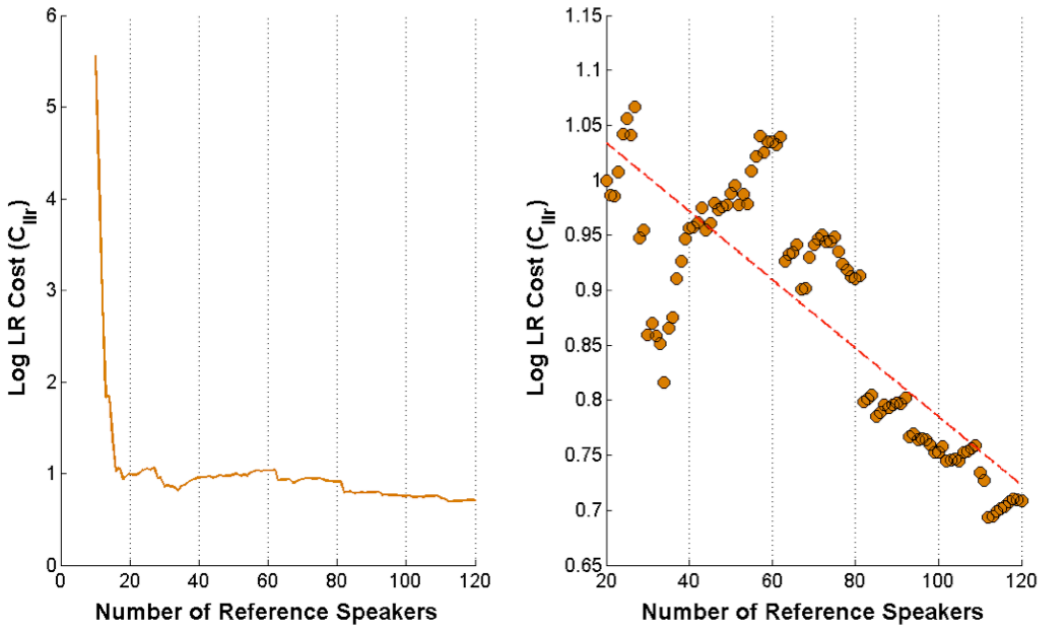

Figure 8.4: C_{llr} based on /uː/ as a function of the number of reference speakers in all conditions (left) and with between 20 and 120 speakers with linear trend (right)

8.3.1.2 /aɪ/

Figure 8.5 displays the distributions of SS scores using /aɪ/ as a function of the number of reference speakers (between ten and 89). The system based on all available reference data (89 speakers) achieved a median score of +2.84 (*moderately strong* support for the prosecution). However, when using the smallest amount of available data (10 speakers), the median was considerably weaker, offering *moderately strong* contrary-to-fact support for the defence (i.e. negative). With ten reference speakers a higher proportion of contrary-to-fact SS scores was also found. Although the general patterning in Figure 8.5 is similar to that for /uː/ (Figure 8.1), SS output for /aɪ/ was found to be much more sensitive to the increase in sample size. Between 25 and 45 speakers, the

median score continued to increase, such that with 45 speakers the median was one order of magnitude higher (+3.56) than with 89 speakers (+2.84). This is equivalent to the difference between *moderately strong* (89 speakers) and *strong* support (45 speakers). Only with the inclusion of 55 or more speakers did the distributions of SS scores appear to become stable.

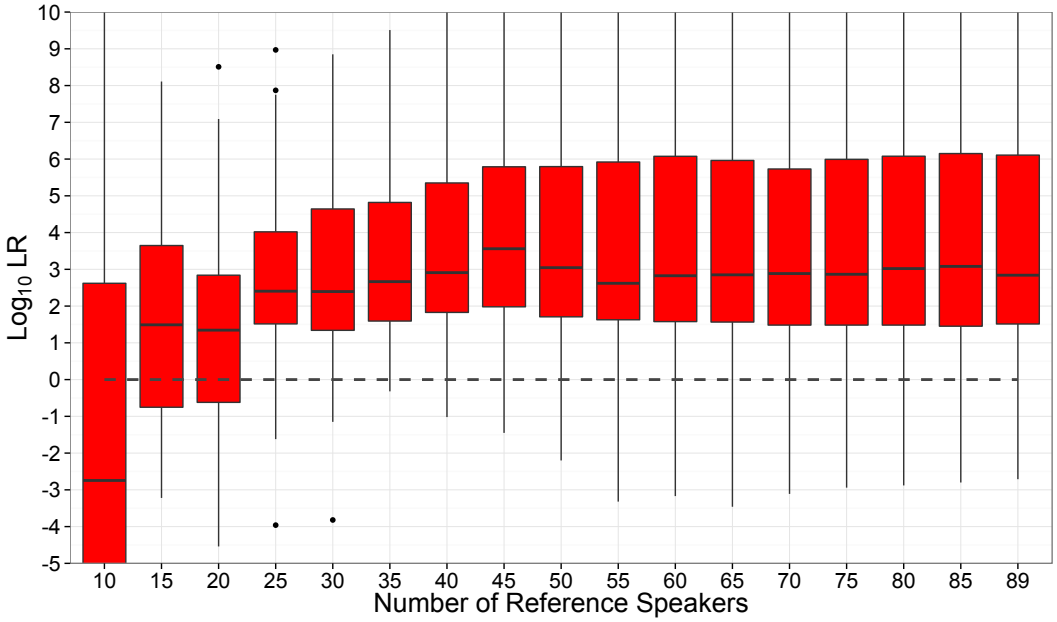

Figure 8.5: Boxplots of SS scores based on /aɪ/ as a function of the number of reference speakers with the y-axis scaled to between 10 and -5 (outliers with 10 speakers extent from c. +13 to -21)

Figure 8.6 shows the distributions of DS scores for /aɪ/ as a function of sample size. The DS median was seven orders of magnitude stronger with ten reference speakers than with 89 speakers. The overall range of DS scores was also considerably wider when using ten speakers with values ranging from +10 to -80, compared with +12 to -55 using 89 speakers. As for the SS scores, DS output based on /aɪ/ was more sensitive to sample size variation than /uː/. With the inclusion of between 15 and 50 speakers, the median strength of evidence was up to two orders of magnitude weaker than with 89 speakers. The distribution of DS scores only stabilised with the inclusion of more than 55 speakers. In terms of verbal equivalents, such sensitivity to sample size is less problematic for DS scores than for SS scores, since the majority of scores were consistently equivalent to *very strong* support for the defence, irrespective of absolute numerical differences.

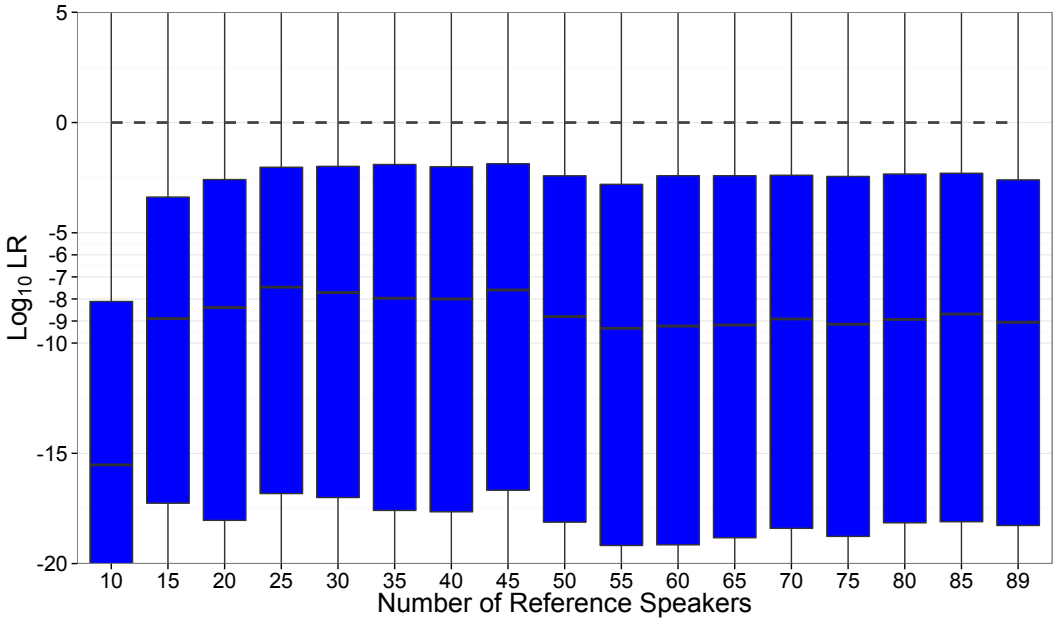

Figure 8.6: Boxplots of DS scores based on /aɪ/ as a function of the number of reference speakers with the y-axis scaled to between +5 and -20 (outliers with ten speakers extent from c. +10 to -80)

The highest EER (25%) was found using the smallest number of speakers, reflecting the high proportion of contrary-to-fact SS scores (Figure 8.7). As in Figure 8.3, EER became relatively stable after the inclusion of 30 reference speakers, although some random variation was found after this point within a range of around 3%. The reason for this is that a proportion of the SS and DS scores were situated around zero (threshold) such that small changes in the make-up of the reference data can cause positive values to become negative and vice-versa. In numerical terms, the differences between positive and negative scores located around zero can be extremely small, despite their relatively large effects on categorical validity.

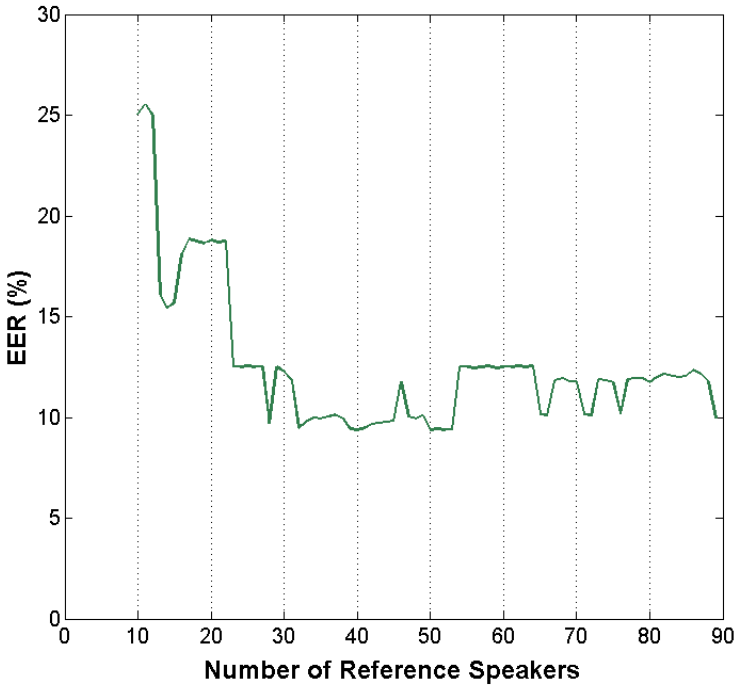

Figure 8.7: EER (%) based on /aɪ/ as a function of the number of reference speakers

With fewer than 22 reference speakers, C_{llr} was considerably greater than one (Figure 8.8), consistent with the high magnitude of the contrary-to-fact SS scores when using small numbers of reference speakers. This reflects extremely bad system validity. There was a marked improvement in C_{llr} between ten and 30 speakers, consistent with the pattern for /uː/. However, the lowest C_{llr} (best validity) was achieved using 34 reference speakers (0.499), reflecting the reduced range of SS scores and therefore reduced magnitude of contrary-to-fact scores between 25 and 35 speakers. Unlike /uː/, no linear improvement in C_{llr} was found for /aɪ/ as the number of reference speakers increased. Rather, there was an increase in C_{llr} between 34 and 42 speakers, followed by random variation around 0.7 up to 89 speakers.

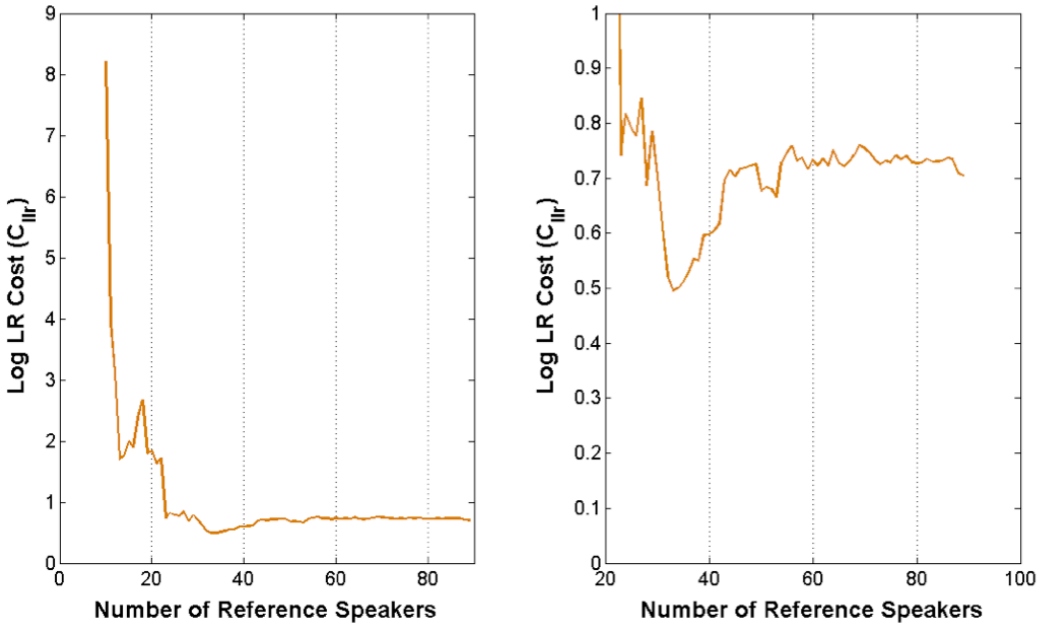

Figure 8.8: C_{llr} based on /aɪ/ as a function of the number of reference speakers in all conditions (left) and with between 20 and 89 speakers (right)

8.3.2 Experiment (2): Number of tokens per reference speaker

8.3.2.1 /uː/

The distributions of SS scores for /uː/ according to the number of tokens per reference speaker are shown in Figure 8.9. The median SS score was somewhat weaker with two tokens per reference speaker compared with 13 tokens, equivalent to the difference between *limited* and *moderate* support for the prosecution. The range of scores in the two-tokens condition was also considerably greater than in any other condition due to the proportion of contrary-to-fact values. Between three and 13 tokens the median SS score remained relatively stable, fluctuating within a range of +1 to +2. There was some instability in the interquartile and overall ranges as the amount of data increased. Most notably, the effects can be seen in the most negative outlying value, which increased by three orders of magnitude (from *strong* to *limited* contrary-to-fact evidence) between the five- and 13-tokens conditions.

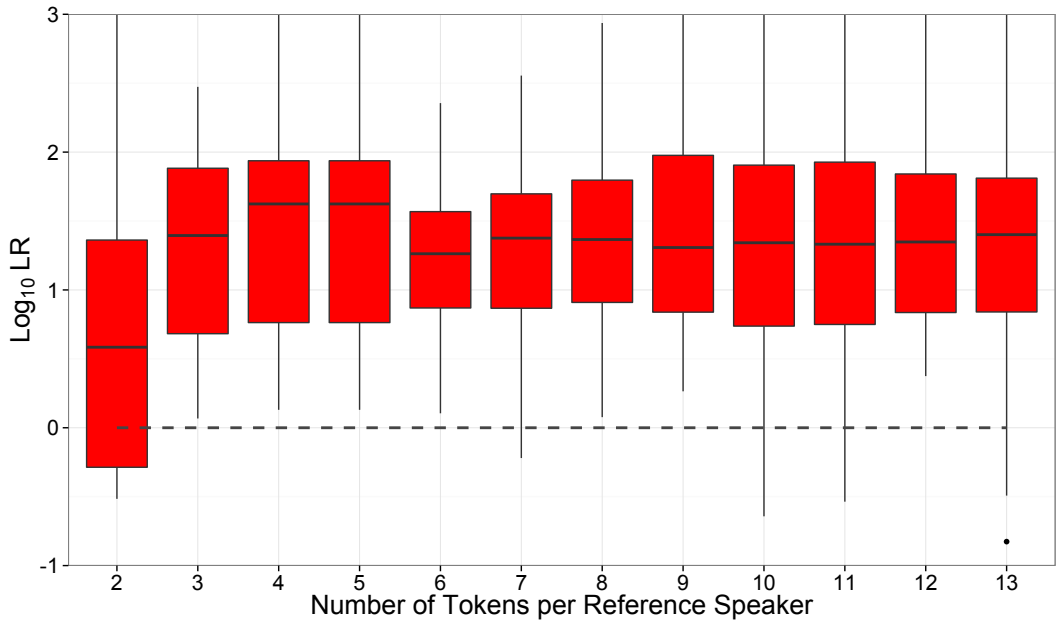

Figure 8.9: Boxplots of SS scores based on /uː/ as a function of the number of tokens per reference speaker with the y-axis scaled to between +3 and -1 (outliers extend to c. -10 using two tokens per reference speaker)

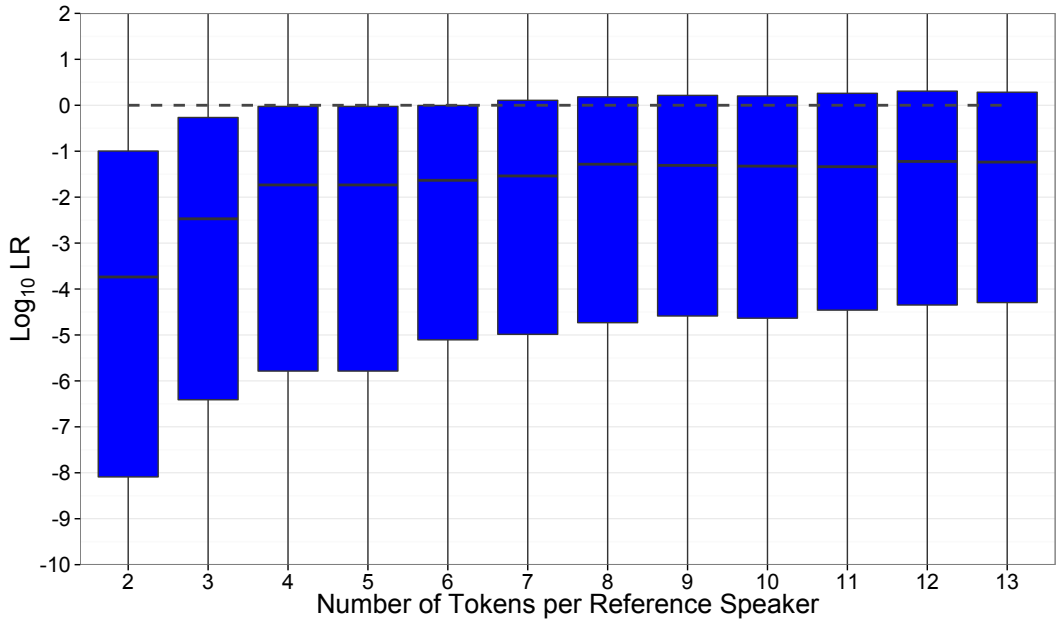

Figure 8.10: Boxplots of DS scores based on /uː/ as a function of the number of tokens per reference speaker with the y-axis scaled to between +2 and -10 (outliers across all conditions extend to > -10, with outliers of up to -30 using two tokens per speaker)

Variation in DS scores (Figure 8.10) as a function of the number of tokens per reference speaker was found to be more systematic. Median DS scores offered considerably

greater support for the defence with small amounts of data per speaker. Between two and 13 tokens, the strength of evidence decreased by the equivalent of two orders of magnitude from *strong* to *moderate* support for the defence. The range of DS scores was greatest when using smaller numbers of tokens per speaker, with outlying DS scores decreasing in strength by as many as ten orders of magnitude between two and 13 tokens. However, DS scores became more stable with the inclusion of more than eight tokens, where the distribution of scores was very similar to that using 13 tokens. While the magnitude of the strongest contrary-to-fact score decreased as sample size increased, the percentage of false hits increased slightly as a function of the number of tokens.

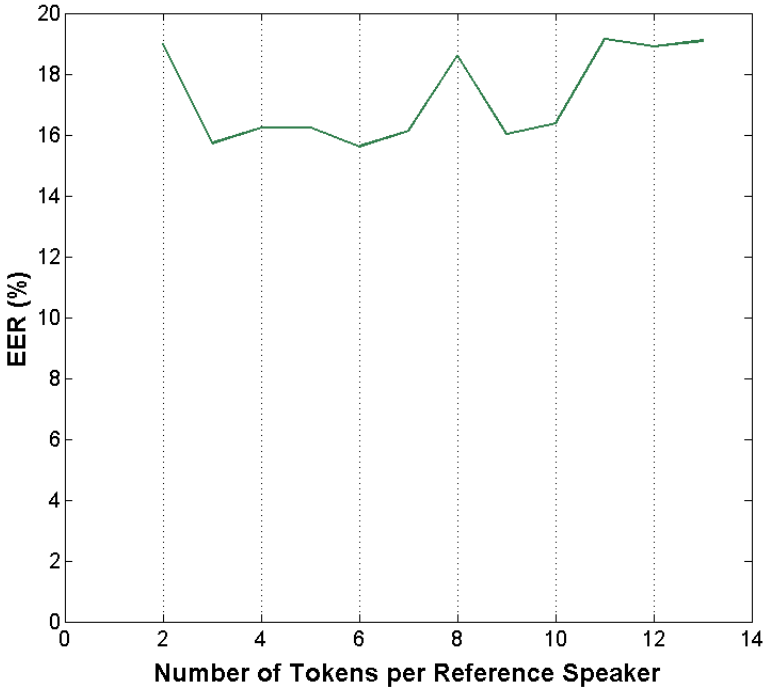

Figure 8.11: EER (%) based on /uː/ as a function of the number of tokens per reference speaker

EER was relatively unstable as the number of tokens per speaker increased, displaying no systematic pattern as a function of sample size (Figure 8.11). This is consistent with the fluctuation in the proportion of SS scores offering support for the defence across conditions in Figure 8.9. Figure 8.12, however, displays a systematic pattern of variation in C_{llr} as a function of sample size. Validity was worst with two tokens per speaker ($C_{\text{llr}} > 2$). As the number of tokens increased, C_{llr} improved such that the

best validity was achieved using all 13 tokens per speaker. The biggest improvement in C_{llr} occurred between two and six tokens per speaker, with only marginal improvement after this point.

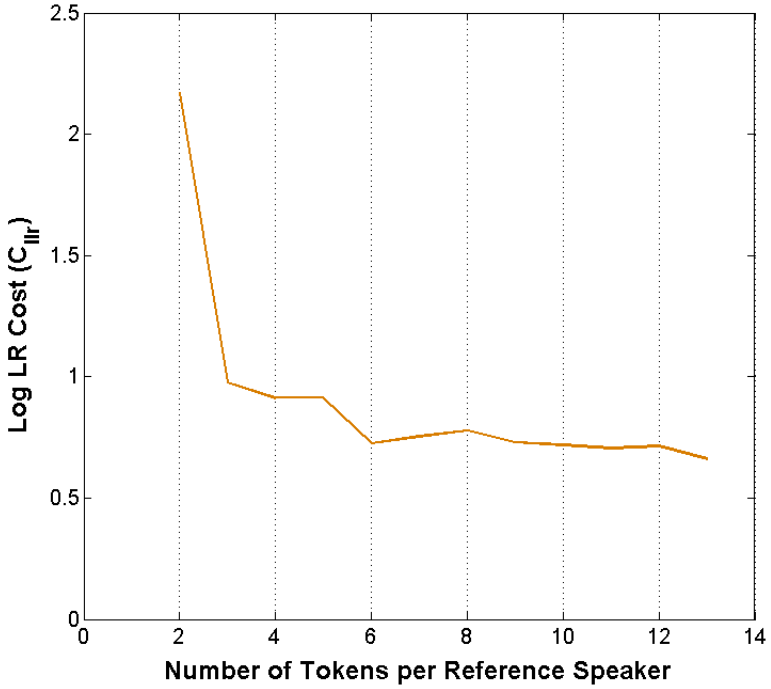

Figure 8.12: C_{llr} based on /uː/ as a function of the number of tokens per reference speaker

8.3.2.2 /aɪ/

Unlike in Figure 8.9, SS scores using two tokens per speaker based on /aɪ/ input were considerably stronger in magnitude than when using three or more tokens (Figure 8.13). Compared with the median SS score of +2.84 using ten tokens, the median using two tokens was almost 20 times stronger. As the number of tokens per speaker increased, the overall magnitude of SS scores decreased. However, at no point between three and nine tokens per speaker was the median within the same order of magnitude as that based on ten tokens. In terms of the overall range, the distribution of scores was relatively stable with the inclusion of more than six tokens per speaker.

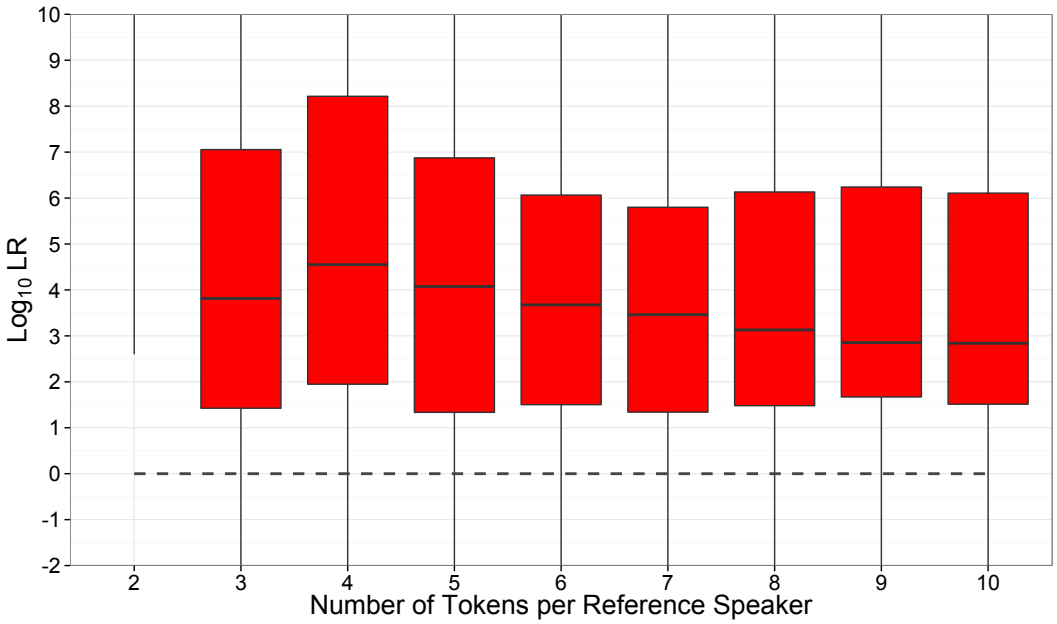

Figure 8.13: Boxplots of SS scores based on /aɪ/ as a function of the number of reference speakers with the y-axis scaled to between 10 and -5 (outliers with 10 tokens per speaker extent from c. +13 to -21)

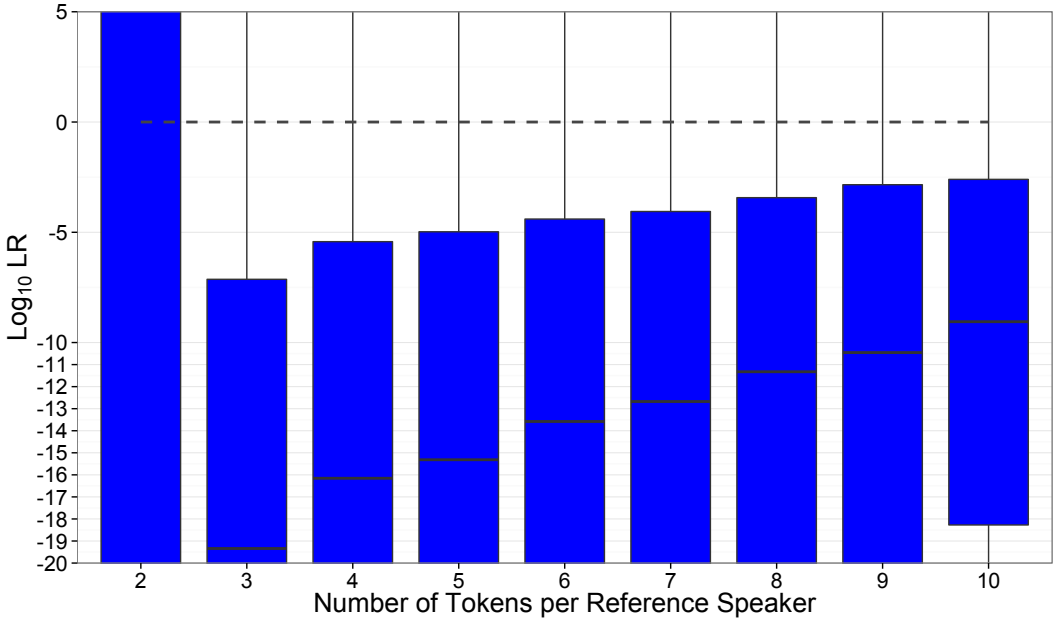

Figure 8.14: Boxplots of DS scores based on /aɪ/ as a function of the number of reference speakers with the y-axis scaled to between +5 and -20 (outliers with two tokens per speaker extent from c. +167 to -136)

A different pattern from that in Figure 8.10 was found for DS scores using /aɪ/ (Figure 8.14). At no point across conditions was the distribution of DS scores particularly

stable. The median DS score using two tokens per speaker was +9.875, reflecting the fact that the majority of DS comparisons produced high magnitude contrary-to-fact support for the prosecution. Between three and ten tokens, there was a decrease in the median strength of evidence, although at no point between three and nine tokens was the median within the same order of magnitude as when using ten tokens. When using two tokens per speaker, the overall range of scores was considerably wider compared with the ten-tokens system. Although the range decreased as sample size increased, the strength of contrary-to-fact scores remained relatively stable across conditions (maximally around +12).

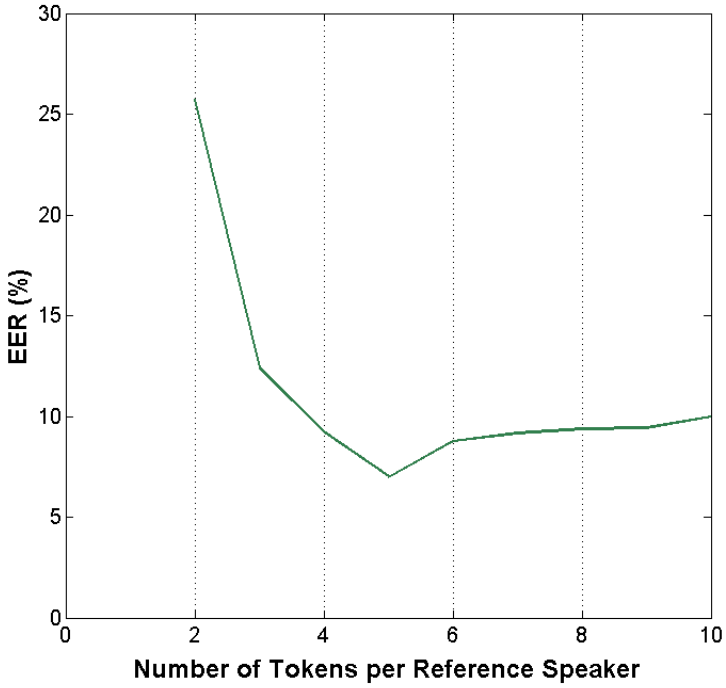

Figure 8.15: EER (%) based on /aɪ/ as a function of the number of tokens per reference speaker

Figure 8.15 shows that EER was highest (c. 25%) when using the smallest amount of reference data (two tokens per speaker). There was a marked improvement in EER as the number of tokens increased, although the lowest EER was found with five tokens. However, EER can be considered stable with the inclusion of more than six tokens per speaker. Similarly, an extremely high C_{llr} was found when using two tokens per speaker (31.54) (Figure 8.16). This is consistent with the large proportion of extremely large magnitude contrary-to-fact DS scores. There was marked improvement in C_{llr} as the

number of tokens increased, with values of less than one achieved by the inclusion of four tokens. C_{llr} was also found to stabilise with the inclusion of six or more tokens per speakers, although the lowest C_{llr} (0.62) was achieved with seven tokens per reference speaker. This is considered a random vagary the dataset, rather than a systematic pattern.

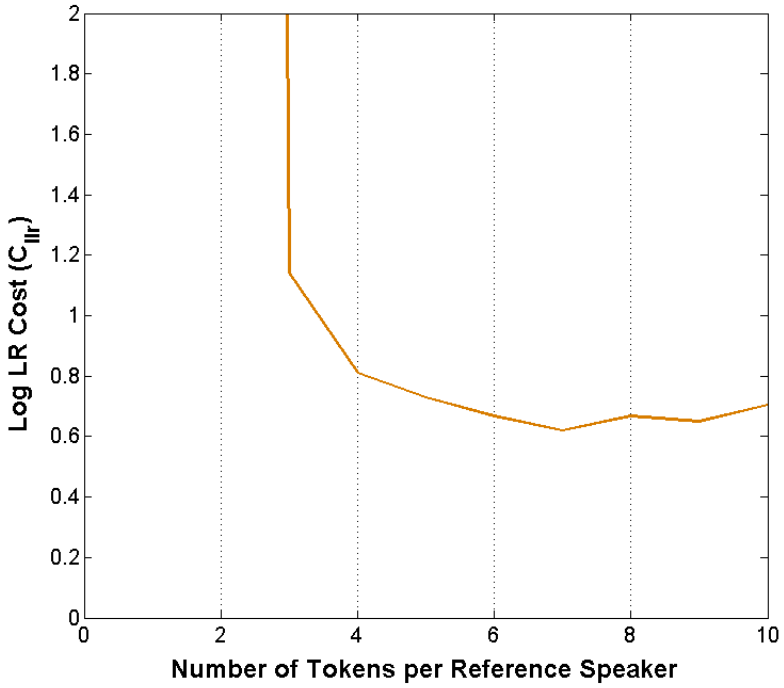

Figure 8.16: C_{llr} based on /aɪ/ as a function of the number of tokens per reference speaker

8.4 Discussion

The results presented in Experiment (1) (§8.3.1) support Ishihara and Kinoshita's claim that LR precision is "heavily compromised if the population data (are) limited to a small number of speakers" (2008: 1941). For both datasets, scores were misrepresentative and unstable with between ten and 20 reference speakers, relative to the scores using all of the available data. In particular, SS scores were generally found to be weaker and DS scores stronger with considerably wider ranges when using small amounts of reference data. The magnitude of SS scores also increased, while the magnitude of DS

scores decreased as sample size increased. The direction of these effects are consistent with Ishihara and Kinoshita (2008: 1942).

There were, however, differences between /uː/ and /aɪ/ in terms of the sample size sensitivity. As in Ishihara and Kinoshita (2008), SS and DS scores for /uː/ were found to stabilise with the inclusion of more than 30 speakers. However, scores for /aɪ/ were found to be considerably more sensitive to the number of reference speakers. With more than 30 speakers, large differences were still found in the median SS and DS scores and the overall range of scores relative to the 89-speakers system. For both SS and DS pairs, stability in the distribution of scores was only achieved using more than 55 reference speakers.

In terms of validity, for both /uː/ and /aɪ/, EER and C_{llr} were markedly higher with very small numbers of speakers. With between ten and 30 speakers, some improvement in EER and C_{llr} was found. For /uː/, continued improvement in validity was also found as the number of speakers increased between 30 and 120. This is consistent with Ishihara and Kinoshita (2008), who also found continual EER improvement as sample size increased. These results highlight an important issue relating to the trade-off between the amount of reference data and system validity. Although the most valid system was that with the largest amount of data the absolute improvement of the 120-speakers system is marginal compared with the scores using 30 speakers. For /aɪ/, EER stabilised after the inclusion of more than 30 speakers, with no evidence of a linear improvement in performance as a function of sample size. Similarly, no linear trend was found for C_{llr}. Figure 8.7 shows considerable improvement in C_{llr} between ten and 34 speakers, with optimum C_{llr} achieved using 34 speakers. The random vagaries of this dataset highlight that the best system performance is not necessarily achieved using the most amount of data.

These results show that linguistic-phonetic variables behave differently with regard to their sensitivity to sample size. This is predicted by the dimensionality of the two variables used in these experiments. The differences between the variables are consistent with the predicted relationship between sample size sensitivity and dimensionality. That is, the 12 dimensional /aɪ/ data were more sensitive to sample size variation that the eight dimensional /uː/ data. From a practical perspective, the results are extremely positive.

The fact that stable distributions of scores were achieved based on an eight dimensional variable /uː/ using 30 reference speakers means that it may be possible to achieve robust LR output for variables with fewer dimensions (as is typical in linguistic-phonetic FVC) using fewer than 30 speakers (see Chapter 9).

