The effect of variability on the outcome of likelihood ratios

Vincent Stephen Hughes

September 2011

Supervisor: Prof Paul Foulkes

Submitted in partial fulfilment of the degree of MSc at the Department of Language and Linguistics, University of York

THE UNIVERSITY *of York*

Word count: 9973

ABSTRACT

The likelihood ratio (LR) is the "logically and legally correct" (Rose and Morrison 2009:143) framework for the estimation of strength-of-evidence under two competing hypotheses. In forensic voice comparison these considerations are reduced to the similarity and typicality of features across a pair of suspect and offender samples. However, typicality can only be judged against patterns in the *relevant population* (Aitken and Taroni 2004:206). In calculating numerical LRs typicality is assessed relative to a sub-section of that population.

This study considers issues of variability relating to the delimitation of reference data with regard to the number of speakers and number of tokens per speaker. Using polynomial estimations of F1 and F2 trajectories from spontaneous GOOSE (Wells 1982), LR comparisons were performed against a reference set of up to 120 speakers and up to 13 tokens per speaker. Results suggest that mean same-speaker LRs are robust to such variation until the reference data is limited to small numbers of speakers and tokens. However, variance and severity of error may be continually reduced with the inclusion of more data.

The definition of the *relevant population* with regard to regional variety is also assessed. Results for LRs are presented across four sets of test data where only one set matches the reference population for accent. In the absence of differences in levels of within-speaker variation, the magnitude of same-speaker LRs and severity of error are shown to be considerably higher for the 'mismatch' test sets. However, results indicate that the removal of regionally-defining acoustic information may reduce the effect of accent divergence between the evidential and reference data. This has positive implications for the application of the numerical LR approach.

CONTENTS

PAGE

LIST OF TABLES

PAGE

LIST OF FIGURES

PAGE

ACKNOWLEDGEMENTS

I wish to thank my supervisor Paul Foulkes for all of his encouragement, support and guidance not only during the ‘dissertation months’, but throughout this year of ups and downs. I am also indebted to him and Dom Watt for initially sparking my interest in forensic phonetics in the final year of my undergraduate degree. A huge debt of gratitude is also owed to Ashley Brereton (Dept. of Maths, University of Liverpool). Ash’s willingness to explain mathematical concepts, answer my questions and essentially teach me how to use MatLab goes above and beyond the level of friendship and without his help this dissertation would literally have been impossible. Special thanks go to Phil Harrison for writing the MatLab LR loop function. I can’t even imagine how I would have gone about performing the 273,408 LR calculations in this study without it! It is also necessary to thank Jen Hay and others involved in ONZE for allowing me to use their 96,372 formant measurements. The prospect of extracting this much data by hand was a particularly daunting one. Similarly, thanks to Bill Haddican for providing the Manchester, York and Newcastle data, and to Hazel Richards for putting up with my continual requests for more sound files. For making me look at my results in a different light I thank Rich Rhodes, and although we only met for the first time relatively recently, I am also grateful to Frantz Clermont for his encouragement and inspiration.

On a personal level, the unwavering love and support (both emotional and financial) of my mum and dad has been invaluable and for that I am truly grateful. Thanks also go to my brother Phil. His words of support have perked me up when I’ve been down and I hope that I can be there for him in the same way that he has been there for me. Finally, I wish to thank Fiona for putting up with me and looking after me when times have been tough. Her patience and love have been instrumental not only in the completion of this dissertation, but throughout this year.

1.0 INTRODUCTION

In the UK voice comparison (FVC) accounts for the vast proportion of casework undertaken by forensic speech scientists (c.70%, French p.c.). Experts are typically presented with two samples (one incriminating sample containing the voice of the offender, the other containing the voice of the suspect) and asked to compare the speech patterns to assess the possibility that the same-speaker is present in both.

However, the *paradigm shift* (Saks and Koehler 2005) across forensic disciplines reflects a move towards the evaluation of such evidence within a framework which is more "scientific" (Morrison 2009a:1). The *shift* reflects the Court of Appeal's concern over "the logically correct evaluation and presentation" (Morrison 2009a:1) of expert evidence in R-v-Doheny and Adams [1996]. The court ruled that DNA evidence presented as posterior probability, i.e. an assessment of the hypotheses given the evidence $p(\text{H}|\text{E})$, committed the prosecutor's fallacy (Thompson and Schumann 1987) and gave undue weight to the expert's testimony.

With DNA "setting the standard" (Baldwin 2005:55) in forensic science, the *Doheny* Court's assertion that "the scientist should not be asked his opinion on the likelihood that it was the Defendant who left the crime stain" (Rose 2007a) emphasises the validity of considering the probability of the evidence rather than the hypotheses. In line with the Court's judgement, the Bayesian framework, and specifically the likelihood ratio (LR), are now widely accepted as the "logically and legally correct" (Rose and Morrison 2009:143) approach for the estimation of strength-of-evidence. Despite the Court of Appeal ruling in R-v-T [2010], Bayes' theorem forms "the model for a scientifically defensible approach in forensic identification science" (Gonzalez-Rodriguez et al 2007:2104) towards which the current *shift* is moving.

The Bayesian approach provides a framework for the assessment of evidence across a criminal trial and its facility for the incorporation of "multiple piece(s) of evidence" makes it "a very attractive measure for (FVC)" (Kinoshita 2002:300). The odds form of Bayes' theorem is:

$$\boxed{\begin{array}{c}\dfrac{p(\mathrm{H_p})}{p(\mathrm{H_d})}\\[6pt] \textbf{Prior Odds}\end{array}} \quad \mathrm{x} \quad \boxed{\begin{array}{c}\dfrac{p(\mathrm{E}|\mathrm{H_p})}{p(\mathrm{E}|\mathrm{H_d})}\\[6pt] \textbf{Likelihood Ratio}\end{array}} \quad = \quad \boxed{\begin{array}{c}\dfrac{p(\mathrm{H_p}|\mathrm{E})}{p(\mathrm{H_d}|\mathrm{E})}\\[6pt] \textbf{Posterior Odds}\end{array}}$$

adapted from Rose (2004:3)

Where	**p**	= probability
	E	= evidence
	|	= ‘given’
	H_p	= prosecution hypothesis (i.e. same-speaker)
	H_d	= defence hypothesis (i.e. different-speaker)
	🟩	= trier-of-fact
	🟥	= expert

The prior odds reflect the trier-of-fact’s assessment of the probability of the hypotheses before the introduction of evidence (see Cohen 1982; Redmayne 1998). The prior odds are modified by planks of evidence expressed within a LR framework to establish posterior odds. The posterior odds are concerned with what Lynch and McNally refer to as the “ultimate issue” (2003:96) of innocence or guilt: an assessment of the probability of the hypotheses given the weight of evidence.

Central to Bayes’ theorem is the LR: an estimation of strength-of-evidence based on its probability of occurrence given H_p divided by the probability of its occurrence given H_d. In FVC, the numerator is equated to the similarity of samples, while the denominator is concerned with their typicality in the *relevant population* (Aitken and Taroni 2004:206). Numerical LRs are calculated using acoustic data, where suspect and offender samples are compared against a sampled sub-section of the relevant population. The outcome is a value centred on one, such that LRs of >1 offer support for H_p whilst LRs of <1 offer support for H_d (Rose 2004:4). The magnitude of the LR determines how much more likely the evidence would be given H_x than H_y (Evett et al 2000). A LR of five is interpreted as: ‘the evidence is five times more likely assuming the hypothesis that the same-speaker was involved than assuming the hypothesis that different-speakers were involved’.

Robertson and Vignaux maintain that “expert evidence should be restricted to the (LR) given by the test or observation of its components” (1995:21). Similarly, Rose describes the role of Bayes’ Theorem in FVC as “non-negotiable” (2004:3). Logical and legal arguments for these claims are numerous.

The distinction between assessing the hypotheses and the evidence ensures that the roles of trier-of-fact and expert are separated. In restricting the expert to the LR, the *ultimate issue* remains the preserve of judge and jury. This also prevents the expert from expressing “inappropriate” $p(\mathrm{H}|\mathrm{E})$ conclusions based on “information and assumptions from sources other than an objective scientific evaluation of the known and questioned samples” (Morrison 2009a:4). Moreover, it is not only inappropriate but logically impossible for the expert to provide posterior probability. This is because $p(\mathrm{H}|\mathrm{E})$ is dependent on prior odds, which are determined by the trier-of-fact and therefore inaccessible. Finally, the LR conforms with the US Supreme Court ruling in Daubert [1993] which requires that theories and techniques have been tested and “actual or potential error rates (…) considered” (Rose 2002:121).

However, the LR approach itself does not ensure a reliable estimation of strength-of-evidence. Numerical LRs are necessarily affected by the input data, such that the removal of an individual from the reference population will vary the outcome. However, very little is known about the robustness of LRs to systematic variability in the reference population and to truly satisfy *Daubert*, it is essential that error, and LR performance more generally, are assessed under such conditions. Therefore, this study presents an investigation into the issue of variability in the definition and delimitation of the reference population with a focus on population size and accent mismatch.

2.0 BACKGROUND

2.1 Expression of conclusions in forensic speech science (FSS)

Gold (2011) surveyed 36 forensic speech scientists to investigate how experts frame conclusions in FVC casework. Results reveal a lack of consensus. The highest proportion of experts (39%) currently use a classical probability framework of the kind described in Baldwin and French (1990:10). However, there has been an increasing impetus for FSS to move away from such $p(\mathrm{H}|\mathrm{E})$ statements. Initial concerns over classical probability scales were raised in Broeders (1999) and subsequently Champod and Evett (2000) and Champod and Meuwly (2000) argued for the assessment of FVC evidence within a LR framework. In countries where probability scales are in operation there is also growing support for the Bayesian approach (Jessen 2011).

However, contrary to the cross-disciplinary *paradigm shift*, just four experts in Gold (2011) have adopted the numerical LR framework. This reflects concerns over the practical implementation of numerical LRs (Nolan 2001). French and Harrison claim that a quantitative approach is primarily precluded by “the lack of demographic data” (2007:142). This is emphasised by Rose as “one of the main factors that make the accurate estimation of LRs problematic” (2004:4). As a solution, Rose (2007a) proposes that experts collect reference data themselves.

To ensure reliable strength-of-evidence from such data two issues must be addressed. Firstly, the relevant population needs to be defined and secondly, the sub-section of the population must be delimited. However, the task of ensuring that reference data is representative, relevant and reliable is not straightforward. Indeed, Morrison claims that “the only principled objections (to LRs) (…) (are) related to defining the relevant population to sample in order to calculate (…) typicality” (2009a:13).

2.2 The ‘defence hypothesis’

According to the LR itself, the reference population is determined by the defence hypothesis (H_d). However, where H_p is likely to be a straightforward submission that two samples contain the voice of the same individual (the defendant), H_d is more complicated. Broeders claims that the LR approach is only feasible “where one or more scientific alternative hypotheses can be formulated” (1999:239). The implications are emphasised by Robertson and Vignaux who affirm that “it is often difficult if not impossible to determine the probability of the evidence with a vague and ill-defined hypothesis" (1995:31). In many jurisdictions the defence will offer simply a 'different-speaker' hypothesis or no alternative at all. Rose claims that in such cases H_d may be assumed to be "another same-sex speaker of the language" (2004:4).

Coleman and Walls define the relevant population as "those persons who could have been involved (ignoring other priors)" (1974:276). Smith and Charrow propose a modification, claiming that typicality should be assessed against "the smallest population known to possess the culprit as a member" (1975:556). Lenth (1986) justifies the need for a more specific hypothesis than 'it was a different speaker', arguing that the LR model assumes that "the alleged source of the evidence is a random selection from those persons having the *required characteristics*" (in Aitken 1991:58). Therefore, only when "there is no evidence to separate the perpetrator from the (…) population" or "results can be regarded as independent of variations in sub-groups" (Robertson and Vignaux 1995:36) should a 'general' population be used. Given the inferences which may be made about an individual on the basis of his speech patterns, in most cases the relevant population may be defined more narrowly than the default assumption in Rose (2004). Such inferences relate to regional background, age and class amongst others.

However, in reality it is not possible to define the reference population on the basis of all social groups to which the perpetrator belongs. In other areas of forensic science this issue may be resolved by *logical relevance* (Kaye 2004). In DNA casework H_d is determined in part by ethnicity since the frequency of certain strands is variable according to ethnic groupings. As offender ethnicity cannot be inferred from DNA alone, a multiple-H_d approach is adopted in which the jury is presented with LRs according to the ethnic grouping of the reference population (Kaye 2008). Since ethnicity affects the magnitude of the LR it is considered logically relevant.

2.3 LR-based studies in FSS

Previous FSS research reveals a lack of consensus regarding the definition of the relevant population. Kinoshita (2001, 2002) represent the first studies to consider the LR performance of traditional phonetic features. In the (2002) study, intrinsic same-speaker (SS) and different-speaker (DS) comparisons (where speakers function simultaneously as test and reference data) based on ten speakers of Japanese were conducted using F1-F4 of five vowel phonemes. Results offered useful strength-of-evidence, despite the small number of speakers. Rose, Osanai and Kinoshita (2003) investigated LR performance based on a multivariate analysis of cepstral coefficients and formants from a nasal, voiceless fricative and vowel. Again intrinsic testing was performed based on 60 male Japanese speakers from 11 prefectures. The test data was also uncontrolled for age.

Alderman (2004) assessed the viability of the Bernard data (Bernard 1967, 1970) as a reference distribution for FVC. By focussing on the role of the reference data, the study represents the first step towards tackling this "deficiency in the (LR) method" (Rose 2002:320). Whilst the age of test speakers is restricted, Alderman claims that three accent groupings based on 'broadness' in Australian English are adequate "for a number of difference variations of (H_d) based on accent" (2004:511). However, the viability of the Bernard data is determined on the strength-of-evidence and LR_{test} (the ratio of SS pairs achieving LRs of >1 to DS pairs achieving LRs of <1) achieved for individual phonemes, rather than the comparative performance of accent groupings.

Extrinsic testing, involving an independent reference set of 166 speakers was performed by Rose et al (2006) based on formant trajectories of /aɪ/. Sound change in the form of a lowering of F2 at the onset and increase in F1 at the offset during the 30 years which separate the reference and test recordings is claimed to be "important" (2006:330). However, no predictions are made as to the expected effects of such change and no reference is made to this when discussing the results.