However, the differences in the results for /uː/ and /aɪ/ may also, in part, be due to the regional divergence between the speakers used as test data. As shown in the results in Chapters 4 and 5, regional differences for /aɪ/ are considerably greater than for /uː/. Given such variation, data from comparisons involving the 24 regionally mismatched (relative to the reference data) are expected to be further onto the tails of the reference distribution than the matched data. Therefore, it is likely that smaller fluctuations in the make-up of the reference data have a much bigger effect of the outcome of LRs for these comparisons. To overcome the limitations of using regional mixed test data here, the experiments in Chapters 9 and 10 are based on more sociolinguistically homogeneous sets of development, test and reference data.

The findings of Experiment (2) are consistent with Experiment (1) in that LR output was unstable and misrepresentative when using small numbers of tokens per speaker. However, marked differences were found between /uː/ and /aɪ/. For /uː/, SS scores were generally weakest using two tokens per speaker. As sample size increased the SS median and proportion of misses varied randomly. DS scores for /uː/ were stronger in magnitude using small numbers of tokens, although the distributions of scores stabilised after the inclusion of seven tokens. For /aɪ/, the magnitude of SS scores was greatest using two tokens per speaker, followed by continual decrease in median strength of evidence as sample size increased. For DS scores, extremely high magnitude contrary-to-fact scores were generated using two tokens per speaker, followed by a continual decrease in the median with increased amounts of data.

Some differences were also found in system validity. EER for /uː/ fluctuated randomly as the number of tokens increased, reflecting no systematic pattern of EER as a function of sample size. For /aɪ/, EER was highest using small amounts of data and improved with the inclusion of more data, such that stable EER was found with more than six tokens per speaker. In terms of C_{llr}, the results across the two datasets were far more comparable. Extremely high C_{llr} values were recorded using two tokens per speaker.

C_{llr} improved markedly between two and six tokens per speaker, such that stability was achieved, for both variables, with greater than six tokens per speaker.

These results suggest that small amounts of data per reference speaker should be avoided when computing numerical LRs, since the models of within-speaker variability cannot be estimated precisely. This finding raises concerns about the magnitude of the LRs reported in research based on small amounts of data per speaker (e.g. two tokens of five vowel in Rose 2011a). Indeed, the lack of stability in the distributions of scores suggests that considerably more than 13 tokens may be required to precisely model within-speaker variation, at least for these variables. Further investigation into the issues of within-speaker sample size is therefore warranted, in order to assess the extent to which these findings are generalisable to other variables.

8.5 Chapter summary

Experiment (1): Number of reference speakers

- Weaker SS scores, high miss rate and considerably wider range using very small numbers of speakers (< 15 speakers).
- Stronger DS strength of evidence and considerably wider range when using small numbers of speakers (< 15 speakers).
- Greater sensitivity to N speakers for /aɪ/ than for /uː/.
  - /uː/ scores stable with more than 30 speakers.
  - /aɪ/ scores stable with more than 55 speakers.
- System validity (EER and C_{llr}) worst with small samples (< 15 speakers).
  - EER stable with more than 30 speakers.
  - Continual improvement in C_{llr} as sample size increased for /uː/.
  - C_{llr} for /aɪ/ stable with more than 42 speakers.

Experiment (2): Number of tokens per reference speaker

- /uː/:
  - Weakest SS scores with two tokens per speaker.
  - Stable EER between two and 13 tokens per speaker, but considerable improvement in C_{llr} as sample size increased.
- /aɪ/:
  - Strongest SS scores with two tokens per speaker.
  - Strongest DS scores with two tokens, but no stability in the distributions of scores as sample size increased.
  - EER and C_{llr} highest with small samples but stability achieved with more than six tokens per speaker.

Chapter 9

Reference Sample Size: Univariate Monte Carlo Simulations

This chapter explores the effects of reference sample size on LR output based on local articulation rate (AR). Monte Carlo simulations (MCS) were performed to generate a large set of synthetic data (1000 speakers/ 200 tokens per speaker) from a sample of existing data. The synthetic data were used to replicate the experiments in Chapter 8 to investigate the effects of (1) the number of reference speakers and (2) the number of tokens per reference speaker on calibrated LLRs and system validity (EER and C_{llr}). Given the availability of a large amount of (synthetic) data, these experiments develop from Chapter 8 to assess the point at which strength of evidence and performance become asymptotic.

9.1 Introduction

The previous chapter considered the issue of sample size (number of reference speakers and tokens per reference speaker) in LR-based testing using two vowel phonemes as input. Consistent with previous research, the findings of these experiments suggest that LRs are generally unstable and misrepresentative when using small samples, although sensitivity to sample size appears to be proportional to the dimensionality of the input variable. This is consistent with a general principle in statistics, that the smaller the

amount of available representative data the more imprecise the model of that (relevant) population. However, according to the law of diminishing returns, at some point the addition of more representative data will have minimal effect on the overall distribution, and subsequent LR output.

With the exception of Rose (2012), no LR-based FVC research has considered such an upper limit at which the inclusion of reference data has an asymptotic effect on LR output. Yet, the efficiency and cost-effectiveness of the numerical LR approach is, to some extent, dependent on knowing how much reference data is required to produce robust LR estimates. The relative lack of research in this area is in part due to a lack of sufficiently large amounts of raw data for testing. As highlighted in §2.4.2, sociolinguistic corpora are generally rather small (up to 30 speakers) and while forensically-realistic databases may be larger, it remains an empirical question whether even these will be sufficiently large to generate precise LRs. Further, the extraction of acoustic-phonetic data from a large number of speakers is labour intensive, compromising the efficiency with which issues of sample size can be tested.

MCS offer a potential solution to this problem. MCS involve generating synthetic values from known properties of the distributions of within- and between-speaker variation of a given variable in a given population. Synthetic data can be built from population statistics from previous research (e.g. mean and SD when the distribution can be assumed to be normal, although the assumption of normality is not a prerequisite; as in Rose 2012) or using an existing set of raw data. While MCS avoid the need for extremely large amounts of raw data, there is a non-trivial *a priori* assumption that the true distribution of the variable in the population is known (or can be well estimated). This is because the distribution of the synthetic data is defined by the properties of the input.

Initial exploration of MCS for testing issues of sample size in FVC is offered by Rose (2012). For a more detailed overview of this paper and a discussion on its limitations see §2.5. The experiments in this chapter use MCS based on an existing dataset of local AR to investigate the effect of (1) the number of reference speakers and (2) the number of tokens per reference speaker on elements of LR output. These experiments expand on Rose (2012) by considering the performance not only of SS pairs but also DS pairs.

This allows for testing of system validity as a function of sample size. Further, given the availability of a considerable amount of input data, the effects of sample size on both calibrated LLRs and uncalibrated scores are considered. Finally, unlike Rose (2012), correlations between elements of the input variable are explicitly tested and included in the modelling procedures when generating synthetic data using MCS.

9.2 Method

9.2.1 Data

The variable analysed in this chapter is local AR, quantified as the number of phonological syllables per second within memory stretches (Jessen 2007). While previous studies of AR (Goldman-Eisler 1998; Laver 1994) provide ranges of between-speaker variation in the population, they rely on a *global* average rate (i.e. from across the entire sample) calculated as:

$$\text{Global AR (syllables/ s)} = \frac{\text{Total number of syllables}}{\text{Total time (s)}} \tag{9.1}$$

Local AR is based on measurements extracted from subsections of the recording. This approach was chosen over *global* AR, since it is essential that multiple tokens per speaker are available to generate a suspect model for assessing similarity. Further, local AR is a more meaningful forensic resource since speakers vary their tempo considerably across utterances (Miller *et al.* 1984).

Local AR was chosen over other FVC variables primarily because it is univariate and can be synthesised relatively straightforwardly (i.e. without having to model multiple, complex correlations between features of a variable). Based on the findings of Chapter 8, unidimensionality predicts that AR will be more robust to reference sample size than /uː/ or /aɪ/, since fewer speakers should be required to adequately model the reference data. However, a potential limitation of AR for testing sample size is that it has been shown to encode relatively little speaker-specific information due to the fact that within-speaker variability is generally higher than between-speaker variation (Gold 2014). The implications of this are considered in analysing the results in §9.4.

An existing dataset of local AR measurements extracted from DyViS Task 2 (§3.1.1) recordings for all 100 speakers collected as part of Gold (2014) was used as raw data for the experiments in this chapter. Gold (2014) identified between 26 and 32 memory stretches per speaker, defined as a period of "fluent speech containing a number of syllables that can easily be retained in short-term memory" (Jessen 2007: 54). The decision to use memory stretches here was a pragmatic one based on the availability of data and does not reflect a theoretical preference for quantifying AR using memory stretches compared with any other approach. However, Gold (2014) found no significant differences in performance compared with inter-pause stretches, claiming that memory stretches are better for FVC as they can be extracted without requiring precise segmentation of individual pauses.

Following Künzel (1997), "fluent speech" was defined as the absence of pauses (of over 100ms), hesitation phenomena and repair processes. Each token was calculated as the total number of phonological syllables divided by the duration (in seconds) of the memory stretch. For each speaker, the first 26 tokens were used in the analysis (the largest number of tokens shared by all speakers). Mean and SD of AR values were calculated by-speaker and converted to z-scores to identify univariate outliers. On the basis of an outlying SD with $z > \pm 3.29$ ($p < 0.01$) (Tabachnick and Fiddell 2007: 73), one speaker (DyViS speaker 95) was removed from the analysis. Of the remaining 99 speakers, 20 were selected at random as development data and a further 20 as test data. The remaining 59 speakers were used as reference data from which synthetic reference speakers and tokens were generated.

9.2.2 Modelling

A univariate implementation of MVKD (§3.2.2.1) was used to compute scores. Based on the fact that the between-speaker KD is a speaker-dependent model made up of Gaussians from each reference speaker, a two-stage process for synthesising data was developed. Normal distributions for each synthetic speaker were initially generated by sampling synthetic means and SDs from the raw data. From the synthetic normal distributions $N(\mu, \sigma)$, a second round of simulations were conducted to generate synthetic tokens for each synthetic speaker. However, before conducting the simulations,

two issues with the raw data were addressed. The first relates to the choice of distribution from which synthetic means and SDs are sampled. Figure 9.1 displays the histograms of raw means and SDs by-speaker fitted with normal distributions.

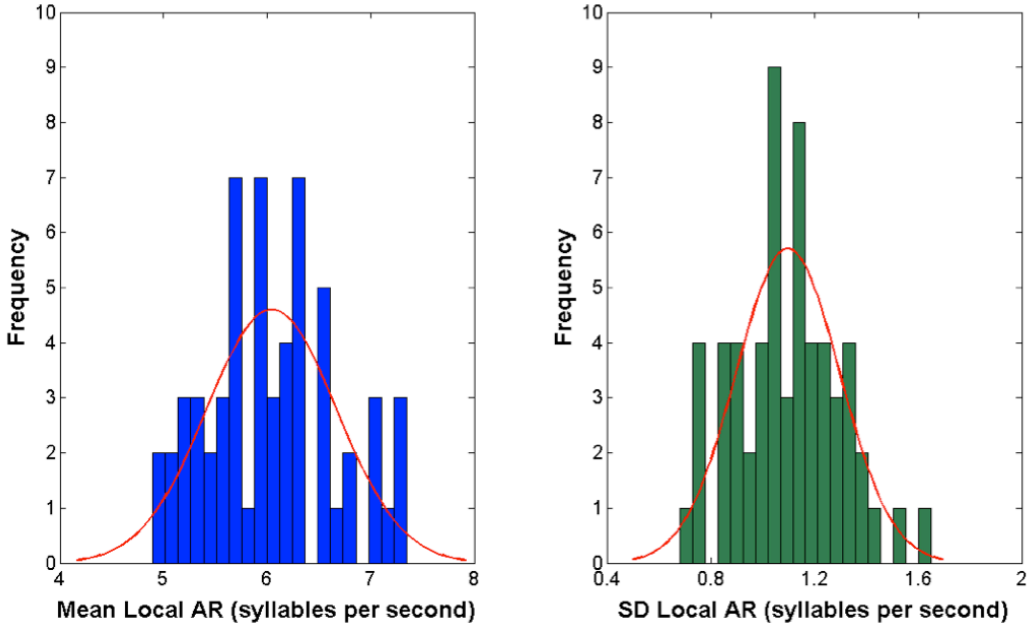

Figure 9.1: Histograms of AR means (left) and SDs (right) for each of the 59 raw speakers fitted with normal distributions

Skew and kurtosis were calculated to assess how well normality models the data. Following Tabachnick and Fidell (2007: 79), skew was analysed by dividing the skewness (S) by the standard error (S_s), defined as $S_s = \sqrt{\frac{6}{N}}$, where N is the number of observations (in this case 59), to give a z-score. A z-score of ± 3.29 indicates significance at the 1% level. Skew was non-significant for both means and SDs ($p >$ 0.4). Kurtosis was analysed by dividing the kurtosis statistic by twice the standard error (Tabachnick and Fiddell 2007) to generate a z-score. For both the means and SDs, kurtosis was also found to be non-significant ($p > 0.24$). Given the statistical assessment of normality and visual inspection of Figure 9.1, it was considered that the normal distribution was appropriate for modelling these data.

The second issue is whether the 59 raw speakers provide a sufficiently precise estimate of patterns in the relevant population. To assess how well the sample of raw data approximates the distribution of values in the relevant population, it is necessary to know *a priori* the properties of that distribution. In the absence of this knowledge,

the effects on the distributions of means and SDs were analysed as data were added. Independent samples t-tests were calculated for means and SDs using the values for all 59 speakers compared against values for minimally ten speakers. Values for each speaker were then added consecutively to the smaller dataset and the t-test re-run at each stage. Welch's Correction (Welch 1947) was applied to account for unequal sample size. The results are analysed with regard to the p-value where a value of one is equivalent to the two samples having the same normal distribution.

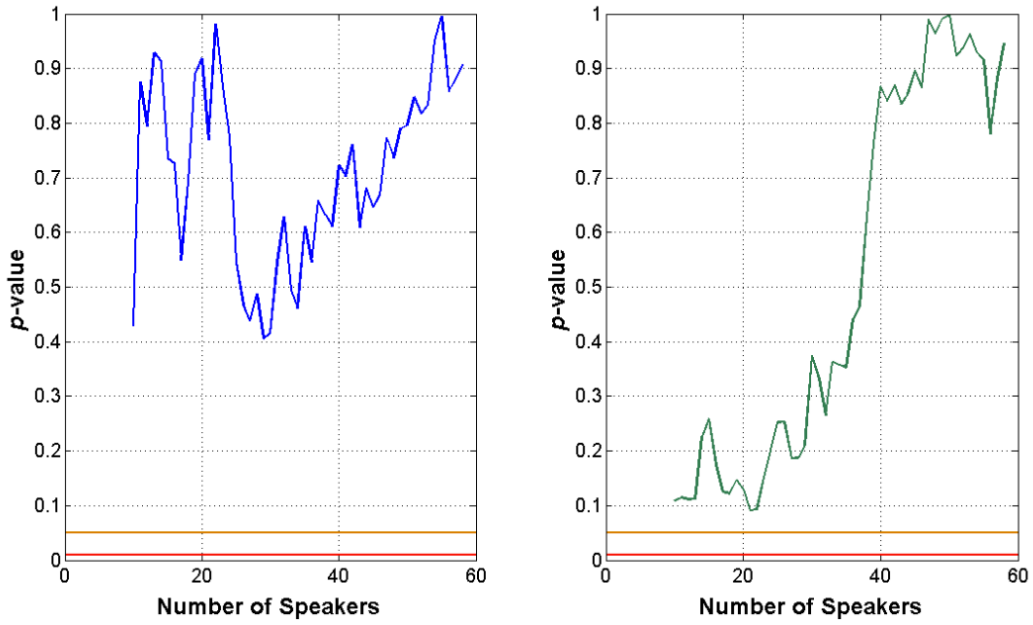

Figure 9.2: p-values based on t-tests comparing the distributions of means (left) and SDs (right) for the number of speakers on the x-axis against that with all 59 raw speakers with 1% (red) and 5% (orange) significance marked

Figure 9.2 shows that there were no significant differences for AR means in the distribution of values using as few as ten speakers compared with the distribution based on 59 speakers. Despite an initial dip with small numbers of speakers, p increased towards one with more than 25 speakers. For SDs, p was relatively low (0.1) with small numbers of speakers, although at no point was the difference between the distributions significant. There was considerable similarity in the distributions of SDs after the inclusion of 40 speakers. Further, the means and SD are consistent with expectations about the range of potential variation reported in Goldman-Eisler (1998). Therefore, it was considered that the distributions based on 59 speakers provided a sufficiently precise estimate of the population.

9.2.3 Monte Carlo simulations

This section details the procedures used for performing MCS to generate synthetic local AR data.

9.2.3.1 Synthetic means

Mean local AR is denoted by x, where x_i is a value for a single speaker (i is speaker number). Based on the testing of normality in §9.2.2, the distribution of raw x_i values was converted to a normal probability density function (PDF) with mean of zero and SD of $\frac{1}{\sqrt{2}}$, $N(0, \frac{1}{2})$:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{9.2}$$

where μ_x is the mean of the raw means and σ_x is the SD, by applying the transformation:

$$z = \frac{x - \mu_x)}{\sqrt{2}\sigma_x} \tag{9.3}$$

This transforms values in the raw x-space to normalised values within the z-space. MCS were then used to generate synthetic z_i values from the preferentially scaled PDF. This is done using the inverse of the cumulative distribution function (CDF). The CDF uses integration to calculate the area under the PDF between $-\infty$ and z_i such that:

$$CDF(z) = \int_{-\infty}^{z} N\left(z, 0, \frac{1}{2}\right) dz \tag{9.4}$$

Given that the normal distribution is so widely used, a special function called the *error function* (erf) (Wang and Guo 1989: 333) has been assigned to the integral ($\int$) meaning that it is possible to generate a CDF based on a normal PDF in the following way:

$$\int_{-\infty}^{z} N\left(z, 0, \frac{1}{2}\right) dz = CDF(z) = \frac{1 + erf(z)}{2} \tag{9.5}$$

where:

$$erf(z) = \frac{2}{\sqrt{\pi}} \int_{0}^{z} e^{-t^2} dt \tag{9.6}$$

Based on Equation 9.5, normally distributed z_i values can be synthesised using a random variable $Z_i = \varepsilon[0, 1]$ (i.e. a random number between 0 and 1). Using the inverse CDF ($CDF^{-1}(z)$), a single synthetically generated z_i value is defined as:

$$CDF(CDF^{-1}(z)) = z = \frac{1 + erf(CDF^{-1}(z))}{2} \tag{9.7a}$$

$$2z - 1 = erf(CDF^{-1}(z)) \tag{9.7b}$$

$$CDF^{-1} = erf^{-1}(2z - 1) \tag{9.7c}$$

As demonstrated in Figure 9.3, using a random value for Z_i and with explicit knowledge of the inverse CDF, a synthetic z_i can be generated in the following way:

$$CDF^{-1}(Z_i) = z_i \tag{9.8}$$

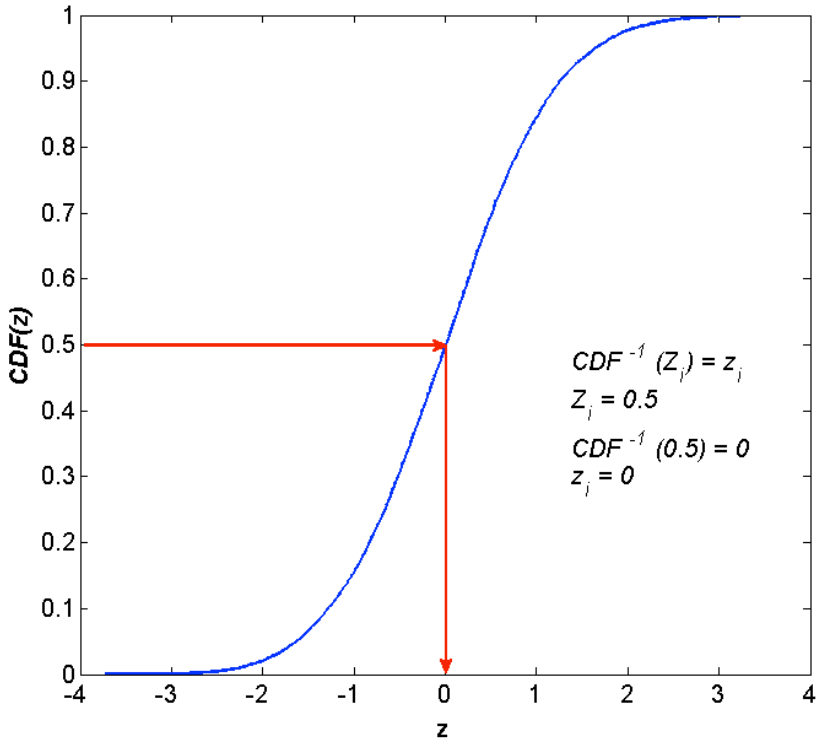

Figure 9.3: Example of the inverse CDF of mean local AR used to generate a synthetic z_i of 0 based on a random Z_i of 0.5 (z_i = 0 equates to x_i = 6.044; i.e. the mean of the raw data)

Synthetic z_i values are then transformed back into the linguistically meaningful x-space by:

$$x_i = (\sqrt{2}\sigma_x \times z_i) + \mu_x \tag{9.9}$$

and used as the mean value for the normal distribution of a single synthetic speaker.

This process was repeated over a number of simulations (n). By the law of large numbers (Wackerly *et al.* 2008: 451), the distribution of $z = (z_1, z_2 \ldots z_n)$ will converge on $N(0, \frac{1}{2})$ as $n \to \infty$. Therefore, with large n the synthetically generated values will have approximately the same normal distribution as the raw values.

9.2.3.2 Synthetic SDs

The SD of local AR is denoted by y such that y_i is the SD for a single speaker. To generate synthetic y_i values, it is necessary to account for any correlation between the means and SDs in the raw data. Figure 9.4 reveals a significant (Pearson's rho = 0.3964; p = 0.0019), positive correlation such that speakers with higher average AR generally displayed greater within-speaker variability. Potentially, this is because speakers with higher mean AR are able to exploit a wider range of variability, particularly in higher rates. Since the mean and SD were seemingly not independent a further projection was incorporated into the simulation of SDs. Rather than sampling from a normal PDF (as in §9.2.3.1), $N(ax_i + b, \beta)$ was used where the mean ($ax_i + b$) is determined by the linear trend line (Figure 9.4) and the SD (β) is given by:

$$\beta_i = \sqrt{\frac{1}{N}\int_{i=1}^{N}(x_i - \tilde{x})^2} = \sqrt{\frac{1}{59} \times 1.9466} = 0.1816 \tag{9.10}$$

where N is the number of speakers (59) and $x_i - \tilde{x}$ is the distance between the trend line and each data point (residuals). Therefore, the mean of the normal distribution from which synthetic SD values were generated ($ax_i + b$) varied as a function of the associated synthetic mean value (x_i).

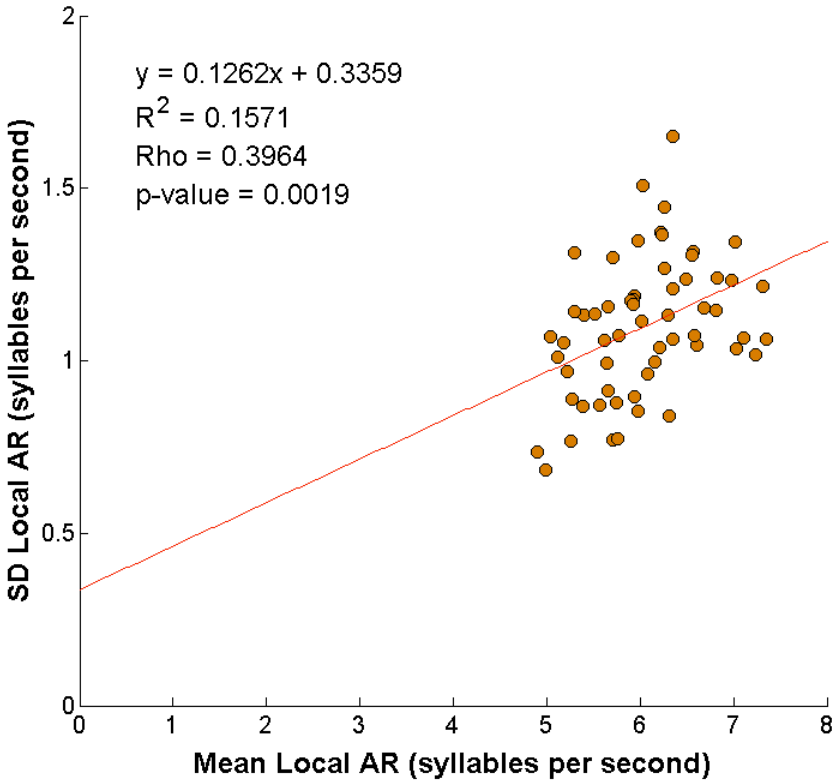

Figure 9.4: Mean local AR plotted against SD of local AR (syllables/ s) for each of the 59 raw speakers

Following the same procedures as in §9.2.3.1, synthetic y_i values were generated by converting $N(ax_i+b, \beta)$ to a normal PDF for each synthetic x_i. The inverse CDF was used to transform a random variable $Z_i^* = \varepsilon[0, 1]$ into normalised z_i^* values (Equation 9.8), before transforming back to the y-space (Equation 9.9). The synthetic mean and SD values represent the normal distribution $N(x_i, y_i)$ for a new synthetic speaker. From this distribution, individual AR tokens were synthesised using the same procedures as in §9.2.3.1. The process of generating synthetic means and SDs was performed 941 times. These synthetic speakers were pooled with the existing 59 raw speakers to create a reference sample of up to 1000 speakers. For the synthetic speakers, up to 200 tokens per speaker were generated. For each of the 59 raw speakers, MCS based on the mean and SD of the 26 raw tokens per speaker were used to generate an additional 174 tokens per speaker. The Appendix provides an example of how data for a single synthetic speaker were generated.

9.2.3.3 Synthetic data

The distributions of means and SDs in the raw data, synthetic data and all reference data combined (raw and synthetic) based on 26 tokens per speaker were compared to assess how well the MCS approximated patterns in the raw data. Table 9.1 reveals minimal difference in the mean of the means (μ_x). The SD of the means in the synthetic data was higher than that in the raw data, although the difference was negligible (0.015). p-values were generated from a comparison of the raw data and the synthetic data, as well as the raw data and all of the reference data, using independent t-tests. The differences between distributions were non-significant, with p approaching 1 in both cases (Table 9.1).

Table 9.1: Mean and SD of mean local AR (syllables/ s) for the raw data, synthetic data and all reference data

	Mean	**SD**	**p-value (t-test)**
Raw data (59 speakers)	6.04	0.63	-
Synthetic data (941 speakers)	6.02	0.64	0.80
Pooled data (1000 speakers)	6.02	0.64	0.81

Table 9.2: Mean and SD of SD local AR (syllables/ s) for the raw data, synthetic data and all reference data

	Mean	**SD**	**p-value (t-test)**
Raw data (59 speakers)	1.10	0.20	-
Synthetic data (941 speakers)	1.10	0.19	0.90
Pooled data (1000 speakers)	1.10	0.19	0.90

There were extremely small differences in the distributions of SD (y) values, with μ_y just 0.0029 higher for the raw data than for the synthetic data (Table 9.2). The differences between the sets in terms of σ_y were also marginal, with SD in the raw data just 0.008 greater than in the synthetic data. Again, paired independent t-tests were performed using the raw data and synthetic data, and the raw data and all reference data combined. In both cases, the differences were non-significant with p-values much closer to one than for the means.

Finally, the means and SDs for each synthetic speaker were plotted and a linear trend line fitted to assess the correlation structure of the synthetic data. Figure 9.5 reveals a linear correlation between the means and SDs consistent with that in the raw data. Beyond the linear correlation, there was again considerable variability around the line of best fit (β). For the synthetic data (941 speakers), β is given by:

$$\beta_i = \sqrt{\frac{1}{N}\int_{i=1}^{N}(x_i - \tilde{x})^2} = \sqrt{\frac{1}{941} \times 29.2857} = 0.1764 \tag{9.11}$$

Compared with β = 0.1816 calculated for the raw data (Equation 9.10), the SD of the residuals in the synthetic data was only marginally lower (0.0052). These comparative results suggest that the procedure for generating synthetic SDs (y_i) was appropriate relative to the correlation structure of the raw data. Further, the synthetic data as a whole was considered sufficiently representative of the raw data.

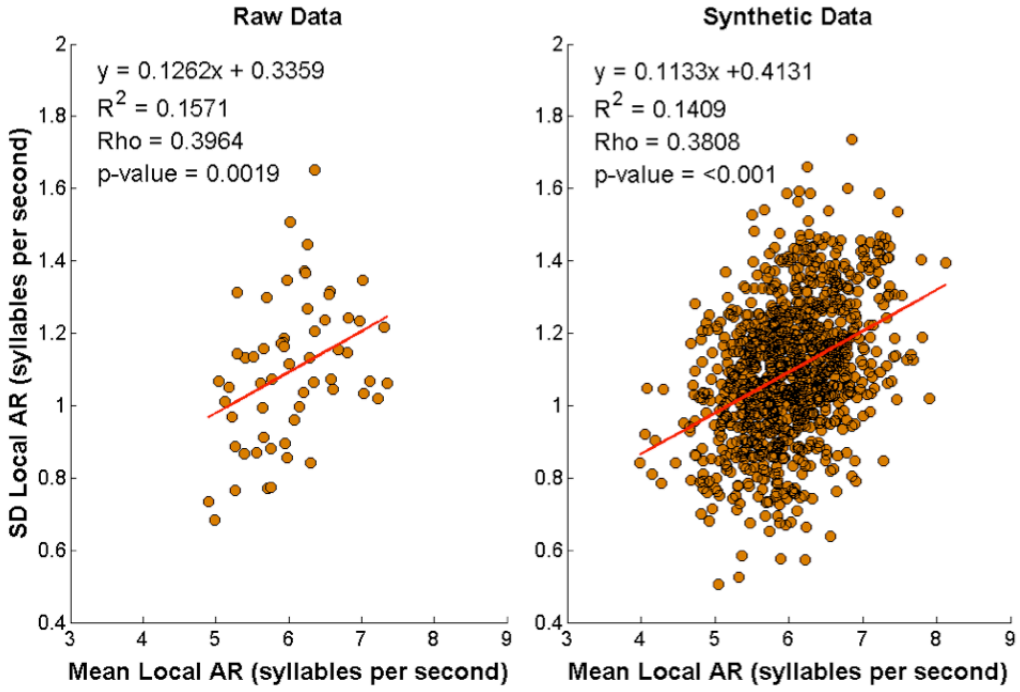

Figure 9.5: Mean local AR values (syllables/ s) plotted against SD of local AR (syllables/ s) for the 59 speakers from the raw data (left) and the 941 synthetic speakers (right) with linear trend lines fitted

9.2.4 Experiments

In this chapter, the two experiments from Chapter 8 are replicated using the local AR data. In both experiments, the same development and test data, each containing 20 speakers with 26 tokens per speaker, were used. In Experiment (1), scores (20 SS/ 380 DS) were computed for the development and test sets as the number of speakers in the reference data was systematically increased by one starting with ten and ending with 1000 reference speakers (26 tokens per speaker). An upper limit of 1000 reference speakers was used because it was considered that the reference distribution would be effectively stable and so the addition of speakers beyond this point would be unlikely to affect LR output.

In Experiment (2), scores (20 SS/ 380 DS) for the development and test data were computed as a single token per speaker was added to the reference data up to a maximum of 200 tokens, using a random reference sample of 200 reference speakers. The decision to include 200 reference speakers in Experiment (2) was made on the basis of the stability in the results of Experiment (1) (§9.3.1). Further, it was felt that 200 speakers reflected a suitably large quantity of reference data for the results to be meaningful, but not unrealistically large in terms of the number of reference speakers that may be extracted from a real forensic database (e.g. Morrison *et al.* 2010-2013).

As outlined in §9.2.2, the univariate KD (§3.2.2.1) approach was used to compute scores for the development and test data. At each stage across both experiments, test scores were calibrated using logistic regression coefficients from the development scores (§3.2.4.1). The distributions of calibrated LLRs, uncalibrated scores and system validity (EER and C_{llr}) were analysed as a function of sample size. Following Rose (2012), the system using the most data was assumed to be the most precise and output from this system is referred to as the *true* output (e.g. *true* LLRs, *true* EER, *true* C_{llr}). For Experiment (1), the *true* values were based on 1000 reference speakers with 26 tokens per speakers, while for Experiment (2) the *true* values were based on 200 reference speakers and 200 tokens per speaker.

9.3 Results

9.3.1 Experiment (1): Number of reference speakers

Figure 9.6 displays the distributions of calibrated LLRs as a function of the number of reference speakers. Figure 9.6 (above) reveals consistency in the distribution of SS LLRs as the number of reference speakers increased. Across conditions, SS comparisons predominantly achieved LLRs equivalent to *limited* support for the prosecution, with absolute numerical values only marginally greater than zero. Of the 20 SS comparisons, only one consistently achieved contrary-to-fact support for the defence, although this value was never less than -0.1. The distribution of SS LLRs based on 50 reference speakers displayed the largest divergence from the distribution of the *true* SS LLRs (based on 1000 speakers). However, even in this case the differences were extremely small (difference in medians = 0.011, difference in ranges = 0.1399).

DS LLRs (Figure 9.6, below) were also extremely robust to the effects of differences in reference sample size. Across all conditions, the median fluctuated maximally within a range of 0.02. The median was marginally weaker in magnitude (i.e. closer to zero), with marginally narrower interquartile and overall ranges, with only ten speakers compared with the *true* LLRs. However, given that the overall range of LLRs was consistently between *limited* support for the prosecution and *limited* support for the defence, it is considered that the LLRs from the ten-speakers condition adequately captured the *true* distribution of DS LLRs for this dataset.

Figure 9.7 (left) displays EER as a function of the number of reference speakers, with the *true* EER (based on 1000 reference speakers) plotted as a means of comparison. The EER of the *true* LLRs was 35.1%. Such performance reflects the very high proportion of DS pairs offering support for the prosecution. There was some fluctuation in performance as the number of reference speakers increased. However, the variation appears to be random since the *true* EER was achieved with as few as 17 speakers. Indeed, the maximum extent of the fluctuation in EER performance was just 0.3% across all conditions, suggesting that categorical validity was relatively stable across sample sizes.

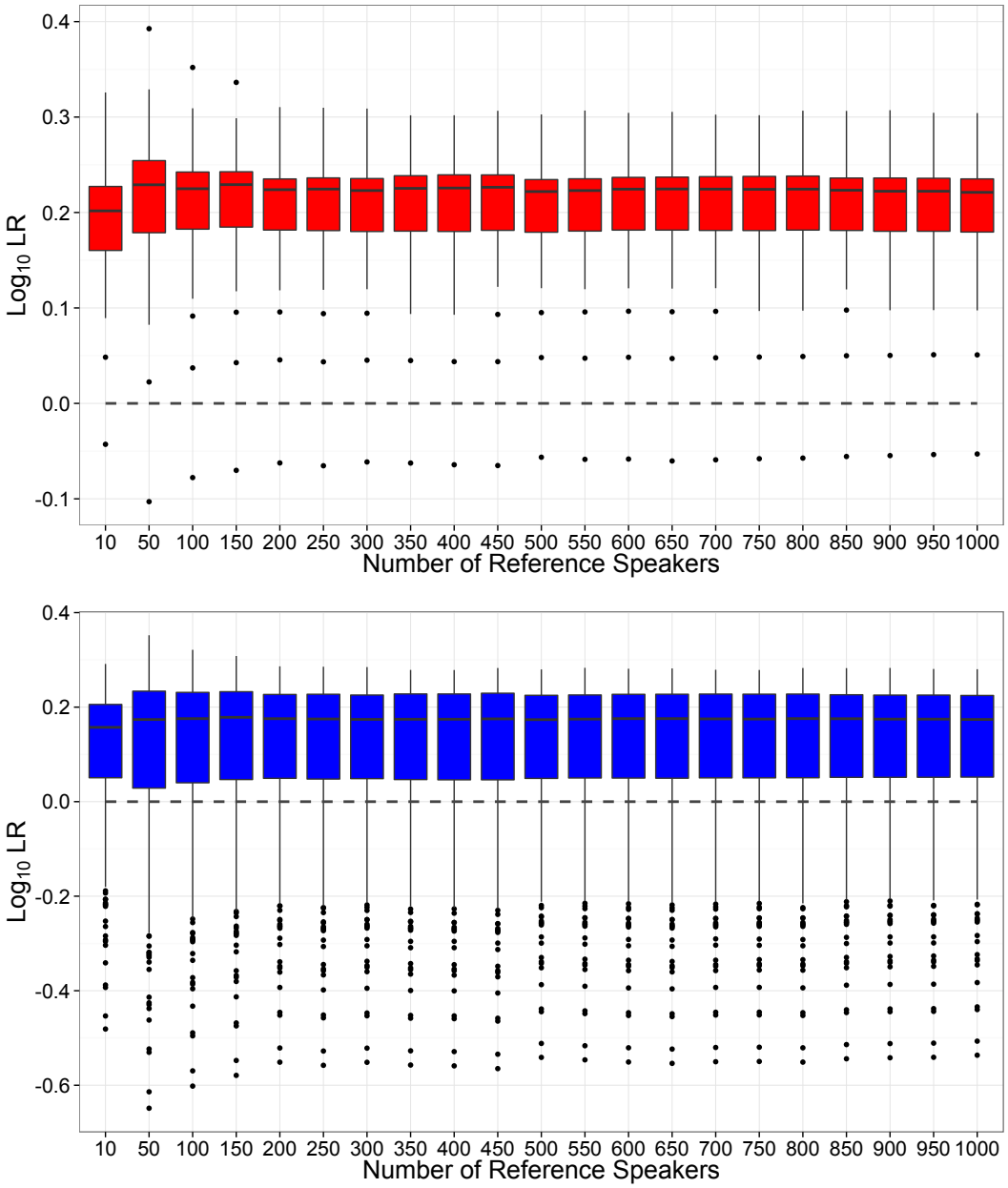

Figure 9.6: Boxplots of SS (above) and DS (below) LLRs as a function of the number of reference speakers

The *true* C_{llr} was 0.971 (Figure 9.7, right). As with EER, this reflects very bad system validity for AR, providing almost no useful information in terms of speaker discrimination. Performance based on C_{llr} as a function of the number of reference speakers was more systematic than for EER. Compared with the *true* value, C_{llr} was marginally better when using fewer than 200 speakers, such that the best validity was achieved with 57 speakers (0.963). With greater than 200 speakers performance appeared asymptotic. However, the overall range of C_{llr} across conditions was very small since the C_{llr}

values from the systems with very small numbers of speakers (ten to 20) were almost equivalent to the *true* C_{llr} (range = c. 0.03).

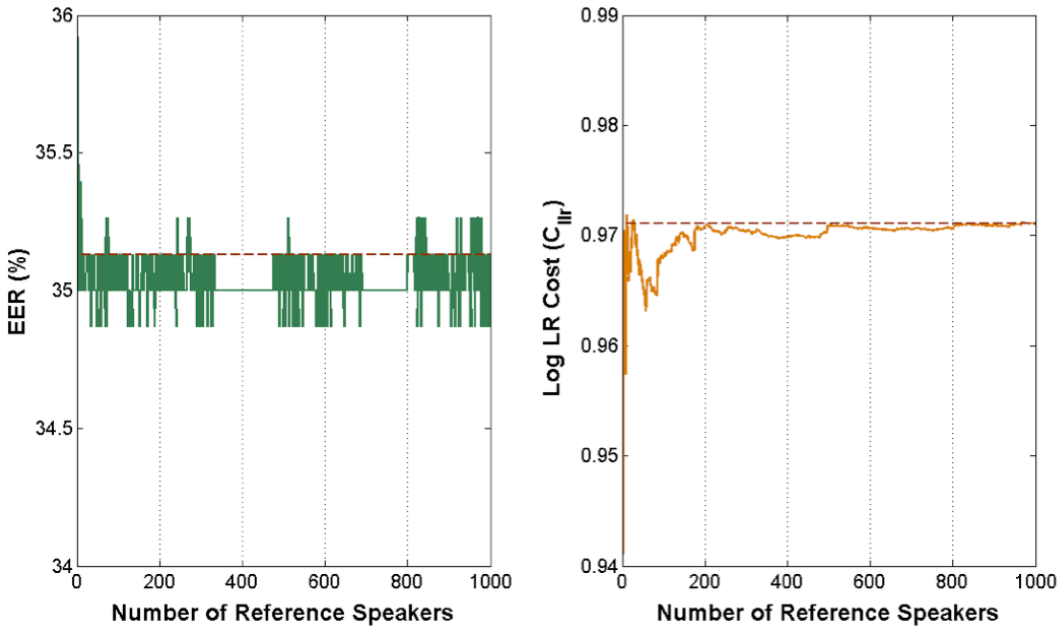

Figure 9.7: EER (%) (left) and C_{llr} (right) as a function of the number of reference speakers with the *true* value (based on 1000 speakers plotted with a dashed maroon line)

Uncalibrated scores

The uncalibrated scores are shown in Figure 9.8. The uncalibrated scores displayed more sensitivity to the size of the reference data than the calibrated LLRs. For SS pairs, the interquartile ranges of scores were always within the range of zero and +1 (*limited* evidence). The median strength of SS evidence was weaker than the *true* median score with small numbers of speakers, such that the lowest median SS score was achieved with ten reference speakers. More significant was the effect of different numbers of reference speakers on individual SS pairs. This was evident in the variability in the furthest outlying contrary-to-fact score. With between ten and 50 reference speakers this score was around -0.5, equivalent to *limited* support for the defence. By the inclusion of 150 reference speakers, this score had decreased by the equivalent of one order of magnitude (to *moderate* support for the defence).

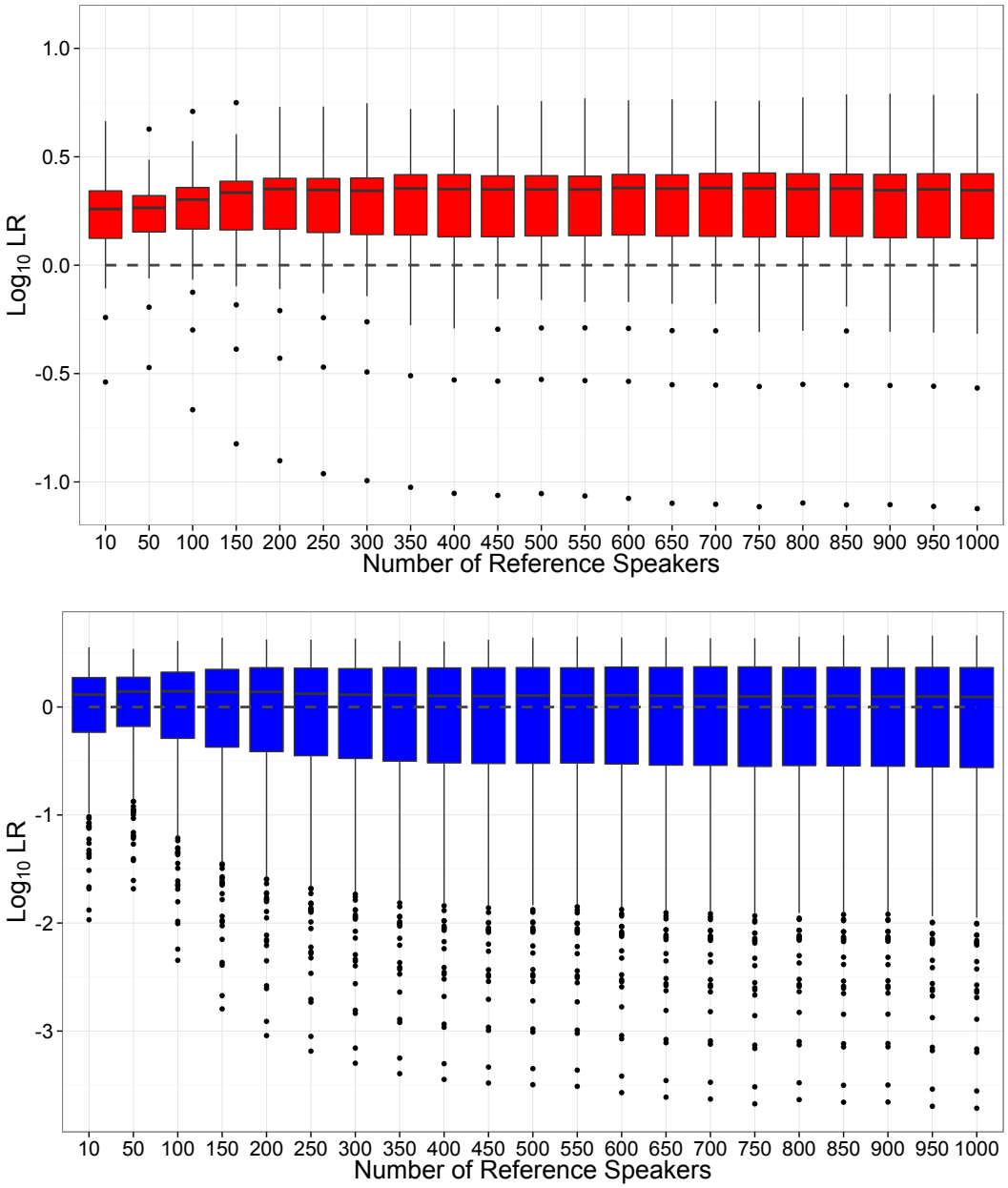

Figure 9.8: Boxplots of SS (above) and DS (below) scores as a function of the number of reference speakers

A similar pattern was found for DS pairs, although the effects were greater. While the DS median was relatively robust to sample size, the interquartile and overall ranges were considerably narrower with smaller samples. There was also greater variability in the distributions of scores when using smaller amounts of data. The most significant effects were again found for the strongest outlying values. For the two most extreme negative DS scores, the strength of evidence increased from greater than -2 (*moderate* evidence) to less than -3 (*moderately strong* evidence), equivalent to a difference of two orders of

magnitude between ten and 1000 reference speakers. For other outlying comparison, scores also increased in terms of the support for the defence by the equivalent of one order of magnitude as sample size increased.

9.3.2 Experiment (2): Number of tokens per reference speaker

Figure 9.9 shows the distributions of LLRs as a function of the number of tokens per reference speaker. When using small number of tokens, the SS medians and ranges were marginally greater than those from the distribution of *true* LLRs. The strongest median value (+0.239) was reached with five tokens per speaker and the highest range reached with six tokens. As in §9.3.1, the extent of variation as a function of the number of tokens per speaker was very minimal with all bar one of the SS comparisons consistently achieving a LLR within a range of zero to +0.4 (*limited* support). The single contrary-to-fact SS LLR across conditions was consistently between zero and -1. The magnitude of the calibrated SS LLRs across all conditions again reflects the fact that AR offers relatively little speaker discriminatory power.

Similar patterns were found in terms of DS LLRs. The median remained essentially stable across all conditions, even when using very small numbers of tokens. The interquartile and overall ranges were marginally wider with small numbers of tokens compared with the distribution of *true* LLRs. This was reflected in the decrease in the strength of evidence for the two most extreme negative DS LLRs, although in numerical terms these LLRs increased by less than 0.1 between the ten- and 200-tokens conditions. In all conditions, DS LRs were maximally spread over a range of two orders of magnitude (between *limited* support for prosecution and *limited* support for defence). Further, the middle 50% of DS LLRs consistently offered contrary-to-fact support for the prosecution, although their magnitude was relatively low (never greater than +0.3).

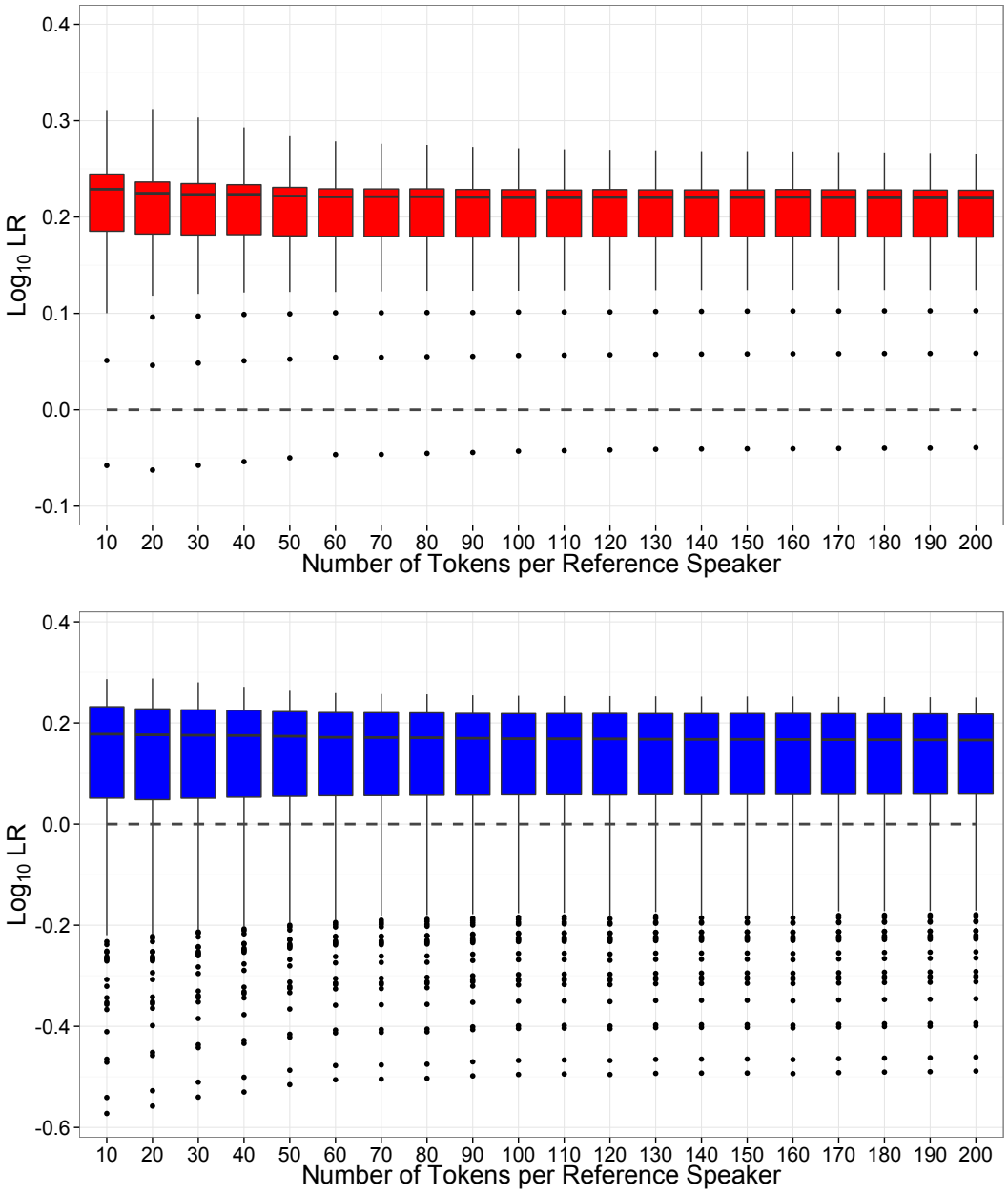

Figure 9.9: Boxplots of SS (above) and DS (below) LLRs as a function of the number of tokens per reference speaker

EER was relatively robust to the number of tokens per reference speaker (Figure 9.10, left). EER based on the maximum amount of reference data was 35%. With the inclusion of more than 96 tokens per speaker, EER was consistently equal to the *true* value. There was very little variability in EER across all conditions (maximally 0.26%) suggesting that increasing the number of tokens does not offer improve categorical system validity for these data. Figure 9.10 (right) also displays C_{llr} as a function of the number of tokens per reference speaker. Relative to the *true* C_{llr}, performance was

marginally better when using small amounts of data. The system with the lowest C_{llr} was based on just six tokens per speaker. After this point, C_{llr} increased marginally until performance became asymptotic with greater than 100 tokens per speaker. However, the range of C_{llr} variability was again extremely small (maximally 0.005).

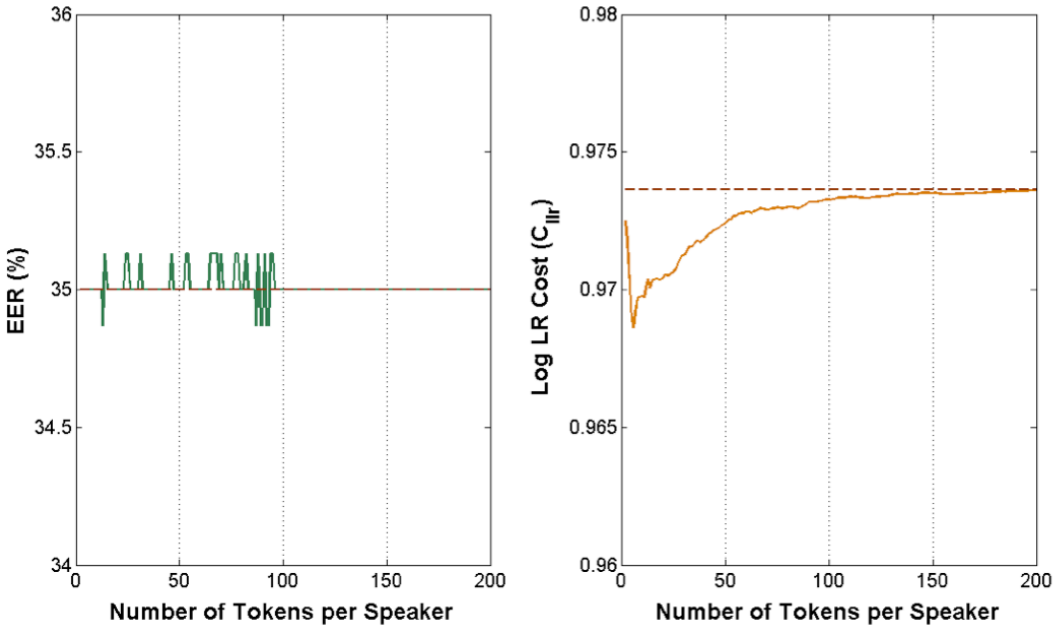

Figure 9.10: EER (%) (left) and C_{llr} (right) as a function of the number of reference speakers with the *true* value (based on 100 speakers plotted with a dashed maroon line)

Uncalibrated scores

As in Experiment (1), calibration appears to play a role in minimising the effects of small amounts of reference data. Figure 9.11 reveals larger differences in the distributions of scores using small samples compared with the *true* scores. The SS median was weakest with fewer than 50 tokens, although it was consistently between zero and +1. The interquartile range was narrower with smaller numbers of tokens (and narrowest with 20 tokens). Again, the outlying contrary-to-fact values were affected to the largest extent. Considering the outlier with the largest negative value, strength of evidence increased by one order of magnitude from *limited* to *moderate* support for the defence between the minimum and maximum number of tokens.

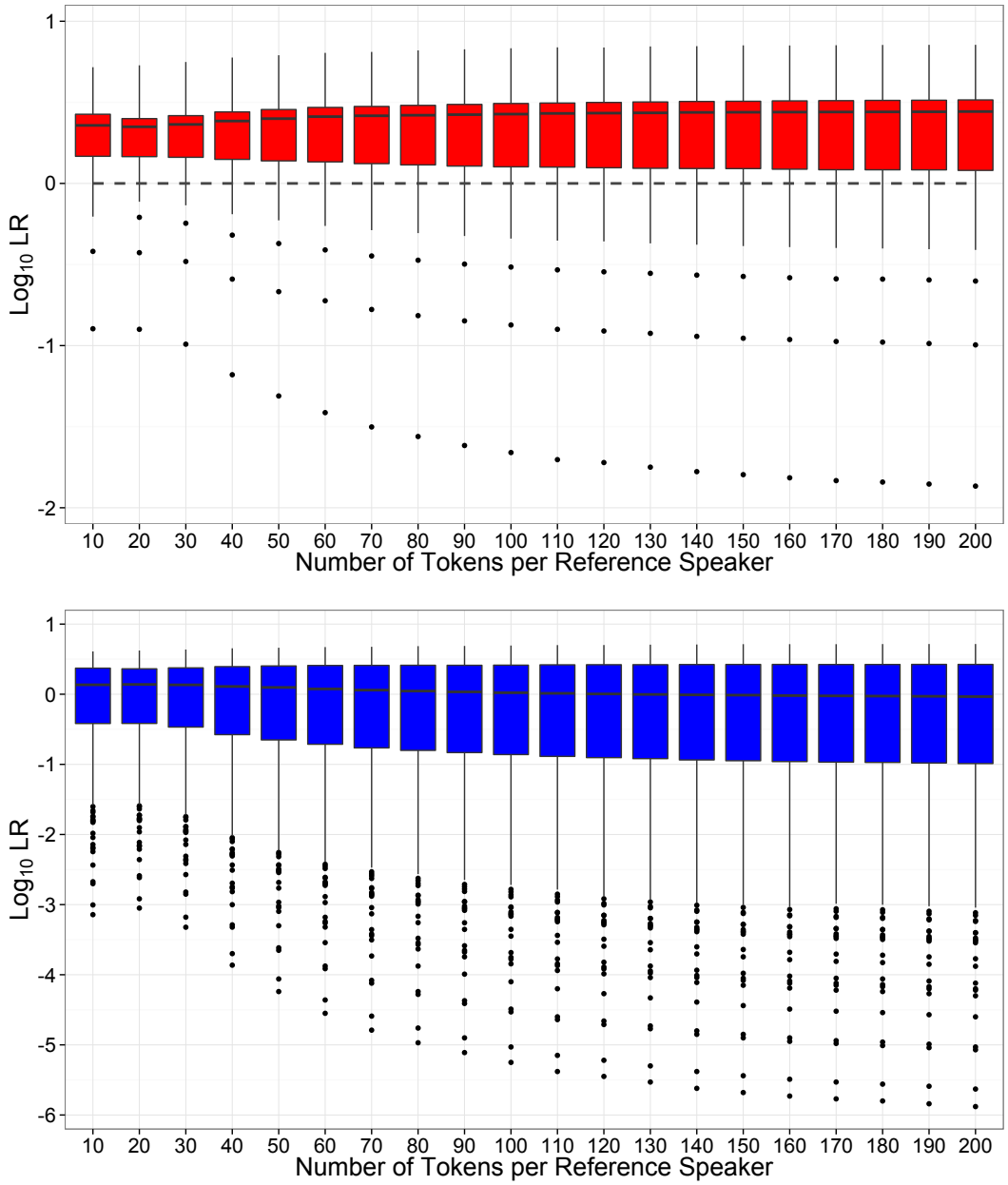

Figure 9.11: Boxplots of SS (above) and DS (below) scores as a function of the number of tokens per reference speaker

The effects of sample size variability were again more dramatic for DS scores. The median decreased marginally as the number of tokens per speaker increased, such that the median based on ten tokens was positive while the median based on 200 tokens was negative. However, in absolute terms the differences in the medians were relatively small (0.17). The interquartile range of DS scores was narrower when using smaller numbers of tokens per speaker, only stabilising after the inclusion of 100 tokens. Generally, strength of evidence increased (more support for the defence) with larger

amounts of data per reference speaker. This was reflected in the magnitude of the most outlying negative scores. For one outlying DS score, there was an increase in strength of evidence between the 20 and 200 tokens conditions equivalent to the difference between *moderately strong* and *very strong* support for the defence.

9.4 Discussion

The results of §9.3.1 and §9.3.2 have revealed that calibrated LLRs for this dataset were relatively robust to the number of reference speakers and the number of tokens per reference speaker. A limited amount of variability was found in the distributions of LLRs with smaller amounts of reference data. However, the medians, interquartile ranges and overall ranges of LLRs with the smallest amounts of data were consistently within the same order of magnitude as the distribution of *true* LLRs. This suggests that precise estimates of the magnitude of calibrated LLRs can be achieved using just ten reference speakers and two tokens per speaker using AR. Although not considered directly here, there are also potential interactions (or trade-offs) between the number of speakers and the number of tokens which may be relevant.

EER remained relatively stable as the number of speakers and the number of tokens per speaker increased, although some *box-like* random variation was found within a very narrow range. Such variability can be explained by the inherent lack of speaker discriminatory power of AR. Since the calibrated LLRs were very close to zero, slight changes in the distribution of the reference data can cause marginally positive values to become marginally negative and vice versa. Since EER deals only in categorical accept-reject decisions, such minor fluctuations have a direct effect on validity. This is an inherent limitation of using EER as a measure of performance, particularly for variables with low speaker discriminatory potential. C_{llr} was found to be lowest when using smaller numbers of speakers and tokens per speaker, such that C_{llr} was systematically better with small amounts of data relative to the *true* LLRs. However, across both experiments, the range of C_{llr} variability was extremely small. The overall stability of calibrated LR output to sample size is extremely positive for the practical application of the LR framework to FVC involving AR, although the analysis of AR in casework

may not be warranted (see Gold 2014) given its low speaker discriminatory value.

Cross-comparison of these results with those in Chapter 8 offers further evidence to support the predicted relationship between the sensitivity of a variable to sample size and its inherent dimensionality. That is, univariate AR was considerably more robust to sample size than /uː/ or /aɪ/. However, it is questionable the extent to which these results are comparable with those in Chapter 8 and the previous research (e.g. Ishihara and Kinoshita 2008; Rose 2012) since the experiments in this chapter considered calibrated LLRs, rather than scores. This is particularly important in this chapter since there is evidence that calibration plays an important role in reducing the sensitivity of LRs to small amounts of reference data. The uncalibrated results are consistent with previous studies in that scores are misrepresentative and unstable when using small numbers of speakers and tokens. However, while previous studies found larger magnitude scores distributed over a wider range when using small samples, the scores in the experiments presented in this chapter were weaker and within a narrower range with small amounts of reference data.

The importance of calibration may be specifically related to AR. The logistic-regression calibration procedure is configured to minimise C_{llr}. For both experiments, the ranges of uncalibrated scores increased as a function of the amount of reference data resulting in more contrary-to-fact scores of a higher magnitude when using larger amounts of reference data compared with smaller samples. Therefore, calibration coefficients generated for systems based on more reference data are greater than those based on less reference data. Despite calibration improving C_{llr} to different degrees for different conditions, the results here suggest that AR performance cannot be improved beyond a ceiling with LLRs close to zero, due to its inherently poor discriminatory value. For better speaker discriminants the role of calibration relative to the size of the reference sample may be different.

The results in this chapter highlight three important general issues for FVC. Firstly, there appears to be an interaction between calibration and the overall sensitivity of LRs to sample size. Whilst calibration counteracted the effects of small sample sizes, calibrated LLRs were spread over a narrower range and were much closer to zero compared with the uncalibrated scores, at least for this variable. Secondly, certain pairs

of samples are more susceptible to the effects of sample size than others. This may be related to the magnitude of the LR itself. Thirdly, the uncalibrated results in Figures 9.8 and 9.11 suggest that scores are not well estimated when the background model consists of small numbers of speakers or tokens, even for univariate data. Therefore, in the absence of calibration, considerable caution should be exercised when interpreting the absolute or relative value of scores generated using a small reference sample.

9.5 Chapter summary

Experiment (1): Number of reference speakers

- Distributions of LLRs equivalent to *true* LLRs even with the smallest number of reference speakers (ten).
- Random fluctuations in EER as the number of reference speakers increased.
  - Small changes to the reference population caused random categorical shifts from positive to negative LLRs (and vice versa).
- Calibrated C_{llr} generally robust to the number os reference speakers.
- Greater instability according to sample size in uncalibrated scores.

Experiment (2): Number of tokens per reference speaker

- Distributions of LLRs essentially the same with two tokens per speaker as with 200.
- Random fluctuations in EER with different sized samples, caused by low magnitude LLRs close to zero.
- C_{llr} robust to the number of tokens per reference speaker.
- Much more sensitivity to sample size with uncalibrated scores.
  - Scores weaker by as much as thee orders of magnitude using two tokens compared with 200 tokens.

General conclusions

- Considerable caution needed when making inferences based on uncalibrated scores generated using small samples.
- Calibration counteracts the effects of small sample sizes, at least for this variable.
- Evidence to support the relationship between sample size sensitivity and the dimensionality of the input variable.

Chapter 10

Development, Test and Reference Sample Size: Multivariate Monte Carlo Simulations

This chapter expands on the experiments in Chapter 9 to explore a broader range of sources of sample size variation in LR computation using a multidimensional variable with greater speaker discriminatory power. Monte Carlo simulations (MCS) were used to synthesise multiple sets of mid-point F1∼F3 data for the hesitation marker UM (*erm*) from a set of raw data. The synthetic data were used to run multiple replications of three experiments to investigate: (1) how many development speakers are required to perform adequate calibration, (2) the number of test speakers needed to ensure that system performance is robust, and (3) the effects of varying the number of reference speakers.

10.1 Introduction

As highlighted in §2.5, relatively little work has considered issues of sample size for LR computation. Further, of the previous work, none has considered the effects of sample size on calibrated LLRs. Chapter 9 provided an initial exploration into this issue using MCS to assess the sensitivity of LR output based on univariate AR to variation in

reference sample size. Calibrated LLRs and system validity were found to be relatively robust to sample size variation even with very small amounts of reference data. However, the extent to which the stability of calibrated LLRs for AR is generalisable to other FVC variables is questionable. There are a number of reasons for this.

First, comparison of the results in Chapters 8 and 9 provides evidence to support the claim (outlined in §8.1) that sample size sensitivity is determined by the dimensionality of the input variable. Across the two chapters, scores based on /aɪ/ (12 dimensions) were most sensitive to sample size, followed by scores for /uː/ (eight dimensions) and finally calibrated LLRs for AR (one dimension). However, the extent to which this pattern will hold for calibrated LLRs remains unclear. Second, AR displays little inherent speaker discriminatory power, producing LLRs close to zero even when using the largest amount of available reference data. Therefore, the range of potential variation as a function of sample size for AR was relatively small. For variables with greater speaker discriminatory power there may be greater sensitivity to sample size since the range of potential variation in LLRs is larger. Finally, in Chapter 9 calibration was found to minimise the effects of small amounts of reference data for AR. The role of calibration was tentatively attributed to the lack of speaker discriminatory power for AR. Therefore, it is unclear whether the role of calibration in minimising the effects of small samples for AR will be the same for other FVC variables .

To test these issues, the present chapter assesses a multidimensional variable with far greater speaker discriminatory power than AR, namely mid-point F1∼F3 values from the vowel portion of the hesitation marker UM. The justification for choosing this variable is given in §10.2.1. As in Chapter 9, MCS were used to generate a large dataset for testing (see §9.1 for an overview of the principles of MCS). The Monte Carlo methods applied to the experiments in this chapter expand on those in Chapter 9 and particularly on Rose (2012). As outlined in detail in §2.5, Rose (2012) provides no explicit discussion on how to model potential correlations between elements of multivariate data, or how to deal with non-normal population distributions. These issues are explored in the simulations in this chapter.

Previous research in this area has focused almost exclusively on the amount of reference data used to assess typicality. As in Chapters 8 and 9, and as in previous research,

the effects of the number of reference speakers used to assess typicality in the feature-to-score stage of LR testing will again be evaluated here. However, this chapter also expands on previous work to assess variability in LR output based on the number of development and test speakers. LR output is necessarily affected by decisions made regarding the size of the development and test sets. The development speakers are used to determine the calibration coefficients applied to the score generated for each comparison pair (i.e. the suspect and offender). The LLRs for the test set are then used to determine the validity and reliability of the system. Unlike previous studies, multiple replications of each experiment were conducted to assess the imprecision in LR output with different sample sizes. The results are considered in terms of the effects of sample size variation in the individual sets, as well as the potential trade-offs across sets.

10.2 Method

10.2.1 Choice of variable

For the purposes these experiments, a multidimensional variable with good speaker discriminatory power was required. LR-based pre-testing was conducted to compare the performance of a number of potential input variables using existing data. Importantly for cross-comparison, the existing data had all been extracted from a single corpus, namely DyViS Task 1 (§3.1.1). The available data (Table 10.1) consisted of mid-point F1∼F3 values from /ɪ ɛ a ɒ ʌ/ collected by Simpson (2008), dynamic F1∼F3 data from /iː ɔː uː/ collected by Atkinson (2009), and dynamic data from the hesitation markers UH and UM collected by Wood (2013) using TextGrids from King (2012) (see §10.2.2). In the case of UM, the formant data had been extracted from the vocalic portion only. Only mid-point (+50%) values of F1∼F3 were used for all variables. Since the maximum number of speakers shared across the datasets was 20, all tokens for the first 20 speakers per set were used for computing LRs. Cross-validated (§3.2.2.3) scores were computed using MVKD (§3.2.2.1). With a small number of exceptions, the same DyViS speakers were used in all ten tests. In the absence of available development data, logistic-regression calibration (§3.2.4.1) was applied using cross-validation. Performance was assessed using C_{llr} and EER.

Table 10.1: Number of available speakers and tokens per speaker (maximum, minimum and mean) for each of the variable

Phoneme	**N Speakers**	**Max N Tokens**	**Min N Tokens**	**Mean N Tokens**
/ɛ/	25	72	27	50
/iː/	20	14	11	12
/uː/	20	13	8	10
/ɪ/	25	59	15	36
/ɒ/	25	47	12	28
/ɔː/	20	12	7	9
/ʌ/	25	36	8	16
/a/	25	34	10	19
UH	38	20	20	20
UM	34	20	20	20

As shown in Table 10.2, the best system validity was achieved using UM. The finding that hesitation markers outperform lexical vowels is, to some extent, predictable. Since hesitations are non-linguistic (i.e. do not encode relevant *speech* or *group* information), they are not expected to be stratified by external sociolinguistic factors to the same extent as lexical vowels; although there are stereotypical realisations of hesitation markers associated with varieties of BrEng (e.g. Liverpool English [eː]). Further, occurring primarily in isolation, hesitation markers are less susceptible to coarticulatory effects (Foulkes *et al.* 2004; Wood *et al.* 2014). Further, Künzel claims that "individuals tend to be quite consistent in using 'their' respective personal variant" (1997: 51). Empirical evidence for this is found in Foulkes *et al.*'s (2004) study of vocalic hesitation markers in Newcastle English. Therefore, on the basis of pre-testing and expectations about the extent of within- and between-speaker variation from previous research, mid-point data for UM were used as input in this chapter.

Table 10.2: C_{llr} and EER performance for each variable

Variable	C_{llr}	EER (%)
/ɛ/	0.547	19.6
/iː/	0.469	15.1
/uː/	0.938	25.0
/ɪ/	0.673	20.0
/ɒ/	0.484	15.9
/ɔː/	0.842	25.1
/ʌ/	0.691	19.7
/a/	0.772	20.7
UH	0.479	15.7
UM	**0.325**	**10.1**

10.2.2 Data

Existing segmented TextGrids for the vowel portion of UM were available for the first 21 DyViS speakers (Task 1) from King (2012). Wood (2013) used the TextGrids to extract dynamic measurements of F1∼F3 for these 21 speakers. Wood (2013) also extracted dynamic F1∼F3 data from a further 22 DyViS speakers, to create a larger dataset of 43 speakers. Across the combined dataset, the number of tokens per speaker ranged from eight to 20. To ensure as precise an estimate of the population distribution for MCS, data for a further 49 speakers were extracted. The procedures used by Wood (2013) for segmentation and data extraction were the same as those used in this chapter.

UM tokens were identified using existing orthographic transcription TextGrids two minutes into each sample. The onset and offset of each token were delimited using the criteria outlined in §3.3.1.1 (see Figure 10.1). Maximally 20 tokens per speaker were included in the analysis. Four speakers with fewer than ten tokens were removed, leaving 88 speakers available for analysis. Tokens were saved to separate sound files to preserve the sampling rate (44.1 kHz) and dynamic formant data were extracted using the procedures outlined at §3.3.1.2. The script was set to find between five and six formants within a range of 0 to 5 kHz range, determined on a token-by-token basis.