Morrison's (2008) study of /aɪ/ explicitly acknowledges potential sources of variation which may affect LR output. Morrison highlights the age range of 19 to 64 years, small number of speakers (intrinsic testing using 27 speakers) and the presence of "some dialect variation" (2008:251) as potential shortcomings of the method employed. Zhang et al's (2011) study of Chinese /iau/ displays a more active consideration of the issues in Morrison (2008). Despite the small reference population (20 speakers), there is greater control over regional background and age. Therefore, the speakers in Zhang et al are a more homogeneous set than those previously investigated in numerical LR studies. Finally, regional variety is also raised by Rose as a factor which makes a "minimal contribution to the good results" (2011:1721) based on formants and cepstral coefficients from two tokens of five vowel phonemes in Japanese.

Previous studies reveal largely only an implicit awareness of the sources of variation which may affect the reliability of strength-of-evidence. However, increasingly researchers are acknowledging these issues and occasionally controlling for them. Loakes (2006) represents the most forceful call for greater controls over the definition of the reference population. Loakes claims that if the sample is not representative, "the resulting LR will in turn be misrepresentative" (2006:197) and suggests that along with "speaker sex and accent (…) tighter constraints on social variables might also need to be applied to population selection" (2006:198).

Systematic research into the effect of variability in the reference data is offered by Ishihara and Kinoshita (2008) and Hawkins and Clermont (2009). Both found that the number of reference speakers can dramatically affect LR output, especially when this number is limited. Further, Hawkins and Clermont (2009) show a broadening of 99% confidence intervals as the number of reference speakers is reduced. In automatic FVC, regional variety has also received some limited attention. Harrison and French (2010) assessed the outcome of LRs generated by BATVOX as a function of the make-up of the reference data. Results reveal that whilst the system is not accent dependent, there is "sensitivity to regional accents". Such sensitivity is likely to be exacerbated in the calculation of LRs based on traditional phonetic features which are expected to contain higher levels of accent-defining information than long-term spectral characteristics.

2.4 The present study

Despite the body of research in FSS conducted within a numerical LR framework, questions relating to the definition of the relevant population remain largely unanswered. Therefore, this study offers a preliminary exploration into the logical relevance of certain sources of variation on the outcome of LRs. The results of two studies into the effect of population size are presented. The first concerns the number of speakers and the second concerns the number of tokens per speaker. To address issues of the relevance of the population, this study also investigates the effect of mismatch between suspect and offender data and the reference data with regard to regional variety. SS and DS LRs are calculated on quadratic and cubic polynomial coefficients of F1 and F2 trajectories for GOOSE.

Given that variable input necessarily affects the numerical output, the primary concern here is the magnitude of LR differences as a consequence of variability and whether such patterns are systematic. The results are not intended to categorically determine how the reference population should be defined and sampled, but rather to highlight issues of logical relevance in FVC. The limitations of the study are discussed at §3.7.

3.0 METHODOLOGY

Extrinsic LR testing is adopted, whereby mock suspect and offender samples (test) are assessed against separate reference data. Extrinsic evaluation allows factors in test and reference sets to be varied independently and "generally provide(s) more realistic and defensible data" (Rose et al 2006:329).

3.1 Segmental material

GOOSE was analysed due to the availability of existing acoustic data. The limitation of GOOSE is that the four test varieties (New Zealand (NZE) (Canterbury), Manchester, Newcastle and York) are predicted to display regionally-defined variation. Hughes et al (2011) found GOOSE-fronting at the onset in Manchester English while Easton and Bauer (2000) claim that GOOSE in NZE is undergoing change involving fronting and diphthongisation. Watt's (1998) auditory analysis of Newcastle GOOSE suggests a maximally-fronted realisation of [ʉ], but that more commonly /u/→[uː,o̝]. Such differences mean that the impact of accent mismatch is, to an extent, predictable.

However, GOOSE is not a 'stereotype' (Labov 1971) of any of the varieties investigated. Individuals are expected to display within-group variation, such that the patterns predicted by the literature are unlikely to be consistent across all speakers. Since GOOSE-fronting in English varieties (RP: Torgersen and Kerswill 2004, Hawkins and Midgley 2005; American English: Clarke et al 1995, Fridland 2008) is closely correlated with age, the use of younger speakers was intended to reduce regionally-defined F2 variation (acoustic correlate of fronting). Further, phonological patterns of increased F2 following /j/ and reduced F2 preceding /l/ (Ash 1996, Hall-Lew 2005, Flynn 2011) are expected to be consistent across test sets. Jones claims that in RP a "diphthongal pronunciation is particularly noticeable in final position" (1966:42). Therefore, tokens were coded for adjacent /j/ and /l/ and open or closed syllable status allowing greater control over within- and between-speaker variation.

3.1.1 The dynamic approach

Research suggests that a ‘dynamic’ approach characterising spectral properties of vowels across their duration offers greater speaker-discriminatory potential than ‘static’ measurements from the steady-state of formant trajectories (Greisbach et al 1995; Ingram et al 1996; Rodman et al 2002; Eriksson et al 2004). Nolan claims that whilst phonetic targets are defined by the speech community, transitions are “acquired through a process of trial and error” (1997:749). Formant trajectories have been investigated extensively within a numerical LR framework (Morrison and Kinoshita 2008; Morrison 2009b).

Data consisted of time-normalised measurements at +10% steps of F1 and F2 for GOOSE (McDougall 2004, 2006; Hughes et al 2009). As such, formant contours were defined by nine raw Hz values.

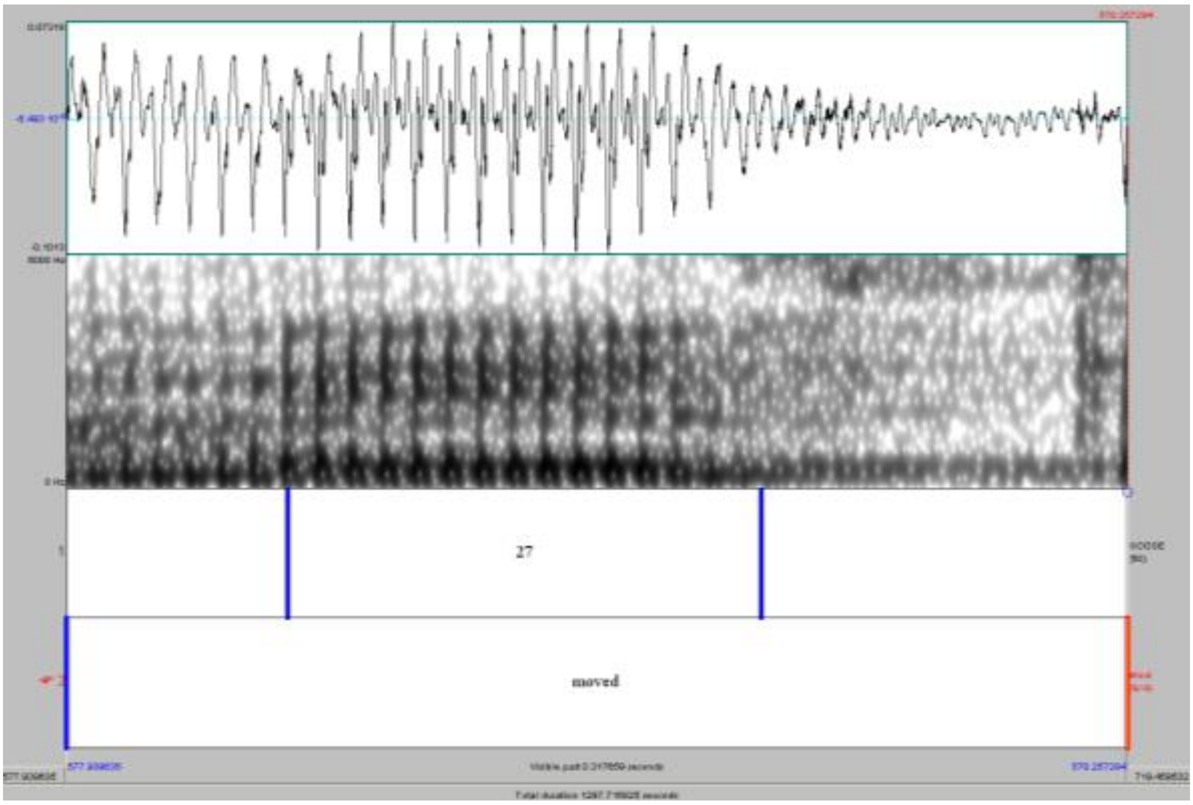

Figure 1 – *TextGrid of the word ‘moved’ with GOOSE delimited on tier 1 produced by speaker 1 in the Manchester test set (06:37) (A_D_ethno.wav)*

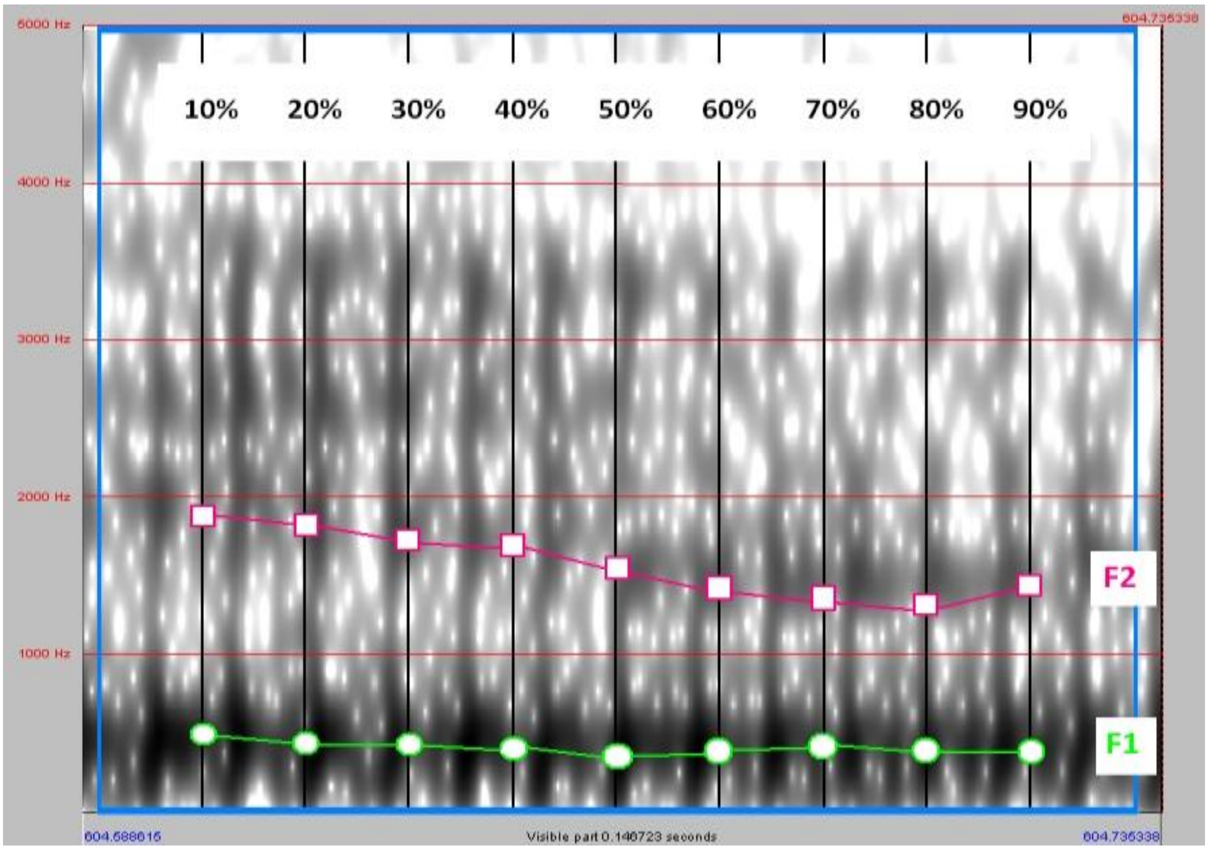

Figure 2 – *Spectrogram of GOOSE isolated from the lexical item 'doing' produced by speaker 1 in the Manchester test set (10:04) (A_D_ethno.wav) showing the location of +10% step markers at which F1 and F2 measurements were taken*

Manual extraction of formant data for Manchester, Newcastle and York was performed using a Praat script. Two-tiered TextGrids were created with tokens and words isolated on separate tiers. Procedures were employed to define the onset and offset of vocalic segments (appendix 1) and boundaries were moved to the nearest zero crossing. Errors were reduced by varying the maximum number of tracked formants (between 5.0-6.0 below 5.5kHz) and hand-correction.

For the NZE test and reference samples, formant data was auto-generated by running an adapted version of the script on force aligned (Sjölander 2003) audio and TextGrid files. However, the author had no access to these files, rendering visual inspection and manual error correction impossible.

3.2 Test sets

Four sets containing eight male speakers aged between 17 and 30 formed the test data. Sets were defined according to regional variety: NZE-Canterbury (ONZE), Manchester, Newcastle and York. LR comparisons were performed on SS and DS pairs, where the 'correct' outcome was known. The acoustic data used in this study is extracted from spontaneous speech.

3.2.1 Manchester

Manchester data was collected as part of the 'Comparative Study of Language Change in Northern Englishes' project (Haddican 2008-2013). The eight male speakers (aged 19-30/mean=21) were recorded in peer-group pairs for between 12 and 33 minutes (23-37 tokens/mean=31). The recordings had been digitised at a sampling rate of 44.1kHz and a 16-bit depth. Due to memory issues re-sampling at a rate of 11.025kHz was performed by the author using Sony Sound Forge 9.0.

3.2.2 Newcastle

Four recordings from the 'Phonological Variation and Change in Contemporary British English' project (Milroy, Milroy and Docherty 1994-1997) were used as Newcastle data. The recordings contained 48 to 64 minutes of conversation between pairs of young male speakers and had been digitised at a sampling rate of 16kHz and a 16-bit depth. Between 37 and 44 tokens per speaker (mean=41) were extracted to a separate audio file using Audacity 1.2.6 to avoid re-sampling.