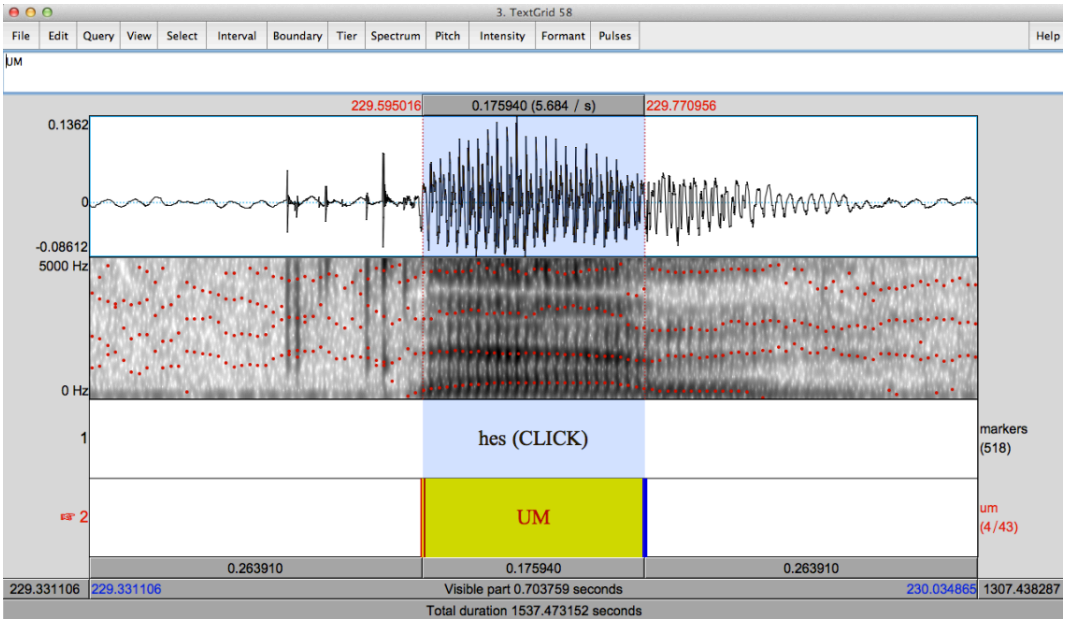

Figure 10.1: Example of a segmented token of UM on a PRAAT TextGrid from DyViS speaker 58

10.2.2.1 Formant correction

The raw dynamic data were inspected and obvious measurement errors (e.g. F3 measured as F2) corrected by hand. Since only the midpoint values were being used for the MCS, missing values at the +10-40% and +60-90% steps were not corrected. Where values were missing at the +50% step (mid-point), a value was calculated based on the difference between the two adjacent values (+40% and +60%). Where erroneous mid-point values for a given formant could not be resolved, the whole token was removed from the analysis. For two speakers, this meant that there were insufficient data for analysis and so these speakers were removed. The data extracted by King (2012) and Wood (2013) had already been manually corrected using the same procedures. The three datasets (King's 21 speakers, Wood's 22 and my 43) were combined to create a dataset containing 86 speakers with between ten and 20 tokens per speaker, although only 11 speakers had fewer than 20 tokens.

10.2.2.2 Mid-points vs. dynamics

The availability of dynamic data for all 86 DyViS speakers raises the question as to why only the mid-point values were used. This decision was made because of the

considerably greater degree of complexity and mathematical uncertainty introduced into MCS when dealing with highly multivariate data such as formant dynamics. There are a number of reasons for this. Raw values within a single formant trajectory are necessarily highly autocorrelated (i.e. correlations between adjacent time points). That is, the +20% value is dependent on the +10% and +30% values (plus possible correlations with later steps in the trajectory). This means that when trying to simulate raw formant trajectories following the procedures in Chapter 9, it would be necessary to test, and potentially simulate, $(54^2 - 54)/2 = 1431$ correlations (2 values (mean and SD) $\times$ 9 measurements $\times$ 3 formants).

Even if simulating lower order polynomial coefficients (e.g. quadratic), there would be 153 potential correlations (2 values (mean and SD) $\times$ 3 measurements (coefficients) $\times$ 3 formants), of which, at the very least, coefficients from the same formant are expected to be correlated. The complexity of the correlation structure of this type of multivariate data means that it is difficult to precisely model the population distribution. The benefit of using mid-point data is that there are far fewer correlations to consider and since the simulations only involve generating a single value per formant there is no *a priori* expectation for any correlations between the elements of UM. Further, it is assumed that vocalic hesitations are broadly monophthongal given their usual IPA transcription. Therefore, the mid-point provides an appropriate representation of their quality.

10.2.3 Modelling

The general procedures for simulating UM were the same as those employed in Chapter 9, whereby means and SDs were firstly generated to created a normal distribution for each synthetic speaker, from which individual tokens were then sampled. As in §9.2.2, prior to simulating data it was necessary to establish the appropriate distribution for modelling the raw means and SDs and to assess whether the raw data provided a sufficiently precise representation of the population distribution. Firstly, means and SDs of F1∼F3 were calculated for the 86 speakers. Figure 10.2 displays the histograms of raw means and SDs for each formant in terms of probability density fitted with a KD estimate. Visual inspection of Figure 10.2 suggests that the normal distribution provides a fairly accurate model of the distributions of the means. The distributions of SDs in

all three cases, however, appeared to be positively skewed. To test the assumption of normality for the means and the skew in the SDs, z-scores for skew and kurtosis were calculated following the approach in §9.2.2.

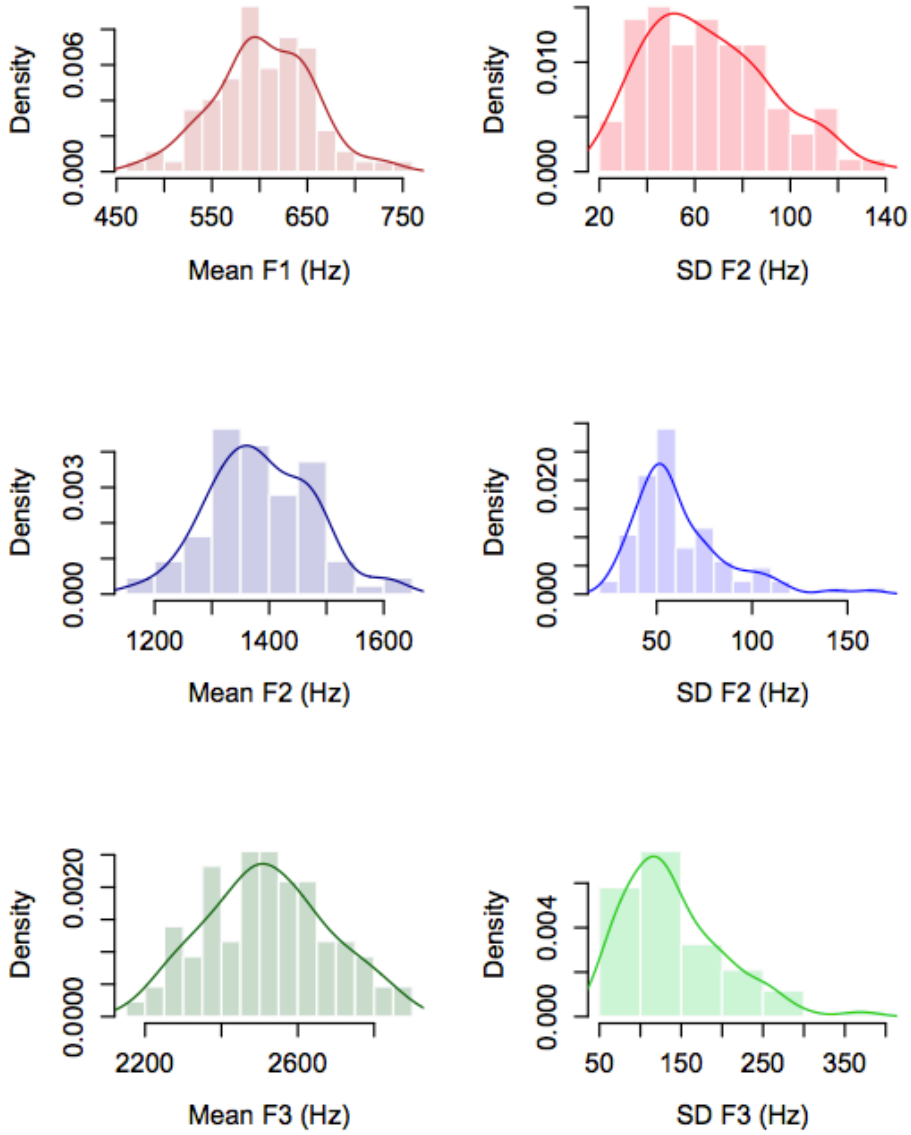

Figure 10.2: Histograms of raw means (left) and SDs (right) for F1 (red), F2 (blue) and F3 (green) based on 86 speakers fitted with a kernel density

For the distributions of means, neither skew nor kurtosis was found to be significant (Table 10.3). Therefore, the assumption of normality was considered valid for the distributions of F1∼F3 means. Consistent with earlier predictions, the distributions of SDs were all positively skewed (significant at least at the 5% level). For F2 and F3 SDs, kurtosis was also significant at the 1% and 5% levels respectively, but not significant

for the F1 SDs. Table 10.3 therefore suggests that an assumption of normality for the distributions of SDs is inappropriate and that an alternative model is required.

Table 10.3: z-scores for skew and kurtosis based on the raw means and SDs for F1, F2 and F3 (with significant values highlighted in red ($p < 0.01$) and orange ($p < 0.05$))

		Skew	**Kurtosis**
F1	Means	-1.8077	0.284
	SDs	**2.0038**	-0.6518
F2	Means	0.6769	0.0973
	SDs	**5.9731**	**6.2471**
F3	Means	0.4462	-0.9825
	SDs	**4.1423**	**2.7237**

The raw SDs are consistent with expectations for lognormal distributions. Where the logarithm of a random variable α is normally distributed, the distribution of α is "said to be lognormal" (Johnson *et al.* 1994: 207), such that:

$$\beta = log(\alpha) = N(\mu, \sigma) \tag{10.1}$$

$$\alpha = exp(\beta)$$

where:

$$\alpha > 0 \tag{10.2}$$

and:

$$-\infty < \mu > \infty \tag{10.3}$$

$$\sigma > 0$$

In this way, the mean and SD of the normal distribution in the log space are directly related to the mean and variance of values in the lognormal space. The properties of the normal distribution of $log(\alpha)$ are defined as:

$$\mu = log\left(\frac{m^2}{\sqrt{v + m^2}}\right) \tag{10.4a}$$

$$\sigma = \sqrt{log(\frac{v}{m^2} + 1)} \tag{10.4b}$$

where m is the mean and v the variance (SD^2) of the lognormal values. The properties of the lognormal distribution are:

$$m = exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{10.5a}$$

$$v = exp(2\mu + \sigma^2)(exp(\sigma^2) - 1) \tag{10.5b}$$

from Patel and Read (1982: 24)

Figure 10.3 displays the distributions of the natural log of the SDs plotted as a histogram with estimated KDs. Figure 10.3 shows that the normal distribution provides a good approximation of the logged data. Across all three formants, the distributions of log values were relatively symmetrical, with the mean approximately at the point of maximum density. Following the same procedure as above, skew and kurtosis were analysed statistically to provide an objective estimate of how appropriate the normal distribution is for the log values (Table 10.4). For all formants, neither skew nor kurtosis were significant. This provides evidence that the lognormal distribution is appropriate for modelling the raw SD values for these data.

Table 10.4: z-scores for skew and kurtosis based on the log SDs for F1, F2 and F3

		Skew	**Kurtosis**
F1	log(SD)	-1.1962	-0.8366
F2	log(SD)	1.5692	0.4416
F3	log(SD)	0.2077	-1.0097

Having established that the lognormal distribution is appropriate for modelling the raw data, it was important to consider whether this assumption can be extrapolated reliably to the population distribution (i.e. if we had 1000 speakers would the distribution of raw SDs become normal?). There are two reasons why the lognormal distribution was considered valid for this population. Firstly, the assumption was based on a large amount of raw data (86 speakers) allowing more robust assumptions about the population to be made. Secondly, the lognormal distribution is justifiable on linguistic grounds. This is because the majority of speakers are expected to display moderate levels of within-speaker variability (c. 40-100 Hz). However, while there is inherently a lower limit of potential within-speaker variability, there is far more potential for high levels of variability for a small proportion of speakers.

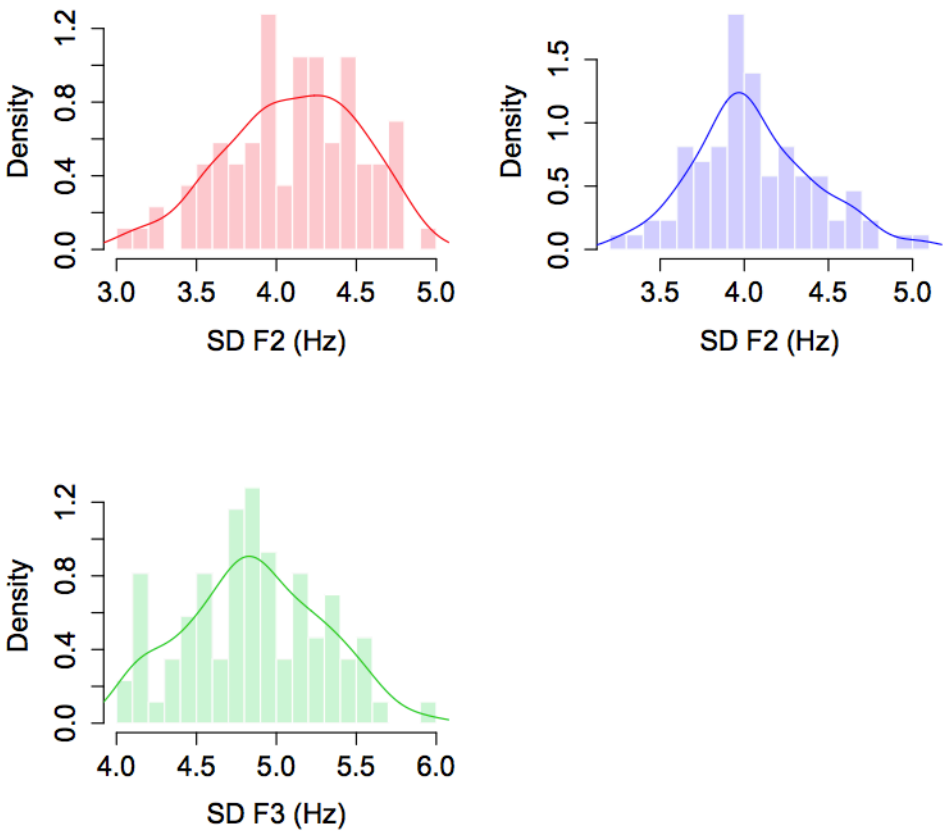

Figure 10.3: Histograms of the natural logarithms of raw SDs for F1 (red), F2 (blue) and F3 (green) based on 86 speakers fitted with a kernel density

10.2.3.1 Precision of population estimate

As in §9.2.2, it was necessary to assess how precisely the raw data approximate the patterns in the population. To do this, the mean and 95% CIs (based on the noninformative priors in §7.2.5; Equation 7.1) of the means and the logged SDs for each formant were calculated initially using values from two speakers, then consecutively with a random additional speaker until all 86 speakers were included. The 95% CI is a probabilistic region within which one can be 95% certain that the true value is located. The CI therefore captures the imprecision in the mean of the means and SDs as the number of speakers increases. If the means and CIs remain stable as the number of speakers is increased it can be inferred that they would also be robust to the addition of further speakers beyond the 86 available here.

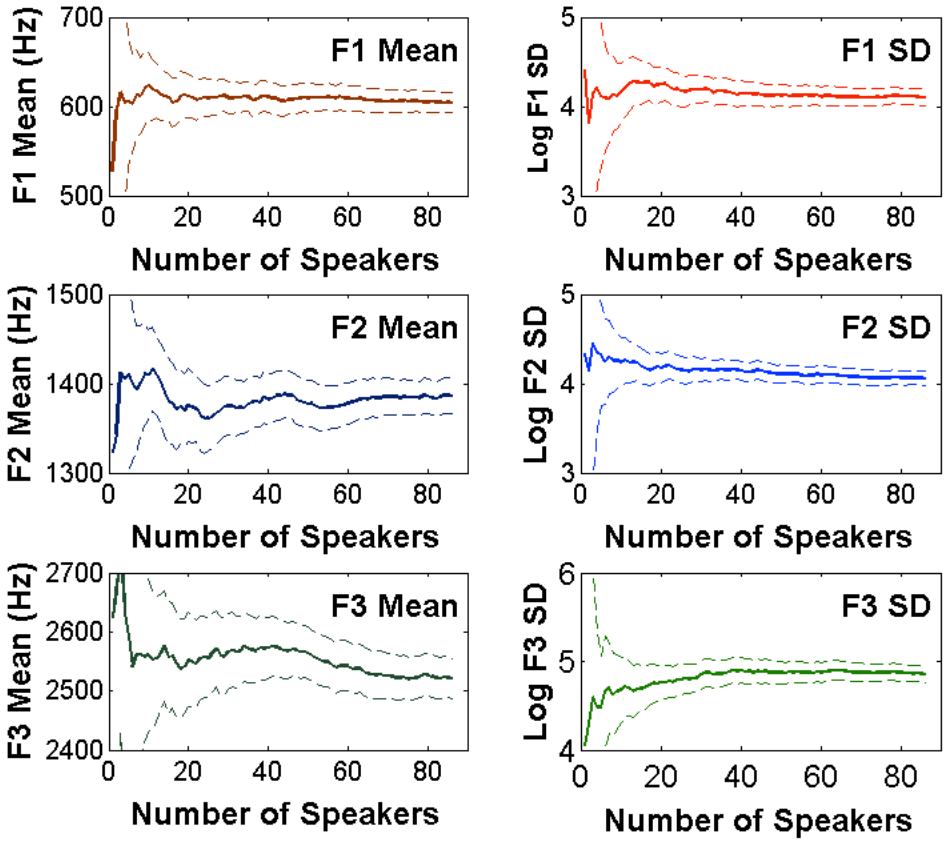

Figure 10.4: Means (solid) and 95% CIs (dashed) for F1 (red), F2 (blue) and F3 (green) means (left) and logged SDs (right) based on the number of speakers included

Figure 10.4 displays the mean and 95% CIs for F1, F2 and F3 means and logged SDs by number of speakers. For all formants, there was considerable imprecision using small numbers of speakers. In all cases, the mean and CIs were stable by the inclusion of all 86 speakers. The point at which this occurred, however, varied. For F1, the mean of the means and logged SDs reached stability by around 20 speakers. F2 and F3 means were more sensitive to the number of speakers, displaying fluctuations in both means and CIs even with relatively large amounts of raw data. However, by around 60 speakers for F2 and 65 speakers for F3 the distributions of means became stable. For F2 and F3 logged SDs, the means and CIs were relatively well estimated after the inclusion of more than 30 speakers.

10.2.3.2 Correlations

The correlation structure of the raw data was analysed prior to the simulations. Table 10.5 displays the partial correlation matrix for UM, based on pairwise Pearson correlation tests using the raw data. Only two significant correlations were found: between F1 and F2 SDs (rho = 0.27, p = < 0.01), and F2 and F3 SDs (rho = 0.3, p = < 0.01).

Table 10.5: Partial correlation matrix based on pairwise Pearson correlation test with rho (left) and p-values (right, italics) for F1, F2 and F3 means and SDs for UM based on input data form 86 speakers (significant correlations in red)

	F1 SD	**F2 Mean**	**F2 SD**	**F3 Mean**	**F3 SD**
F1 Mean	-0.01 *(0.96)*	-0.11 *(0.31)*	-0.01 *(0.95)*	0.17 *(0.12)*	0.01 *(0.96)*
F1 SD	-	0.04 *(0.73)*	**0.26 *(0.01)***	0.18 *(0.10)*	0.08 *(0.45)*
F2 Mean	-	-	0.17 *(0.11)*	0.14 *(0.20)*	0.07 *(0.51)*
F2 SD	-	-	-	0.13 *(0.24)*	**0.30 *(0.01)***
F3 Mean	-	-	-	-	0.14 *(0.19)*

However, the correlation between F1 and F2 SDs appeared to be driven by a single speaker with atypically high variability for F1 (121 Hz) and F2 (162 Hz) (DyViS speaker 3). Figure 10.5 displays scatterplots of the data both including (left) and excluding (right) the outlying speaker. With the removal of this speaker, the correlation between the SDs was no longer significant (p = 0.08). Therefore, for the purposes of simulation the correlation was ignored, since there was insufficient evidence that the trend holds for the population. The outlying speaker was nonetheless included in the independent simulation of F1 and F2 SDs, since the formant measurements for this speaker were considered accurate. As in Figure 10.5, the correlation between F2 and F3 SDs was also inspected visually. Although the interaction was not clearly driven by a single speaker, visual inspection of the data and the rho value of 0.30 suggested that the correlation was relatively weak. Therefore the correlation was not considered robust enough to infer that it would hold amongst the population at large.

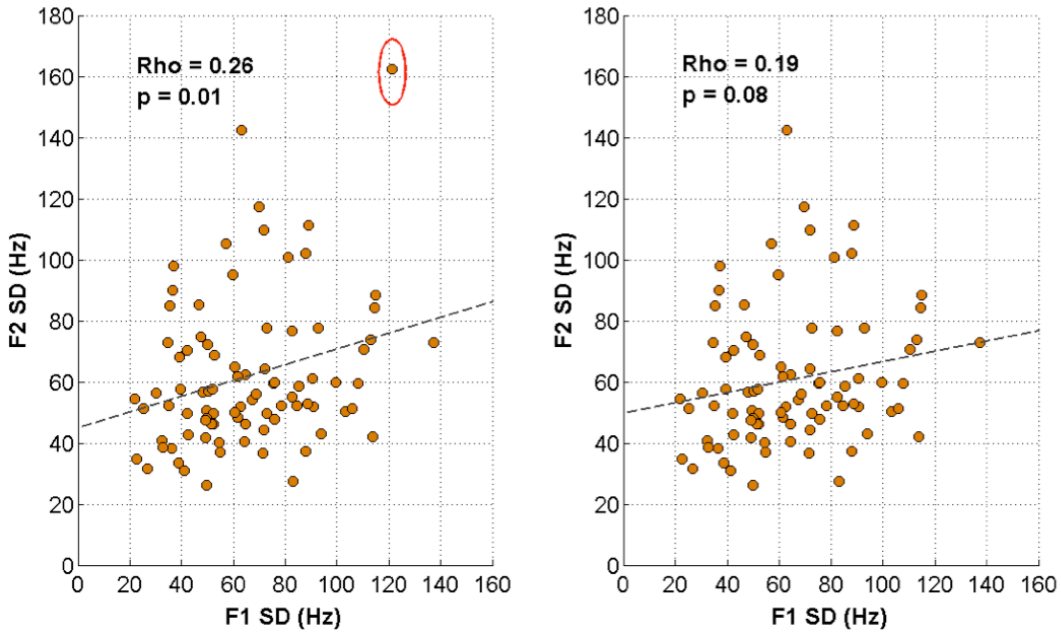

Figure 10.5: Scatterplots of F1 SDs and F2 SDs fitted with a linear trend line and using data from all 86 speakers (left; the outlying speaker is marked with a red ellipse) and with the outlying speaker removed (right)

10.2.4 Monte Carlo simulations

The absence of meaningful correlations in the raw data is convenient for the simulation process, since means and SDs can be generated independently. This is beneficial, firstly, in terms of computational efficiency. Secondly, the lack of correlation minimises the degree of statistical uncertainty in the simulation procedure. Having established that F1∼F3 means are normally distributed, the procedures outlined in §9.2.3.1 for sampling from the normal distribution were also followed here to create synthetic speaker means.

In principle, the procedures used to generate synthetic SDs are the same to those used for the normally distributed means. The difference lies in how the PDF is defined for lognormal data. The SD of a given formant is denoted by y, where y_i is the SD for an individual speaker. The raw y_i values from all 86 speakers are transformed using a natural logarithm, and the mean (μ) and SD (σ) of the logged values calculated. These properties of the distribution of logged values are used to define the lognormal PDF such that:

$$\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(log\ x-\mu)^2}{2\sigma^2}} \tag{10.6}$$

from Patel and Read (1982: 24)

Based on Equation 10.6, the lognormal distribution is normalised by applying the transformation:

$$z = \frac{(log\ y - \mu_y)}{\sqrt{2}\sigma_y} \tag{10.7}$$

This conversion transforms linguistically meaningful values (in the y-space) to normalised z-space values, such that the area under the normalised distribution is equal to one.

As with normally distributed data, it is necessary to define the inverse CDF to convert pseudo-random area values (Z_i) into synthetic z_i values. Having defined z for lognormal data in Equation 10.7, the CDF of the lognormal PDF can be calculated using Equations 9.5 and 9.6. The inverse CDF can then also be derived in the same way as for normally distributed data, as described in Equation 9.7. With explicit knowledge of the inverse CDF, synthetic z_i values are generated using a random variable $Z_i = \varepsilon[0, 1]$ and then converted back to the y-space by:

$$z\sqrt{2}\sigma_y = log(y) - \mu_y \tag{10.8a}$$

$$z\sqrt{2} + \mu_y = log(y) \tag{10.8b}$$

$$y = e^{(z\sqrt{2}\sigma_y + \mu_y)} \tag{10.8c}$$

Synthetic F1∼F3 means and SDs are randomly assigned to a synthetic speaker (since they are independent of each other). Tokens for synthetic speakers can then be sampled from the normal distributions of F1∼F3 (based on the synthetic mean and SD) following the same process as in §9.2.3.1. Figure 10.6 provides a comparison of the CDFs of means and SDs for the raw data and those of an example set of 1000 synthetic speakers. The degree of overlap between the two sets of data indicates that the simulations provide an accurate model of the distributions in the raw data.

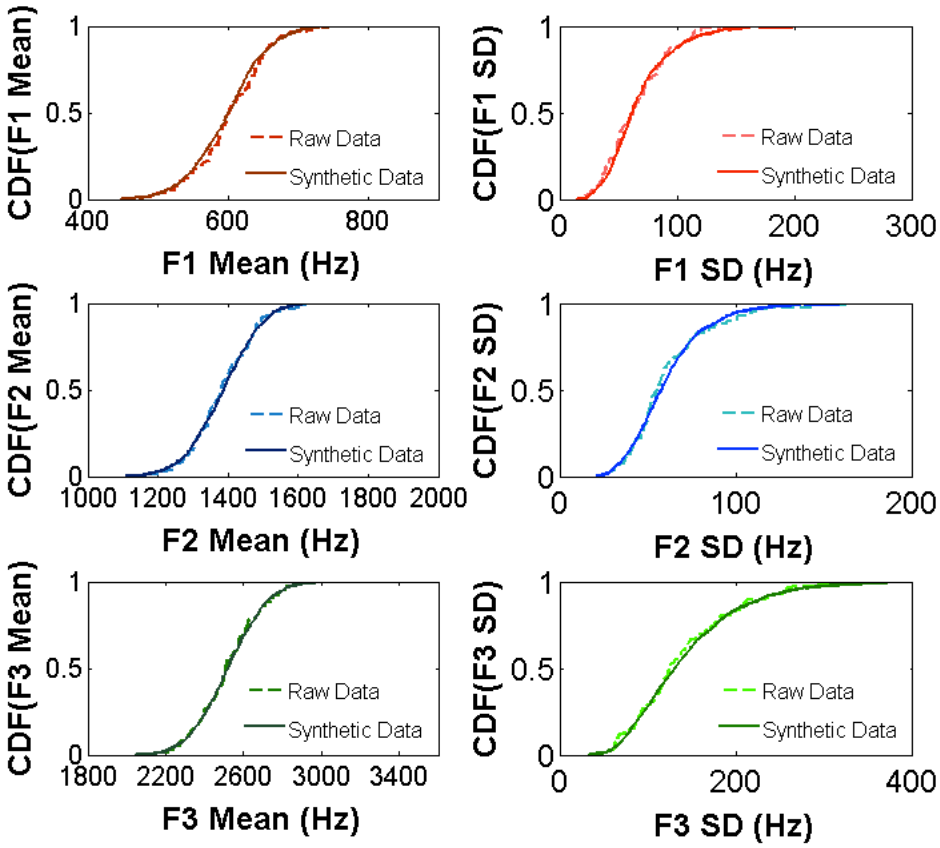

Figure 10.6: CDFs of F1 (red), F2 (blue) and F3 (green) means (left) and SDs (right) based on the raw data (86 speakers) and an example set of 1000 synthetic speakers

10.2.5 Experiments

The experiments in this chapter expand on those in Chapters 8 and 9 by using multiple replications of the same experiment to assess the sensitivity of LR output to variation in the number of (1) development speakers, (2) test speakers, and (3) reference speakers. The results of these studies therefore have in-built replicability testing to assess precision for each number of speakers tested. Development, test and reference data were created from the synthetic data only (raw data were used only for simulation). For each experiment, only the number of speakers in the target set (i.e. development, test or reference) was varied to remove confounding sources of sample size variability. Across all experiments, scores for the development and test speakers were computed using MVKD (§3.2.2.1), using the reference data to assess typicality. The test scores were

then calibrated based on coefficients from the development scores (§3.2.4.1). As in Chapter 9, the results of each experiment are compared with the baseline results based on the maximum amount of available data, referred to as the *true* values (Rose 2012).

Experiment (1) considers how many speakers are required for adequate system calibration and how the calibration coefficients affect the performance of a large set of test data. Initially, 100 test speakers and 100 reference speakers with 100 tokens each were generated following the procedures in §10.2.4. The choice of the number of speakers was considered sufficiently large as to provide robust LR estimates without compromising the efficiency of the experiment. An independent set of 1000 development speakers with 100 tokens per speaker was then created. LR scores were computed initially using two development speakers. Calibration coefficients were then calculated and applied to the scores for the 100 test speakers. This process was looped, increasing the number of development speakers by one each time.

For each N development speakers, calibration shift and scale values (see Equation 3.6) were recorded. The effects on LR output for the test set are analysed using the median calibrated LLRs (as an indication of the overall distribution of LLRs) and C_{llr}. Calibrated EER was not considered in this experiment because the logistic-regression calibration procedure is optimised to reduce C_{llr}. Therefore the EER based on the test data remained the same irrespective of the calibration coefficients applied to the scores. The whole experiment was run with 20 different sets of development speakers (replications), using the same sets of test and reference data across all replications.

Experiment (2) assesses the number of test speakers required to reliably estimate system performance. LR scores were initially computed for 100 synthetic development speakers (based on the results in §10.3.1) using 100 synthetic reference speakers (100 tokens per speaker), and calibration coefficients calculated. A set of 1000 test speakers (100 tokens per speaker) was then created and scores computed initially for between two and 1000 speakers, increasing by one each time. At each stage, calibration coefficients from the development data were applied. Measures of validity (C_{llr} and EER), as well as median SS and DS LRs, were calculated at each N test speakers stage. The experiment was run over 20 replications, using the same sets of development (therefore the same calibration coefficients) and reference data.

Finally, Experiment (3) replicates the experiments in §8.3.1 and §9.3.1 investigating the effects of variability in the number of reference speakers. Scores were computed for a set of 50 development speakers and 50 test speakers, each with 50 tokens, using up to 100 reference speakers (50 tokens per speaker). The choice of 50 development and test speakers was based on a number of factors. Firstly, Experiment (3) was conducted after Experiments (1) and (2), and so the sizes of the development and test sets were informed by the results of these experiments. Secondly, Chapters 8 and 9 present the results of experiments based on very large sets of reference data. Across these experiments the most interesting effects were found with relatively small amounts of reference data. Thirdly, in terms of the practical application of the numerical LR to casework, it is more insightful to explore the effects of sample size when the number of speakers is relatively small.

Testing in Experiment (3) was performed initially with ten reference speakers. The analysis consisted of two elements. Firstly, the effects of the number of reference speakers on calibration coefficients from the development data were assessed. Secondly, the distributions of calibrated LLRs and system validity based on the test data were analysed. To compare with §9.3.1, median LRs and C_{llr} were calculated for the test set before and after calibration to assess the extent to which calibration can minimise the effects of small reference samples. EER was omitted from the analysis of scores because the results were the same as those for the calibrated LLRs. The experiment was run over 20 replications, using single sets of development and test data.

Across the three experiments, results are considered primarily in terms of variation in sample size up to 50 or 100 speakers. This is intended to focus on the variability within the range of speakers which could viably be collected in research or casework. Results for more than 100 speakers are only presented where there was considerable fluctuation in LR output with very large samples.

10.3 Results

10.3.1 Experiment (1): Number of development speakers

The effects of the number of development speakers are considered firstly in terms of the calibration coefficients generated using logistic regression. As outlined in §3.2.4, logistic regression calibration is a means of optimising C_{llr}. The logistic regression model can be described in terms of the scale (slope) and shift (intercept) coefficients of the linear trend line in the log-odds space (see Figure 3.7). As shown in Equation 3.6, to generate a calibrated LLR, a score (s) from the test data is multiplied by the scale coefficient (β) and then added to the shift coefficient (α). The scale value determines the range of the resulting LLRs and typically reduces the magnitude of extreme scores (both consistent-with- and contrary-to-fact scores). The lower the value of the scale coefficient the greater the reduction in the magnitude of extreme scores. The shift coefficient determines where the logistic regression line crosses the y-axis, and shifts all scores towards more positive values or more negative values. The larger the shift coefficient, the greater the change in the magnitude of the scores when converted to LLRs. Given that the logistic regression calibration procedure requires natural log input, calibration shift and scale are analysed in terms of their natural log values.

Figure 10.7 displays the mean calibration scale values with 95% CIs (calculated using noninformative priors as outlined in Equation 7.1) across the 20 replications as a function of the number of development speakers. There was considerable variability in the mean and imprecision across the replications with small amounts of development data. The overall range of scale values was greatest when using very small numbers of development speakers. However, scale values were found to stabilise relatively quickly, such that the *true* mean scale value of 1.20 was reached by the inclusion of 20 speakers. Between 20 and 1000 speakers the CIs also remained relatively stable.

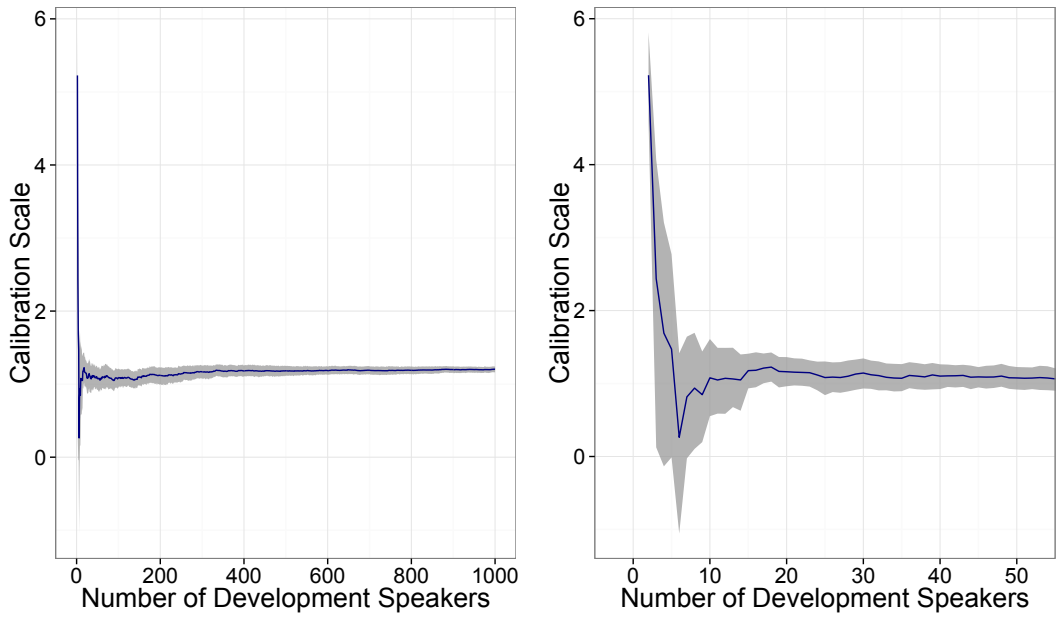

Figure 10.7: Mean (blue) and 95% CIs (grey) of calibration scale values as a function of the number of development speakers (left = two to 1000 speakers, right = two to 50 speakers)

Figure 10.8 displays the mean calibration shift values with 95% CIs across the 20 replications as a function of the number of development speakers. As in Figure 10.7, there was more variation in the mean calibration shift value with wide 95% CIs when using very small numbers of development speakers. After the inclusion of ten speakers there was a decrease in the mean shift value towards the *true* value and gradual narrowing of the CIs, reflecting greater precision across replications. However, the distribution of *true* calibration shift values was effectively reached by the inclusion of 20 development speakers and remained relatively stable until 1000 development speakers.

The effect of the scale and shift values on the calibrated LLRs was then considered. Figure 10.9 displays the distributions of median SS LLRs (with 95% CIs) across each replication based on calibration coefficients from between two and 100 development speakers. Using as few as two development speakers, SS median values were as many as three orders of magnitude stronger than the *true* value (the difference between *moderate* and *very strong* support for the prosecution). Consistent with Figures 10.7 and 10.8, the greatest imprecision in SS medians was found using small numbers of development speakers. The SS medians began to stabilise after the inclusion of 20 speakers. While the overall range of medians across replications continued to narrow as the number of

development speakers increased, the distribution of *true* SS medians was approximately achieved with the inclusion of more than 30 development speakers.

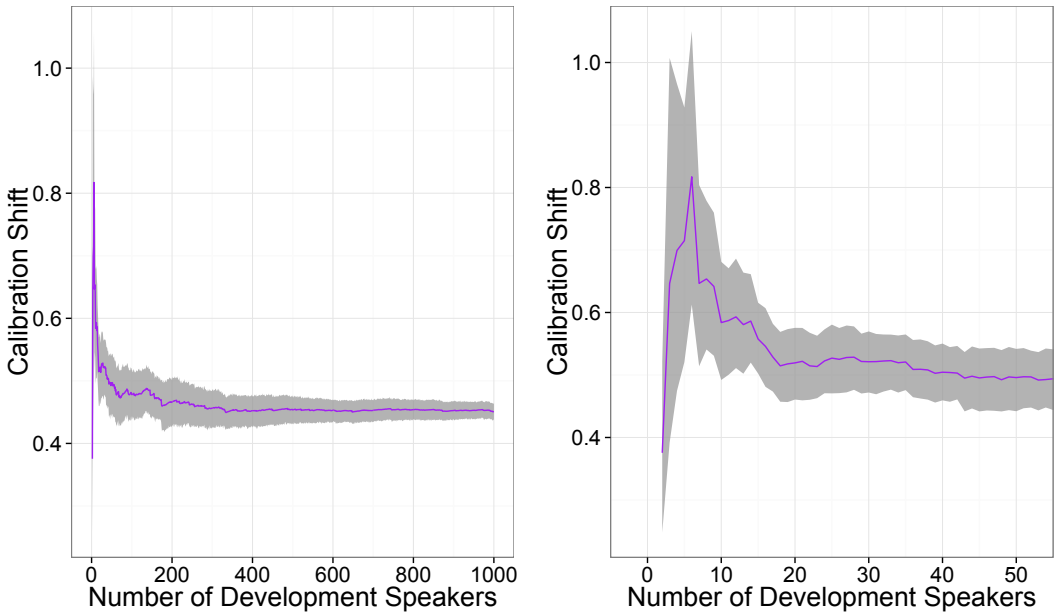

Figure 10.8: Mean (purple) and 95% CIs (grey) of calibration shift values as a function of the number of development speakers (left = two to 1000, right = three to 50)

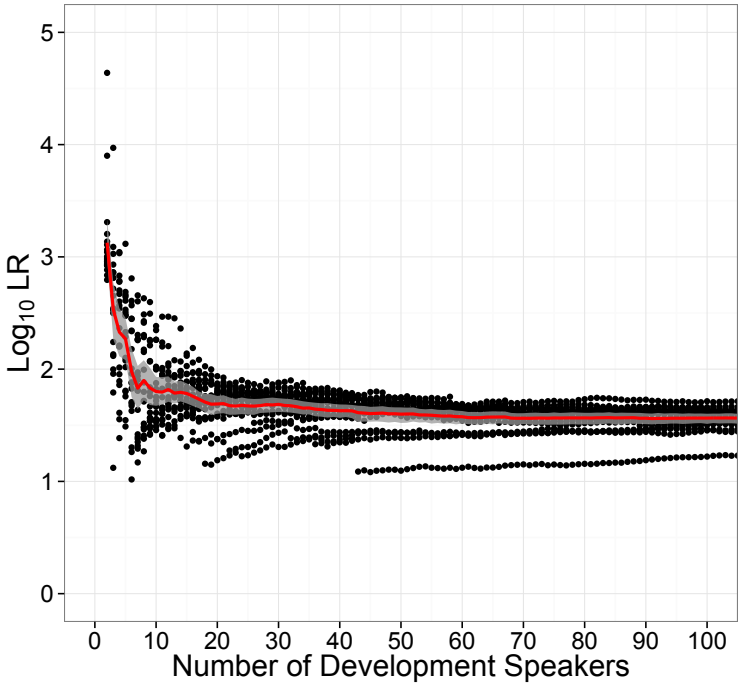

Figure 10.9: Scatterplot of median SS LLRs using between two and 100 development speakers fitted with mean (red) and 95% CIs (grey)

Much bigger effects were revealed in the distributions of median DS LLRs (Figure 10.10). The widest range of DS LLRs across replications was found using just two development speakers (from -90.64 to +1.94), compared with a range of -9 to -12 (three orders of magnitude) for the distribution of *true* median DS LLRs. As the number of development speakers increased the imprecision in the median values decreased. By 35 speakers, the mean of the medians was within one order of magnitude of the *true* value. Yet only with the inclusion of more than 68 speakers was the mean of the medians within the same order of magnitude (between -10 and -11) as the *true* value. However, given that after the inclusion of more than 15 speakers the median DS LLR was never weaker than -5, in verbal terms the medians were consistently equivalent to *very strong* support for the defence, irrespective of the absolute differences in their numerical values.

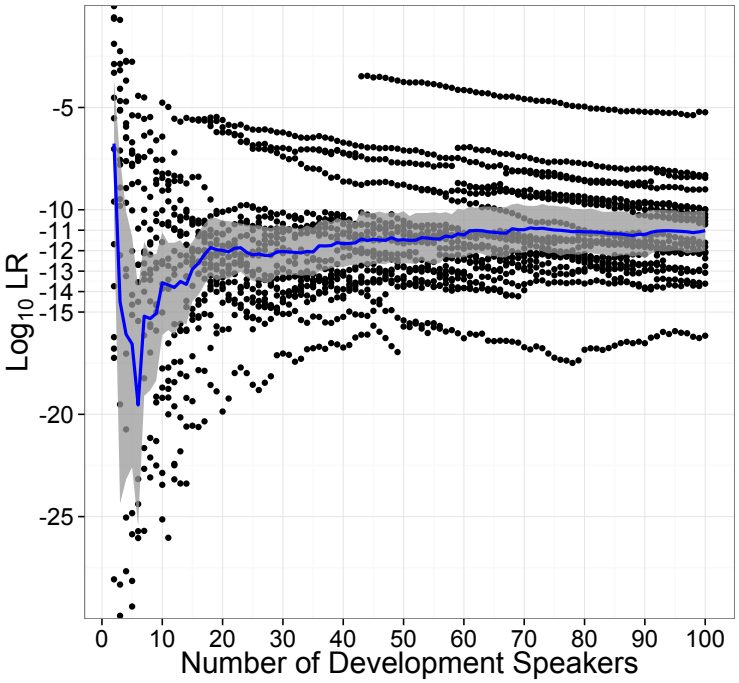

Figure 10.10: Scatterplot of median DS LLRs using between two and 100 development speakers fitted with mean (blue) and 95% CIs (grey) (outliers extend from +1.94 to -90.64 using two development speakers)

Finally, the effects of the number of development speakers are considered in terms of C_{llr} (Figure 10.11). As with the distributions of medians, there was considerably more variability in C_{llr} with fewer than ten development speakers. The range of C_{llr} values across replications extended from 0.13 to maximally 2.90 when using two speakers,

compared with a *true* range of between 0.106 and 0.116 (based on 1000 development speakers). For the majority of replications, C_{llr} was higher when using small numbers of development speakers compared with the *true* values. However, with the inclusion of more than 20 speakers the distributions of C_{llr} values were essentially stable, despite the large differences in the magnitudes of DS median LLRs in Figure 10.10. This is probably due to the fact that UM is a good speaker discriminant which produces high magnitude LLRs. Since the variability in DS LLRs as a function of development sample size primarily occurred in higher magnitudes (i.e. stronger support for the defence), it had little effect on overall system validity.

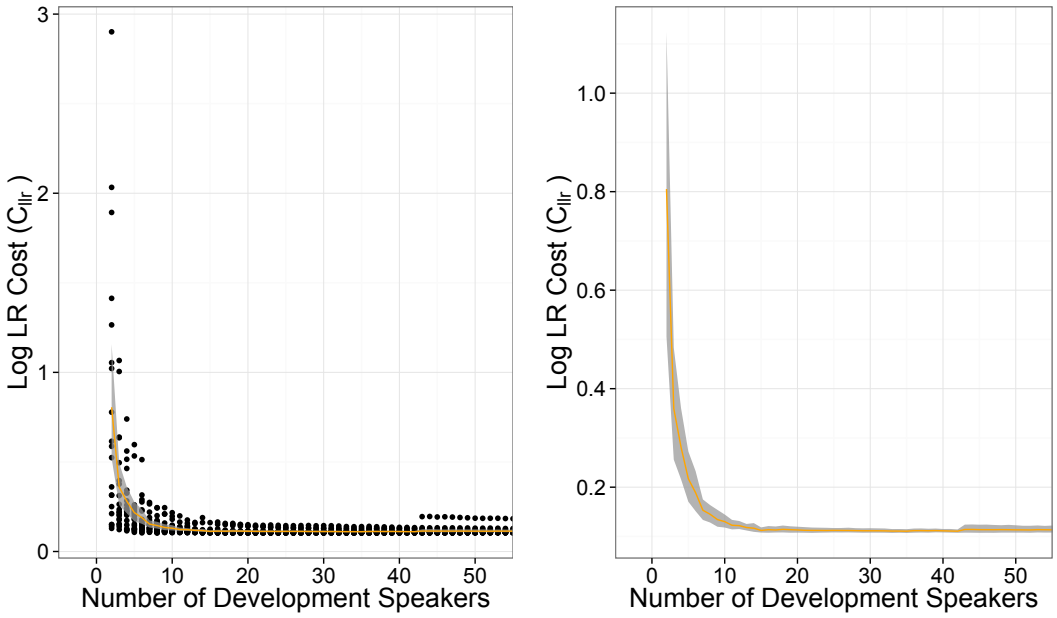

Figure 10.11: Scatterplot of C_{llr} values (left; scale = 0-3) using between two and 100 development speakers fitted with mean and 95% CIs (right; scale = 0-1)

10.3.2 Experiment (2): Number of test speakers

Figure 10.12 displays the distributions of calibrated median SS LLRs across the 20 replications using between two and 50 test speakers. As in §10.3.1, the greatest imprecision in median SS LLRs relative to the distribution of *true* medians was found using the smallest number of test speakers (in this case five). Between two and ten test speakers, medians were both weaker and stronger than the *true* values. This imprecision was reflected in the width of the CIs. However, across all replications with between

five and 1000 test speakers, the median never extended outside the range of *moderate* support for H_p. With the inclusion of more than 20 test speakers the distribution of SS medians appeared to stabilise, such that there was essentially no difference in the mean, 95% CIs and overall range of median values between 20 and 1000 test speakers.

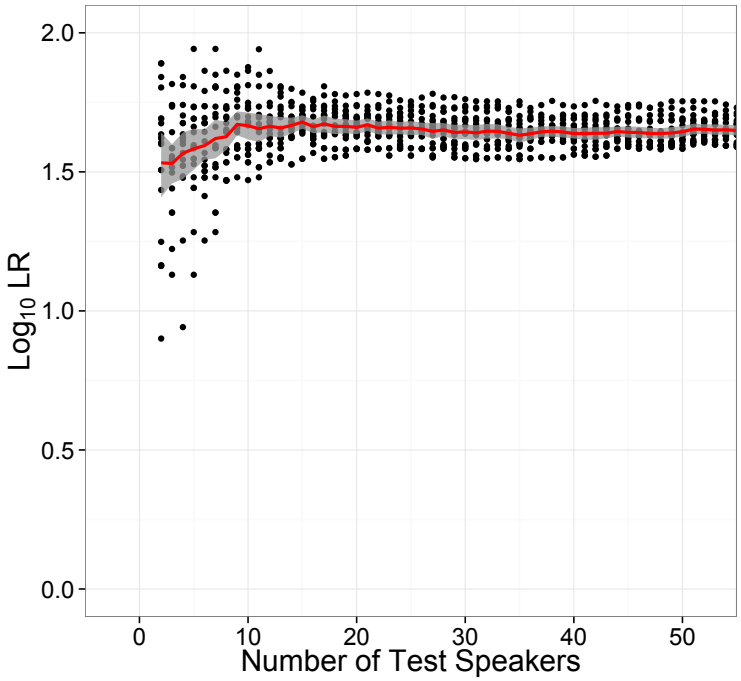

Figure 10.12: Scatterplot of median SS LLRs using between two and 100 test speakers fitted with mean (red) and 95% CIs (grey)

As in Experiment (1), considerably greater variability as a function of the number of test speakers was found for DS LLRs. Figure 10.13 displays DS medians using between two and 1000 test speakers. Medians were both weaker and stronger than the *true* medians (based on 1000 test speakers) by as many as 15 orders of magnitude when using the smallest number of test speakers. Consistent with this, the largest range of DS medians across replications was found using five speakers. However, Figure 10.13 suggests that there was still considerable fluctuation in the distribution of medians with relatively large numbers of test speakers. While the central tendency stabilised with more than 50 speakers, the interquartile range decreased markedly between 50 and 100 speakers, after which there was still some decrease in the overall range of medians.

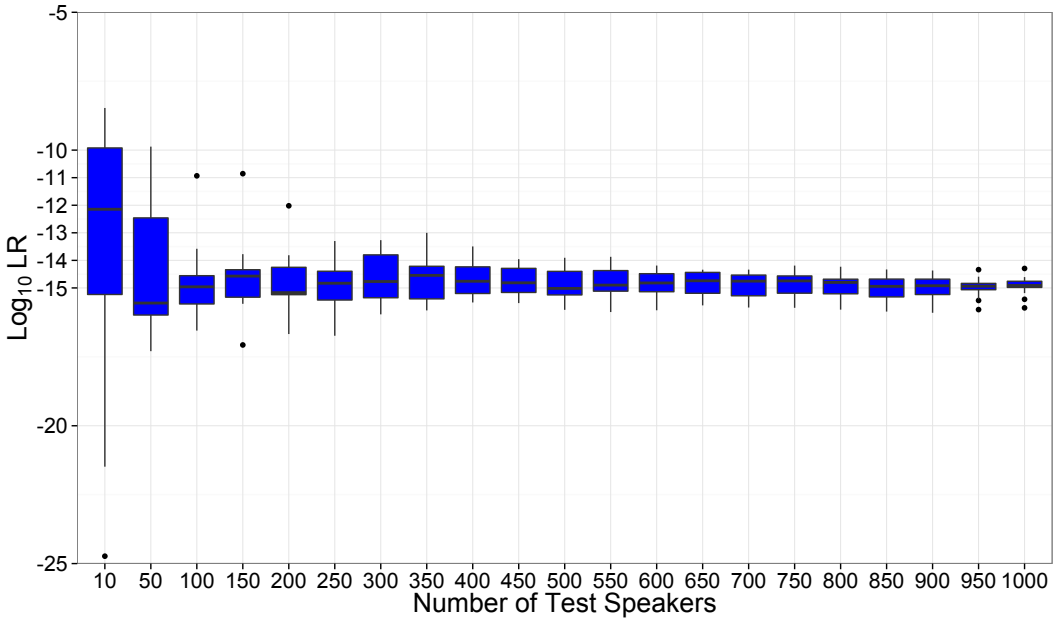

Figure 10.13: Boxplots of median DS LLRs as a function of the number of test speakers

Of more interest for FVC casework is the issue of validity variability as a function of the number of test speakers. Figure 10.14 displays EER and C_{llr} values using between two and 50 test speakers. EER was overly optimistic (i.e. lower than the *true* values) when using small amounts of test data. In fact, when using only two test speakers, all 20 replications produced an EER of 0%. This is unsurprising given that with two test speakers there were only two SS and four DS comparisons. As the number of test speakers increased between five and 30 speakers the overall range of EER values across replications increased, reflecting greater imprecision in validity across replications. Although the mean EER stabilised at around 3% by the inclusion of more than 25 test speakers, there was continual narrowing of the overall range of EERs as the size of the test sample increased. Such narrowing continued with very large sample sizes since the range of EER values based on 50 speakers was greater (2.12%) than that of the *true* EERs using 1000 test speakers (0.48%).

A similar pattern was found for C_{llr}. In 16 of the 20 replications, C_{llr} was lower using two test speakers compared with the *true* C_{llr} values based on 1000 test speakers. Following the initially overly optimistic C_{llr} performance, the bulk of C_{llr} values began to stabilise and cluster within a range of around 0.05 after the inclusion of ten test speakers. However, in terms of the CIs, there was a marked increase in imprecision as the number of test speakers increased between two and 12 speakers. This was due to

the C_{llr} values from two replications which achieved higher C_{llr} values relative to the other replications. After this point, with the inclusion of more test speakers the level of imprecision decreased. The overall range of C_{llr} values continued to decrease with the inclusion of more than 50 speakers towards the distribution of the *true* C_{llr} values using 1000 speakers.

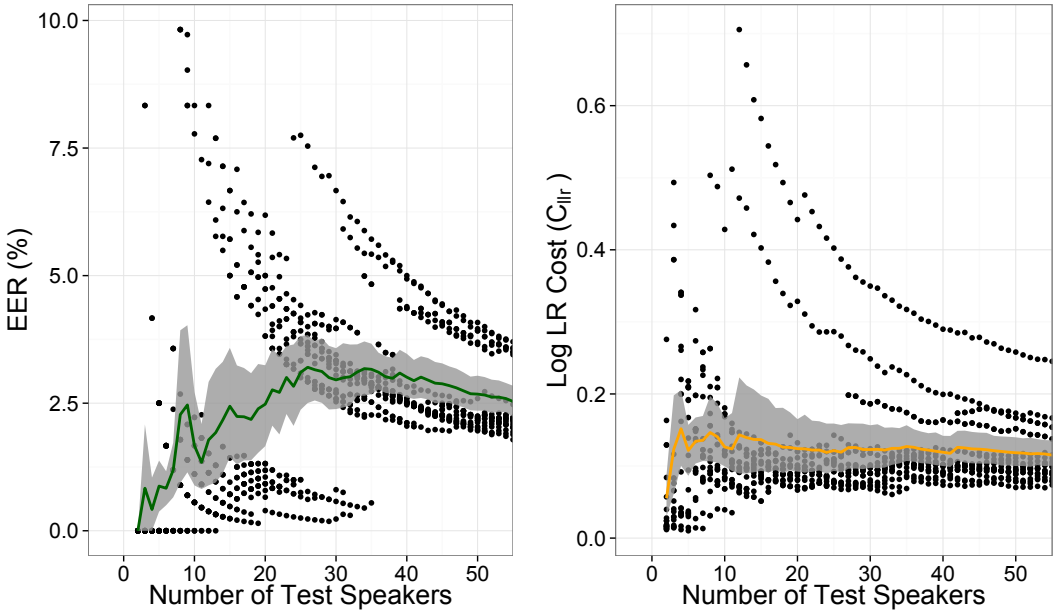

Figure 10.14: Scatterplot of EER (left) and C_{llr} (right) as a function of the number of test speakers fitted with the group mean and 95% CIs (grey)

10.3.3 Experiment (3): Number of reference speakers

This section explores the effects of the number of reference speakers firstly on calibration coefficients and calibrated LLRs, and then discusses the effects on uncalibrated scores.

Calibrated LLRs

Figure 10.15 displays the mean calibration shift and scale values with 95% CIs (in terms of natural log values) across replications as a function of the number of reference speakers used in the feature-to-score stage for the development data. With the smallest number of reference speakers the mean calibration shift was lower by 0.09 relative

to the *true* mean shift value of 0.763. Despite some random fluctuation, there was a continual increase in the mean shift value as the number of reference speakers increased. However, this variability occurred within a very narrow range.

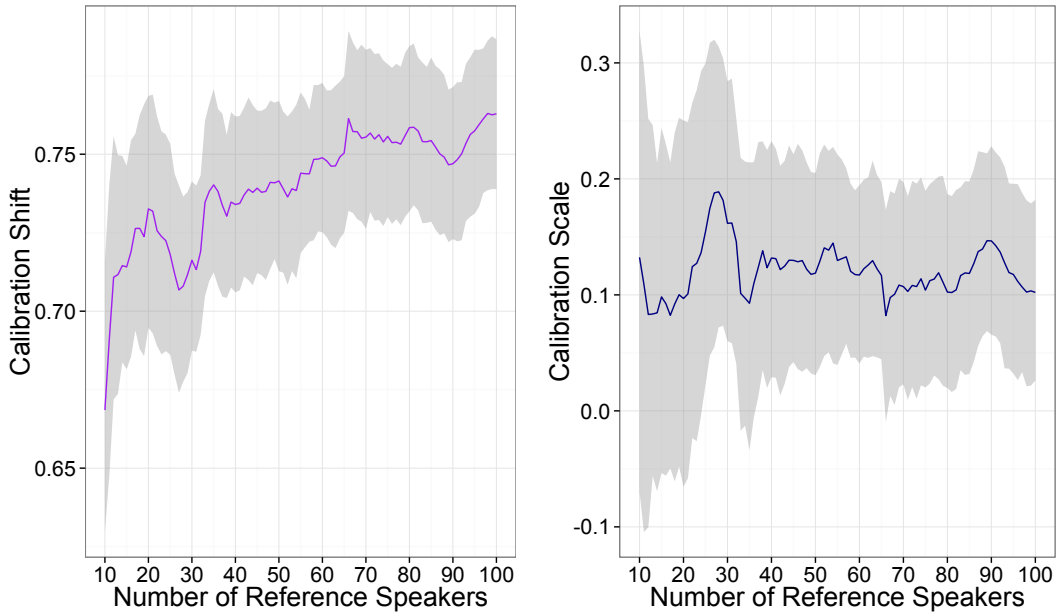

Figure 10.15: Mean and 95% CIs of calibration shift (left) and scale (right) values using between 10 and 100 reference speakers

Unlike the calibration shift values, there was no clear correlation between mean calibration scale and the number of reference speakers. There was random variation in the mean within a very narrow range, such that the mean based on ten reference speakers (0.132) was essentially the same as the *true* mean (0.102) using 100 speakers. There was however a marginal narrowing of the 95% CIs as the number of reference speakers increased, reflecting a slight decrease in the imprecision across replications. Nonetheless, Figure 10.15 suggests that calibration coefficients were essentially robust to the effects of small reference sample size for these data.

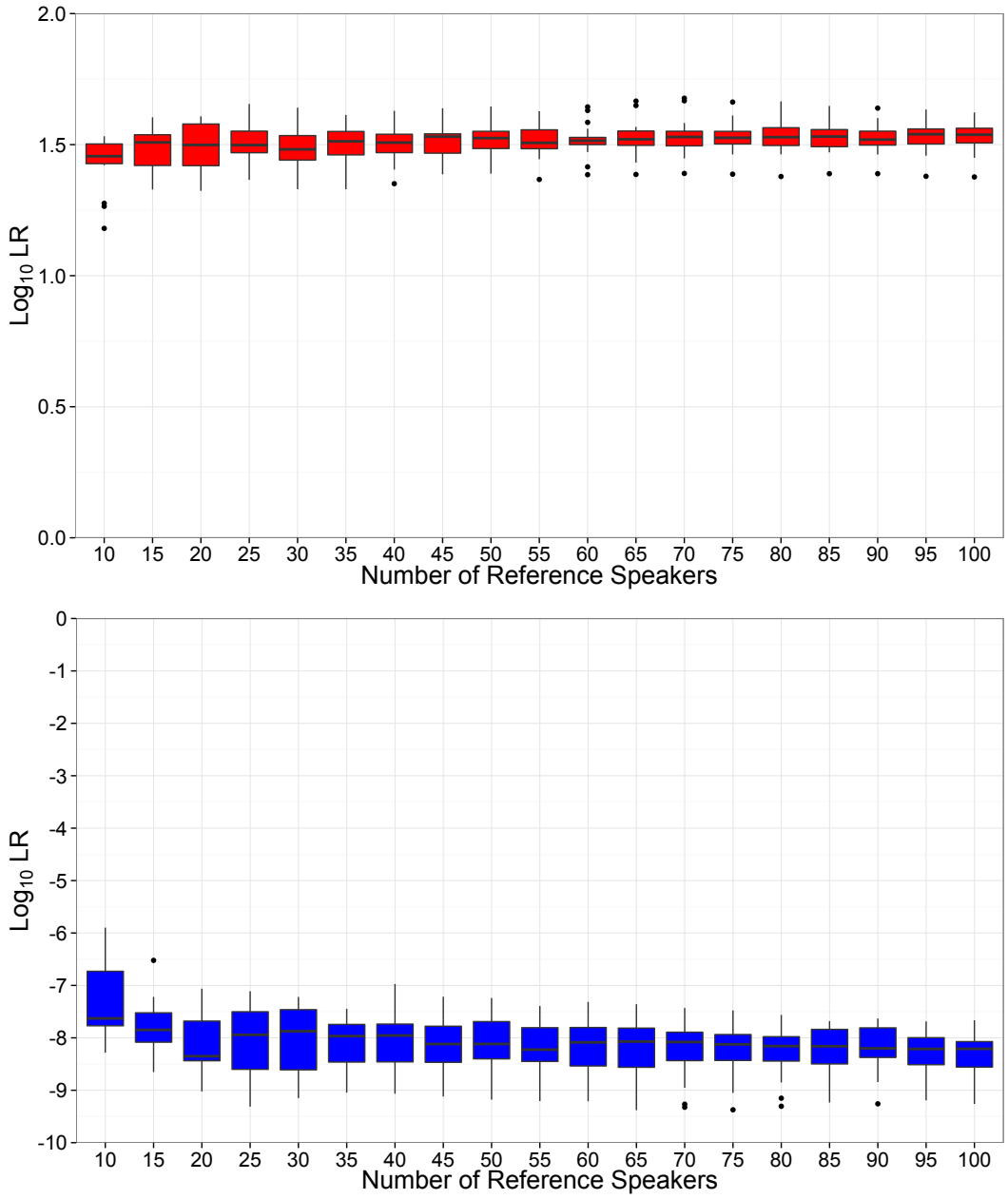

Figure 10.16: Boxplots of median calibrated SS (above) and DS (below) LLRs as a function of the number of reference speakers

Given the stability of the calibration shift and scale values, it is unsurprising that the associated LLRs were also relatively robust to the number of reference speakers (Figure 10.16). There was very little difference in the distribution of SS medians based on ten reference speakers and the *true* values using 100 speakers. In fact, all SS medians across all replications were consistently within the range of +1 and +2, equivalent to *moderate* support for the prosecution. There was greater variability in the distributions of DS medians. Firstly, there was greater imprecision across replications even in terms of the

distribution of the *true* medians, with values extending over two orders of magnitude from -7.665 to -9.262. Secondly, the DS medians displayed more sensitivity to small numbers of reference speakers. The median was generally weaker (by up to three orders of magnitude) when using ten reference speakers compared with the *true* medians. The overall range of DS medians was also largest when using ten reference speakers. By the inclusion of 35 speakers, the distributions of DS medians became relatively stable.

As shown in Figure 10.17, EER was relatively robust to the number of reference speakers across replications. Mean EER was consistently less than 6% with some random variation within a range of 0.5% as the size of the reference sample varied. Further, the upper and lower bounds of the 95% CIs were consistently within the range of 5% and 6%, although there was evidence of a slight decrease in imprecision as sample size increased. The EER results, therefore, indicate that categorical validity was largely unaffected by using different groups of N reference speakers and differences in sample size itself for these data.

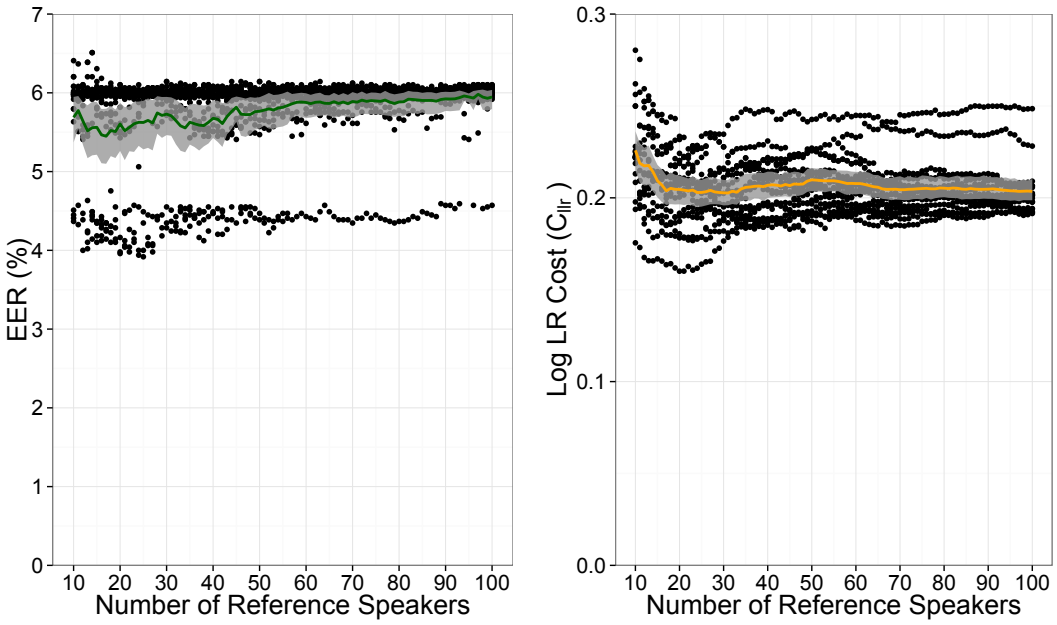

Figure 10.17: Scatterplot of EER (left) and C_{llr} (right) based on LLRs as a function of the number of reference speakers fitted with the group mean and 95% CIs

More variability was found in terms of C_{llr}. As with the DS medians, there was some variability across replications when considering only the *true* C_{llr} values based on all 100 reference speakers. While 18 of the 20 replications produced C_{llr} values of between

0.19 and 0.21, two replications produced more extreme values approaching 0.25. This highlights that even with large amounts of reference data, there is still potential for quite considerable variability in system validity depending on the specific speakers used. Across conditions, C_{llr} was relatively robust to reference sample size after the inclusion of more than 30 speakers. With fewer than 30 speakers, there was considerably more imprecision in C_{llr} values across replications.

Uncalibrated scores

To compare with the results in §9.3.1, the effects of reference sample size on the distributions of uncalibrated scores and respective C_{llr} values were also analysed. Figure 10.18 displays the distributions of median scores as a function of the number of reference speakers. Compared with the calibrated results in Figure 10.16, both SS and DS median scores were considerably larger in magnitude (i.e. offered stronger evidence). Considering just the results based on 100 reference speakers, calibration reduced the magnitude of the SS medians by approximately three orders of magnitude and reduced the magnitude of the DS medians by approximately 17 orders of magnitude.

In terms of sample size sensitivity, the results for the scores were similar to those for the calibrated results in Figure 10.16. The distribution of *true* median SS scores across replications ranged from +4.17 to +4.62, consistently equivalent to *very strong* support for the prosecution. Only when using fewer than 15 speakers did the SS medians extend outside the +4 to +5 range. As in all of the experiments in this chapter, greater imprecision in the distribution of SS scores was found using smaller numbers of reference speakers. With more than 20 speakers, the median SS scores remained relatively stable, although the overall ranges continued to decrease as the number of reference speakers increased. As for the calibrated LLRs, greater sensitivity to sample size was revealed for the DS scores. The range of median DS scores (-18 to -35; range = 17) based on ten reference speakers was considerably greater that of the *true* median DS scores (range = 8). With the inclusion of more than 35 speakers, the range of DS median scores became much more stable.

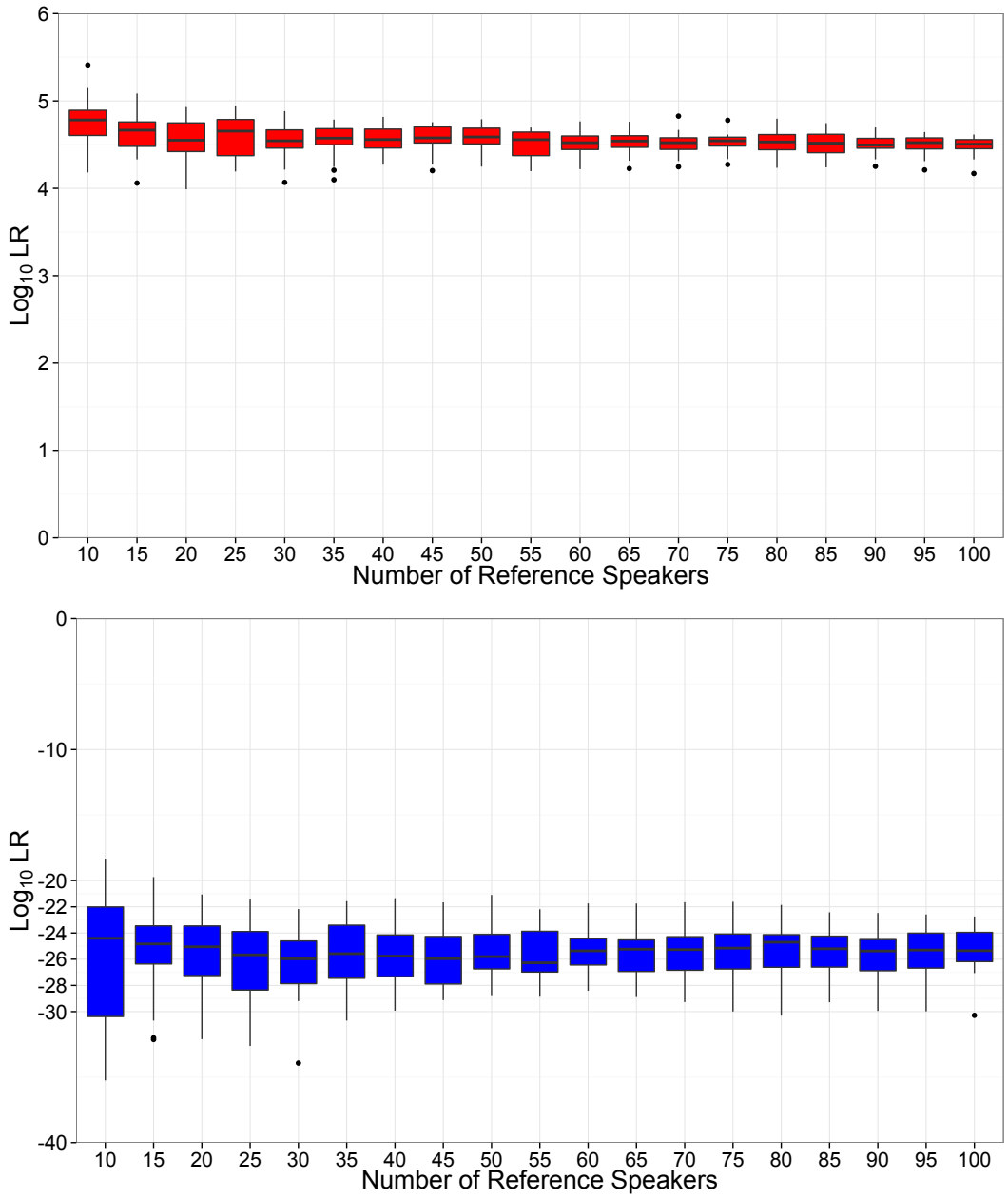

Figure 10.18: Boxplots of median SS (above) and DS (below) scores as a function of the number of reference speakers

Figure 10.19 displays C_{llr} based on scores as a function of reference sample size. Greater imprecision in C_{llr} values was found across all conditions based on scores, compared with the calibrated LLRs. However, the patterns of C_{llr} variability based on the scores were similar to those based on calibrated LLRs (Figure 10.17). Imprecision was greatest with the smallest number of reference speakers, with C_{llr} generally higher compared with the *true* C_{llr} values. Further, stability in the distribution of C_{llr} values was achieved with the inclusion of more than 30 reference speakers.

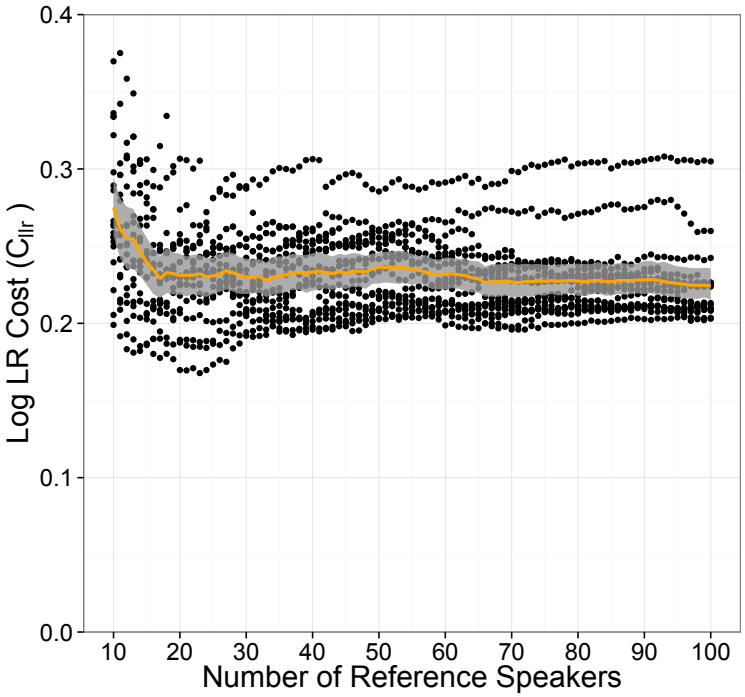

Figure 10.19: Scatterplot of C_{llr} based on scores as a function of the number of reference speakers fitted with the group mean and 95% CIs

10.4 Discussion

The results of Experiment (1) showed that while calibration shift values were essentially robust to the number of development speakers, calibration scale values were more sensitive to sample size variation. Specifically, scale values were generally higher and displayed considerably more imprecision across replications with small amounts of development data (i.e. fewer than 15) compared with the distribution of *true* scale values. However, with the inclusion of more than 15 speakers the distribution of calibration scale values stabilised. The effects of such variability on the resulting LLRs were also analysed. Based on these data, minimally 20 development speakers are considered adequate for achieving stable distributions of LLRs, at least for this variable. However, stable C_{llr} values were achieved with the inclusion of more than 12 speakers.