3.2.3 York

The York data consisted of five speakers recorded in 1998 as part of the 'Roots of Identity' project (Tagliamonte 1996-1998) (York98) and three speakers from Haddican's (2008-2013) corpus (York08). The speakers were aged between 17 and 26 (mean=20). The author was provided with edited audio files containing isolated GOOSE tokens (37-40 tokens/mean=39). The York98 recordings had been digitised at a sampling rate of 22.05kHz and a 16-bit depth. The York08 data was sampled at 44.1kHz.

3.2.4 ONZE

The Origins of New Zealand English project (ONZE) consists of three corpora containing recordings of speakers born between 1850 and 1987. The present study utilises the Canterbury Corpus (CC) (Maclagan and Gordon 1999; Gordon et al 2007) which has been collected since 1994. CC contains 169 males from the Canterbury region grouped as younger (20-30) or older (45+) speakers.

Dynamic formant data for spontaneous GOOSE was auto-generated for all speakers in CC. As only date of birth information was provided, a lower cut-off for inclusion in the test set of 1970 was chosen to ensure that all speakers were between 20 and 30 years old when recorded. With this restriction in place, 74 speakers were eligible (10-92 tokens/mean=32). A screening process was developed to remove formant tracking errors and to identify the eight speakers with the lowest between-speaker variation.

In order to include a range of phonological conditions, speakers with fewer than 20 tokens after each screening-stage were omitted. A pass-band of between 250Hz-600Hz was implemented for F1. Values at any +10% step outside this range were considered measurement errors and tokens removed. The restrictions allow for considerable F1 variation, since Hay et al claim that NZE has a central GOOSE variant which is "linked with an off-glide" such that /uː/→[əʉ] (2008:24). Since the average male F1 for schwa is around 500Hz (Johnson 2003:96), an upper limit of 600Hz was considered sufficient to capture variation in vocal tract length, without accepting erroneous values.

The reliability of adjacent values within formant contours was assessed visually. Where deviations between +10% steps were considered questionable, the token was removed. Finally, univariate outliers were identified by calculating *between-speaker* z-scores, such that values ±3.29 standard deviations from the mean for each +10% step were removed (Tabachnick and Fidell 2007:73). The final eight speakers with the lowest mean z-scores were used as the ONZE test set (32-70 tokens/mean=56).

3.2.5 Within- and between-speaker variability

In LR calculations, similarity between SS and DS pairs is assessed in terms of between-sample variation. Therefore, both within- and between-speaker variation must be controlled across test sets to reduce the effect of quantifiable similarity and difference between suspect and offender data which may obscure results. To minimise within-speaker variation, *within-speaker* z-scores for all test speakers were calculated. Tokens containing values greater than ±3.29 were removed. Z-scores for the remaining tokens were categorised according to five phonological contexts:

Table 1 – *Phonological categorisation of GOOSE tokens and the maximum number of tokens in such contexts shared by every speaker in each of the test sets*

Phonological Context	Maximum Number of Tokens shared by all Test Speakers
j ___	6
___ l	1
non-j ___ non-l	4
j ___ #	2
non-j ___ #	4
	Total = 16 (17)

Given the need for an even number of tokens in each context in order to perform reliable comparisons, all ___l tokens were omitted. For each of the remaining tokens, z-scores were added together and ranked within phonological grouping. The six ‘j___’ tokens, four ‘non-j___non-l’ tokens, two ‘j___#’ tokens and four ‘non-j___#’ tokens per speaker with the lowest combined z-scores were used in LR calculations. Tokens for each speaker were assigned equally by phonological category to either the suspect or offender condition. This ensured that pairs of samples were comparable in terms of the number of tokens and range of phonologically predictable variation.

The lowest levels of variation are found in the ONZE set. This is a result of the availability of a considerable amount of acoustic data, providing greater freedom to reduce the group of potential speakers to those with minimal levels of variation. The highest between-speaker variation is found in the York set. Processes of sound change in the ten years between York98 and York08 may account for this.

Table 2 – *Mean within- and between-speaker variation across the duration of the both F1 and F2 trajectories according to test set together with % difference with ONZE for Manchester, Newcastle and York data (a breakdown of these values by +10% step is provided at appendix 2)*

	ONZE	**Manchester**	*%diff with* ONZE	**Newcastle**	*%diff with* ONZE	**York**	*%diff with* ONZE
Mean within-speaker SD (Hz)	111	128	+15.37	124	+11.27	123	+10.33
Between-speaker SD (Hz)	127	148	+16.07	135	+6.08	195	+53.28

3.3 Reference data

Auto-generated GOOSE data from CC was also used as reference data. With the exception of the eight ONZE test speakers, 161 males born between 1932 and 1987 were eligible for inclusion (5-111 tokens/mean=30). As before, the raw data contained numerous formant tracking errors.

Speakers with fewer than 10 tokens were removed from the analysis at each stage of the screening process. Restrictions on F1 of 250Hz-600Hz were implemented along with F2 restrictions of 750Hz-2400Hz. The range of permitted F2 variation accounted for maximally fronted and retracted realisations. Univariate outliers were identified using *between-speaker* z-scores.

Finally, all ___l tokens were removed. Given the inconsistency between speakers with regard to the number of tokens in each context, it was not possible to control for phonological conditioning and simultaneously ensure that speakers had the same number of tokens overall. Instead combined z-scores were used to rank tokens by speaker, such that tokens were included on the basis of minimal between-speaker variation rather than phonological context. Therefore, there is a divergence between the test and reference data in the proportion of tokens in each context. The resultant reference data consists of 120 speakers with a minimum of 10 tokens per speaker.

Table 3 – *Percentage of tokens in test sets and reference data in each of the four phonological contexts coded for*

Phonological Context	% of tokens in test sets	% of tokens across reference data
j ___	37.5	23.7
non-j ___ non-l	25.0	26.8
j ___ #	12.5	18.0
non-j ___ #	25.0	31.5

Table 4 – *Number of speakers in the reference set grouped according to the number of tokens per speaker*

Number of tokens per speaker	10+	15+	20+	30+
Number of speakers	120	85	43	19

3.4 Multidimensional speaker-space

The F1~F2 plot of mean trajectories across all phonological contexts (Figure 3) displays general regionally-defined patterns. However, the range of between-set variation is low with mean F1 spread over 100Hz and mean F2 spread maximally over 300Hz. For almost all speakers these values are within the range of intra-speaker variability. Therefore, despite the broad regionally-defined patterns, acoustic differences between sets are considered minimal.

To assess statistically how acoustic differences between test and reference data affect LR output, Euclidean distance was calculated in PASW 18. Euclidean distance quantifies the proximity between the test and reference data in the multidimensional speaker-space. The distance (D) between two speakers (x,y) is calculated by dividing the square root of the combined difference between the speakers' mean F2 and mean F1 by the number of input variables (N) (2 formants x 9 measurement points). This may be formalised as:

$$D = \frac{\sqrt{\sum_{i=1}^{N} (\bar{y}_{(F2)i} - \bar{x}_{(F2)i})^2 + (\bar{y}_{(F1)i} - \bar{x}_{(F1)i})^2}}{N}$$

adapted from Young (1985:651)

(Brereton p.c.)

For the test speakers, all 16 tokens were included as input. Distances relative to the reference data were calculated on the basis of the 10 tokens per speaker with the lowest combined z-scores for all 120 reference speakers.

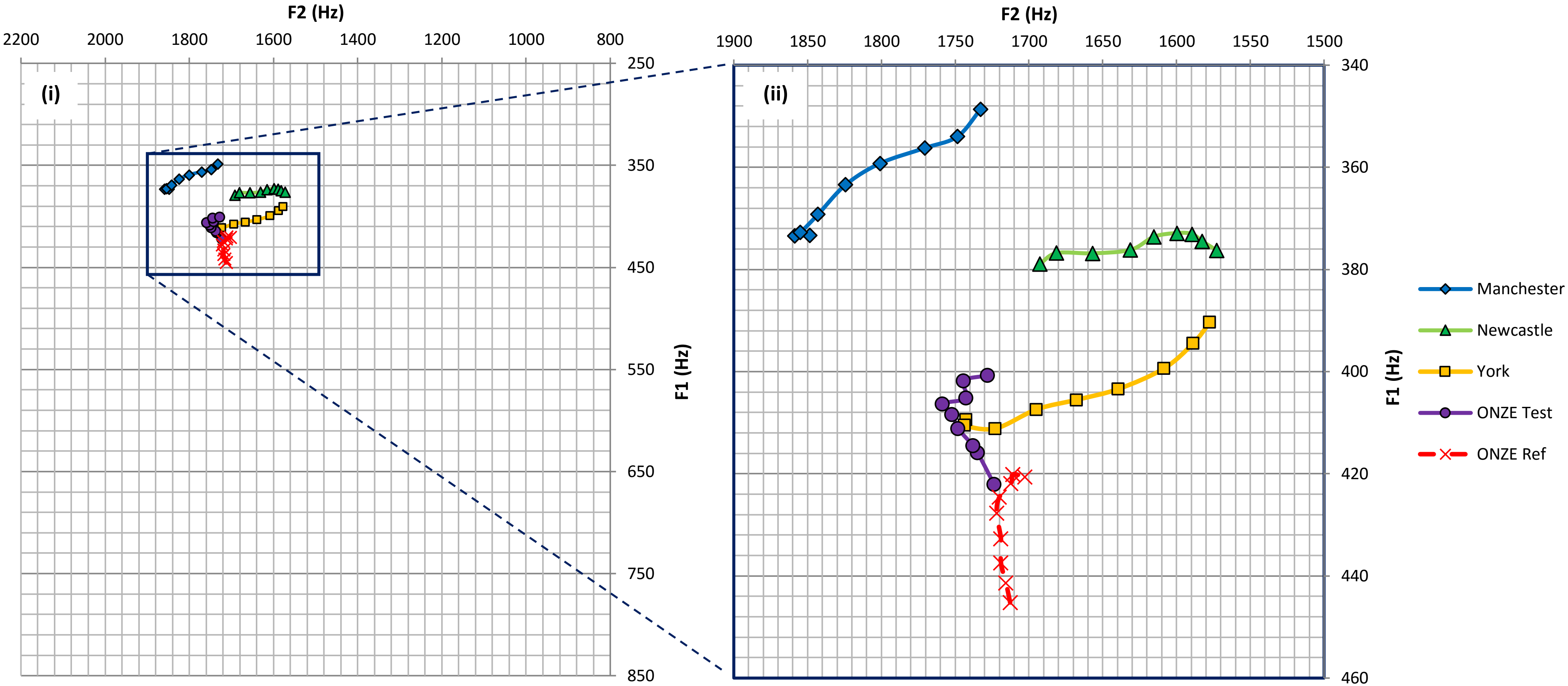

Figure 3 – *F1~F2 plots of mean GOOSE trajectories based on raw acoustic data for all tokens grouped according to regional variety (with ONZE test and reference data separated. Plot (i) is scaled according to expectated upper and lower limits for F1 and F2 across the complete vowel plane, whilst (ii) is the same information plotted on a much narrower F1 and F2 range so as to emphasise between-group differences*

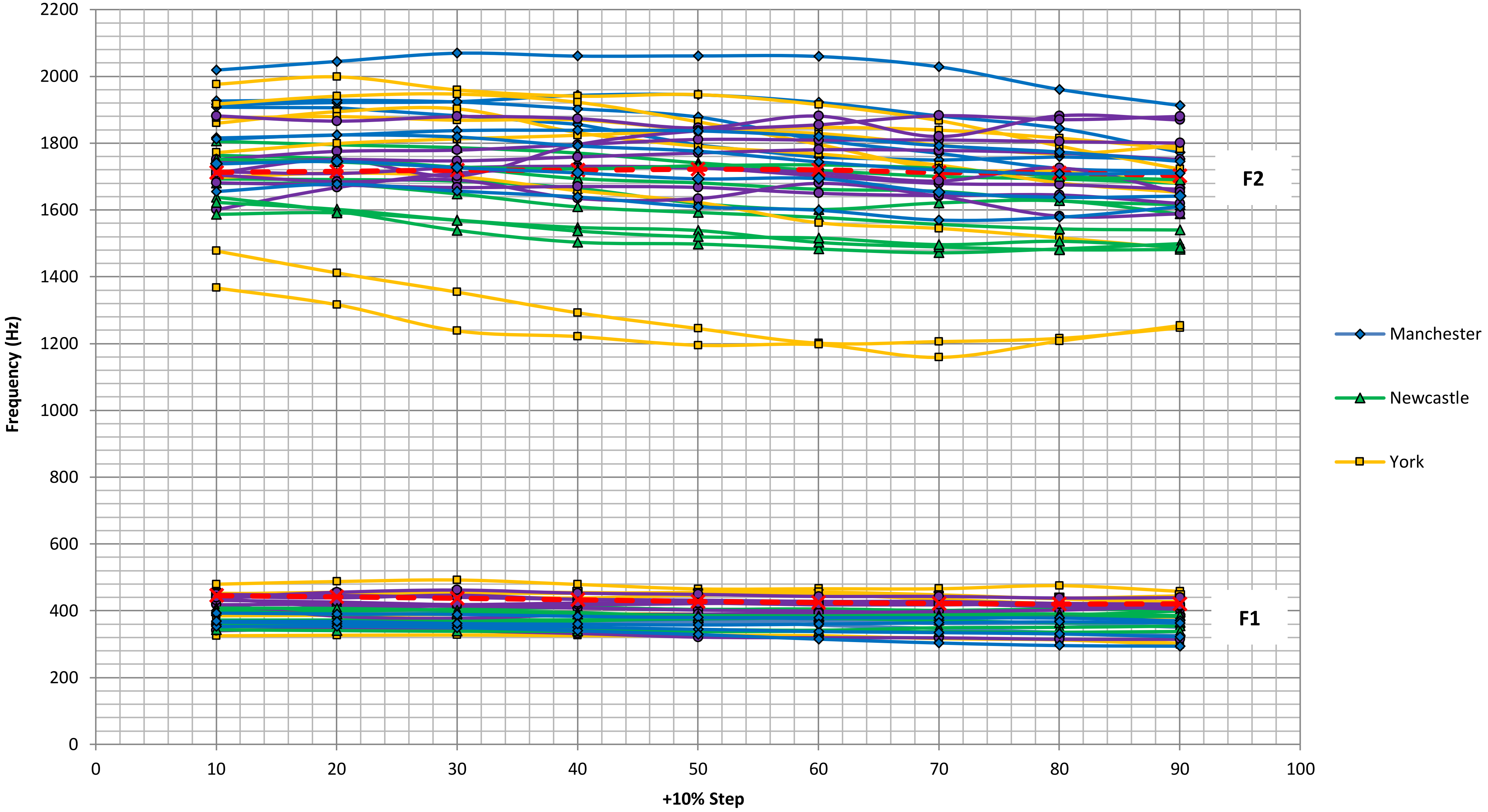

Figure 4 – *Mean F1 and F2 formant contours of GOOSE for the eight speakers in each of the regionally defined test data sets together with mean contours for the 120 speakers functioning as the reference set (based on 10 tokens per speaker)*

Distances were normalised by variable within a -1 to +1 range (appendix 3) and plotted on a two-dimensional visualisation of the speaker-space using multidimensional-scaling (MDS) (appendix 4) (Nolan et al 2007; Ferragne and Pelligrino 2010). Figure 5 displays the MDS plot for all test and reference data. Again, it is possible to group test sets broadly according to regional variety (except for York). The high level of between-speaker variation across the York data is reflected in the spread of individuals in the speaker-space.

Since typicality is assessed relative to the reference population, the magnitude of LRs is expected to increase with the distance between paired samples and the reference data. Further, in the absence of differences in within- and between-speaker variation, the MDS plot predicts that LR output will be grouped broadly according to regional variety. However, given that the distances reflect marginal acoustic divergence between groups, LR differences are not expected to be substantial.

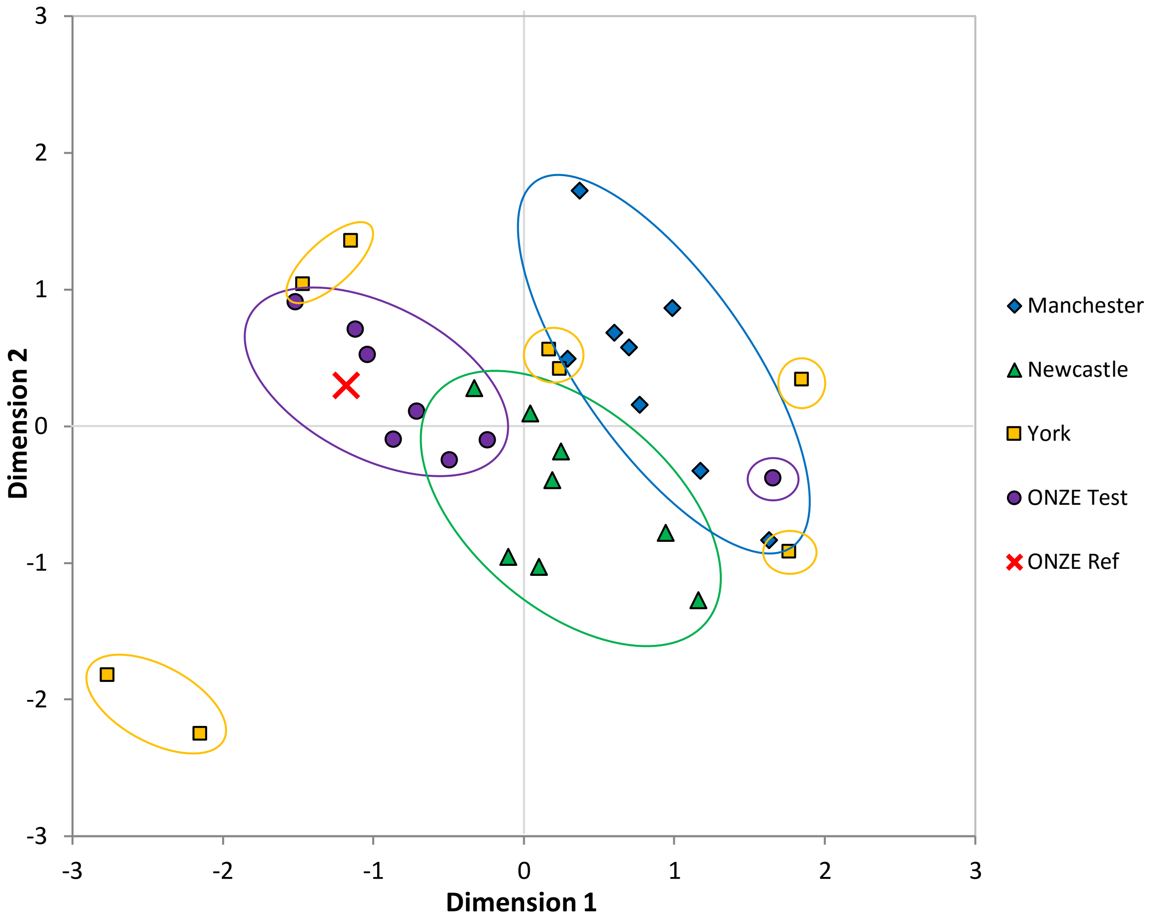

Figure 5 – *MDS plot of Euclidean distance between each of the speakers in the test sets and the reference data where co-ordinates are generated on the basis of combined acoustic measurements for that individual or group (in the case of the reference data) relative to each of the other speakers. Ellipses indicate broad variety groupings and are not statistically significant*

3.5 Polynomial curve fitting

Polynomial regression is a technique which approximates the non-linear relationship between two data sets. By fitting nth order polynomials, raw data is reduced to a series of coefficients (Seber and Wild 1989). A version of the MatLab polyfit function (adapted by Ashley Brereton) was used to fit quadratic and cubic polynomials to raw F1 and F2 curves in the form $\tilde{y} = a_0 + a_1x + a_2x^2$ and $\tilde{y} = a_0 + a_1x + a_2x^2 + a_3x^3$ where the a_i coefficient is calculated using a least-squared method, which reduces the sum of the squared residuals (ε) (Whittle 1983) (i.e. the distance between raw (y) and fitted data (yfit/ỹ)). The goodness of fit is determined by R^2 which increases towards one as a function of higher order polynomial complexity:

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{N} \varepsilon_i^2}{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2} \right)$$

Brereton (p.c.)

Non-linear regression is adopted primarily as a data reduction technique which is able to capture the shape of formant trajectories with between three (quadratic) and four (cubic) coefficients. Curve fitting also minimises the effect of formant tracking errors. Quadratic and cubic coefficients were used as input data for LR calculations.

McDougall (2006) used polynomial curves to model F1-F3 contours of Australian English /aɪ/ and found the speaker-discriminatory potential of cubic polynomials to be marginally higher than quadratic polynomials. Further, in McDougall and Nolan (2007) quadratic and cubic polynomials of /uː/ were shown to outperform quartic approximations in discriminant analysis. These results highlight the potential for “over-fitting” (Morrison 2008:253) and reducing discriminatory potential using higher order polynomials. Visual inspection of the raw data suggests that formant trajectories of GOOSE rarely display greater complexity than parabolic-shaped curves. Therefore, there is no expectation for better performance of cubic over quadratic polynomials.

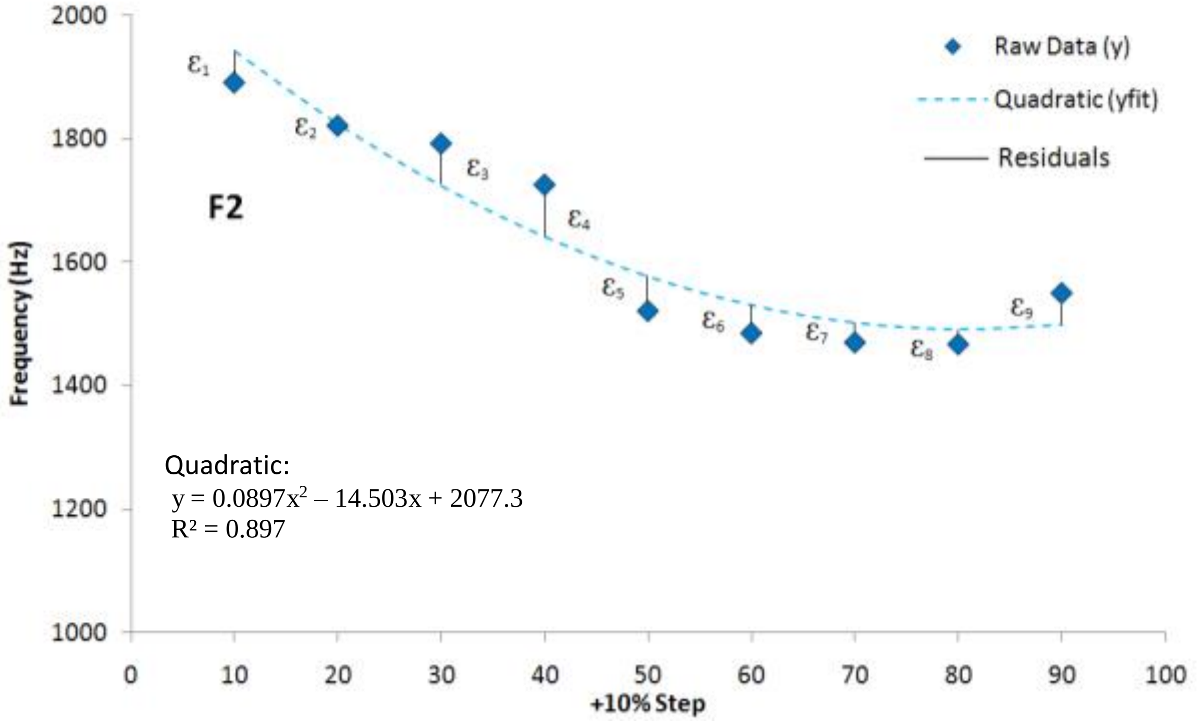

Figure 6 – *Scattergram of the raw F2 contour for GOOSE in 'doing' produced by Manchester test speaker 1 relative to the fitted quadratic polynomial curve and the residuals (ε). The least-squared method reduces the sum of the squared residuals*

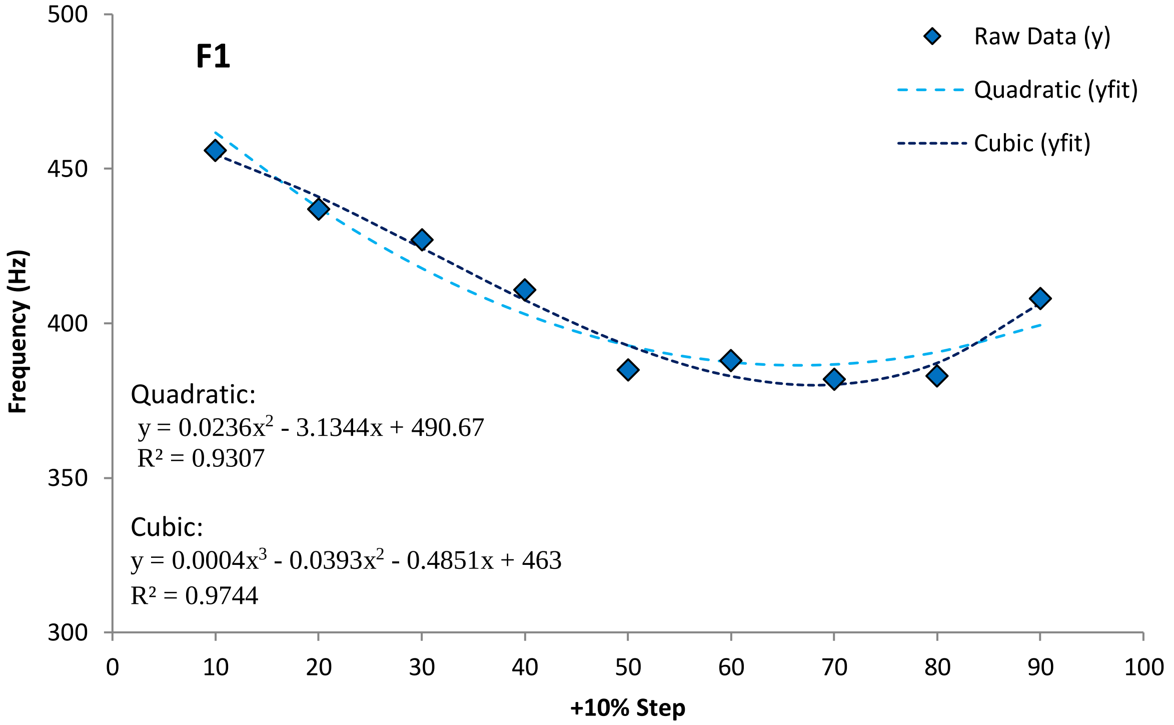

Figure 7 – *Scattergram of raw F1 contour for GOOSE in 'doing' produced by Manchester test speaker 1 relative to quadratic and cubic estimations of contour shape. Both model the trajectory well, with the cubic polynomial approach achieving a marginally higher R^2 value*

3.6 Calculation of LRs

LR comparisons were performed using a MatLab implementation of Aitken and Lucy's (2004) Multivariate Kernel-Density (MVKD) formula (Morrison 2007). Rose (2006b) claims that the advantage of the MVKD formula over Lindley's (1977) univariate LR is that it can account for within-segment correlations. This is significant given the interdependencies of frequency values across formant trajectories. Further, the MVKD formula is two-levelled in that it accounts for within- and between-speaker variation. Morrison claims that the inability of Lindley's LR to account for "occasion dependent within-speaker variation" (2008:97), makes the MVKD approach more suitable (although not ideal) for analysing speech evidence.

The MVKD formula assumes that within-speaker variability is normally distributed. Between-speaker variation is not assumed to be distributed normally and is estimated using kernel-density, which can account for skewed distributions. Despite the finding in Morrison (2011b) that a universal background model (GMM-UBM) (Reynolds et al 2000), which does not assume the normality of within- or between-speaker values, performs better than MVKD, its application has predominantly been tested on automatic systems. MVKD has been shown to provide useful strength-of-evidence across a number of studies using traditional acoustic-phonetic features (Rose 2006a, 2007b; Morrison 2008, 2009a) and is therefore employed here.