Experiment (2) investigated the size of the test data on LR output. As in Experiment (1), LLRs were generally weaker and displayed considerably greater imprecision across replications with fewer than 20 test speakers compared with the distributions of *true* LLRs. The wide range of variability in the distributions of median LLRs when using

small numbers of test speakers highlights that in system testing small test samples may provide a misleading assessment of the general strength of evidence achieved by a given system. More importantly, both EER and C_{llr} were found to be overoptimistic when using very small numbers of test speakers (i.e. fewer than five), due to the small number of comparisons. With between five and 30 test speakers, there was considerably greater imprecision in EER and C_{llr} across replications compared with the distributions of *true* EER and C_{llr} values based on 1000 test speakers.

Experiment (3) tested the number of reference speakers on the calibration coefficients generated through logistic regression. Both shift and scale values were found to be extremely robust to the size of the reference data. The distributions of the resulting SS LLRs were also found to be essentially robust to reference sample size. DS medians were initially weaker and displayed greater imprecision using small numbers of reference speakers (i.e. fewer than 20) compared with the distribution of *true* medians using 100 speakers. This contrasts with the results for /uː/ (§8.3.1.1) and /aɪ/ (§8.3.1.2) where stronger DS LLRs were found using ten reference speakers. Finally, C_{llr} in Experiment (3) stabilised relatively quickly (by around 20 speakers), contrasting with the linear improvement in C_{llr} with the addition of more speakers for /uː/ and to some extent for /aɪ/.

The relative stability of calibrated LLR output across conditions in Experiment (3) suggests that sensitivity to reference sample size is not dependent on the inherent speaker discriminatory value of the input variable. Rather, the results of Experiment (3) provide further evidence to support the claim that sensitivity to sample size is, at least in part, determined by the dimensionality of the variable. Comparison across Chapters 8, 9 and 10 shows that calibrated LLRs based on UM (three dimensions) were more sensitive to reference sample size than those based on AR (one dimension) and less sensitive to sample size than the scores based on /uː/ and /aɪ/. However, it is important to reiterate that the experiments using /uː/ and /aɪ/ were based on scores which may also account for the sensitivity to sample size displayed in Chapter 8.

The uncalibrated results of Experiment (3) also offer insights into the potential role of calibration in reducing the sensitivity of LR output to reference sample size, as discussed in §9.4. In Experiment (3), the same effects on the distributions of LLRs

and C_{llr} were found for the uncalibrated scores. The magnitude of the effects was also broadly the same across the two sets of LR output, indicating that for UM the uncalibrated scores were no more sensitive to variability in reference sample size than the calibrated LLRs (although inevitably the medians are stronger and the C_{llr} values worse for the uncalibrated scores). This finding contrasts with the results in Chapter 8, where calibrated LLRs for AR were essentially stable regardless of reference sample size, while scores were more sensitive to sample size variation. The contrasting results between AR and UM offer support for the proposition in §9.4 that the role of calibration in reducing the sensitivity to reference sample size may be dependent on the individual variable and its inherent speaker discriminatory power.

The three sets of experiments also raise issues relating to the potential trade-offs between the numbers of development, test and reference speakers used either in system testing. Predictably, effects were found as a function of all three sources of sample size variability. However, the relative importance of the size of each of the development, test and reference datasets, and which, given the inevitable practical constraints in research and casework, should contain the most amount of speakers, is dependent on which element of the LR output the analyst is interested in. To generate meaningful calibration coefficients the size of the development set is of primary concern, owing to the imprecision in shift and scale values when using small amounts of development data. Extrapolating from Experiment (3), calibration coefficients are robust to small numbers of reference speakers provided a large amount of development data is used. Given the large effects of small numbers of development speakers on calibration coefficients, the magnitudes of calibrated SS and DS LLRs are also most dependent on the size of the development set. That is, to achieve more precise calibrated LLRs it is preferable to have a large set of development speakers and a small set of reference speakers rather than vice versa. In the absence of calibration (i.e. where there is no development data), a precise estimate of the distribution of SS and DS scores is dependent on both the number of test and reference speakers, where preference should marginally be given to the number of test speakers.

Predictably, system validity, both in terms of C_{llr} and EER, was most sensitive to the number of test speakers used. This applied equally to calibrated LLRs as to uncalibrated

scores. Importantly, the size of the test set had a range of effects on system validity. When using fewer than five test speakers, validity was over optimistic (e.g. EER = 0%), while with small to moderate numbers of test speakers (5-15) validity was considerably more imprecise. On the basis of these potential trade-offs, in any form of LR-based FVC testing, large sets of development and test data should be considered a priority. With sufficiently large amounts of development and test data (20-30 speakers per set), a moderate amount of reference data (15 speakers) should suffice to provide precise LR output.

Finally, the results of the three experiments in this chapter have highlighted that even with relatively large numbers of development, test and reference speakers there may still be considerable imprecision in elements of LR output. In Experiment (2), the C_{llr} values were spread over a range of around 0.2 even with the largest number of available test speakers. This is a relatively large range of potential variability in terms of the absolute value for system validity which is reported to the court. Potential means of dealing with such inherent imprecision are considered in §11.3.

10.5 Chapter summary

Experiment (1): Number of development speakers

- Calibration shift and scale values stable with more than 15 development speakers.
- Calibrated SS LLRs stable when using more than 20 development speakers.
- Calibrated DS LLRs more sensitive to the number of development speakers, particularly when using small amounts of development data (fewer than 30).
- C_{llr} robust by the inclusion of 15 or more speakers.
- Fewer than 20 speakers should be avoided when calibrating using logistic regression.

Experiment (2): Number of test speakers

- SS median LLRs relatively stable to the number of test speakers, although greater imprecision with fewer than 20 speakers.
- DS median LLRs more sensitive to the number of test speakers with considerably wider overall range when using fewer than 40 speakers.
- EER and C_{llr} overoptimistic using very small numbers of test speakers (between two and five) due to the small number of available comparisons.
  - Greater imprecision in validity between five and 12 speakers followed by a decrease in imprecision as the number of test speakers increased.

Experiment (3): Number of reference speakers

- Calibration coefficients largely unaffected by reference sample size.
- 20 reference speakers considered minimum for robust LLRs (but ideally should be greater than 30 speakers).
  - Increasing the number of reference speakers only improves the precision of the LLRs themselves, rather than improving overall system performance.
- Calibration does not appear to play a role in ensuring output is robust to sample size for UM.

Chapter 11

Discussion and Conclusions

This chapter addresses the research questions outlined in §2.6, in light of the empirical results presented in Chapters 4 to 10. §11.1 provides an overview of the results based on dimensions of regional and social variation (Chapters 4 to 7) in the definition of the relevant population. The limitations of current approaches to defining the relevant population are also considered and three alternative approaches are presented, namely (i) multiple defence propositions; (ii) normalisation of variation in sub-populations; and (iii) expert-judged speaker similarity. §11.2 provides an overview of the effects of sample size on LR output across the experiments in Chapters 8 to 10. §11.3 considers the wider implications of the findings for FVC, while §11.4 considers directions for future research. Finally, a general conclusion is given at §11.5.

11.1 Defining the relevant population

Effects on the magnitudes of LLRs

The results in Chapters 4 to 7 reveal a number of systematic effects of different definitions of the relevant population when using logical relevance on the magnitude of LRs. §4.3.1 suggests that SS scores are generally stronger (by one order of $\log_{10}$ magnitude for /uː/) when using Mismatched reference data compared with Matched data. The higher magnitude of SS scores is a predictable outcome of using Mismatched and Mixed data at the feature-to-score stage. This is because the distribution of a

Mismatched reference dataset will be shifted relative to that of Matched reference data, where both are expected to display similar levels of between-speaker variation. A Mixed set, however, is expected to display greater between-speaker variation since it contains a sociolinguistically heterogeneous group of speakers. Therefore, certain offender values will be located onto the tails of the Mixed distribution, and further onto the tails of the Mismatched distribution, meaning that $p(\mathrm{E}|\mathrm{H}_d)$ is lower than it would be using Matched reference data. The result is higher magnitude scores for the Mixed and Mismatched data than for the Matched data.

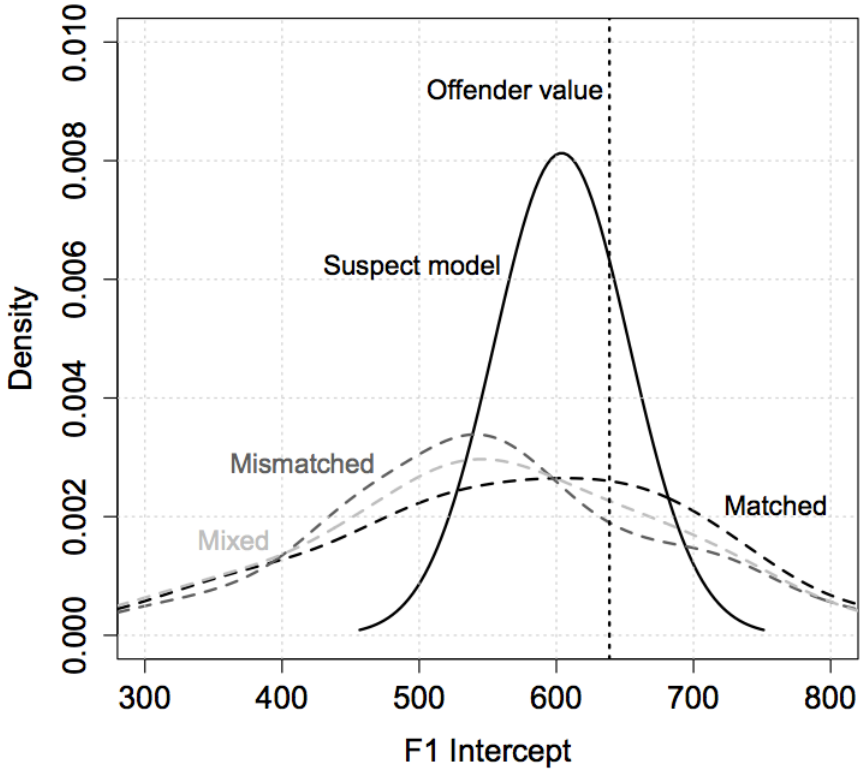

Figure 11.1: Univariate example of a SS comparison (test speaker 11) from §7.3.2 (variation in age) assessing the probability of the offender value (639) at the intersection of the normal suspect model and KD Matched, Mismatched and Mixed models

Analysis of the uncalibrated results across Chapters 4 to 7 reveals overinflation of SS scores (albeit to different extents) when using Mismatched or Mixed data, similar to that in §4.3.1. A univariate KD example of this based on a SS comparison from §7.3.2 (variation in speaker age) using the F1 intercept of /eɪ/ is displayed in Figure 11.1. In this case, the Matched (raw) score would be approximately 2.41, compared with 2.82 in the Mixed condition and 3.42 in the Mismatched condition. Consistent with this, the SS

scores from §7.3.2 and the replications in §7.3.4 were weakest for the Matched system, followed by the Mixed system, while the Mismatched system generally produced the strongest SS scores. This is exemplified by the fact that 40% of the SS scores from §7.3.2 were over one order of $\log_{10}$ magnitude stronger using the Mismatched system compared with the Matched system.

Despite this, relatively little difference was found between the distributions of calibrated SS LLRs across systems for any of the experiments. This may be explained by the use of Matched, Mismatched and Mixed data throughout training as well as testing. As outlined in §3.2.4.1, logistic regression calibration is a means of minimising C_{llr} based on knowledge of how the system performs using development data. The scale (slope) coefficient of the logistic regression model is primarily responsible for minimising the magnitude of strong contrary-to-fact LRs, such that the smaller the scale coefficient the greater the reduction in magnitude when converting from score-to-LR. A flatter logistic regression model also has the effect of reducing the magnitude of strong consistent-with-fact scores, thus reducing the overall range of the calibrated LLRs compared with the scores. In this way, the scale coefficient affects the magnitude of most extreme SS and DS scores.

Given that the development and reference sets in these experiments contained speakers of the same demographic background (irrespective of the demographic background of the test data), the logistic regression models across the three systems are roughly equivalent (i.e. no overinflation of SS scores), and in particular scale coefficients are roughly equivalent. When the calibration model is then applied to the test scores, the magnitudes of the overinflated scores produced by using Mismatched/Mixed reference data are reduced. Calibration therefore appears to scale the distributions of Mismatched/Mixed SS scores towards the distribution of the Matched scores, meaning that the distributions of the resulting calibrated LLRs across systems are more similar to each other. These experiments are the first to report on the role of calibration in quasi-normalising LR output from different definitions of the relevant population.

The role of calibration is of particular interest in terms of the results in Chapter 7. Despite the fact that there was less class-based variation in the raw data (§7.2.5) than age-based variation, bigger differences were found in the SS LLRs across the Matched,

Mismatched and Mixed systems when varying class rather than age. When considering the uncalibrated results, however, overinflation of SS scores from the Mismatched and Mixed systems was considerably greater for age than for class (consistent with differences in the raw data). Calibration therefore reduced the magnitude of these overinflated scores, ameliorating the effects of using Mismatched and Mixed data, to the extent that differences across systems in SS LLRs for age were smaller than those based on class.

The shifting of the reference distribution when using Mismatched and Mixed data also affects DS scores. Given the differences in the location of the distributions for each system, certain offender values will again be situated further onto the tails of the Mismatched or Mixed distribution, meaning that $p(\mathrm{E}|\mathrm{H}_d)$ will be lower than in the Matched condition. In such cases, the resulting scores will be lower in magnitude when using Mismatched or Mixed reference data. For example, based on the suspect and background models in Figure 11.1, a DS offender value of 765 from speaker 19 would produce the highest magnitude score using the Matched system, followed by the Mixed system and finally the Mismatched system would produce the weakest score.

The differences in the uncalibrated DS scores across systems were also found in the resulting calibrated LLRs albeit to a lesser extent. LLRs were generally weaker using Mixed systems and weakest using the Mismatched systems, compared with the distributions of DS LLRs produced by the Matched systems. Since LLRs were shifted closer to zero using the Mismatched and Mixed systems, there was also generally a higher proportion of contrary-to-fact DS LLRs compared with the Matched systems. Comparison of the uncalibrated and calibrated LRs suggests that calibration is able to reduce the effects of using Mismatched and Mixed system data for DS comparisons, but not to the same extent as for SS pairs, since the systematic patterns found in the feature-to-score stage were also found in the calibrated DS LLRs.

It is important to emphasise that, irrespective of the general patterns, LLRs from individual comparisons are affected by the use of Matched, Mismatched and Mixed systems to different extents. In particular, pairs which produced higher magnitude support generally displayed greater variability in LLRs across systems. This is highlighted by the fact that the width of the 95% CIs in Figures 6.9, 6.14 and 7.11 increased as the

strength of the mean SS and DS LLRs increased. Magnitude may also explain why the differences in the distributions of DS LLRs across systems were generally much bigger than those for SS LLRs. As stated by Rose *et al.* (2006), "two samples (from the same speaker) cannot get more similar for a feature than identical" (p. 334), whereas for different speaker pairs the range of potential dissimilarity is considerably greater. Further, the fact that cepstral input (Chapter 6) generated the strongest LLRs explains why it also produced greater variability in individual LLRs (i.e. highest mean 95% CI) than linguistic-phonetic input.

Effects on system validity

With the exception of Chapter 6, systematic effects were also found in terms of system validity. Consistent with the empirical demonstration in Morrison *et al.* (2012), validity was generally better using data selected based on an appropriate assumption about the relevant population (i.e. Matched), rather than using Mixed data. Considerably larger differences in terms of EER and C_{llr} were found between the Matched and Mismatched systems, where the Matched systems almost consistently produced the best performance and the Mismatched systems consistently produced the worst performance. In Chapter 6, essentially no difference was found across the Matched, Mismatched or Mixed systems in terms of EER or C_{llr}, despite considerable variability in LLRs from individual comparisons. There are two reasons for this. First, the use of forensically unrealistic data involving read speech and high quality recordings means that performance is essentially at ceiling (EER and C_{llr} are extremely close to zero). Secondly, since the variability in the individual LLRs occurs so far away from the zero threshold, it has little effect on validity.

Individual variables and sociolinguistic factors

The extent of the differences in LR output across different definitions of the relevant population depends, of course, on the variable under analysis and the source of between-speaker variation. In general terms the degree of sensitivity to the definition of the relevant population was predictably dependent on the amount of *group* information (Garvin and Ladefoged 1963) encoded in the variable. Consistent with the high level

of regional variation in BrEng predicted in §5.2.2, considerably greater variation in LR output across systems was found for /aɪ/ (Chapter 5) than for general English /uː/ (Chapter 4) or AmEng cepstral input (Chapter 6). Similarly, the level of social (class and age) variation encoded in NZE /eɪ/ was not predicted to be as great at the regional variation in /aɪ/. This was again confirmed by the fact that variation in LR output across systems for /aɪ/ based on regional groups was more severe than that for /eɪ/ based on class or age. Further, in Chapters 4 and 5, the removal of F1 and F2, variables which encode *speech* information and *speaker* information relating to the *group* (i.e. primarily responsible for the perceptual cues which account for phonetic contrast and regional and social patterns; see §3.3.1.1), reduced the differences between the Matched and Mixed systems in terms of the distributions of the resulting LLRs.

However, a number of exceptions to this pattern were also found which are problematic for predicting the effects of different definitions of the logically relevant population. First, in Chapter 5, the EER and C_{llr} differences between the Matched and Mixed systems were greatest using F3-only from /aɪ/ compared with F1∼F3 or F2 and F3, despite the fact that F3 was shown empirically to encode more speaker-specific (§5.3.3) than region-specific information (§5.3.2). The validity results may be explained by the fact that the addition of F1 and F2 in §5.3.3 increased the magnitude of SS and DS LLRs, compared with any individual formant. Therefore, when using F3-only LLRs were generally closer to zero, meaning that the smaller differences in the magnitudes of LLRs between the Matched and Mixed systems had much bigger effects on validity. Secondly, the opposite pattern was found using cepstral input in Chapter 6. Cesptral input was expected to display less regional variation that linguistic-phonetic input; however, very large differences were found across the LLRs from individual comparisons, especially those of a very high magnitude. As outlined above, the range within which this variability was found meant that it did not affect validity.

Finally, age-based differences found in the raw data in Chapter 7 were larger than class-based differences. However, the variation in terms of the distributions of LLRs and the variability in LLRs from individual comparisons across systems were greater when varying the definition of class. This is primarily attributed to the effects of calibration (see above). These exceptions suggest that predictions about the effects of different

definitions of the relevant population are also dependent on the element of LR output the analyst is interested in. A concerning issue with these exceptions is that they affect the precision of individual LLRs and system validity; the two elements of LR output that are of primary importance in FVC casework.

Therefore, in defining the logically relevant population, the analyst must consider the important regional and social forces which determine between-speaker variation for the variable(s) under analysis. This allows the expert to determine which factors should be controlled when defining the relevant population and, in particular, when collecting data for LR testing based on their potential effects on LR output. Such predictions may be made based on experience and expertise, as well as with reference to published research in sociolinguistics and sociophonetics. Further, the analyst must also be aware of the potential method-internal effects (e.g. LLR magnitude, the speaker discriminatory potential of the variable and calibration) of such variation on different elements of LR output.

Issues with logical relevance

Even with increased awareness of the complexity of between-speaker variation and the potential effects of different definitions of the relevant population, there remain a number of difficulties with the application of logical relevance to FVC. Firstly, the results of Chapters 4 to 7 suggest that, where possible, a narrow definition of the relevant population in terms of regional background, class and age is preferable for LR testing. However, as outlined in §2.3.1.1, the clear paradox in FVC casework is that since the identity of the offender is unknown it is not possible to know, for certain, the population of which he is a member. It may be possible for the sociolinguistically aware analyst to make judgements about regional background, sex, age etc. based on the offender sample, but the analyst cannot know whether such judgements are correct. Secondly, the experiments in this thesis have considered the effects of a single sociolinguistic factor on the outcome of LRs from a single variable. However, as outlined in §2.2.5, there are numerous potentially relevant factors which may need to be controlled. It may be possible to account for the complexity of the competing forces of sociolinguistic variation and their potential effects on the OLR using a small number of variables.

However, in a fully componential auditory-acoustic phonetic FVC analysis (§1.1.3) it is likely to be much more difficult to predict the effects of different definitions of the relevant populations on the OLR.

Thirdly, while the results in these experiments have focused on issues with using general (Mixed) or inappropriate (Mismatched) data in LR testing, the use of a narrowly defined relevant population may also be problematic. This is because there is an interaction between the definition of the relevant population according to grouping factors such as regional background, class and age, and the prior odds. Rose (2013b) highlights that in a given legal case the defence team may propose a narrowly defined alternative proposition which "may well increase the LR denominator and thus decrease the LR" (p. 285), thus reducing the evidential support for the prosecution. However, Rose (2013b) also states that such a strategy "will also change the prior odds in an unwanted direction" (p. 285), which in turn will affect the posterior probability of the propositions given the evidence.

For example, the population of Britain in July 2014 was estimated at 63,742,977, of whom approximately 23,969,294 are males over the age of 20.[22] Therefore, for the sake of exposition, the approximate prior odds based on the definition the relevant population as BrEng speaking males (following Rose 2004) is 1 in 23,969,294 (although this, of course, does not constitute anything like a linguistically homogeneous group due to complexities of defining *language*, outlined in §2.3.1.1) Assuming that by using this general definition of the relevant population an overinflated (by one order of magnitude) SS LR of 100 is generated, consistent with the patterns in §4.3.1. The posterior probability is therefore:

$$\frac{\left(\frac{1}{N}\right)}{\left(\frac{N-1}{N}\right)} \times LR = \frac{\left(\frac{1}{23{,}969{,}294}\right)}{\left(\frac{23{,}969{,}294-1}{23{,}969{,}294}\right)} \times 100 = 4.17 \times 10^{-6} \tag{11.1}$$

Now, assume that the relevant population is defined more narrowly as young (15-24 years) males from Manchester and the prior odds become 1 in 145,487.[23] Assuming a

[22] `http://www.indexmundi.com/united_kingdom/demographics_profile.html` (accessed: 30th September 2014).

[23] `http://www.indexmundi.com/united_kingdom/demographics_profile.html` (accessed: 30th September 2014).

more conservative LR of 10 based on Matched data, the posterior probability is:

$$\frac{\left(\frac{1}{N}\right)}{\left(\frac{N-1}{N}\right)} \times LR = \frac{\left(\frac{1}{145{,}487}\right)}{\left(\frac{145{,}487-1}{145{,}487}\right)} \times 100 = 6.87 \times 10^{-5} \tag{11.2}$$

In this example, the posterior probability based on a specific defence proposition (Equation 11.2) is approximately 16 times stronger in favour of the prosecution than the posterior probability based on a general defence proposition (Equation 11.1). Thus, all else being equal, the evidence is likely to contribute more towards a guilty verdict, despite the defence apparently using a specific alternative proposition to reduce the evidential support for the prosecution. If, as in Chapter 5, there are essentially no differences in the calibrated SS LR using a general or specific relevant population, then the differences between the posterior probabilities will be even greater. Assuming the same priors as above and having generated a LR of 10 for both definitions of the relevant population, the posterior probability using the specific defence proposition would be approximately 165 times stronger than that using the general defence proposition. It important for the analyst to be aware of the implications of the definition of the relevant population on posterior probability. However, Rose (2013b) highlights that such issues are unlikely to be considered by the trier-of-fact.

Clearly, the application of logical relevance to speech remains problematic. The only alternative thus far proposed is lay listener-judged speaker similarity (Morrison *et al.* 2012). However, as outlined in detail in §2.3.2.1, there are several reasons to reject this proposal. The underlying logic is limited since the initial judgement to submit recordings for expert analysis is likely made based on other evidence in the case. The replicability of the data it generates, and the transparency of the decisions made by the lay listeners are questionable, due to the way in which lay listeners are expected to judge similarity. Further, this approach shifts the decision making away from the expert and onto the untrained lay listener. Given the limitations of logical relevance and lay listener-judged speaker similarity, the following sections present three alternative approaches for defining the relevant population. These alternatives emulate current practices in forensic DNA analysis and ASR.

11.1.1 Emulating DNA

11.1.1.1 Multiple defence propositions

Logical relevance may be applied to forensic DNA analysis by presenting multiple LRs based on different assumptions about the offender (Kaye 2004). Since it is not possible to infer offender race from a sample of questioned DNA, in the UK the suspect and offender samples may be compared relative to each of the three available racially defined UK databases to generate three LRs (Gill and Clayton 2009: 34; see §2.3.1). Applied to speech, this approach would involve offering multiple OLRs for different definitions of the relevant population based on multiple individual LRs from each of the variables analysed. The use of multiple defence propositions resolves the paradox described above, since it captures the inherent uncertainty involved in judging the demographic background of the offender. Further, it allows the analyst to use specific alternative propositions which have been shown to generally outperform Mixed data in this thesis.

However, the practicality of implementing this approach is considerably more difficult for speech than for DNA analysis. As explained in §2.4.2, the number of available corpora suitable for use in FVC casework is extremely small. Given that there are so many potentially relevant factors which may be controlled in a multiple H_d approach, it would likely not be possible to compute numerical LRs for certain defence propositions using existing data. In this case, it may be possible to collect case-by-case data (§2.4.1); however given the time and cost expected for computing a single LR based on a single variable, the inefficiency of this approach will mean that it is not practically viable. An alternative, of course, is to estimate typicality based on different definitions of the relevant population using experience and published research. However, as with the use of a single logically relevant defence proposition, multiple propositions account only for the factors which are controlled by the analyst. That is, any uncontrolled sources of mismatch between the evidential and system data may still affect LR output.

11.1.1.2 Correction factor (F)

As outlined in §2.3.1, in forensic DNA analysis a number of studies have investigated genetic differences in sub-populations (such as geographical locations) within racial groups (Foreman and Evett 2001; Gill and Evett 1995). The primary focus has been on genetic correlations (Balding *et al.* 1996) due to potential shared ancestry (coancestry) between the suspect and other potential offenders, since the use of a general database may produce an overoptimistic LR for populations which are highly inbred. Due to the fact that levels of shared ancestry have generally been found to be low in cosmopolitan populations (Gill and Clayton 2009: 34), genetic correlations are often ignored in DNA analyses. However, such variation can be accounted for by including a coancestry coefficient (F_{ST}) into the LR computation (Balding and Nichols 1994; see also Balding and Nichols 1995 and Balding and Donnelly 1995 for other correction factors).

It may be possible to apply such a correction factor (F) to variation in sub-populations in speech. F is a value between zero and one which reflects whether there is less variation for a given variable (allele frequencies in the case of DNA analysis) in sub-populations than in the population at large. F increases towards one as the diversity within the sub-population for the given variable decreases (Barnshad *et al.* 2004). Following the DNA approach, it would be necessary to have large established databases controlled for gross between-speaker differences (e.g. regional background and sex), from which LRs could be computed with a value for F applied. In the case of speech, an overall value for F would need to comprise multiple correction factors to account for the multitude of potential sources of between-speaker variation considered relevant in the database (e.g. class, ethnicity, age). The benefit of such an approach is that general databases could be used, while also accounting for the sociolinguistic dimensions of variation which have been shown to affect LR estimates.

The first limitation of this approach is the availability of a large general database containing sufficient, representative variation in sub-populations to provide meaningful values for F. Further, in the collection of such a database *a priori* decisions regarding the most important sources of between-speaker variation would need to be defined. Perhaps most significantly, the complexity of variation in speech (§2.2.5) means that it would be extremely difficult to calculate a value for F, even if a suitable database were

available. This is because, where F_{ST} accounts for a single source of variation in allele frequency, an overall value for F in FVC would need to incorporate potentially multiple sources of between-speaker variation, accounting for the necessary interrelatedness of these factors. The multidimensionality of between-speaker variation would therefore introduce considerable statistical uncertainty into the calculation of F, which may generate meaningless LRs. Further, unlike in DNA analysis, any overall correction factor in FVC would need to account for within-speaker variability as well as between-speaker variation.

11.1.2 Emulating ASR

Although the alternative approaches suggested in §11.1.1.1 and §11.1.1.2 resolve some of the limitations of logical relevance, their limitations mean that it would be difficult to implement them practically in FVC. Given that these approaches were developed for DNA analysis, their limitations derive primarily from the attempt to apply them to more complex patterns of variation in speech. Therefore, it is considered preferable to develop a default definition of the relevant population which specifically accounts for the complexity of speech evidence.

11.1.2.1 Expert-judged speaker similarity

This section proposes such an alternative which derives from current approaches in ASR systems. As in Morrison *et al.* (2012), it is considered logically appropriate that the relevant population should consist of speakers who are similar sounding to the offender. This is also proposed by Rose (2002: 57-58) who states that the defence proposition can generally be assumed to be that the suspect and offender samples contain the voices of different but similar sounding speakers. Inherently, any judgement about the similarity of two speakers is a subjective one. However, it is considered preferable that such judgements are made by the expert where decisions are more defensible and explainable to the court, rather than by lay listeners. Following this approach, and in the absence of a more specific alternative, the defence proposition may be expressed as:

> The voice on the offender sample does not belong to the suspect, but to another similar sounding speaker as judged by an expert (or more specifically, the voice on the offender sample does not belong to the suspect, but to another sufficiently similar sounding speaker such that an expert would consider the recordings for FVC analysis).

There are a number of logical and practical reasons why the decision relating to similarity should be made by the expert. Firstly, this approach captures the fact that in most cases the decision to proceed with expert analysis is made by the expert themselves based on similarity rather than by the police officer potentially on the basis of other evidence from the case. Secondly, the decisions about similarity will be more linguistically principled, meaning that the set of resulting data should be more homogeneous with regard to sociolinguistic factors than that based on lay listener judgements. Therefore, this approach can, to some extent, capture the relevant sources of between-speaker variation without the expert having to be explicit about the offender's regional background, age, class etc. as in logical relevance.

It is, of course, possible to judge speakers as sounding similar without them being in any sense from the same demographic backgrounds. For example, in cases where the offender's sex is unclear (French *et al.* 2010: 145; Foulkes and French 2012: 569), the data could consist of both males and females, since there are linguistic reasons for potentially judging males and females as sounding similar to the offender (e.g. f0, VQ). However, unlike lay listener-judged speaker similarity, this approach would still mean that the relevant population consists of "those persons who could have been involved (in the crime)" (Coleman and Walls 1974: 276). Thirdly, this approach is arguably more replicable and the decisions more transparent for the trier-of-fact than lay listener-judged speaker similarity.

There are a number of ways in which expert-judged speaker similarity could be assessed. Two potential alternatives are tentatively proposed here, although considerably more research would be needed to investigate the validity and reliability of such procedures. In both cases a large, general corpus of forensically realistic recordings from a wide range of speakers would be needed. One approach would be for the expert to assign similarity scores to each pairwise comparison of each speaker in a corpus and the

offender. Such a score may be based on auditory analysis, but could also include acoustic analysis if necessary. The reference data would then consist of the N speakers from the database judged most similar sounding to the offender. Although preferable in one sense, this method also has a number of limitations. Firstly, the replicability of the data would be dependent on high inter-analyst agreement. Secondly, as with the Morrison *et al.* (2012) approach, potentially problematic *cliff-edge* distinctions between the inclusion/exclusion of speakers would need to be made. Finally, it would be likely be prohibitively expensive and time-consuming for an expert to conduct what are essentially multiple voice comparisons to generate a set of data for LR testing data in each FVC case.

A second, more viable, approach would be to use continuous acoustic measures to calculate distances within the multidimensional acoustic space between the offender and all other speakers in the database. The resulting data would then consist of the N speakers closest to the offender in the acoustic space. The most efficient way to do this would be to use CCs extracted holistically from across each sample (equivalent to the approach used in the empirical demonstration of lay listener-judged similarity in Morrison *et al.* 2012), although it would also be possible to use continuous acoustic-phonetic variables. This approach is equivalent to the in-built reference population selection algorithm in commercially available ASR software such as BATVOX. BATVOX chooses speakers for testing based on Kullback-Leibler distances between the speakers in the database and the suspect using holistic MFCC analysis.

The primary difference between BATVOX and the approach outlined above is that expert-judged speaker similarity is based on distances between the speakers in the database and the offender, since the defence proposition should be defined by the offender rather than the suspect. Of course, the extent to which this approach is in any way objective and replicable is again dependent on the decisions made by the analyst in terms of how the acoustic data are extracted (see Harrison 2013 for issues with formant measurements) and then modelled statistically. However, the approach does allow for data based on a principled assumption about the relevant population to be assembled relatively quickly and efficiently, improving the viability of the numerical LR approach for FVC.

11.2 Collecting development, test and reference data

The results of Chapters 8, 9 and 10 have revealed a number of effects of sample size on LR output. Consistent with the results of Ishihara and Kinoshita (2008) and Rose (2012), LR output was found to be imprecise and misrepresentative with small numbers of reference speakers. Specifically, the smallest (ten speakers) numbers of reference speakers generated the widest ranges of SS and DS LRs, and validity was generally considerably worse compared with the *true* LRs computed using the maximum number of available reference speakers. However, the point at which stable LRs and system validity were achieved differed across input variables.

Using univariate AR data in Chapter 9, calibrated LLRs appeared relatively stable to variation in reference sample size, even with as few as ten reference speakers. Similarly, aside from minor fluctuations, EER and C_{llr} were generally robust to reference sample size using AR. In Chapter 10, LLRs and system validity based on mid-point F1∼F3 values for UM were somewhat more sensitive to sample size, with imprecision across replications stabilising with the inclusion of between around 20 reference speakers. The comparison of formant trajectory input for /uː/ and /aɪ/ in Chapter 8 revealed greater sensitivity to sample size. The distributions of scores were found to stabilise for /uː/ only with more than 30 reference speakers (as in Ishihara and Kinoshita 2008), although there was a linear trend for an improvement in C_{llr} as sample size increased up to the maximum number of speakers (120). For /aɪ/, stable LR output was only achieved with more than around 50 speakers (C_{llr} was stable with more than 42 speakers).

These results highlight a number of important issues for LR-based FVC. First, there is considerable evidence to support the predicted relationship between sample size sensitivity and the dimensionality of the input variable (Rose 2013a; outlined in §8.1). The most stable LR output was achieved using univariate AR data (one dimension), while the least stable LR output was generated using the multivariate /aɪ/ trajectory data (12 dimensions). As outlined in Rose (2013a), this is because more data are required to precisely model high dimensional density functions. Secondly, there is some evidence that calibration may reduce the sensitivity of LR output to variation in sample size. In §9.3.1, sample size effects present in the uncalibrated scores were absent from the resulting calibrated LLRs. In §10.3.3, although calibration did not affect the point at

which stable LR output was achieved (around 30 speakers), it did reduce the magnitude of the effects of sample size when the number of speakers was small.

Thirdly, the inherent speaker discriminatory power of the variable appears to play a role in sample size sensitivity. For variables which produce low magnitude LLRs, there is a narrower range of potential variation. This may account for the stability in the distributions of LLRs for AR, since inherently AR offers very limited speaker discriminatory power (optimally C_{llr} approaching 1 and EER of around 35%). As outlined in §9.4, speaker discriminatory power may also explain the different effects of calibration in reducing the sensitivity of LR output to sample size for AR and UM. Finally, sample size affects LLRs from individual comparisons in different ways. The results in this thesis suggest that the effects on individual LLRs are, to some extent, dependent on magnitude. For comparisons which generate high magnitude LLRs, there is considerably greater variability as a function of sample size. This is highlighted by the variation in values generated by the comparisons which produce the most outlying LRs across Chapters 8, 9 and 10. This is because changes to the reference distribution as speakers are added have a much more substantial effect on the LRs for offender values on the tails of the reference distribution. This also explains why DS LLRs are generally more sensitive to sample size than SS LLRs.

The results of Chapter 10 also have important implications for FVC. The use of multiple replications highlights that even with an extremely large number of speakers there is potentially a large amount of variability across systems with datasets of the same size but containing different speakers. Such inherent uncertainty in the absolute value of the LR should be acknowledged and explained to the court in LR-based FVC. Chapter 10 also revealed trade-offs between the number of development, test and reference speakers according to different elements of LR output. Calibration coefficients were almost entirely dependent on the size of the development set, and with a sufficiently large development set (15+ speakers) were stable even when using very small amounts of reference data (e.g. ten speakers). Validity metrics (EER and C_{llr}) were, predictably, dependent on the number of test speakers, displaying a considerably wider range of variability and overly optimistic performance with the smallest number of test speakers (two).

Finally, Chapters 8 and 9 presented the first investigation into the effects of the number of tokens per reference speaker on LR output. As with the number of speakers, LR output was severely compromised when using very small numbers of tokens (two). Different patterns were again found for the different input variables. For /uː/ and /aɪ/, there is some evidence of stable SS scores and validity metrics with more than six tokens. However, for /aɪ/ no stability in the distribution of DS scores was found. For AR, LR output was consistently stable even when using the smallest amount of data per speaker (two tokens). As with the number of speakers, these patterns may be explained by the dimensionality of the input variables. Similarly, there is also evidence that the speaker discriminatory power of the variable and the magnitude of individual LLRs again contribute towards sample size sensitivity.

11.3 Practical implications

The results of the experiments in this thesis have a number of practical implications for FVC casework. Given the range of sources of systematic between-speaker variation in speech, across all forms of FVC analysis it is important to consider the appropriate population against which to assess typicality (as suggested by Morrison *et al.* 2012; Morrison and Stoel 2014). For LR-based FVC using logical relevance, this is all the more important given the potential effects (due to both between-speaker variation and method-internal factors) of different definitions of the relevant population on numerical LR output. On the basis of the findings in this thesis, a narrowly defined relevant population should be preferred where there is no dispute over given sociolinguistic factors (which may be determined by the court). In cases where elements of the offender's demographic background are uncertain, the analyst should prefer a general definition of the relevant population. However, from a practical perspective expert-judged speaker similarity (§11.1.2.1) may be a more viable means of accounting for the complexity of between-speaker variation and generating robust LRs than logical relevance.

The results of sample size testing suggest that, predictably, more data in the development, test and reference sets is better for LR testing. For all of the variables tested, stable LR

output (in terms of the magnitudes of LRs and system validity) was achieved using 50 or more reference speakers, although it may be possible to use far fewer if the variable has a relatively small number of dimensions (fewer than nine). For variables with more than 12 dimensions it may be necessary to use far more than 50 reference speakers. However, the trade-offs explored in Chapter 10 suggest that for a relatively low dimensional variable (three dimensions), if the number of available speakers is small preference should be given to the size of the development and test sets. In Chapter 10, stable calibrated LLR output was achieved using minimally 20-30 development speakers, 20-30 test speakers and as few as 15 reference speakers, although this depended on which element of LR output was analysed.

Given that sample size sensitivity has been shown to differ for individual variables, it may be necessary in casework to perform pre-testing to establish the point at which LR output becomes stable and assess the overall degree of precision in LR estimates as a function of sample size (similar to that in Rose 2012). In the absence of suitably large available databases, MCS have provided a valuable resource for investigating these issues in this thesis and could be used as part of system pre-testing (as in Rose 2013b). In practical terms, MCS are easy to implement and can be used to generate a large amount of data quickly and efficiently. Crucially, however, MCS are dependent on the assumption that the underlying distribution of within- and between-speaker variation in the relevant population is known, either through previous research or raw data. Further, complexity and uncertainty is introduced into the MCS procedures when simulating multiple correlations between elements of a variable. Therefore, caution is advised when implementing MCS procedures using already small sets of raw data or highly multivariate variables.

Across all experiments, differences have been found between uncalibrated scores and calibrated LLRs, with results suggesting that calibration is able to help reduce some of the effects of inappropriate or general definitions of the relevant population and some of the effects of small samples in LR testing. Therefore, caution is advised when interpreting the strength of evidence or system performance based on uncalibrated scores, especially where the amount of data for testing is small or the definition of the relevant population is general. The role of calibration in determining sample size

sensitivity has not been reported previously in terms of the number of speakers used in LR testing, but is suggested in Ishihara and Kinoshita's (2012) study of the number of tokens per test speaker (see §2.5 for an overview).

Finally, given the inherent uncertainty in LR estimates due to variation in the definition of the relevant population and due to sample size, it may be preferable to express the strength of evidence as a range rather as a single LR value. This is currently done using the 95% CI to account for the imprecision in LR estimates across multiple non-contemporaneous samples. However, this should be expanded to incorporate the imprecision across the many subjective decisions made by the analyst in building and testing a FVC system. Alternatively, the numerical LR may be expressed in the form of a verbal equivalent. In Chapter 6, despite the large differences in the magnitude of individual DS LLRs across systems based on cepstral input, following Champod and Evett's (2000; see Table 3.3) scale, almost all values would be classified verbally as offering *very strong* support for the defence. Similarly, for the DS results based on /uː/ in Chapter 8, much of the substantial sample size variation would be normalised by classifying all values of less than -4 as *very strong* support. The use of verbal scales may also, to some extent, resolve the courts concerns about the interpretability of LR-based evidence raised in *R v Doheny and Adams* [1996] (see further §2.1.4.1). Although the court in *R v T* [2010] expressed reservations over the use of verbal equivalents, such scales have subsequently been received by the courts in *R v South* [2011].

However, the verbal LR is not capable of resolving all of the issues relating to the relevant population and sample size. Firstly, the introduction of *cliff-edge* effects could result in small numerical differences between systems being exaggerated (e.g. 1.99 vs. 2.01 = *moderate* vs. *moderately strong* support). This is particularly problematic for SS LLRs since they are generally situated within the region of verbal differences (i.e. between zero and +4). Secondly, and perhaps more importantly, verbal LRs do not resolve differences in system validity. In light of these limitations, it would appear best to present the results of an LR-based analysis in terms of both a numerical interval which accounts for the imprecision across analyst decisions, an estimate of the validity of the system within a given interval which also accounts for the subjectivity in analyst decisions, and a verbal expression of the strength of evidence (with obvious caveats

regarding potential *cliff-edge* effects). Indeed, Robertson and Vignaux (1995b: 57) argue that a combined numerical and verbal LR is the best way to ensure that the probabilistic detail of the analysis is maintained, but that the strength of evidence is also interpreted correctly.

11.4 Future work

The findings and implications of the experiments in this thesis offer considerable scope for future research into the definition of the relevant population and the collection of data for LR testing. It would be useful to replicate the experiments in Chapters 4 to 7 using non-contemporaneous, forensically realistic data. In particular, given the limitations of TIMIT in Chapter 6, more research is warranted to explore the social stratification of ASR variables using more forensically realistic speech samples. It would also be interesting to evaluate the sensitivity of cepstral input to regional variation for varieties with known differences in vocal settings. The sensitivity of semi-automatic variables such as LTFDs to systematic sources of between-speaker variation is also a potentially interesting avenue for future research. Moreover, building on Harrison and French (2012), it is necessary to test the sensitivity of commercially available ASR software to different definitions of the relevant population. Thus, it may be possible to apply sociolinguistic knowledge to further improve ASR systems.

Considerably more work is required to empirically test different approaches to defining the relevant population. In particular, more research is needed to develop different versions of lay listener- and expert-judged speaker similarity and evaluate what similarity means in different contexts. Specifically, future research should consider the level of agreement between lay listeners, experts and fully numerical acoustic measures in determining how similar pairs of speakers are. These results will contribute towards evaluating lay listener- (Morrison *et al.* 2012) and expert-judged (§11.1.2.1) speaker similarity approaches in LR-based FVC. The results of such studies will also contribute towards better understanding the linguistic information to which lay listeners are sensitive in making judgements of similarity (see McDougall 2013; Nolan *et al.* 2013).

Beyond between-speaker variation, further research is required to investigate how the numerous sources of systematic within-speaker variability (see §2.2.5) affect LR output. It would also be interesting to compare the direction and magnitude of the effects on LRs of such sources of within-speaker variability relative to the effects of using contemporaneous and non-contemporaneous samples (Enzinger and Morrison 2012). Through such work, it will be possible to assess which are the most important sources of forensically relevant variability to control in FVC research and casework. Further, despite the mass of literature in sociolinguistics and sociophonetics into the effects of stylistic factors on linguistic-phonetic variables, very little work has considered the potential extent of within-speaker variability in ASR variables such as CCs.

Although the issue of the number of speakers in the development, test and reference data has been investigated extensively in Chapters 8 to 10, there a number of avenues for future research with regard to sample size. As above, it would be preferable to replicate the experiments in this thesis using non-contemporaneous, more forensically realistic samples. Further, given the differences in the results across individual variables, it would be useful to replicate these studies using highly multivariate ASR-type variables. Future work should also build on the relatively small-scale studies in this thesis and on the results of Ishihara and Kinoshita (2012) to investigate more systematically the extent to which LR output is compromised by the amount of data per development, test and reference speaker.

This thesis has also highlighted the inadequacies of currently available corpora for LR-based FVC. Across the experiments, a range of corpora collected for different purposes have been used, all of which reflect some degree of compromise in order to address the research questions. Therefore, it is considered essential that a large set of forensically realistic data is collected for use in FVC research and casework, which incorporates the best elements of currently available sociolinguistic, ASR and forensic corpora. As in sociolinguistic research, such a corpus would need to contain a diverse set of speakers controlled for a number of social dimensions and with multiple recordings per speaker reflecting different stylistic conditions. As in ASR research, such a corpus would need to be extremely large with multiple mismatched samples with regard to technical effects (e.g. telephone transmission). Finally, as in FVC research, such a corpus would need

to include samples of speakers involved in forensically realistic tasks such as those in DyViS (§3.1.1). Potential procedures for collecting such a database are outlined in Morrison, Rose and Zhang (2012).

Beyond the specific issues addressed in this thesis, there remain a number of barriers to the widespread application of the fully numerical LR approach in FVC casework. On a micro level, as highlighted in Gold and Hughes (2014), Gold (2014) and §2.2.5, there remain a number of method-internal issues to resolve. In particular, it is important that the field of FSS, in collaboration with forensic statisticians, develops models to account for all of the variables that may be analysed in a given case (see Aitken and Gold 2013; Foulkes *et al.* 2013-2015). LR formulae are needed to compute numerical LRs for the many discrete and binary variables analysed in FVC, in particular for VQ given that experts consider this to be one of the most speaker discriminatory variables in FVC (Gold and French 2011). Further work is also required to appropriately account for the complex correlation structure of speech evidence when combining LRs from individual variables into an overall estimate of the strength of evidence (Gold and Hughes 2013-2014).

On a macro level, the interpretability of a fully numerical LR-based analysis and the resulting expression of the strength of the evidence by the trier-of-fact is problematic. Relatively little work (with the exception of Cudmore 2011; Martire *et al.* 2014) has considered how well (or rather how badly) lay people understand expert evidence expressed in the form of a LR. However, as outlined in *R v Turner* [1975] the role of the expert is to "furnish the court with ... information which is likely to be outside the experience and knowledge of a judge or jury." An essential part of this role is to ensure that the expression of the strength of evidence is correctly interpreted. This is highlighted by the FRE 702 requirement (reaffirmed in Daubert) that expert evidence "will assist the trier-of-fact to understand the evidence." However, as highlighted by Berger (2010), "a logically incorrect conclusion that is 'understood' is no alternative to a logically correct conclusion which needs explanation" (p. 784). Therefore, considerably more work is needed to ensure that the expert's conclusion is based on logically and legally correct reasoning and that such reasoning is well understood by the trier-of-fact. As argued by Ledward (2004) and Fenton and Neil (2012), such progress is dependent

on interdisciplinary collaboration between statisticians, forensic scientists, lawyers and the courts.

11.5 Conclusion

The overarching aim of this thesis has been to consider the dimensions of potential variability in analyst decisions involved in data-driven LR-based FVC. Specifically, the issues of the definition of the relevant population and the collection of data for system testing have been considered in view of the complexity and multidimensionality of speech as a form of evidence. The results have shown that LR output is heavily dependent on which sources of systematic between-speaker variation the analyst controls when defining the relevant population. More specific definitions based on a range of sociolinguistic factors (regional background, class and age) have largely been shown to produce more valid systems than the use of a general definition such as Rose (2004). The use of small samples has been shown to compromise the validity and reliability of LRs produced by a given system. The level of sensitivity to sample size is determined largely by the dimensionality and speaker-specificity of the variable under analysis. Further, in cases where the amount of available data is small, there may be trade-offs in the size of the development, test and reference sets to generate reliable LR output.

It is hoped that the results of this thesis will contribute towards to improving the viability of the numerical LR approach for FVC casework and help to improve the extent to which LR methods account for the linguistic-phonetic complexity of speech. By accounting for this complexity in future work to address other issues with using the LR for analysing speech, the quality of FVC evidence will necessarily improve and the acceptance of the LR as a practical tool will also hopefully increase within the field of FSS.

Appendix

Generating a synthetic speaker distribution: an example using univariate data

This section provides a worked example of how synthetic means and SDs were generated using MCS based on a set of raw univariate AR data from 59 speakers. The means and SDs were used in Chapter 9 to function as Gaussian distributions for each synthetic speaker, from which individual tokens of local AR were extracted.

The distribution of mean local AR values (x) from the raw data is firstly converted to a normal PDF through a process of normalisation to the z-space using the mean of the means ($\mu_x = 6.044$) and the SD of the means ($\sigma_x = 0.627$):

$$z = \left(\frac{x - \mu_x}{\sqrt{2}\sigma_x}\right) = \left(\frac{x - 6.044}{\sqrt{2} \times 0.627}\right).$$

A random area value for $f(z) = Z_i = 0.4478$ is output by the *rand* function in Matlab, which generates a pseudo-random value between 0 and 1 from a normal distribution $N(0.5, 0.341)$. The inverse CDF of Z_i is therefore equal to z_i (Equation 9.7 such that:

$$z_i = CDF^{-1}(0.4478) = -0.1312.$$

z_i is then transformed to the linguistically meaningful x-space by:

$$x_i = (\sqrt{2}\sigma_x \times z_i) + \mu_x = (0.8867 \times -0.1312) + 6.044 = 5.9277.$$

The synthetic mean local AR value for this speaker is therefore 5.9277.

The synthetic SD (y_i) can then be generated from $N(ax_i + b, \beta_i)$ based on the linear correlation, and variability around the trend line, between the means and SDs in the

raw data where $a = 0.1262$, $b = 0.3359$ and $x_i = 5.9277$ is the synthetic mean that has just been created. In this case, $ax_i + b = (0.1262 \times 5.9277) + 0.3359 = 1.084$. The β_i element (based on $N = 59$ speakers) is calculated as:

$$\beta_i = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{59} \times 1.9466} = \sqrt{0.033} = 0.1816,$$

such that the SD (y_i) associated with the mean (x_i) of 5.9277 is sampled from a distribution $N(1.084, 0.1816)$. This distribution is converted to a PDF where:

$$z = \left(\frac{x - 1.084}{\sqrt{2} \times 0.1816}\right).$$

A pseudo-random $f(z) = Z_i = 0.3733$ is generated using *rand* and again z_i is calculated by $CDF^{-1}(0.3733) = -0.3231$. This is transformed back to the linguistically meaningful y-space such that:

$$y_i = (\sqrt{2}\sigma_x \times z_i) + \mu_x = (0.2884 \times -0.3231) + 1.084 = 1.001.$$

The normal distribution of local AR values for this new synthetic speaker has a mean of 5.9277 syllables per second and a SD of 1.001 phonological syllables per second. Individual tokens of local AR are then sampled from this Gaussian distribution following the same procedure as outlined above for generating synthetic means.

List of Abbreviations

$\vert$	*given*
A	Delta-delta (accelaration) coefficient
AmEng	(North) American English
AR	Articulation rate
ASR	Automatic speaker recognition
AusEng	Australian English
BrEng	British English
CanCor	Canterbury Corpus
CC	Cepstral coefficient
CDF	Cumulative distribution function
C_{llr}	Log likelihood ratio cost function
CI	Credible interval
D	Delta (differential) coefficient
DA	Discriminant analysis
DCT	Discrete cosine transformation
DFT	Discrete Fourier transform
DR	Dialect region
DS	Different speaker
DyViS	Dynamic Variability in Speech (database)
E	Evidence
EER	Equal error rate
erf	Error function
f0	Fundamental frequency

F(1-3)	Formant (1^{st}-3^{rd})
FSS	Forensic Speech Science
F	Correction factor
F_{ST}	Coancestry coefficient
FVC	Forensic voice comparison
GMM-UBM	Gaussian mixture model - universal background model
H_d	Defence proposition/hypothesis
H_p	Prosecution proposition/hypothesis
HTK	Hidden Markov Model Toolkit
Hz	Hertz (frequency)
KD	Kernel density
kHz	Kilohertz (frequency)
LLR	Log likelihood ratio (base 10 unless otherwise stated)
LPC	Linear prediction cepstrum
LPCC	Linear prediction cepstral coefficient
LR	Likelihood ratio
MCS	Monte Carlo simulations
MFC	Mel frequency cepstrum
MFCC	Mel frequency cepstral coefficient
MVKD	Multivariate kernel density
$N(\mu, \sigma)$	Normal distribution with mean μ and standard deviation σ
NE	Northern Englishes (corpus)
NIST	National Institute of Standards and Technology
NZE	New Zealand English
OLR	Overall likelihood ratio
ONZE	Origins of New Zealand English (corpus)
p	Probability
PDF	Probability density function
PVC	Phonological Variation and Change (corpus)
RMS	Root mean square
RoI	Roots of Identity (corpus)
SD	Standard deviation

SS	Same speaker
SSBE	Standard Southern British English
TIMIT	Texas Instruments (TI), Massachusetts Institute of Technology (MIT) (database)
UKPS	UK Position Statement
VQ	Voice quality

Legal Cases

Daubert v Merrell Dow Pharmaceuticals [1993] 509 US 579.

Frye v United States [1923] 293 F. 1013 D.C. Cir.

George v R [2007] EWCA Crim 2722.

R v Robb [1991] 93 Cr App R 161.

R v Adams [1996] 2 Cr App R 467.

R v Deen [1993] (EWCA (Criminal Division)) (reported: Times, January 10, 1994).

R v Doheny and Adams [1996] EWCA Crim 728.

R v O'Doherty [2002] NICB 3173.

R v Flynn [2008] EWCA Crim 970.

R v Sally Clark [2003] EWCA Crim 1020.

R v South [2011] EWCA Crim 754.

R v T [2010] EWCA Crim 2439.

R v Turner [1975] QB 834.

United States v Robert N Angleton [2003] 269 F Supp 2nd 892 S D TX.

Bibliography

Adams, R. N. (1904). *How to Pronounce Accurately on Scientific Principles*. Dunedin: Otago Daily Times.

Aitken, C. G. G. (1995). *Statistics and the Evaluation of Evidence for Forensic Scientists*. Chichester: Wiley.

Aitken, C. G. G. and E. Gold (2013). Evidence evaluation for discrete data. *Forensic Science International* 230 (1-3), 147–155.

Aitken, C. G. G. and D. Lucy (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics* 54, 109–122.

Aitken, C. G. G. and D. A. Stoney (1991). *The Use of Statistics in Forensic Science*. London: Ellis Horwood.

Aitken, C. G. G. and F. Taroni (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists (2nd edition)*. Chichester: Wiley.

Aitken, C. G. G. et al. (2011). Expressing evaluative opinion: a position statement. *Science and Justice* 51, 1–2.

Albert, J. (2009). *Bayesian Computation with R (2nd edition)*. Heidelberg: Springer.

Alderman, T. (2004a). The Bernard data set as a reference distribution for Bayesian likelihood ratio-based forensic speaker identification using formants. In *Proceedings of the 10th Australasian International Conference on Speech Science and Technology*. Macquarie University, Australia, pp. 510–515.

— (2004b). The use of Australian-English vowel formant data sets in forensic speaker identification. In *Proceedings of the 10th Australasian International Conference on Speech Science and Technology*. Macquarie University, Australia, pp. 177–182.

Alexander, A. (2005). *Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions*. Unpublished PhD Thesis, Swiss Federal Institute of Technology, Lausanne.

Alexander, A., F. Botti, and A. Drygajlo (2004). Handling mismatch in corpus-based forensic speaker recognition. In *Proceedings of Forensic Speaker Recognition Workshop, Speaker Odyssey '04*. Toledo, Spain, pp. 69–74.

Alexander, A. and A. Drygajlo (2004a). *Automatic speaker recognition: a simple demonstration using Matlab*. `http://www.anilalexander.org/publications/Speaker_Recognition_Intro_A-A.pdf`. Biometrics Course, Communication Systems, EPFL. Accessed: 13th March 2014.

— (2004b). Scoring and direct methods for the interpretation of evidence in forensic speaker recognition. In *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP), Jeju, Korea*, pp. 510–515. URL: `http://www.anilalexander.org/publications/AlexanderDrygajlo_ICSLP2004.pdf`.

Ash, S. (1996). Freedom of movement: /uw/-fronting in the Midwest. In *Sociolinguistic Variation: Data, Theory and Analysis – Selected Papers from NWAV 23 at Stanford*. (Ed.) J. Arnold et al. Stanford, CA: Centre for the Study of Language and Information (CSLI) Publications, Stanford University, pp. 2–23.

Atkinson, N. (2009). *Formant dynamics of SSBE monophthongs in unscripted speech*. Unpublished MSc Dissertation, University of York.

— (in progress). *Variable factors affecting voice identification in forensic contexts*. Unpublished PhD Thesis, University of York.

Balding, D. J. (2005). *Weight-of-Evidence for Forensic DNA Profiles. Statistics in Practice*. Chichester: John Wiley.

Balding, D. J. and P. Donnelly (1995). Inference in forensic identification (with discussion). *Journal of the Royal Statistical Society* 158, 21–53.

Balding, D. J., M. Greenhalgh, and R. A. Nichols (1996). Population genetics of STR in Caucasians. *International Journal of Legal Medicine* 108, 300–305.

Balding, D. J. and R. A. Nichols (1994). DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International* 64 (2-3), 125–140.

Balding, D. J. and R. A. Nichols (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 96, 3–12.

Baldwin, J. R. and P. French (1990). *Forensic Phonetics*. London: Pinter.

Barnshad, M. et al. (2004). Deconstructing the relationship between genetics and race. *Nature* (5), 598–609.

Bayes, T. (1763). Essay towards solving a problem in the doctrine of chances. *Philosophical transcripts of the Royal Society of London* 53, 370–418.

Beal, J. (2004). The phonology of English dialects in the north of England. In *A Handbook of Varieties of English (vol. 1)*. (Ed.) B. Kortmann and E. W. Schneider. Berlin: Mouton, pp. 113–133.

Becker, T., M. Jessen, and C. Grigoras (2008). Forensic speaker verification using formant features and Gaussian mixture models. In *Proceedings for Interspeech Special Session: Forensic Speaker Recognition – Traditional and Automatic Approaches*. Brisbane, Australia, pp. 1505–1508.

— (2009). Speaker verification based on formants using Gaussian mixture models. In *Proceedings for NAG/DAGE International Conference on Acoustics*. Rotterdam, Netherlands.

Berger, C. E. H. et al. (2011). Evidence evaluation: a response to the court of appeal judement in R v T. *Science and Justice* 51 (2), 43–49.

Bernard, J. R. (1970). Toward the acoustic specification of Australian English. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 23, 113–128.

Bigelow, S. J. (1997). *Understanding Telephone Electronics (3rd edition)*. Oxford: Newnes.

Boersma, P. and D. Weenink (2011). *Praat: doing phonetics by computer [Computer Program] Version 5.2.32*. `http://www.praat.org`. Accessed: 22nd July 2011.

Bogert, B. P., M. J. R. Healy, and J. W. Tukey (1963). The quefrency analysis of time series for echoes: cepstrum, pseudo autocovariance, cross-cepstrum and saphe cracking. In *Proceedings of the Symposium on Time Series Analysis*, pp. 209–243.

Botti, F., A. Alexander, and A. Drygajlo (2004). On compensation of mismatched recording conditions in the Bayesian approach for forensic automatic speaker recognition. *Forensic Science International* 146 (1), 101–106.

Britain, D. (2013). Space, diffusion and mobility. In *Handbook of Language Variation and Change (2nd edition)*. (Ed.) J. Chambers and N. Schilling. Oxford: Wiley-Blackwell, pp. 471–500.

Broeders, A. P. A. (1995). The role of automatic speaker recognition techniques in forensic investigations. In *Proceedings of the 13th International Congress of Phonetic Sciences*. Stockholm, Sweden, pp. 154–161.

— (1999). Some observations on the use of probability scales in forensic identification. *Forensic Linguistics* 6 (2), 228–241.

— (2001). Forensic speech and audio analysis: forensic linguistics. 1998 to 2001: a review. In *Proceedings of 13th INTERPOL Forensic Science Symposium*. Lyon, France.

Brown, G. and J. Wormald (2014). Speaker profiling: An automatic method? In *Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference*. 31 August-3 September 2014. Zürich, Switzerland.

Brümmer, N. and J. du Preez (2006). Application-independent evaluation of speaker detection. *Computer Speech and Language* 20 (2-3), 230–275.

Brümmer, N. et al. (2007). Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST SRE 2006. *IEEE Transactions on Audio Speech and Language Processing* 15, 2072–2084.

Buckleton, J., C. M. Triggs, and S. J. Walsh (2005). *Forensic DNA Evidence Interpretation*. Boca Raton, FL: CRC Press.

Budlowe, B. et al. (1999). Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans and Trinidadians. *Journal of Forensic Sciences* 44 (6), 1277–1286.

Bull, R. and B. R. Clifford (1984). Earwitness voice recognition accuracy. In *Eyewitness Testimony: Psychological Perspectives*. (Ed.) G. L. Wells and E. F. Loftus. Cambridge: Cambridge University Press, pp. 92–123.

— (1999). Earwitness testimony. In *Analysing Witness Testimony: A Guide for Legal Practitioners and Other Professionals*. (Ed.) A. Heaton-Armstrong, E. Shepherd, and D. Wolchover. London: Blackstone Press, pp. 194–206.

Byrd, D. (1992). Preliminary results on speaker-dependent variation in the TIMIT database. *Journal of the Acoustical Society of America* 92 (1), 593–596.

Byrne, C. and P. Foulkes (2004). The mobile phone effect on vowel formants. *International Journal of Speech, Language and the Law* 11 (1), 83–102.

Campbell, J. P. (1997). Speaker recognition: a tutorial. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Munich, Germany.

Campbell, J. P. and D. A. Reynolds (1999). Corpora for the evaluation of speaker recognition systems. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Phoenix, Arizona.

Champod, C. and I. W. Evett (2000). Commentary on A. P. A. Broeders (1999) Some observations on the use of probability scales in forensic identification. *Forensic Linguistics* 7 (2), 238–243.

Champod, C. and D. Meuwly (1998). The inference of identity in forensic speaker recognition. In *ESCA Workshop on Speaker Recognition and its Commercial and Forensic Applications, RLA2C*, pp. 125–134.

— (2000). The inference of identity in forensic speaker recognition. *Speech Communication* 31, 193–203.

Chládkova, K. and S. Hamann (2011). High vowels in Southern British English: /u/-fronting does not result in merger. In *Proceedings of the 17th International Congress of Phonetic Sciences*. Hong Kong, pp. 476–479.

Chollet, G. et al. (1996). Swiss French PolyPhone and PolyVar: telephone speech databases to model inter- and intra-speaker variability. *IDIAP Research Report*.

Clermont, F. (2013). Cepstrum-to-formant mapping of spoken vowels. In *Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference*. 22-24 July 2013. Tampa, Florida.

Clermont, F. and S. Itahashi (1999). Monophthongal and diphthongal evidence of isomorphism between formant and cepstral spaces. In *Proceedings of the Spring Meeting of the Acoustical Society of Japan*. Meiji, Japan, pp. 205–206.

— (2000). Static and dynamic vowels in a 'cepstro-phonetic' subspace. *Journal of the Acoustical Society of Japan* 21 (4), 221–223.

Clermont, F. and P. Mokhtari (1998). Acoustic-articulatory evaluation of the upper vowel-formant region and its presumed speaker-specific potency. In *Proceedings of 7th Australasian International Speech Science and Technology Conference*. Sydney, Australia.

Coe, S. (2012). *The effect of sample contemporaneity on the outcome of likelihood ratios for forensic voice comparison*. Unpublished MSc Thesis, University of York.

Cohen, J. L. (1982). The Problem of prior probabilities in forensic proof. *Ratio* 24 (1), 518.

Coleman, R. F. and H. J. Walls (1974). The evaluation of scientific evidence. *Criminal Law Review*, 276–287.

Coupland, N. (2007). *Style: Language variation and Identity*. Cambridge: Cambridge University Press.

Cox, F. (1999). Vowel change in Australian English. *Phonetica* 56, 1–27.

Cruttenden, A. (2001). *Gimson's Pronunciation of English (6th edition)*. London: Edward Arnold.

Cudmore, A. (2011). *The interpretation of forensic speaker comparison evidence by potential jurors in the UK*. Unpublished MSc Dissertation, University of York.

Daniloff, R. G. and R. E. Hammarberg (1973). On defining coarticulation. *Journal of Phonetics* 1, 239–248.

Duda, R. O., P. E. Hart, and D. G. Stork (2000). *Pattern Classification (2nd edition)*. New York: Wiley.

Easton, L. and L. Bauer (2000). An acoustic study of the vowels of New Zealand English. *Australian Journal of Linguistics* 44 (2), 93–117.

Edwards, W. (1986). Comment. *Boston University Law Review* 66, 623–626.

Elley, W. B. and J. C. Irving (1985). The Elley-Irving socio-economic index: 1981 census revision. *New Zealand Journal of Educational Studies* 20, 115–128.

Ellis, A. J. (1889). *On Early English Pronunciation*. London: Trübner and Co.

Ellis, S. (1994). The Yorkshire Ripper enquiry: part 1. *Forensic Linguistics* 1 (2), 197–206.

Enzinger, E. (2010). Parametric diphthong formant trajectory representations for forensic speaker recognition. In *Proceedings of the 36th annual conference of the German Acoustical Society (DAGA)*. Berlin, Germany, pp. 993–994.

Enzinger, E. and P. Balazs (2011). Speaker verification using pole/zero estimates of nasals. In *Proceedings of the Multi-Conference on Systems and Structures (SysStruc '11)*. Reșița, Romania, pp. 4820–4823.

Enzinger, E. and C. H. Kasess (2013). Experiments on using Vocal Tract Estimates of Nasal Stops for Speaker Verification. In *Proceedings of the 7th International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013)*. Cluj-Napoca, Romania.

Enzinger, E. and G. S. Morrison (2012). The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems. In *Proceedings of the 14th Australasian Conference on Speech Science and Technology*. Sydney, Australia, pp. 137–140.

— (2014). A demonstration of the evaluation of forensic evidence under conditions reflecting those of an actual forensic-voice-comparison case. In *Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference*. 31 August-3 September 2014. Zürich, Switzerland.

Enzinger, E., C. Zhang, and G. S. Morrison (2012). Voice source features for forensic voice comparison – an evaluation of the GLOTTEX software package. In *Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop*. Singapore, pp. 78–85.

Eriksson, A. (2011). Aural/acoustic vs. automatic methods in forensic phonetic case work. In *Forensic Speaker Recognition: Law Enforcement and Counter-terrorism*. (Ed.) A. Neustein and H. A. Patil. New York: Springer-Verlag, pp. 41–69.

Eriksson, E. J. et al. (2004a). Cross-language speaker identification using spectral moments. In *Proceedings of the 17th Swedish Phonetic Conference (FONETIK)*. Stockholm, Sweden, pp. 76–79.

— (2004b). Robustness of spectral moments: a study using voice imitations. In *Proceedings of the 10th Australasian Conference on Speech Science and Technology*. Macquarie University, Australia, pp. 259–264.

Esling, J. H. and B. C. Dickson (1985). Acoustical procedures for articulatory setting analysis in accent. In *Papers from the First International Conference on Methods in Dialectology*. (Ed.) H. J. Warkentyne. British Columbia: University of Victoria, pp. 155–170.

Evett, I. W. (1991). Interpretation: a personal odyssey. In *The Use of Statistics in Forensic Science*. (Ed.) C. G. G. Aitken and D. A. Stoney. London: Ellis Horwood, pp. 9–22.

Evett, I. W. et al. (2000). The impact of the principles of evidence interpretation on the structure and content of statements. *Science and Justice* 40 (4), 233–239.

Fabricius, A. (2000). *T-glottaling between stigma and prestige: a sociolinguistic study of modern RP*. Unpublished PhD Thesis, Copenhagen Business School.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Fenton, N. and M. Neil (2012). *On limiting the use of Bayes in presenting forensic evidence*. `http://www.eecs.qmul.ac.uk/~norman/papers/likelihood_ratio.pdf`. Accessed: 30th September 2014.

Fisher, W. M. et al. (1986). The DARPA Speech Recognition Research Database: specifications and status. In *Proceedings of the DARPA Speech Recognition Workshop, Report No. SAIC-86/1546*. Palo Alto, California.

Foreman, L. A. and I. W. Evett (2001). Statistical analyses to support forensic interpretation for a new ten-locus STR profiling system. *International Journal of Legal Medicine* 114, 147–155.

Foulkes P., Carroll G. and S. Hughes (2004). Sociolinguistics and acoustic variability in filled pauses. In *Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference*. 28-31 July 2004. Helsinki, Finland.

Foulkes, P. and G. J. Docherty (1999). *Urban Voices: Accent Studies in the British Isles*. London: Arnold.

— (2006). The social life of phonetics and phonology. *Journal of Phonetics* 34 (4), 409–438.

Foulkes, P., G. J. Docherty, and M. Jones (2010). Analyzing stops. In *Best Practices in Sociophonetics (2nd edition)*. (Ed.) M. Di Paolo and M. Yaeger-Dror. London: Routledge.

Foulkes, P. and J. P. French (2001). Forensic phonetics and sociolinguistics. In *Concise Encyclopedia of Sociolinguistics*. (Ed.) R. Mesthrie. Amsterdam: Elsevier Press, pp. 329–332.

Foulkes, P. and J. P. French (2012). Forensic speaker comparison: a linguistic-acoustic perspective. In *Oxford Handbook of Language and the Law*. (Ed.) P. Tiersma and L. Solan. Oxford: Oxford University Press, pp. 557–572.

Foulkes, P. et al. (2013-2015). *Modelling Features for Forensic Speaker Comparison*. British Academy/ Leverhulme Trust Small Research Grant.

French, J. P. (1994). An overview of forensic phonetics with particular reference to speaker identification. *Forensic Linguistics* 1 (2), 197–206.

— (2005). Forensic speaker identification evidence and the Prosecutor' s Fallacy. In *Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference*. 3-6 August 2005. Marrakech, Morocco.

French, J. P. and P. Harrison (2006). Investigative and evidential applications of forensic speech science. In *Witness Testimony: Psychological, Investigative and Evidential Perspectives*. (Ed.) A. Heaton-Armstrong et al. Oxford: Oxford University Press, pp. 247–262.

— (2007). Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech, Language and the Law* 14 (1), 137–144.

French, J. P. et al. (2010). The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrison. *International Journal of Speech, Language and the Law* 17 (1), 138–163.

Fromont, R. and J. Hay (2008). ONZE Miner: the development of a browser-based research tool. *Corpora* 3 (2), 173–193.

— (2012). LaBB-CAT: an annotation store. In *Proceedings of the 14th Australasian Conference on Speech Science and Technology*. Sydney, Australia, pp. 113–117.

Garofolo, J. S. et al. (1993). *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*. `http://perso.limsi.fr/lamel/TIMIT_NISTIR4930.pdf`. U.S. Department of Commerce, National Institute of Standards and Technology. Accessed: 21st January 2014.

Garvin, P. L. and P. Ladefoged (1963). Speaker identification and message identification in speech recognition. *Phonetica* 9, 193–199.

Gill, P. and T. Clayton (2009). The current status of DNA profiling in the UK. In *The Handbook of Forensic Science*. (Ed.) J. Fraser and R. Williams. Cullompton: Willan Publishing, pp. 29–56.

Gill, P. and I. W. Evett (1995). Population genetics of short tandem repeat (STR) loci. *Genetica* 96, 69–87.

Gill, P. et al. (2000). Report of the European Network of Forensic Science Institutes (ENFSI): formulation and testing of principles to evaluate STR multiplexes. *Forensic Science International* 108 (1), 1–29.

Gold, E. (2014). *Calculating likelihood ratios in forensic speaker comparison cases using phonetic and linguistic features*. Unpublished PhD Thesis, University of York.

Gold, E. and J. P. French (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law* 18 (2), 293–307.

Gold, E., J. P. French, and P. Harrison (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. In *Proceedings of Meetings on Acoustics (POMA) 19*. Montréal, Canada.

Gold, E. and V. Hughes (2013-2014). *Identifying Correlations Between Speech Parameters for Forensic Speaker Comparisons*. IAFPA Research Grant.

— (2014). Issues and opportunities: the application of the numerical likelihood ratio framework to forensic speaker comparison. *Science and Justice* 54 (4), 292–299.

Goldman-Eisler, F. (1998). *Psycholinguistics: Experiments in Spontaneous Speech*. London: Academic Press.

González-Rodríguez, J. et al. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language* 20 (2-3), 331–355.

González-Rodríguez, J. et al. (2007). Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions of Audio, Speech and Language Processing* 15 (7), 2104–2115.

Good, I. J. (1991). Weight of evidence and the Bayesian likelihood ratio. In *The Use of Statistics in Forensic Science*. (Ed.) C. G. G. Aitken and D. A. Stoney. London: Ellis Horwood, pp. 85–106.