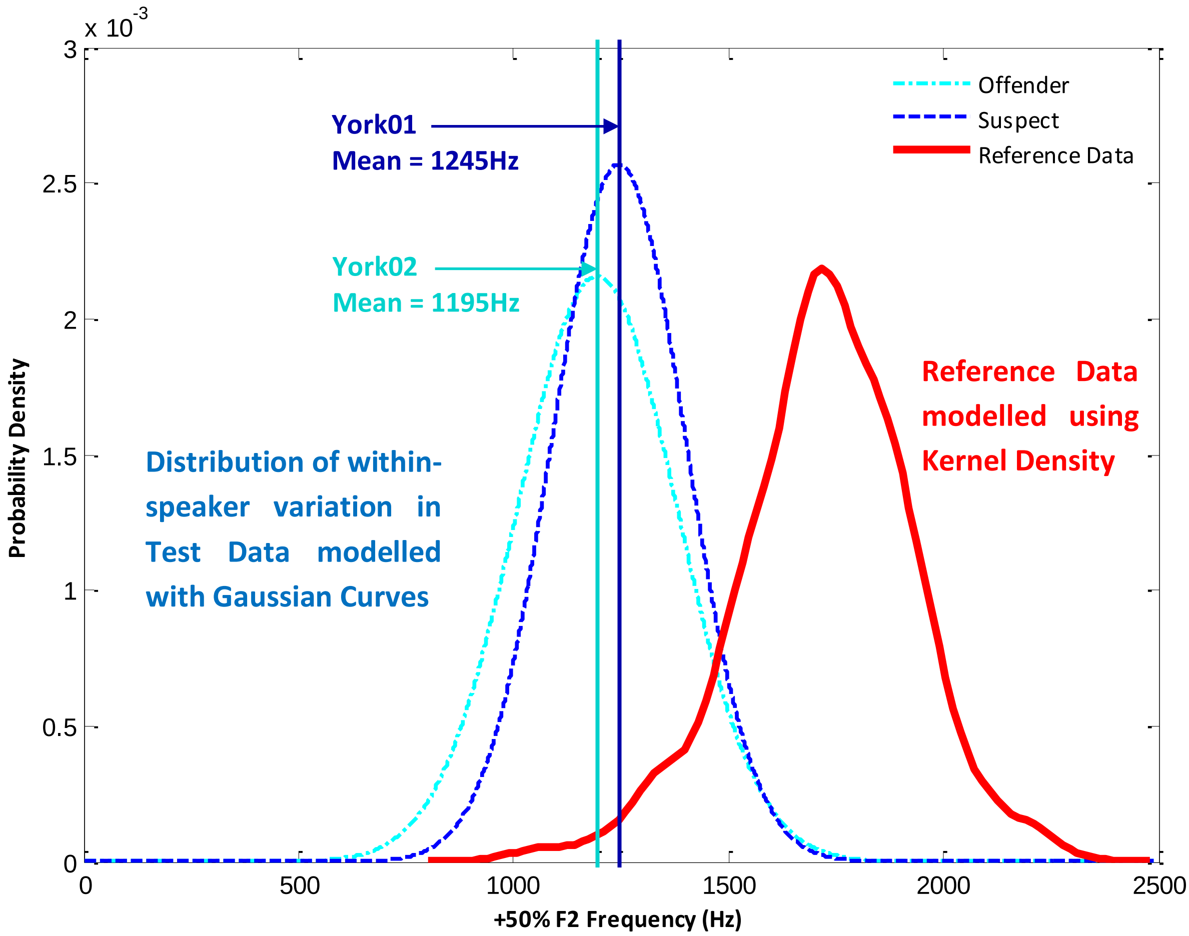

Figure 8 – *Diagrammatical representation of the MKVD formula's modelling of within- and between-speaker variation based on +50% measurement of F2 with the first (dark blue) and second (light blue) speakers from the York test data functioning as mock suspect and offender, assessed against the reference data based on 120 speakers and 10 tokens per speaker (red) (adapted from Rose 2007a)*

A MatLab script (*ss_ds_lrs.m*) developed by Philip Harrison was used to run multiple SS and DS LR calculations. The script divides speaker's data in half such that SS comparisons may be performed, consequently doubling the number of DS comparisons. Thus for each speaker there is a suspect and offender sample containing eight tokens equally matched for phonological context.

Raw LRs were transformed using natural and base$_{10}$ logarithms. This centres the turning point between support for H_p and H_d on zero, improving interpretability and ensuring that positive and negative values are scaled symmetrically. Log transforms also normalise distributions which are skewed by very high, infrequent values.

3.7 Limitations

The lack of non-contemporaneous samples means that suspect and offender data for each speaker was extracted from a single recording. In this regard the study does not match real forensic casework since occasion-to-occasion variability is removed and speaking style is constant. The lack of access to sound and TextGrid files for the ONZE data is also a substantial limitation, since the accuracy of segmental boundaries is reduced for force aligned TextGrids compared with manually delimited data. Further, procedures implemented to remove measurement error do not guarantee reliability. They serve to identify clear outliers, rather than more subtle tracking errors.

The inconsistency of speakers in the reference data with regard to the number of tokens in equivalent phonological contexts is likely to impact on the outcome of the §4.1 experiments and is considered when analysing the results. Finally, although in FVC casework there may only be a limited amount of segmental material available, for identifying systematic effects of variability on LRs it would be preferable to have increased numbers of speakers and tokens in the test and reference sets.

4.0 RESULTS

The following sections detail the results of two sets of experiments into the effect of variation on the outcome of numerical LRs. The first investigates issues of population size. The second considers how accent may affect strength-of-evidence. LRs are considered with reference to Champod and Evett's (2000:240) verbal scale:

Table 5 – *Verbal expressions of raw and log_{10} LRs according to Champod and Evett's verbal scale (2000:240)*

Raw LR	Log_{10} LR	Verbal expression
1-10	1	Limited evidence
10-100	2	Moderate evidence
100-1000	3	Moderately strong evidence
1000-10000	4	Strong evidence
>10000	5	Very strong evidence

4.1 Variation in the reference data

4.1.1 Number of speakers

This experiment uses quadratic and cubic F1-F2 coefficients to assess the effect of variable numbers of reference speakers. Test sets were combined such that the LR function performed 32 SS and 992 DS comparisons per condition. The function was initially run using 120 speakers and 10 tokens per speaker as reference. In the conditions which followed one speaker was removed consecutively, starting with the eldest. The smallest number of reference speakers investigated was 10.

4.1.1.1 Quadratic, F1-F2

Figure 9 displays mean $\log_{10}$ LRs and ± standard deviation (SD) based on quadratic coefficients according to the number of reference speakers. Aside from micro fluctuations, SS means remain robust to variation, especially when the population size is large, displaying a move in categorical strength-of-evidence from limited support for H_p to limited support for H_d only when fewer than 14 speakers are used as reference.

However, marked fluctuation for DS pairs begins at 45 speakers, whilst the SS mean remains stable until 20 speakers. Below this point strength-of-evidence decreases, compared with Hawkins and Clermont (2009) who found that mean SS LRs increased as population size decreased. A potential explanation is that speakers were removed according to age. Therefore test and reference sets will become more homogeneous as the population declines, meaning SS pairs are likely to become more typical.

Further, some pairs of samples are more sensitive to population variation than others. Between 110 and 109 speakers, the SS $\log_{10}$ LR for Speaker1 from Manchester increases from 0.85 to 2.12, the equivalent of limited and strong support for H_p. In terms of raw LRs, this represents the difference between seven and 133 times more likely. The lack of phonological conditioning in the reference data may explain why the removal of certain individuals has greater influence on LRs.

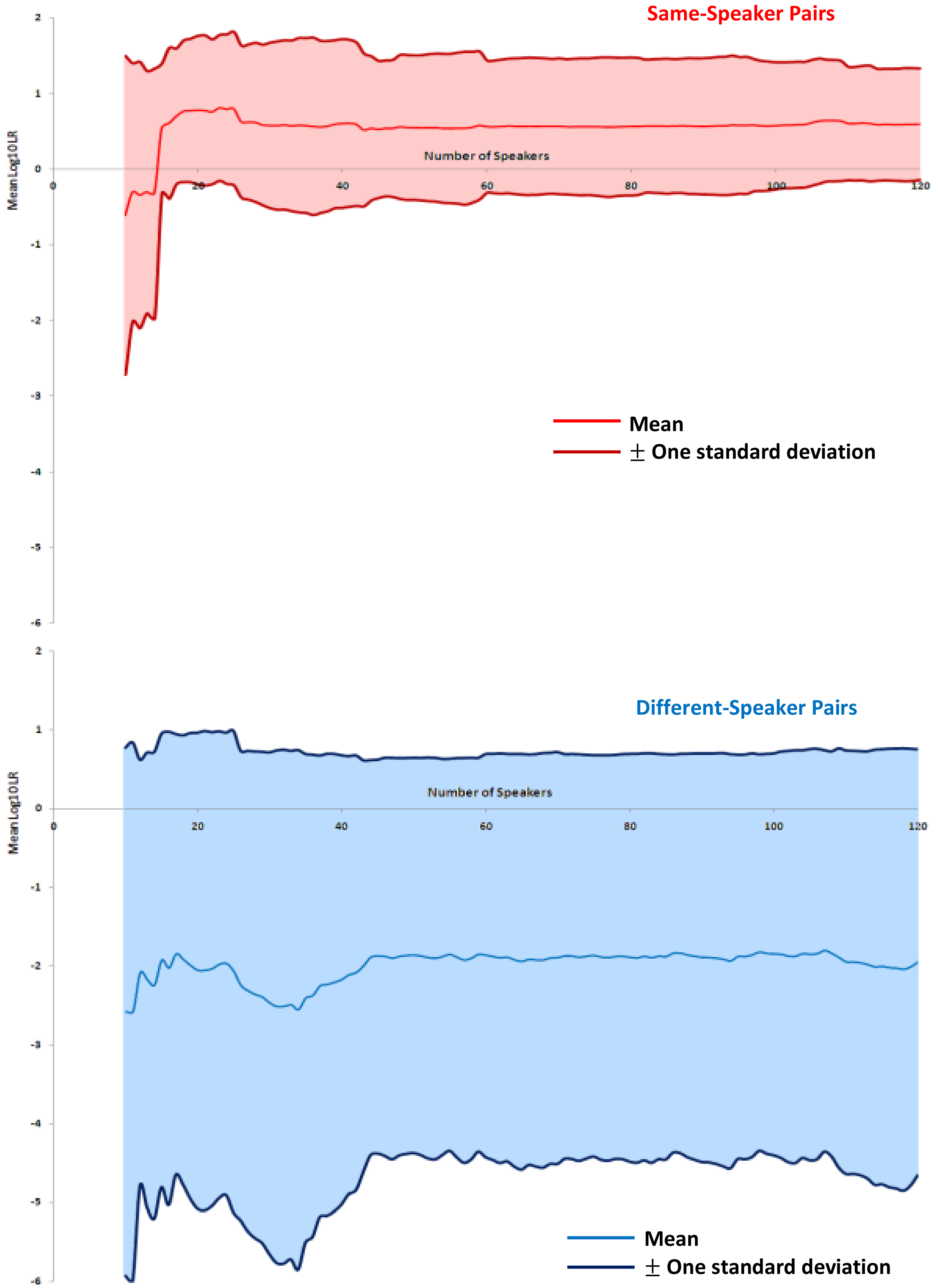

Figure 9 – *Mean* $\log_{10}$ *LRs (lighter lines) based on quadratic polynomials for same- (red) and different-speaker (blue) pairs where all speakers from ONZE, Manchester, Newcastle and York were used as test data according to the number of speakers in the reference data. Dark lines indicate ± one standard deviation from the mean*

Consistent with Hawkins and Clermont (2009), Figure 10 reveals a significant inverse correlation between SD and population size. However, dramatic effects only appear when the number of speakers is small. Whilst a gradual increase in SD is found across SS comparisons until 15 speakers, DS pairs actually display a decrease between 120 and 49 speakers.

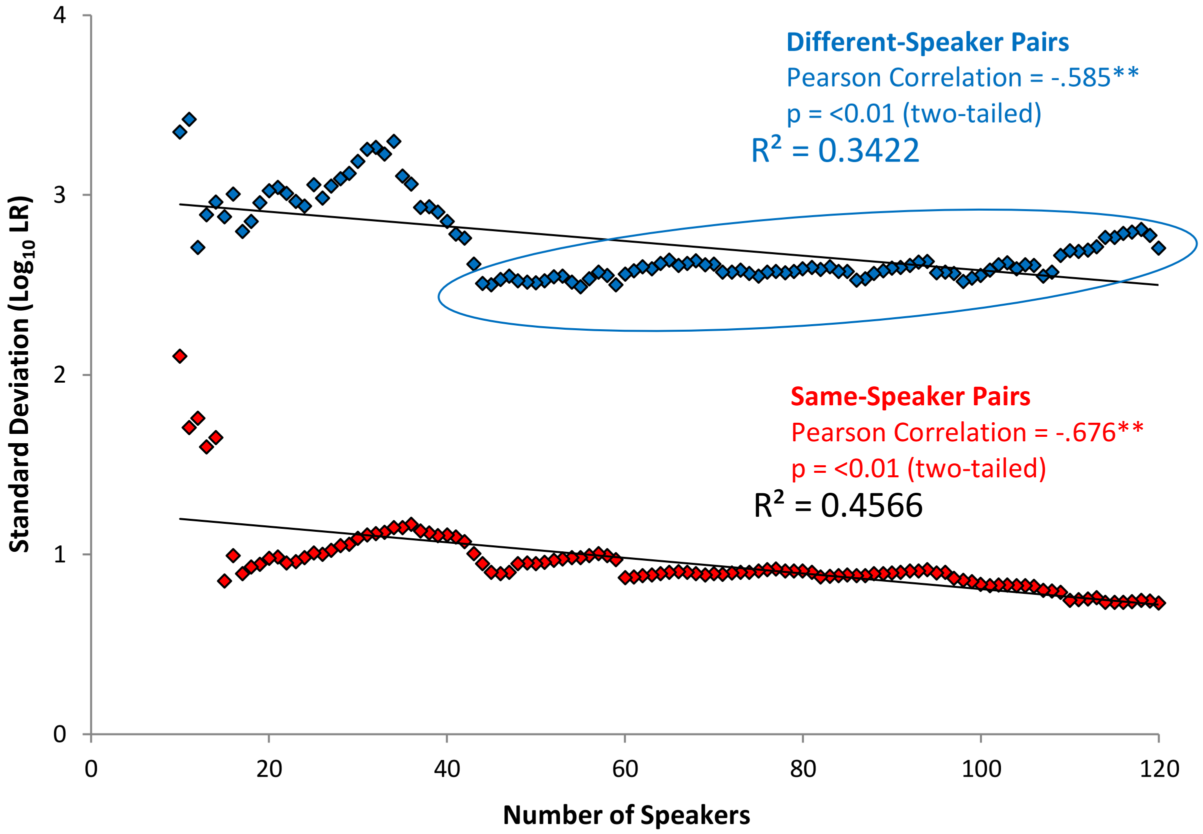

Figure 10 – *Scattergram of standard deviation of $\log_{10}$ LRs for SS and DS pairs relative to the size of the reference population. Linear trend lines have been plotted and significant correlations indicated. The decrease in SD for DS pairs between 120 and 49 speakers is marked in blue*