Gordon, E., M. Maclagan, and J. Hay (2007). The ONZE corpus. In *Models and Methods in the Handling of Unconventional Digital Corpra: Volume 2, Diachronic Copora*. (Ed.) J. C. Beal, K. P. Corrigan, and H. Moisl. London: Palgrave, pp. 82–104.

Gordon, E. et al. (2004). *New Zealand English: Its Origins and Evolution*. Cambridge: Cambridge University Press.

Greenberg, G. et al. (2010). Human assisted speaker recognition (HASR) in NIST SRE2010. In *Proceedings of Odyssey 2010: The Language and Speaker Recognition Workshop*. Singapore, pp. 180–185.

Gregersen, F. and I. Pedersen (1991). *The Copenhagen Study in Urban Sociolinguistics*. Copenhagen: C. A. Reitzels Forlag.

Greisbach, R., E. Osser, and C. Weinstock (1995). Speaker identification by formant contours. In *Studies in Forensic Phonetics. Beiträge zur Phonetik und Linguistik 64*. (Ed.) A. Braun and J. Köstner. Trier: Wissenschaftlicher Verlag Trier, pp. 49–55.

Grigoras, C. et al. (2013). Forensic audio analysis – Review: 2010-2013. In *Proceedings of the 17th International Science Managers' Symposium*. Lyon, France, pp. 612–637.

Gruber, J. S. and F. Poza (1995). Voicegram identification evidence. *American Jurisprudence Trials* 54 (1), §1–§133.

Haddican, B. (2008-2013). *A Comparative Study of Language Change in Northern Englishes*. Economic and Social Research Council (ESRC) of Great Britain. RES-061-25-0033.

Haddican, B. et al. (2013). Interaction of social and linguistic constraints on two vowel changes in northern England. *Language Variation and Change* 25 (3), 371–403.

Hall-Lew, L. (2005). One shift, two groups: when fronting alone is not enough. *University of Pennsylvania Working Papers in Linguistics* 10 (2), 105–116.

Harrington, J. (1997). Acoustic phonetics. In *A Handbook of Phonetic Science*. (Ed.) W. Hardcastle and J. Laver. Oxford: Blackwell, pp. 81–129.

— (2008). Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: an acoustic and perceptual study. *Journal of the Acoustical Society of America* 123 (5), 2825–2835.

Harrison, P. (2013). *Making accurate formant measurements: an empicial investigation of the influence of the measurement tool, analysis settings and speaker on formant measurements*. Unpublished PhD Thesis, University of York.

Harrison, P. and J. P. French (2012). Evaluation of the BATVOX automatic speaker recognition system for use in UK based forensic speaker comparison casework Part II. In *Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference*. 18-21 July 2010. University of Trier, Germany.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *Elements of Statistical Learning: Data Mining, Inference and Prediction (2nd edition)*. Heidelberg: Springer-Verlag.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Modelling* 44 (1), 1–12.

Hawkins, S. and J. Midgley (2005). Formant Frequencies of RP monophthongs in four age-groups of speakers. *Journal of International Phonetic Association* 35 (2), 183–199.

Hay, J., M. Maclagan, and E. Gordon (2008). *Dialects of English: New Zealand English*. Edinburgh: Edinburgh University Press.

Heselwood, B. and L. McChrystal (2000). Gender, accent features and voicing in Panjabi-English bilingual children. *Leeds Working Papers in Linguistics and Phonetics* 8, 45–70.

Hollien, H. (2002). *Forensic Voice Identification*. San Diego, CA: Academic Press.

Huang, C., E. Chang, and T. Chen (2004). Accent issues in large vocabulary continuous speech recognition. *International Journal of Speech Technology* 7, 141–153.

Huckvale, M. (2004). ACCDIST: a metric for comparing speakers' accents. In *Proceedings of the International Conference on Spoken Language Processing*. Jeju, Korea.

— (2007). Hierarchical clustering of speakers into accents with the ACCDIST metric. In *Proceedings of the 16th International Congress of Phonetic Sciences*. Universität des Saarlandes, Saarbrücken, pp. 1821–1824.

Hughes, A., P. Trudgill, and D. Watt (2005). *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles (4th edition)*. London: Hodder Arnold.

Hughes, V. (2009). *Diphthong dynamics in unscripted speech*. Unpublished MSc Dissertation, University of York.

Hughes, V. et al. (2011). Vowel variation in Manchester: a dynamic approach. In *Paper presented at the 8th UK Conference on Language Variation Change (UKLVC8). 12-14 September 2011*. Edge Hill University, United Kingdom.

Hultzen, L. S, Joseph H. D. Allen, and M. S. Miron (1964). *Tables of Transitional Frequencies of English Phonemes*. Urbana: University of Illinois Press.

Ingram, J. C. L., R. Prandolini, and S. Ong (1996). Formant trajectories as indices of phonetic variation for speaker identification. *Forensic Linguistics* 3 (1), 129–145.

Ishihara, S. and Y. Kinoshita (2008). How many do we need? Exploration of the Population Size Effect on the performance of forensic speaker classification. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech)*. Brisbane, Australia, pp. 1941–1944.

— (2012). The effect of sample size on the performance of likelihood ratio based forensic voice comparison. In *Proceedings of the 14th Australasian Conference on Speech Science and Technology*. Sydney, Australia, pp. 1–4.

Iversen, G. R. (1984). Bayesian statistical inference: quantitative applications in the social sciences. *Sage University Paper* 43.

Jessen, M. (2007). Forensic reference data on articulation rate in German. *Science and Justice* 47 (2), 50–67.

— (2008). Forensic phonetics. *Language and Linguistics Compass* 2 (4), 671–711.

— (2012). *Phonetische und Linguistische Prinzipien des Forensischen Stimmenvergleichs*. LINCOM Studies in Phonetics.

Jessen, M. and E. Enzinger (2014). *Comparing MVKD and GMM in the analysis for individual vowel formants*. Guest lecture to Forensic Speech Science group, University of York. 10th March 2014.

Johnson, K. (2008). *Auditory and Acoustic Phonetics (2nd edition)*. Oxford: Blackwell.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1994). *Continuous Univariate Distributions (Vol. 1)*. New York: Wiley.

Jurafsky, D. (2007). *Feature extraction and acoustic modelling*. http://nlp.stanford.edu/courses/lsa352/lsa352.lec6.6up.pdf. LSA 352: Speech Recognition and Synthesis (Lecture 6). LSA Summer Institute 2007. Accessed: 11th April 2014.

Kapadia, S., V. Valtchev, and S. J. Young (1993). MMI training for continuous phoneme recognition on the TIMIT database. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 491–494.

Kasess, C. H. et al. (1993). Estimation of the vocal tract shape of nasals using a Bayesian scheme. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*. Portland, Oregon, pp. 491–494.

Kavanagh, C. (2012). *New consonantal acoustic parameters for forensic speaker comparison*. Unpublished PhD Thesis, University of York.

Kaye, D. H. (2004). Logical relevance: problems with the reference population and DNA mixtures in People v. Pizarro. *Law, Probability and Risk* 3, 211–220.

— (2008). DNA probabilities in People v. Prince: When are racial and ethnic statistics relevant? In *Probability and Statistics: Essays in Honour of David A Freedman*. (Ed.) T. Speed and D. Nolan. Beachwood, OH: Institute of Mathematical Statistics, pp. 289–301.

Keller, E. (2005). The analysis of voice quality in speech processing. In *Nonlinear Speech Modeling and Applications*. (Ed.) G. Chollet et al. Berlin: Springer-Verlag, pp. 54–73.

Kersta, L. G. (1962). Voiceprint identification. *Nature* 196, 1253–1257.

Kerswill, P. (2006). Standard English, RP and the standard/ non-standard relationship. In *Language in the British Isles (2nd edition)*. (Ed.) D. Britain. Cambridge: Cambridge University Press, pp. 34–51.

Khodai-Joopari, M. (2006). *Forensic speaker analysis and identification by computer: a Bayesian approach anchored in the cepstral domain*. Unpublished PhD Thesis, University of New South Wales.

King, J. (2012). *Assessing the discriminatory power of vocalic hesitation markers for forensic speaker comparison casework*. Unpublished MSc Dissertation, University of York.

Kinoshita, Y. (2001). *Testing realistic forensic speaker identification in Japanese: a likelihood ratio-based approach using formants*. Unpublished PhD Thesis, Australian National University.

Kinoshita, Y. (2002). Use of likelihood ratio and Bayesian approach in forensic speaker identification. In *Proceedings of the 9th Australasian Conference on Speech Science and Technology*. Melbourne, Australia, pp. 297–302.

— (2005). Does Lindley' s LR estimation formula work for speech data? Investigation using long-term f0. *International Journal of Speech, Language and the Law* 12 (2), 235–254.

Kinoshita, Y. and T. Osanai (2006). Within-speaker variation in diphthongal dynamics: what can we compare. In *Proceedings of the 11th Australasian Conference on Speech Science and Technology*. University of Auckland, New Zealand, pp. 112–117.

Kondoz, A. M. (2004). *Digital Speech: Coding for Low Bit Rate Communication Systems (2nd edition)*. Chichester: John Wiley.

Kononenko, I. (1990). Comparison of inductive and naïve Bayesian capitalised learning approaches to automatic knowledge acquisition. In *Current Trends in Knowledge Acquisition*. (Ed.) B. Wielinga et al. Amsterdam: IOS Press.

Koops, C. (2010). /u/-fronting is not monolithic: two types of fronted /u/ in Houston Anglos. *University of Pennsylvania Working Papers in Linguistics* 16 (2), 113–122.

Kuchera, H. and W. N. Francis (1967). *Computational Analysis of Present-Day American Engish*. Providence, Rhode Island: Brown University Press.

Künzel, H. J. (1997). Some general phonetic and forensic aspects of speaking tempo. *International Journal of Speech, Language and the Law* 4 (1), 48–83.

— (2001). Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *International Journal of Speech, Language and the Law* 8 (1), 80–99.

— (2013). Automatic speaker recognition with cross-language speech material. *International Journal of Speech, Language and the Law* 20 (1), 21–44.

Kurath, H. and R. I. McDavid (1961). *The Pronunciation of English in the Atlantic States*. Ann Arbor: University of Michigan Press.

Labov, W. (1971). The study of language in its social context. In *Advances in the Sociology of Language (vol. 1)*. (Ed.) J. A. Fishman. The Hague: Mouton, pp. 152–216.

Labov, W. (1988). The judicial testing of linguistic theory. In *Linguistics in Context: Connecting Obersvation and Understanding*. (Ed.) D. Tannen. Norwood, NJ: Ablex Publishing Corporation, pp. 159–182.

— (1991). The three dialects of English. In *New Ways of Analyzing Sound Change*. (Ed.) P. Eckert. New York: Academic Press, pp. 1–44.

Labov, W., S. Ash, and C. Boberg (1997). *A national map of the regional dialects of American English*. `http://www.ling.upenn.edu/phono_atlas/NationalMap/NationalMap.html#fn0`. Accessed: 2nd April 2014.

— (2006). *The Atlas of North American English*. Berlin: Mouton de Gruyter.

Labov, W. and W. A. Harris (1994). Addressing social issues through linguistic evidence. In *Language and the Law*. (Ed.) J. Gibbons. Harlow: Longman, pp. 265–305.

Ladefoged, P. and K. Johnson (2010). *A Course in Phonetics (6th edition)*. Boston: Thomson Wadsworth.

Laver, J. (1994). *Principles of Phonetics*. Cambridge: Cambridge University Press.

Law Commission of England and Wales (2004). *Understand forensic evidence: do lawyers and the judiciary understannd forensic evidence and the Bayesian approach?* `http://www.qebholliswhiteman.co.uk/articles-pdfs/understanding-forensic-evidence.pdf`. Accessed: 30th September 2014.

— (2011). *Expert evidence in criminal proceedings in England and Wales*. `http://lawcommission.justice.gov.uk/docs/lc325_Expert_Evidence_Report.pdf`. Accessed: 14th September 2014.

Lee, P. (2004). *Bayesian Statistics: An Introduction (3rd edition)*. Chichester: John Wiley.

Lei, H. and E. Lopez-Gonzalo (2009). Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition. In *Proceedings of Interspeech 2009*. International Speech Communication Association, pp. 2323–2326.

Lempert, R. (1977). Modelling relevance. *Michigan Law Review* 89, 1021–1057.

Lewis, S. R. (1984). Philosophy of speaker identification. Police applications of speech and tape recording analysis. *Proceedings of the Institute of Acoustics* 6 (1), 69–77.

Lindh, J. and G. S. Morrison (2011). Humans versus machine: forensic voice comparison on a small database of Swedish voice recordings. In *Proceedings of the 17th International Congress of Phonetic Sciences*. Hong Kong, pp. 1254–1257.

Lindley, D. V. (1977). A problem in forensic science. *Biometrika* 64, 207–213.

Linville, S. E. and J. Rens (2001). Vocal tract resonance analysis of aging voice using long-term average spectra. *Journal of Voice* 15 (3), 323–330.

Ljolje, A. and M. D. Riley (1991). Automatic segmentation and labelling of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 473–476.

Loakes, D. (2006). *A forensic phonetic investigation into the speech patterns of identical and non-identical twins*. Unpublished PhD Thesis, University of Melbourne.

Lu, L. et al. (2009). The effect of language factors for robust speaker recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4217–4220.

Lucy, D. (2005). *Introduction to Statistics for Forensic Scientists*. Chichester: John Wiley.

Lynch, M. and R. McNally (2003). 'Science', 'common sense', and DNA evidence: a legal controversy about the public understanding of science. *Public Understanding of Science* 12, 83–103.

Maclagan, M. (1982). An acoustic study of New Zealand English vowels. *The New Zealand Speech Therapists Journal* 37, 20–26.

Maclagan, M. and E. Gordon (1999). Data for New Zealand social dialectology: the Canterbury Corpus. *New Zealand English Journal* 13, 50–58.

Maeda, S. (1990). Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech Production and Speech Modelling*. (Ed.) W. J. Hardcastle and A. Marchal. Dordrecht: Kulwer, pp. 131–150.

Maekawa, K. et al. (2000). Spontaneous speech corpus of Japanese. In *Proceedings of the 2nd International Conference of Language Resources and Evaluation (LREC2000)*. Athens, Greece, pp. 947–952.

Martire, K. A. et al. (2014). On the interpretation of likelihood ratios in forensic science evidence: presentation formats and the weak evidence effect. *Forensic Science International* 240, 61–68.

McDougall, K. (2004). Speaker-specific formant dynamics: an experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law* 11 (1), 103–130.

— (2005). *The role of formant dynamics in determining speaker identity*. Unpublished PhD Thesis, University of Cambridge.

— (2006). Dynamic features of speech and the characterisation of speakers: towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law* 13 (1), 89–126.

— (2013). Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades. *International Journal of Speech, Language and the Law* 20 (2), 163–172.

McDougall, K. and F. Nolan (2007). Discrimination of speakers using the formant dynamics of /uː/ in British English. In *Proceedings of the 16th International Congress of Phonetic Sciences*. Universität des Saarlandes, Saarbrücken, pp. 1825–1828.

Meuwly, D. (2001). *Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique*. Unpublished PhD Thesis, University of Lausanne.

Miller, J., F. Grosjean, and C. Lomanto (1984). Articulation rate and its variability in spontaneous speech: a reanalysis and some implications. *Phonetica* 41, 215–225.

Milroy, J. (1995). Investigating the Scottish Vowel Length Rule in a Northumbrian dialect. *Newcastle and Durham Working Papers in Linguistics* 3, 187–196.

— (1996). A current change in British English: variation in (th) in Derby. *Newcastle and Durham Working Papers in Linguistics* 4, 213–222.

Milroy, L., J. Milroy, and G. J. Docherty (1994-1997). *Phonological Variation and Change in Contemporary British English*. Economic and Social Research Council (ESRC) of Great Britain. R000234892.

Milroy, L. et al. (1999). Phonological variation and change in contemporary English: evidence from Newcastle upon Tyne and Derby. *Cuadernos de Filologia Inglesa* 8, 35–46.

Mokhtari, P. (1998). *A acoustic-phonetic and articulatory study of speech-speaker dichotomy*. Unpublished PhD Thesis, University of New South Wales.

Montgomery, C. (2007). *Northern English dialects: a perceptual approach*. Unpublished PhD Thesis, University of Sheffield.

Moreno, A. et al. (2006). The influence of dialects on automatic speaker recognition systems. In *Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference*. 23-26 July 2006. University of Gothenburg, Sweden.

Morrison, G. S. (2008). Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/. *International Journal of Speech, Language and the Law* 15 (2), 249–266.

— (2009a). Forensic voice comparison and the paradigm shift. *Science and Justice* 49 (4), 298–308.

— (2009b). Likelihood-ratio voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America* 125 (4), 2387–2397.

— (2009c). The place of forensic voice comparison in the ongoing paradigm shift. In *Invited Presentation at the 2nd International Conference on Evidence Law and Forensic Science, 25–26 July 2009*. Beijing, China, pp. 1–16.

— (2010). Forensic voice comparison. In *Expert Evidence (Ch. 99)*. (Ed.) I. Freckelton and H. Selby. Sydney: Thomson Reuters.

— (2011a). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM). *Speech Communication* 53 (2), 242–256.

— (2011b). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science and Justice* 51 (3), 91–98.

— (2012). The likelihood-ratio framework and forensic evidence in court: a response to R v T. *International Journal of Evidence and Proof* 16, 1–29.

— (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences* 45 (2), 173–197.

Morrison, G. S. (2014). Distinguishing between forensic science and forensic pseudoscience: testing of validity and reliability, and approaches to forensic voice comparison. *Science and Justice* 54 (3), 245–256.

Morrison, G. S. and P. F. Assmann (2013). *Vowel Inherent Spectral Change*. Heidelberg: Springer-Verlag.

Morrison, G. S. and Y. Kinoshita (2008). Automatic-type calibration of traditionally derived likelihood ratios: forensic analysis of Australian English /o/ formant trajectories. In *Proceedings of Interspeech 2008 Incorporating SST 2008*. International Speech Communication Association, pp. 1501–1504.

Morrison, G. S., F. Ochoa, and T. Thiruvaran (2012). Database selection for forensic voice comparison. In *Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop*. Singapore, pp. 74–77.

Morrison, G. S., P. Rose, and C. Zhang (2012). Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Australian Journal of Forensic Sciences* 44, 155–167.

Morrison, G. S. and R. D. Stoel (2014). Forensic strength of evidence statements should preferably be likelihood ratios calculated using relevant data, quantitative measurements, and statistical models – a response to Lennard (2013) Fingerprint identification: How far have we come? *Australian Journal of Forensic Sciences* 46, 282–292.

Morrison, G. S., T. Thiruvaran, and J. Epps (2010). Estimating the precision of the likelihood ratio output of a forensic-voice-comparison system. In *Proceedings of Odyssey 2010: The Language and Speaker Recognition Workshop*. Brno, Czech Republic, pp. 63–70.

Morrison, G. S., C. Zhang, and P. Rose (2011). An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic Science International* 208 (1-3), 59–65.

Morrison, G. S. et al. (2010-2013). *Making Demonstrably Valid and Reliable Forensic Voice Comparison a Practical Everyday Reality in Australia*. ARC Linkage Project, project ID. LP100200142.

Mullen, C. et al. (2013). Perception problems of the verbal scale. *Science and Justice* 54 (2), 154–158.

Nair, B., E. Alzqhoul, and B. Guillemin (2012). Optimizing forensic voice comparison using a subset of cepstral coefficients. In *Proceedings of the 14th Australasian Conference on Speech Science and Technology*. Macquarie University, Sydney, pp. 133–136.

— (2014). Determination of likelihood ratios for forensic voice comparison using principal component analysis. *International Journal of Speech, Language and the Law* 21 (1), 83–112.

National Research Council (1996). *The Evaluation of Forensic DNA Evidence*. Washington: National Academy Press.

— (2009). *Strengthening forensic science in the United States: a path forward*. `http://www.nap.edu/catalog.php?record_id=12589`. Accessed: 27th May 2014.

Neocleous, T. et al. (2014). Evidence evaluation for forensic voice comparison. In *Poster presented at the 9th International Conference on Forensic Inference and Statistics (ICFIS)*. 19-22 August 2014. Leiden University, Netherlands.

Neumann, C., Evett. I. W., and J. Skerrett (2012). Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm. *Journal of the Royal Statistical Society* 175 (2), 371–415.

Nolan, F. (1983). *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.

— (1997). Speaker recognition and forensic phonetics. In *The Handbook of Phonetic Sciences*. (Ed.) W. J. Hardcastle and J. Laver. Oxford: Blackwell, pp. 744–767.

— (2001). Speaker identification evidence: its forms, limitations, and roles. In *Proceedings of the Law and Language: Prospect and Retrospect Conference*. 12-15 December 2001. Levi, Finland.

— (2003). A recent voice parade. *International Journal of Speech, Language and the Law* 10 (2), 277–291.

Nolan, F. and E. Grabe (1996). Preparing a voice line-up. *Forensic Linguistics* 3 (1), 74–94.

Nolan, F., K. McDougall, and T. Hudson (2013). Effects of the telephone on perceived voice similarity: implications for voice line-ups. *International Journal of Speech, Language and the Law* 20 (2), 229–246.

Nolan, F. and T Oh (1996). Identical twins, different voices. *Forensic Linguistics* 3 (1), 39–49.

Nolan, F. et al. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16 (2), 31–57.

Nolan, F. et al. (2005-2009). *Dynamic Variability in Speech: A Forensic Phonetic Study of British English*. Economic and Social Research Council (ESRC) of Great Britain. RES-000-23-1248.

— (2009). *Voice similarity and the effect of the telephone: a study of the implications for earwitness evidence*. Full Research Report, ESRC End of Award Report, RES-000-22-2582.

O'Shaughnessy, D. (1987). *Speech Communications: Human and Machine*. Boston: Addison-Wesley.

Pang, J. and P. Rose (2012). Likelihood ratio-based forensic voice comparison with the Cantonese diphthong /ei/ f-pattern. In *Proceedings of the 14th Australasian Conference on Speech Science and Technology*. Sydney, Australia, pp. 205–208.

Patel, J. K. and C. B. Read (1982). *Handbook of the Normal Distribution*. New York: Marcel Dekker.

Patrick, P. L. (2008). The speech community. In *Handbook of Language Variation and Change*. (Ed.) J. K. Chambers, P. Trudgill, and N. Schilling-Estes. Oxford: Wiley-Blackwell, pp. 573–597.

Peterson, G. E. (1959). The acoustics of speech – part II: acoustical properties of speech waves. In *Handbook of Speech Pathology*. (Ed.) L. E. Travis. London: Peter Owen, pp. 137–173.

Pigeon, S, P. Druyts, and P. Verlinde (2000). Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digital Signal Processing* 10, 237–248.

Police and Criminal Evidence Act (PACE) (1984). *Code E: Code of Practice on Audio Recording Interviews with Suspects*. Applies to interviews carried out after midnight on 1 May 2010.

Przybocki, M. A., A. F. Martin, and A. N. Le (2007). In *NIST speaker recognition evaluations utilizing the Mixer corpora 2004, 2005, 2006*. Honolulu, Hawaii, pp. 1951–1959.

Rabiner, L. and B. H. J. Juang (1993). *Fundamentals of Speech Recognition*. Englewood Cliffs: Prentice-Hall.

Ramos-Castro, D. (2007). *Forensic evaluation of the evidence using automatic speaker recognition systems*. Unpublished PhD Thesis, Universidad Autónoma de Madrid.

— (2012). Reliable support: measuring calibration of likelihood ratios. In *Paper presented at 6th European Academy of Forensic Science Conference*. 20-24 August 2012. The Hague, Netherlands.

Redmayne, M. (2001). *Expert Evidence and Criminal Justice*. Oxford: Oxford University Press.

Redmayne, M. et al. (2011). Forensic science evidence in question. *Criminal Law Review* 5, 347–356.

Reubold, U., J. Harrington, and F. Kleber (2010). Vocal aging effects on F0 and the first formant: a longitudinal analysis in adult speakers. *Speech Communication* 52 (7-8), 638–651.

Reynolds, D. A. (1995). Large population speaker identification using clean and telephone speech. *Signal Processing Letter IEEE* 2 (3), 46–48.

Reynolds, D. A., T. F. Quatieri, and R. B. Dunn (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing* 10, 19–41.

Reynolds, D. A. et al. (1995). In *The effects of telephone transmission degradations on speaker recognition performance*. Detroit, Michigan, pp. 329–332.

Rhodes, R. (2009). *Using /aɪ/ to Discriminate between Derby Speakers using Formant Dynamics in Spontaneous Speech*. Unpublished MSc Dissertation, University of York.

Robertson, B. and G. A. Vignaux (1995a). DNA evidence: wrong answers or wrong questions? In *Human Identifiction: The Use of DNA Markers*. (Ed.) B. S. Weir. Dordrecht: Kluwer Academic, pp. 145–152.

— (1995b). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. Oxford: Oxford University Press.

Rodman, R. et al. (2002). Forensic speaker identification based on spectral moments. *Forensic Linguistics* 9 (1), 22–43.

Rose, P. (1998). A forensic phonetic investigation in non-contemporaneous variation in the F-pattern of similar-sounding speakers. In *Proceedings of the 5th Australasian Conference on Speech Science and Technology*. Canberra, Australia, pp. 49–52.

— (1999). Long- and short-term within-speaker differences in the formants of Australian *hello*. *Journal of the International Phonetic Association* 29 (1), 1–31.

— (2002). *Forensic Speaker Identification*. London: Taylor and Francis.

— (2004). Technical Forensic Speaker Identification from a Bayesian Linguist's Perspective. In *Keynote paper, Forensic Speaker Recognition Workshop, Speaker Odyssey ' 04*. Toledo, Spain, pp. 3–10.

— (2006). The intrinsic forensic discriminatory power of diphthongs. In *Proceedings of the 11th Australasian Conference on Speech Science and Technology*. University of Auckland, New Zealand, pp. 64–69.

— (2007a). Forensic speaker discrimination with Australian English vowel acoustics. In *Proceedings of the 16th International Congress of Phonetic Sciences*. Universität des Saarlandes, Saarbrücken, pp. 1817–1820.

— (2007b). Going and getting it - forensic speaker recognition from the perspective of a traditional practitioner/ researcher. In *Paper presented at the Australian Research Council Network in Human Communication Science Workshop: FSI not CSI – Perspectives in State-of-the-Art Forensic Speaker Recognition*. Sydney, Australia.

— (2007-2010). *Catching Criminals by their Voice: Combining Automatic and Traditional Methods for Optimum Performance in Forensic Speaker Recognition*. Australian Research Council Discovery Project grant no. DP0774115.

— (2010). Bernard's 18 – vowel inventory size and strength of forensic voice comparison evidence. In *Proceedings of the 13th Australasian Conference on Speech Science and Technology*. Melbourne, Australia, pp. 30–33.

— (2011a). Forensic voice comparison with Japanese vowel acoustics – a likelihood ratio-based approach using segmental cepstra. In *Proceedings of the 17th International Congress of Phonetic Sciences*. Hong Kong, pp. 1718–1721.

— (2011b). Forensic voice comparison with secular shibboleths - a hybrid fused gmm-multivariate likelihood ratio-based approach using alveolo-palatal fricative cepstral spectra. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5900–5903.

Rose, P. (2012). The likelihood ratio goes to Monte Carlo: the effect of reference sample size on likelihood-ratio estimates. In *Paper presented at the UNSW Forensic Speech Science Conference*. 3 December 2012. Sydney, Australia.

— (2013a). More is better: likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends. *International Journal of Speech, Language and the Law* 20 (1), 77–116.

— (2013b). Where the science begins and the law ends: likelihood ratio-based forensic voice comparison in a $150 million telephone fraud. *International Journal of Speech, Language and the Law* 20 (2), 277–324.

Rose, P., Y. Kinoshita, and T. Alderman (2006). Realistic extrinsic forensic speaker discrimination with the diphthong /aɪ/. In *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*. University of Auckland, New Zealand, pp. 329–334.

Rose, P. and G. S. Morrison (2009). A response to the UK Position Statement on forensic speaker comparison. *International Journal of Speech, Language and the Law* 16 (1), 139–163.

Rose, P. and E. Winter (2010). Traditional forensic voice comparison with female formants: Gaussian mixture model and multivariate likelihood ratio approaches. In *Proceedings of the 13th Australasian Conference on Speech Science and Technology*. Melbourne, Australia, pp. 42–45.

Saks, M. J. and J. J. Koehler (2005). The coming paradigm shift in forensic identification science. *Science* 309, 892–895.

Salvi, G. (2003). Accent clustering in Swedish using the Bhattacharyya distance. In *Proceedings of the 15th International Congress of Phonetic Sciences*. Barcelona, Spain, pp. 1149–1152.

Schwartz, R. et al. (2011). USSS-MITLL 2010 human assisted speaker recognition. In *Proceedings of ICASSP*. Prague, Czech Republic, pp. 5904–5907.

Scobbie, J., N. Hewlett, and A. E. Turk (1999). Standard English in Edinburgh and Glasgow: the Scottish Vowel Length Rule revealed. In *Urban Voices: Accent Studies in the British Isles*. (Ed.) P. Foulkes and G. J. Docherty. London: Arnold, pp. 230–245.

Seber, G. A. F. and C. J. Wild (1989). *Nonlinear Regression*. New York: John Wiley.

Simpson, S. (2008). *Testing the speaker discrimination ability of formant measurements in forensic speaker comparison cases*. Unpublished MSc Dissertation, University of York.

Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech. *PHONUM* 9, 93–96.

Smith, R. L. and R. P. Charrow (1975). Upper and lower bounds for the probability of guilt based on circumstantial evidence. *Journal of the American Statistical Association* 70, 555–560.

Sõskuthy, M. et al. (in press). The closing diphthongs in early New Zealand English. In *Listening to the Past*. (Ed.) R. Hickey. Cambridge: Cambridge University Press.

Stevens, K. N. (2000). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Stevens, L. and J. P. French (2012). Voice quality in Standard Southern British English: distribution of features, inter-speaker variability and the effect of telephone transmission. In *Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference*. 5-8 August 2012. Santander, Spain.

Sun, D. X. and L. Deng (1995). Analysis of acoustic-phonetic variations in fluent speech using TIMIT. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 201–204.

Tabachnick, B. G. and L. S. Fiddell (2007). *Using Multivariate Statistics (5th edition)*. Boston: Pearson.

Tagliamonte, S. (1996-1998). *Roots of Identity: Variation and Grammaticalisation in Contemporary British English*. Economic and Social Research Council (ESRC) of Great Britain. Ref: R000221842.

Thomas, E. (2001). *An Acoustic Analysis of Vowel Variation in New World English*. Durham, NC: Duke Univeristy Press.

Thompson, W. C. and E. L. Schumann (1987). Interpretation of statistical evidence in criminal trials: the prosecutor's fallacy and the defence attorney's fallacy. *Law and Human Behaviour* 11 (3), 167–187.

Tiersma, P. and L. Solan (2012). *Oxford Handbook of Language and Law*. Oxford: Oxford Univeristy Press.

TIMIT (1990). *TIMIT Readme*. `https://catalog.ldc.upenn.edu/LDC93S1`. Accessed: 12th November 2013.

Tosi, O. I. (1979). *Voice Identification: Theory and Legal Applications*. Baltimore: University Park Press.

Turk, A., S. Nakai, and M. Sugahara (2006). Acoustic segment durations in prosodic research: a practical guide. In *Methods in Empirical Prosody Research*. (Ed.) S. Sudhoff et al. Berlin: De Gruyter, pp. 1–28.

Upton, C. and J. D. A. Widdowson (2006). *An Atlas of English Dialects (2nd edition)*. London: Routledge.

van der Vloed, D. et al. (2011). Influence of the size of the population dataset on the results produced by the Batvox software. In *Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference*. 24-28 July 2011. Vienna, Austria.

van Leeuwen, D. and J. S. Bouten (2004). Results of the 2003 NFI-TNO forensic speaker recognition evaluation. In *Proceedings of Odyssey 2004: The Language and Speaker Recognition Workshop*. Singapore, pp. 75–82.

van Leeuwen, D. and N. Brümmer (2007). An introduction to application-independent evaluation of speaker recognition systems. In *Speaker Classification I, LNAI 4343*. (Ed.) C. Müller. Berlin: Springer-Verlag, pp. 330–353.

Vandyke, D. J. (2014). *Glottal waveforms for speaker inference and a regression score post-processing method applicable to general classification problems*. Unpublished PhD Thesis, University of Canberra, Canberra.

Wackerly D. D., Mendenhall III W. and R. L. Scheaffer (2008). *Mathematical Statistics with Applications (7th edition)*. London: Thomson.

Wagner, M. et al. (2010). The Big Australian Speech Corpus (The Big ASC). In *Proceedings of the 13th Australasian Conference on Speech Science and Technology*. Melbourne, Australia, pp. 166–170.

Wang, C. Y. and P. Rose (2012). Likelihood ratio-based forensic voice comparison with Cantonese /i/ f-pattern and tonal f0. In *Proceedings of the 14th Australasian Conference on Speech Science and Technology*. Sydney, Australia, pp. 209–212.

Wang, Z. X. and D. R. Guo (1989). *Special Functions*. London: World Scientific.

Watt, D. (2000). Phonetic parallels between the close-mid vowels of Tyneside English: are they internally or externally motivated? *Language Variation and Change* 12 (1), 69–101.

Watt, D. (2002). 'I don't speak with a Geordie accent, I speak, like, the Northern accent': contact-induced levelling in the Tyneside vowel system. *Journal of Sociolinguistics* 6 (1), 44–63.

Watt, D. and L. Milroy (1999). Variation in three Tyneside vowels: is this dialect levelling? In *Urban Voices: Accent Studies in the British Isles*. (Ed.) G. J. Docherty and P. Foulkes. London: Arnold, pp. 25–46.

Welch, B. L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika* 34 (1/2), 28–35.

Wells, J. C. (1982). *Accents of English (3 vols)*. Cambridge: Cambridge University Press.

West, P. (1999). The extent of coarticulation of English liquids: an acoustic and articulatory study. In *Proceedings of the 14th International Congress of Phonetic Sciences*. San Francisco, U.S.A, pp. 1901–1904.

— (2000). Long-distance coarticulatory effects of British English /l/ and /r/: an EMA, EPG and acoustic study. In *Proceedings of the 5th Seminar on Speech Production: Models and Data*. Kloster Seeon, Germany, pp. 105–108.

Whittle, P. (1983). *Prediction and Regulation by Linear Least-Squared Methods (2nd edition)*. Minneapolis: University of Minnesota Press.

Wilder, C. (1978). Vocal aging. In *Transcripts of 7th Symposium: Care of the Professional Voice. Part II: Life Span Changes in the Human Voice*. (Ed.) B. Weinberg. New York: Voice Foundation.

Williams, A. and P. Kerwill (1999). Dialect levelling: change and continuity in Milton Keynes, Reading and Hull. In *Urban Voices: Accent Studies in the British Isles*. (Ed.) G. J. Docherty and P. Foulkes. London: Arnold, pp. 141–162.

Wood, S. (2013). *Filled pauses: a discriminatory parameter for speaker comparison cases*. Unpublished MSc Thesis, University of York.

Wood, S., V. Hughes, and P. Foulkes (2014). Filled pauses as variables in speaker comparison: dynamic formant analysis and duration measurements improve performance. In *Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference*. 31 August-3 September 2014. Zürich, Switzerland.

Yan, Q., Z. Zhengjuan, and L. Shan (2012). Chinese accent identification with modified MFCCs. *Instrumentation, Measurement, Circuits and Systems: Advances in Intelligent and Soft Computing* 127, 659–666.

Yim, A. C. Y. and P. Rose (2012). Are nasals better? Likelihood ratio-based forensic voice comparison with segmental cepstra from Cantonese and Japanese syllabic/mora nasals. In *Proceedings of the 14th Australasian Conference on Speech Science and Technology*. Sydney, Australia, pp. 5–8.

Young, S. et al. (2006). *The HTK Book (for HTK Version 3.4)*. `http://htk.eng.cam.ac.uk/prot-docs/htkbook.pdf`. Accessed: 11th September 2013.

Zhang, C. and G. S. Morrison (2011). *Forensic database of audio recordings of 68 female speakers of Standard Chinese*. `http://databases.forensic-voice-comparison.net`.

Zhang, C., G. S. Morrison, and P. Rose (2008). Forensic speaker recognition in Chinese:a multivariate likelihood ratio discrimination on /i/ and /y/. In *Proceedings of Interspeech 2008 Incorporating SST 2008*. International Speech Communication Association, pp. 1937–1940.

Zhang, C., G. S. Morrison, and T. Thiruvaran (2011). Forensic voice comparison using Chinese /iau/. In *Proceedings of the 17th International Congress of Phonetic Sciences*. Hong Kong, pp. 2280–2283.

Zhang, C. et al. (2012). Reliability of human-supervised formant-trajectory measurement for forensic voice comparison. *Journal of the Acoustical Society of America* 133, EL54–EL60.

Zhang, C. et al. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices. *Speech Communication* 55 (6), 796–813.