As is common in automatic FVC (Ramos Castro 2007), severity of error was assessed using log-LR cost (C_{llr}). C_{llr} is a Bayesian error metric appropriate for quantifying how well the system offers LR output aligning with prior knowledge of whether samples were produced by the same or different speakers. The ability of C_{llr} to "capture the gradient goodness of a set of likelihood ratios derived from test data" (Morrison 2009a:6) is preferred over "error-based" "hard detectors" (Brümmer and du Preez 2006:230), such as equal error rate (EER), which deal with categorical decisions of (in)correct acceptance and rejection. Previous studies have shown that C_{llr} is appropriate for speech evidence (Morrison and Kinoshita 2008; Morrison 2011a):

$$C_{\mathrm{llr}} = \frac{1}{2}\left(\frac{1}{N_{\mathrm{ss}}}\sum_{i=1}^{N_{\mathrm{ss}}} \log_2\left(1 + \frac{1}{\mathrm{LR}_{\mathrm{ss}_i}}\right) + \frac{1}{N_{\mathrm{ds}}}\sum_{j=1}^{N_{\mathrm{ds}}} \log_2(1 + \mathrm{LR}_{\mathrm{ds}_j})\right)$$

from Morrison (2009:2391)

Where N_{ss} = No. of SS pairs
N_{ds} = No. of DS pairs
LR_{ss} = LR from SS pairs
LR_{ds} = LR from DS pairs

Error was calculated using the *cllr.m* function in Brümmer's FOCAL toolkit[1] with natural log-LRs as input. C_{llr} is closer to zero when error is low. Values nearing one are considered poor, whilst values of above one indicate very poor performance (van Leeuwen and Brümmer 2007:343-344).

Figure 11 displays a significant correlation between C_{llr} and the number of reference speakers, with C_{llr} rising markedly when the population drops below 20 speakers. However, for all conditions where the population size was greater than 15, values of <1 were recorded.

[1] http://sites.google.com/site/nikobrummer/focal (Downloaded: 3rd June 2011)

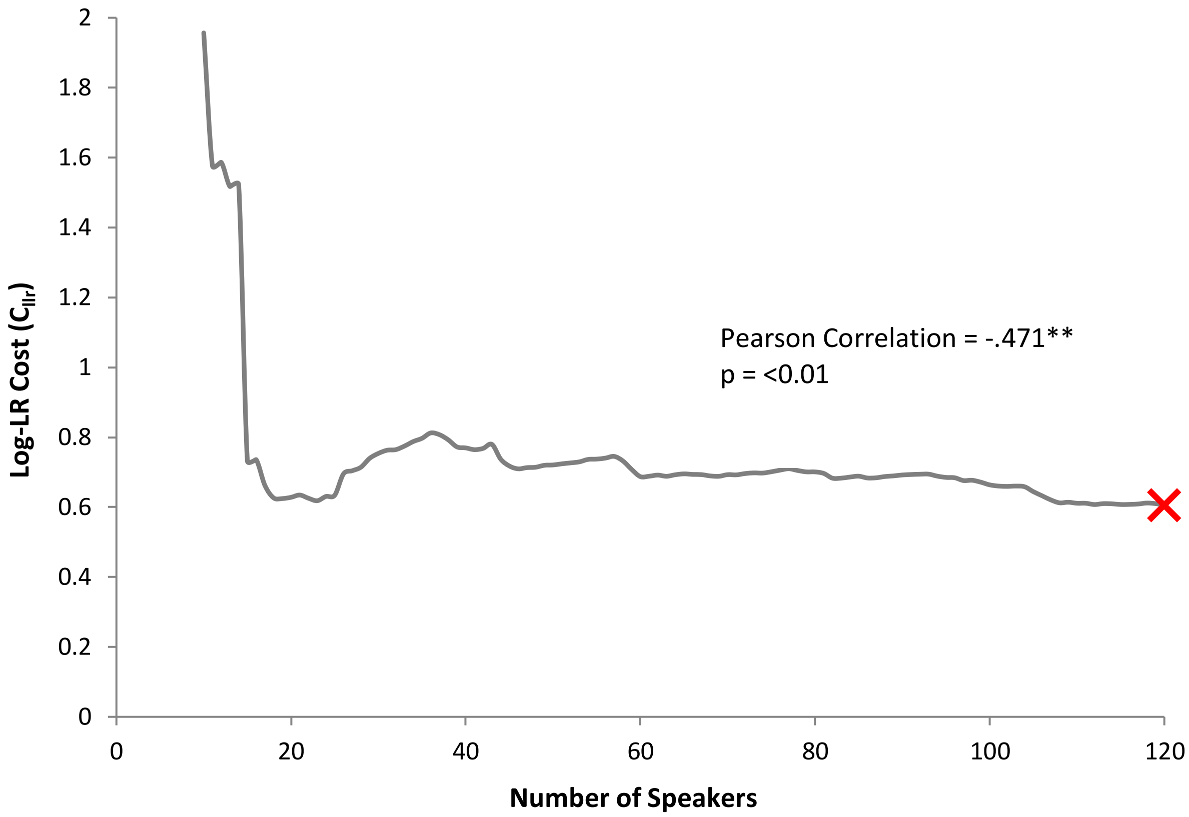

Figure 11 – *Contour of log LR cost plotted against the number of speakers in the reference data set, where the lowest C_{llr} achieved is indicated by a red cross (at 120 speakers)*

Whilst Figure 11 suggests that C_{llr} stabilises as population size increases, inspection between 40 and 120 speakers reveals a strong correlation between the two variables. The severity of error was found to decrease consistently as the number of speakers increased, with the lowest C_{llr} achieved at 120 speakers.

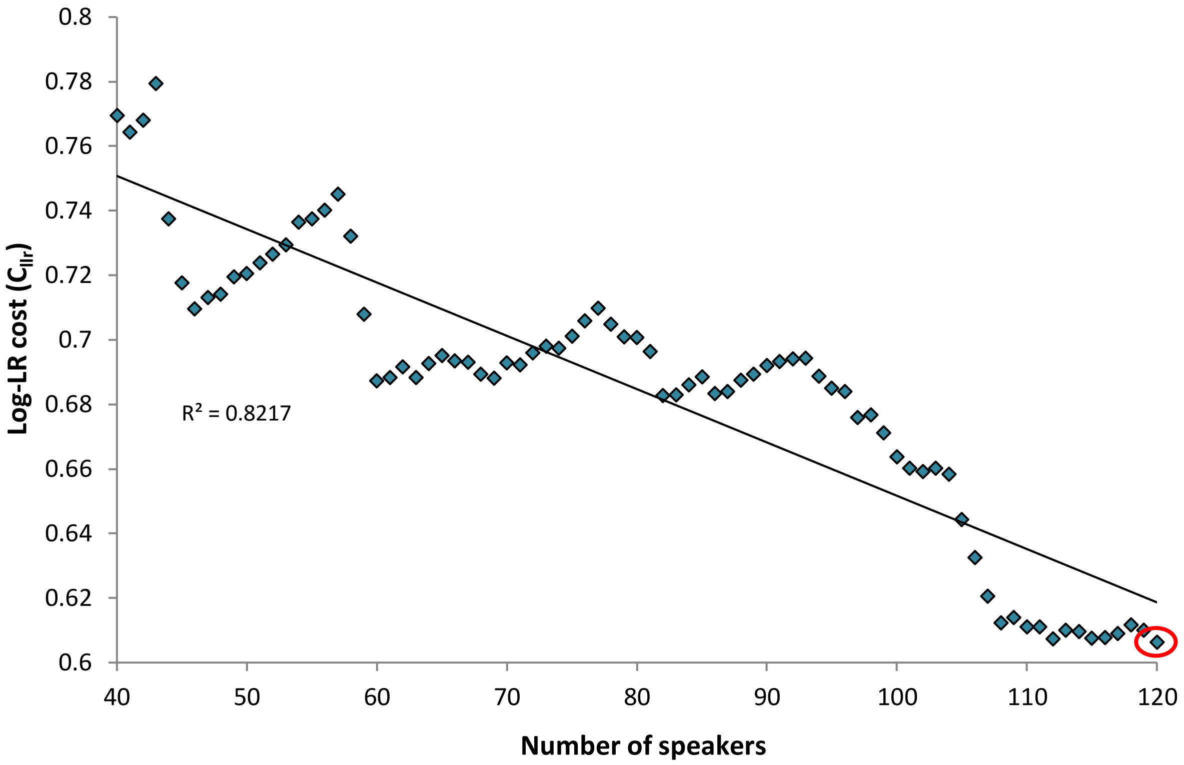

Figure 12 – *Scattergram of log-LR cost plotted against between 40 and 120 reference speakers. The negative correlation is plotted with a linear trend line and the condition achieving the lowest C_{llr} marked in red*

4.1.1.2 Cubic, F1-F2

Cross-comparison with quadratic results suggests that strength-of-evidence is higher in the cubic system, despite similar patterns of variation. SS means are largely stable, although the reduction from 20 to 10 speakers represents the difference between moderate support for H_p and limited support for H_d. The DS mean was more sensitive, initially displaying a divergence from a stable trajectory at 68 speakers. Maximally DS strength-of-evidence increases by two categories from moderate to strong.

Significant increases in SD were displayed as a function of decreasing population size (SS=Pearson's correlation: -.493**, p<0.01; DS=Pearson's correlation: -.880**, p<0.01). The magnitude of SD is consistently greater for cubic coefficients than quadratic.

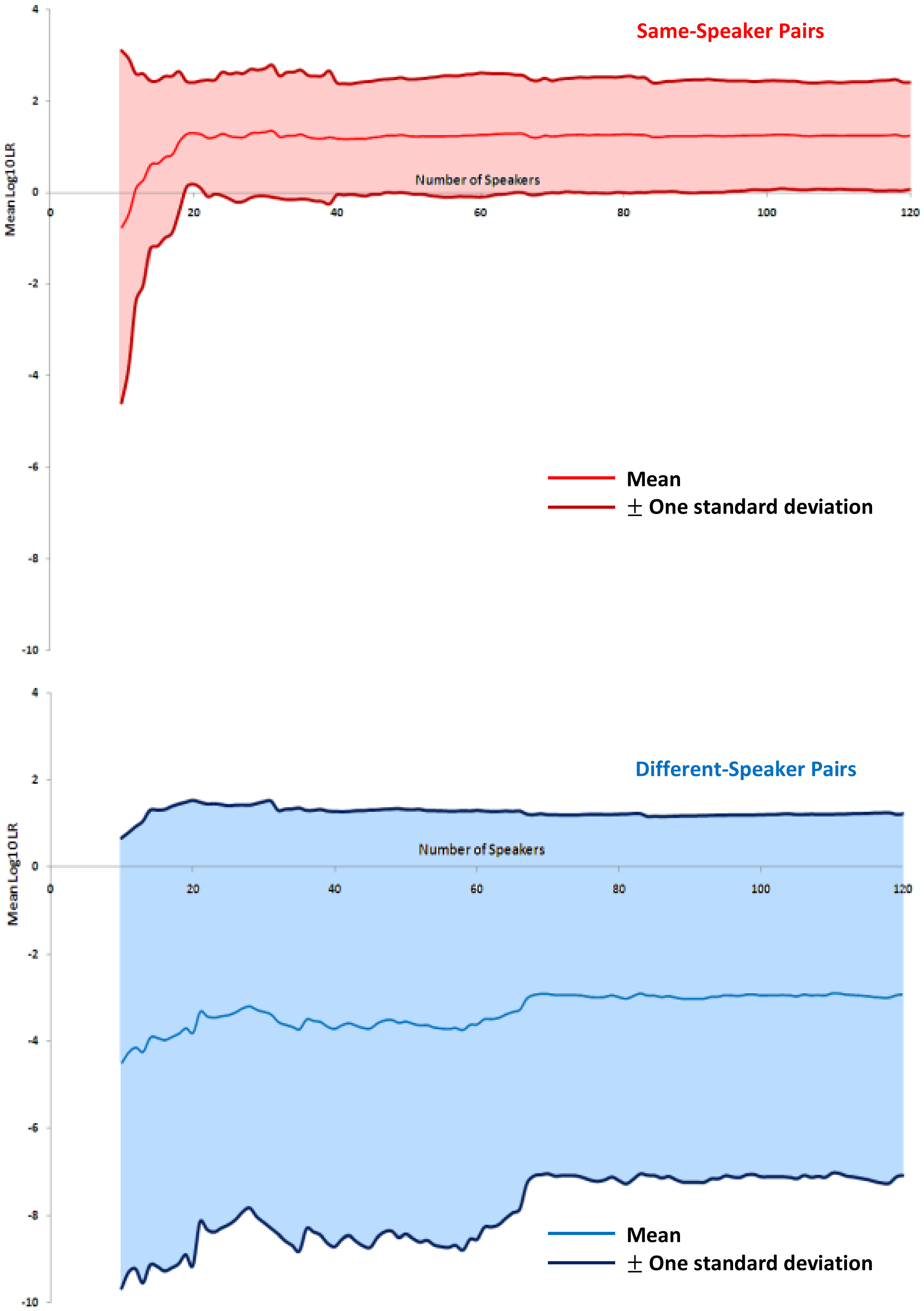

Figure 13 – *Mean $\log_{10}$ LRs (lighter lines) based on cubic polynomials for same- (red) and different-speaker (blue) pairs according to the number of speakers in the reference data. Dark lines indicate ± one standard deviation from the mean*

Again error decreases significantly when population size increases, despite the the lowest C_{llr} being achieved with 20 speakers. Given the instability of mean $\log_{10}$ LRs when the population is small, the lowest C_{llr} between 40 and 120 speakers (102 speakers) was identified for use in subsequent experiments.

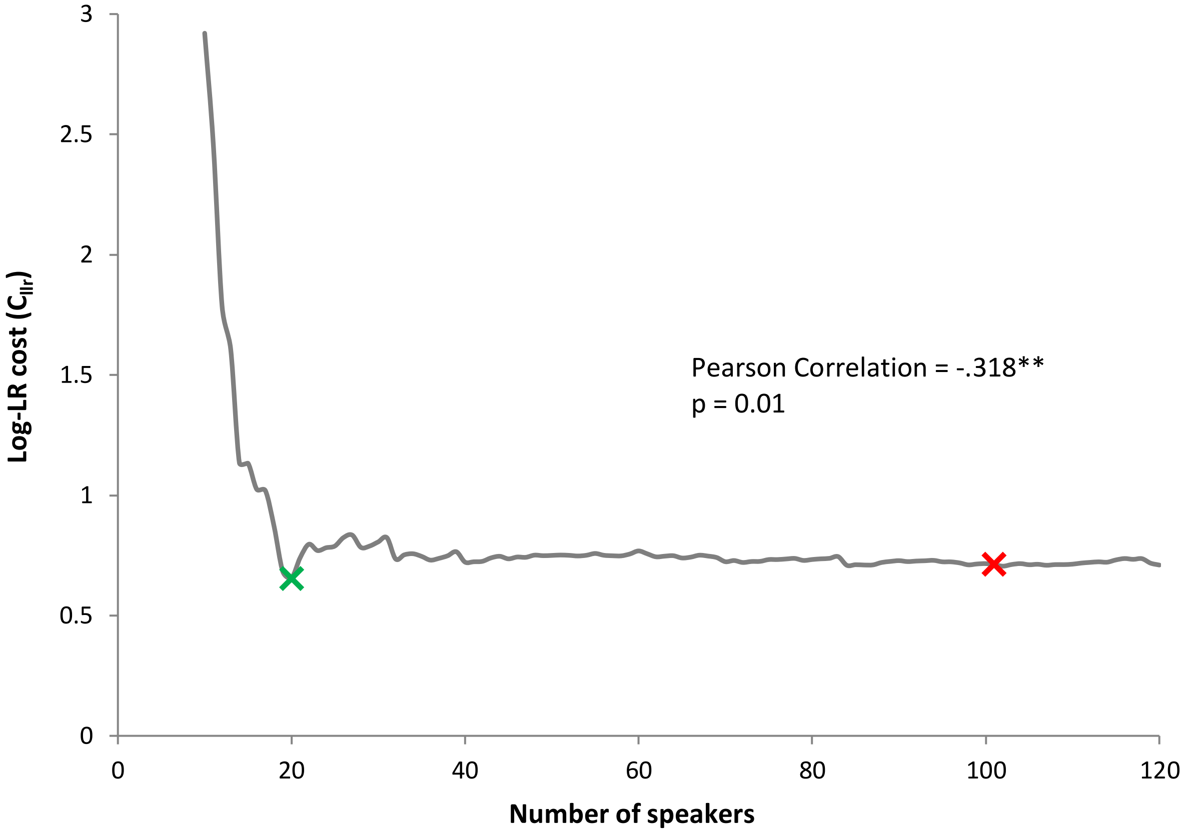

Figure 14 – *Contour of log-LR cost plotted against the number of speakers in the reference data set, where the lowest C_{llr} achieved is indicated by a green cross (at 20 speakers) and the lowest C_{llr} between 40 and 120 speakers indicated by a red cross (102 speakers)*

4.1.2 Number of tokens per speaker

This section discusses the effect of variable numbers of tokens per reference speaker. LRs were calculated on the same 32-speaker test set as §4.1.1. Based on §4.1.1.2 a reference set of 102 speakers was used. By including 102 rather than 120 speakers, the experiment could be run with a greater number of tokens (max 13). SS and DS LRs were calculated as the number of tokens per speaker was reduced by one (min 2). No control was made for the adjacent phonological context of tokens removed at each stage. Tokens were removed according to their position within the original recordings such that those produced later were removed first. This was intended to recreate variable sample length.

4.1.2.1 Quadratic, F1-F2

Mean SS $\log_{10}$ LRs were robust to the removal of reference tokens. DS pairs were more susceptible to variation displaying no period of stability in mean LRs. Rather, there is a continual increase in strength-of-evidence with the 13-token and two-token conditions displaying a divergence of 4.237. Further, SD increases significantly as a function of such variability (SS=Spearman's rho: -.788**, p:0.002; DS=Spearman's rho: -.993**, p<0.01), suggesting that certain pairs are more sensitive to the removal of reference tokens than others.

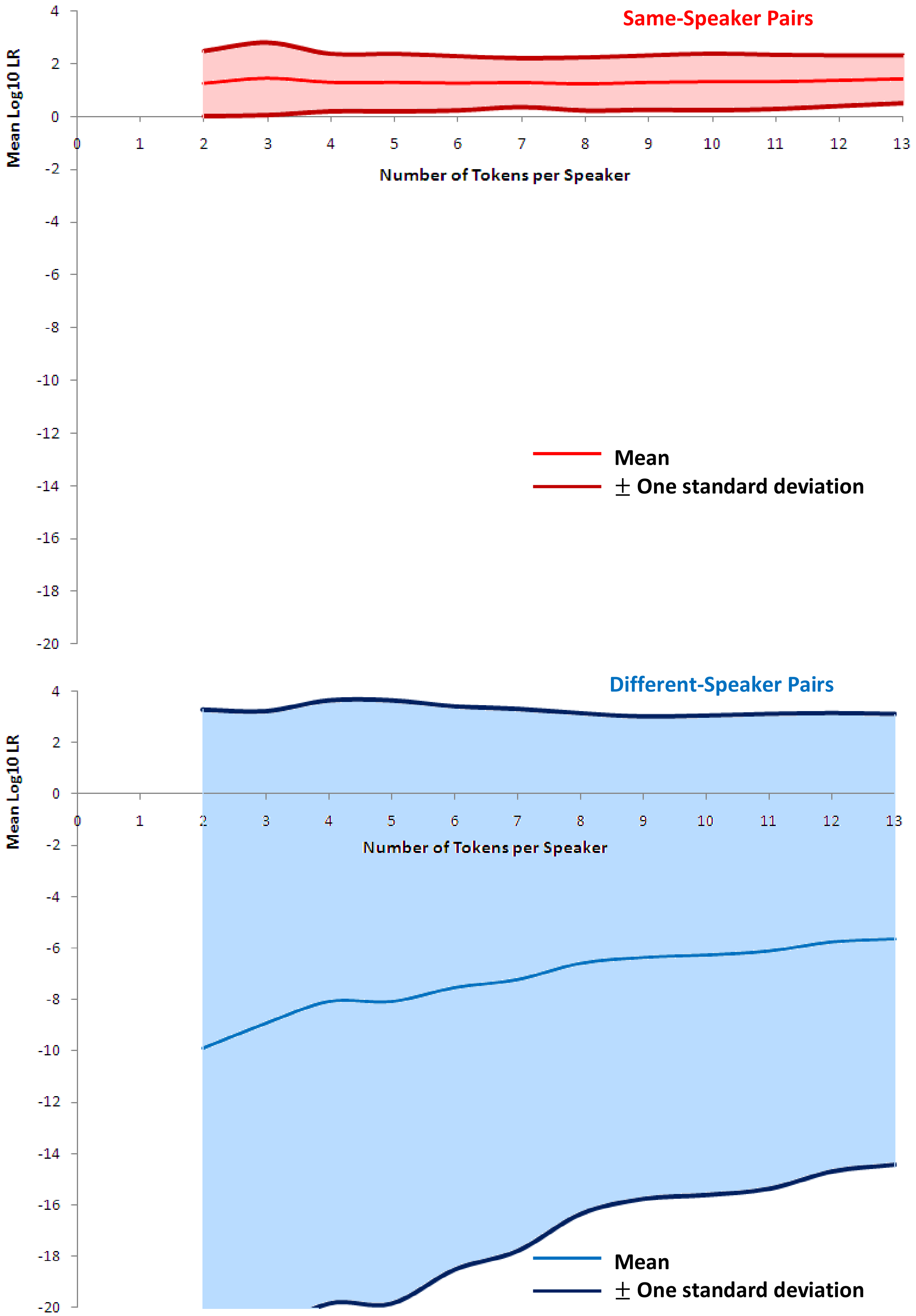

Figure 15 – *Mean log$_{10}$ LRs (lighter lines) based on quadratic polynomials for same- (red) and different-speaker (blue) pairs for 32 test speakers according to the number of tokens per speaker in the reference data. Dark lines indicate ± one standard deviation from the mean*

There is a significant negative correlation between C_{llr} and the number of tokens, with the lowest error achieved at 13 tokens. However the magnitude of C_{llr} increase is minimal, with all conditions achieving values below 0.77. Unlike error in §4.1.1 there is no exponential increase when the population size is small. This suggests that system performance is only marginally improved when more tokens are included in the reference data.

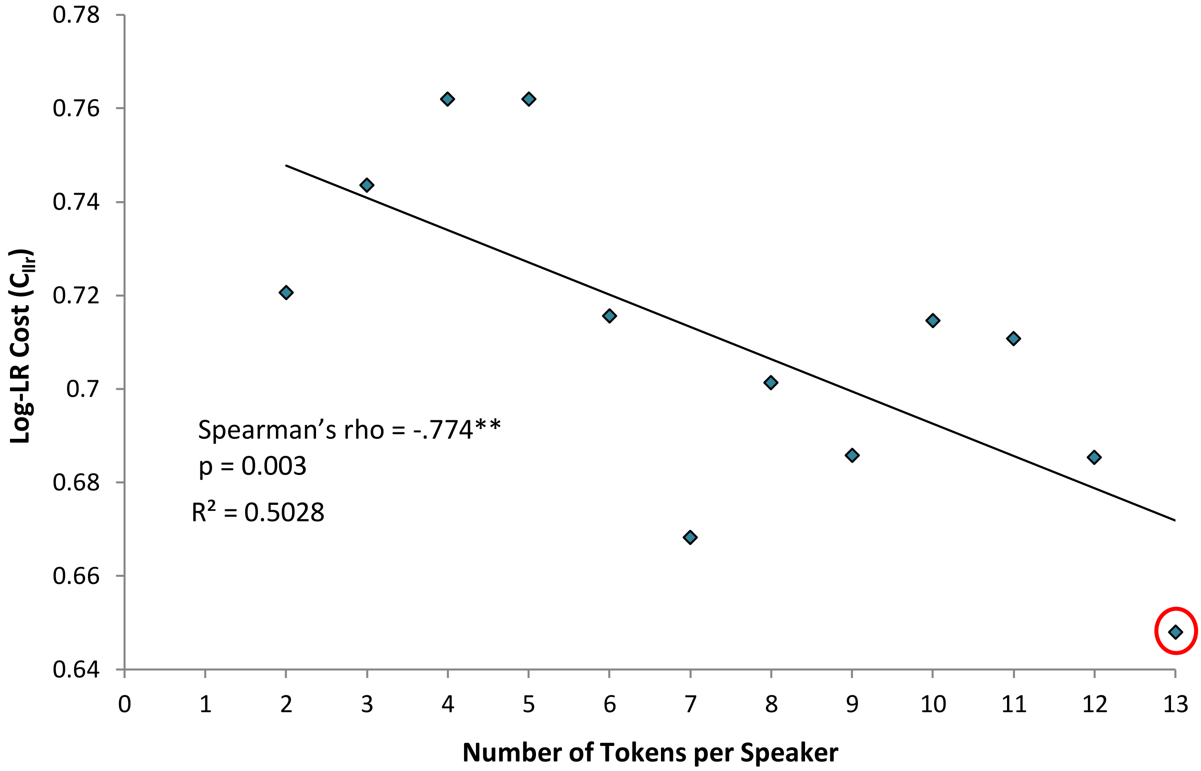

Figure 16 – *Scattergram of log-LR cost according to the number of tokens per speaker in the reference data. The lowest C_{llr} (13 speakers) is marked with a red circle*

4.1.2.2 Cubic, F1-F2

Similar patterns of variation were displayed in the cubic results. Mean LRs for SS pairs were stable as the number of tokens was decreased, however a substantial decline was found when the reference population was modelled using two tokens. Mean DS LRs decreased steadily across each condition, with the strongest support for H_d achieved in the two-token condition. The decrease in mean DS LRs between 13 and two tokens is equivalent to the difference between moderately strong and very strong evidence.

Contrary to §4.1.1, the magnitude of mean LRs is lower in the cubic- than the quadratic-based system. Whilst sensitivity to reference population variation was again shown to increase with decreasing number of tokens (SS=Spearman's rho: -.914**, p<0.01; DS=Spearman's rho: -.998**, p<0.01), SD values were predominantly lower than those for quadratic polynomials.

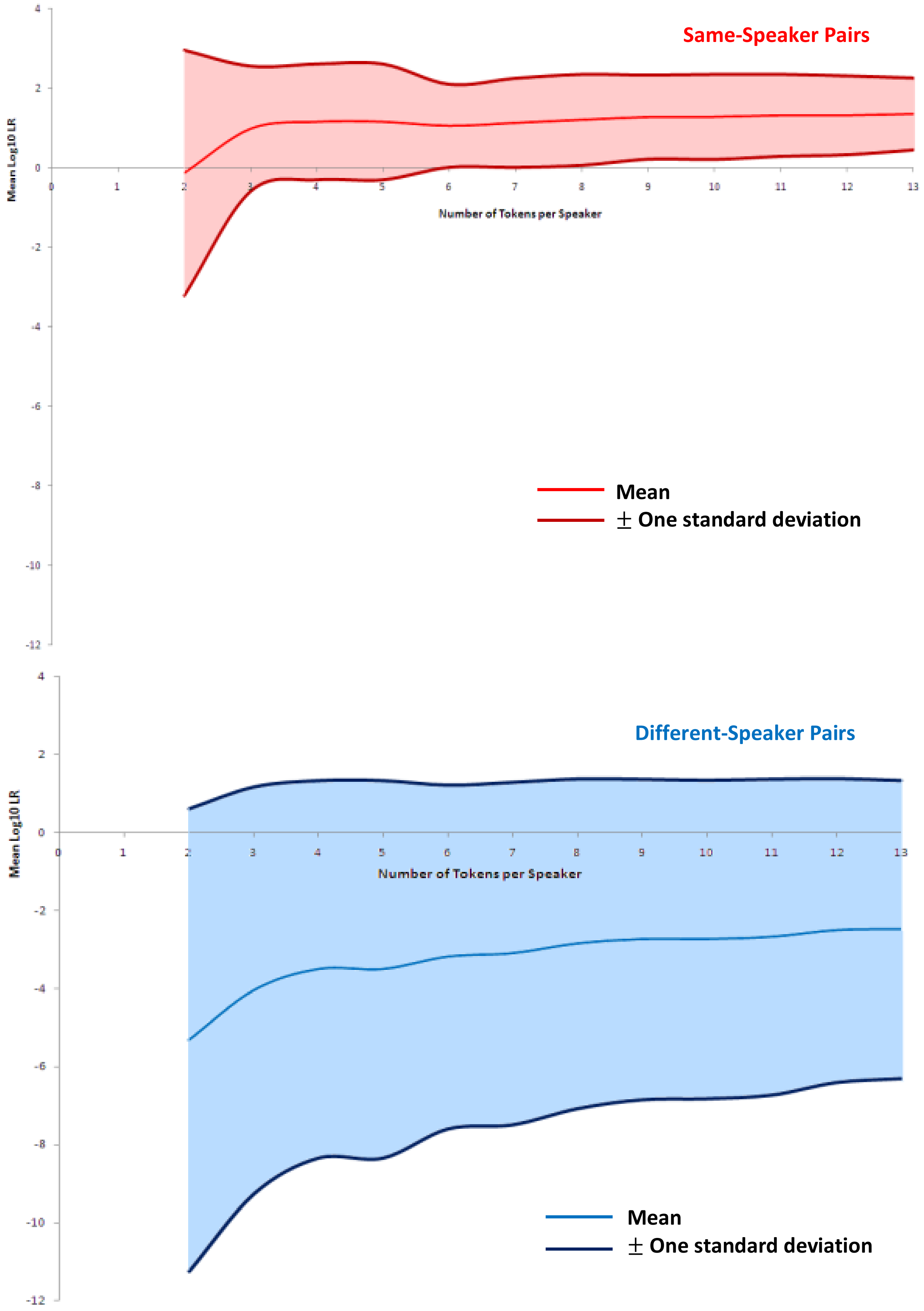

Figure 17 – *Mean $\log_{10}$ LRs (lighter lines) based on cubic polynomials for same- (red) and different-speaker (blue) pairs for 32 test speakers according to the number of tokens per speaker in the reference data. Dark lines indicate ± one standard deviation from the mean*

Figure 18 again suggests that severity of error is improved with more tokens. However, the trend line is likely to be overestimated by the exponential increase in C_{llr} between three and two tokens. Omitting the two-token condition, the relationship between C_{llr} and the number of reference tokens remains statistically significant (Spearman's rho: -.925**, p<0.01).

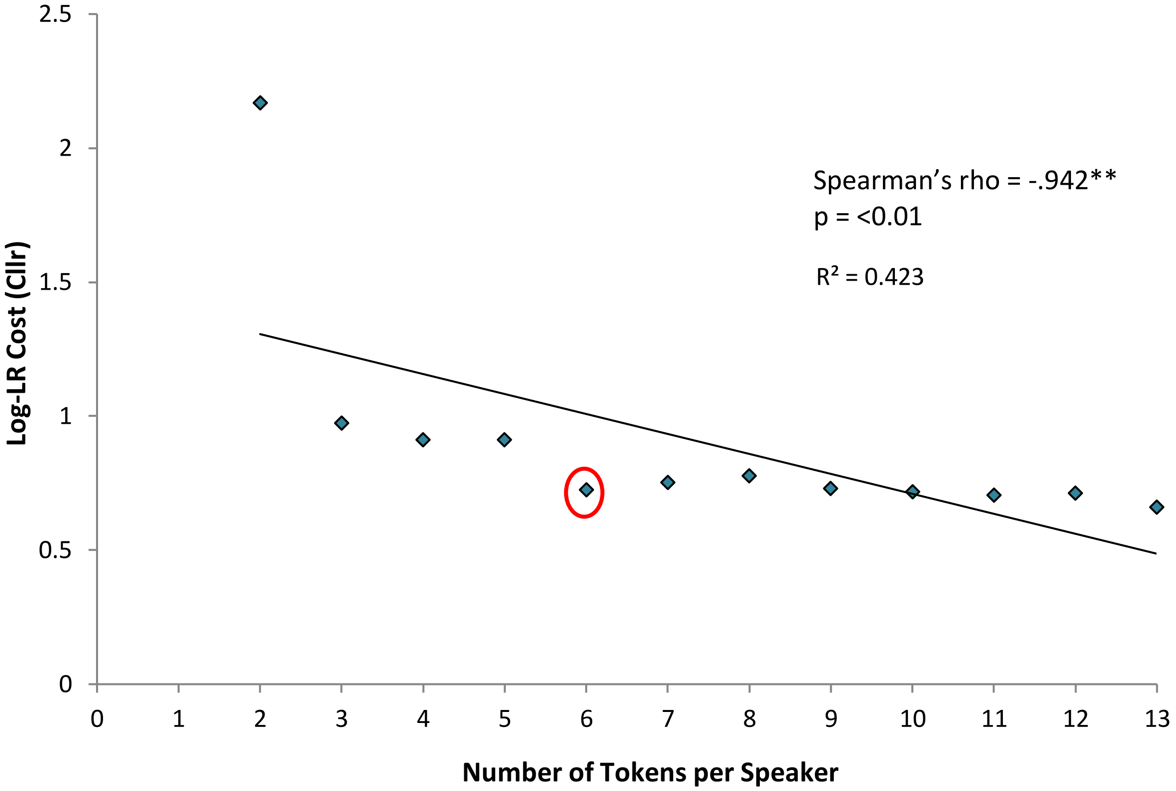

Figure 18 – *Scattergram of log-LR cost plotted against the number of tokens per reference speaker with the condition achieving the lowest C_{llr} (six speakers) marked in red*

4.2 Regional variety

This section assesses the effect of mismatch between the test and reference data with regard to accent. The four regionally-defined test sets were assessed independently. Thus, for each series of tests the LR function performed eight SS and 56 DS comparisons. Based on the findings above, typicality was calculating relative to a reference population of 102 speakers, with 13 tokens each.

Results are visualised using tippett plots. The x axis displays $\log_{10}$ LRs where zero represents the division between support for H_p (>0) and support for H_d (<0). Cumulative proportion is displayed on the y axis, such that the point at which the results contour intersects the x-y axes represents the percentage of pairs which achieve a $\log_{10}$ LR of equal to or less than the value on the x axis. Overall strength-of-evidence is indicated by the gradient of the results contour. Flatter contours indicate a higher proportion of pairs achieving stronger strength-of-evidence, while steeper contours suggest weaker strength-of-evidence.

The results for SS and DS pairs are assessed separately. Tippett plots of combined SS and DS results for each condition are at appendix 5.

4.2.1 F1-F2

In this section results for quadratic and cubic F1-F2 coefficients are presented.

4.2.1.1 Same-speaker pairs

Figure 19 reveals that the weakest strength-of-evidence for SS comparisons based on quadratic polynomials was achieved by ONZE. SS $\log_{10}$ LRs for Manchester and Newcastle were consistently higher than ONZE. The broadest range of SS LRs was found in the York data, minimally achieving values of below one (limited support) and maximally achieving $\log_{10}$ LRs of over three (moderately strong support). Identical patterns are displayed for cubic polynomials, although strength-of-evidence was marginally lower using cubic-based input. This is potentially a consequence of ‘over-fitting’.

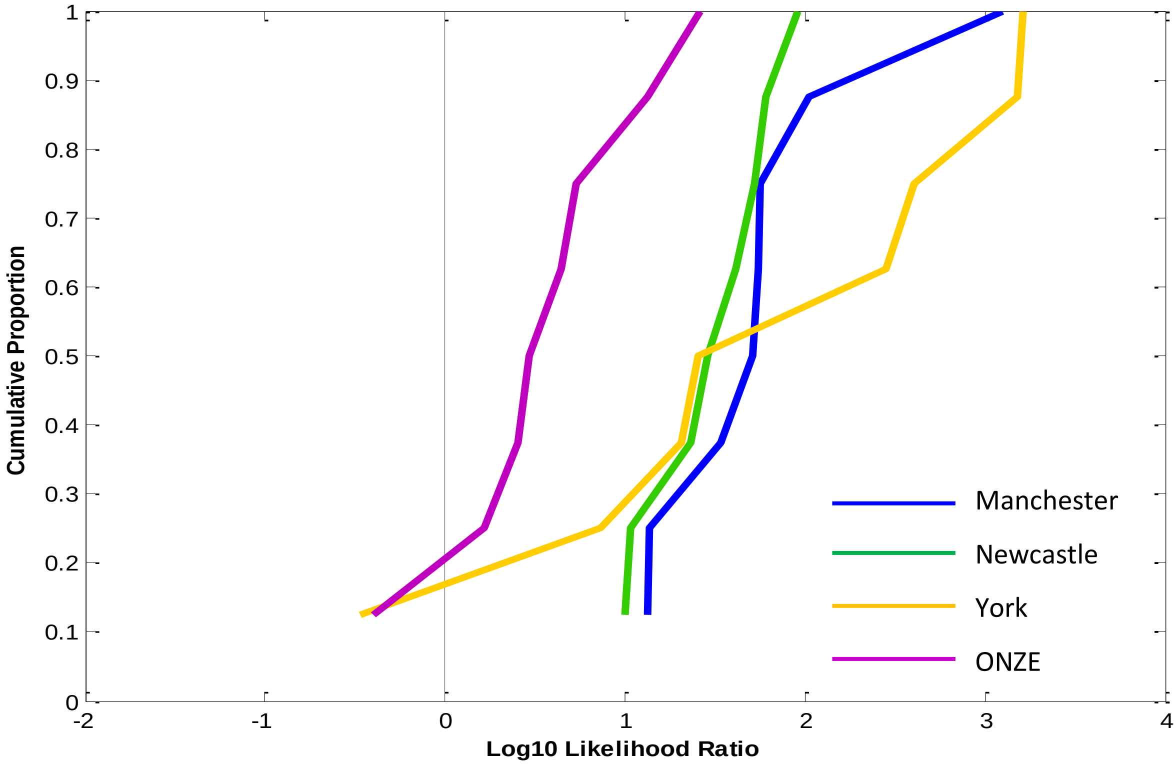

Figure 19 – *Tippett plot of same-speaker comparisons based on quadratic polynomials of F1 and F2 for each of the test sets*

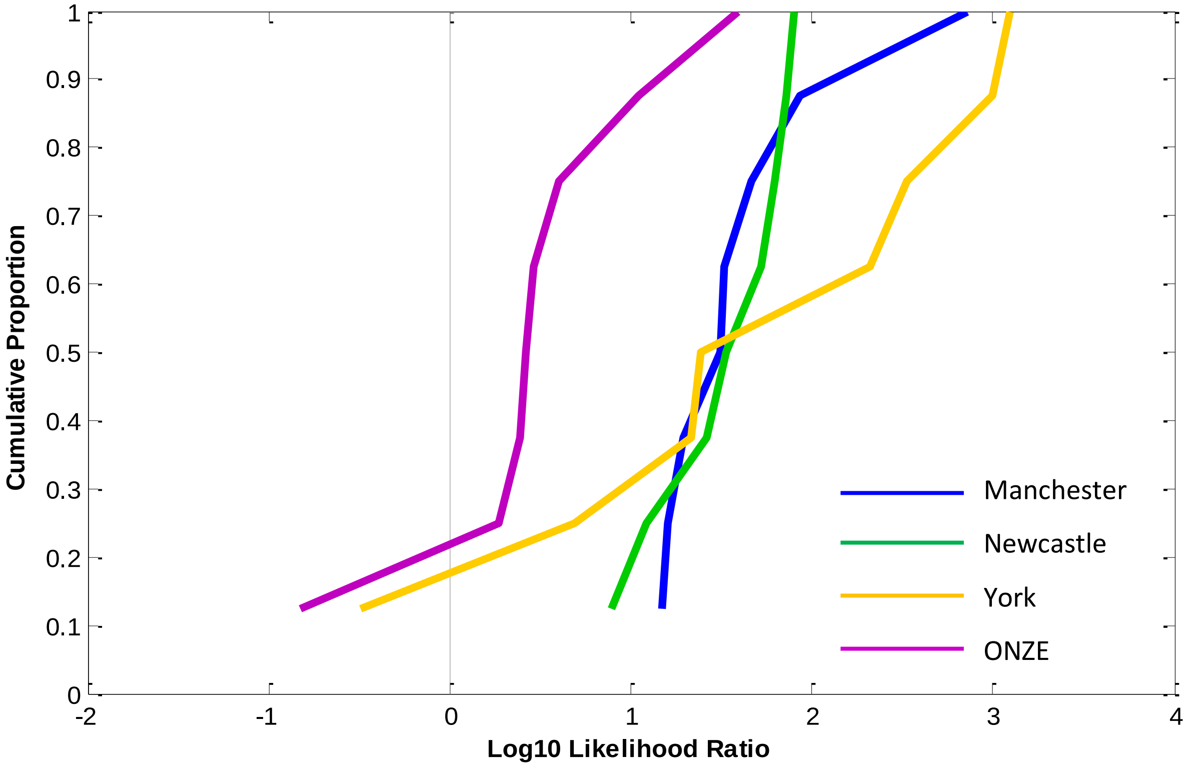

Figure 20 – *Tippett plot of same-speaker comparisons based on cubic polynomials of F1 and F2 for each of the test sets*

However, the results in Figures 19 and 20 do not conform with the levels of within-speaker variation across test sets. In terms of similarity, lower within-speaker variation would predict greater support for H_p. Based on Table 6 the highest mean $\log_{10}$ LRs would be predicted for ONZE, but this is not the case.

Table 6 – *Mean within-speaker variation (Hz) according to regionally-defined set*

Manchester	128
Newcastle	124
York	123
ONZE	111

Considering speakers on an individual basis, no significant correlation was discovered between the level of intra-speaker variability and the outcome of SS LRs. This suggests that between-group differences with relation to within-speaker variation are having little (if any) effect on SS results.

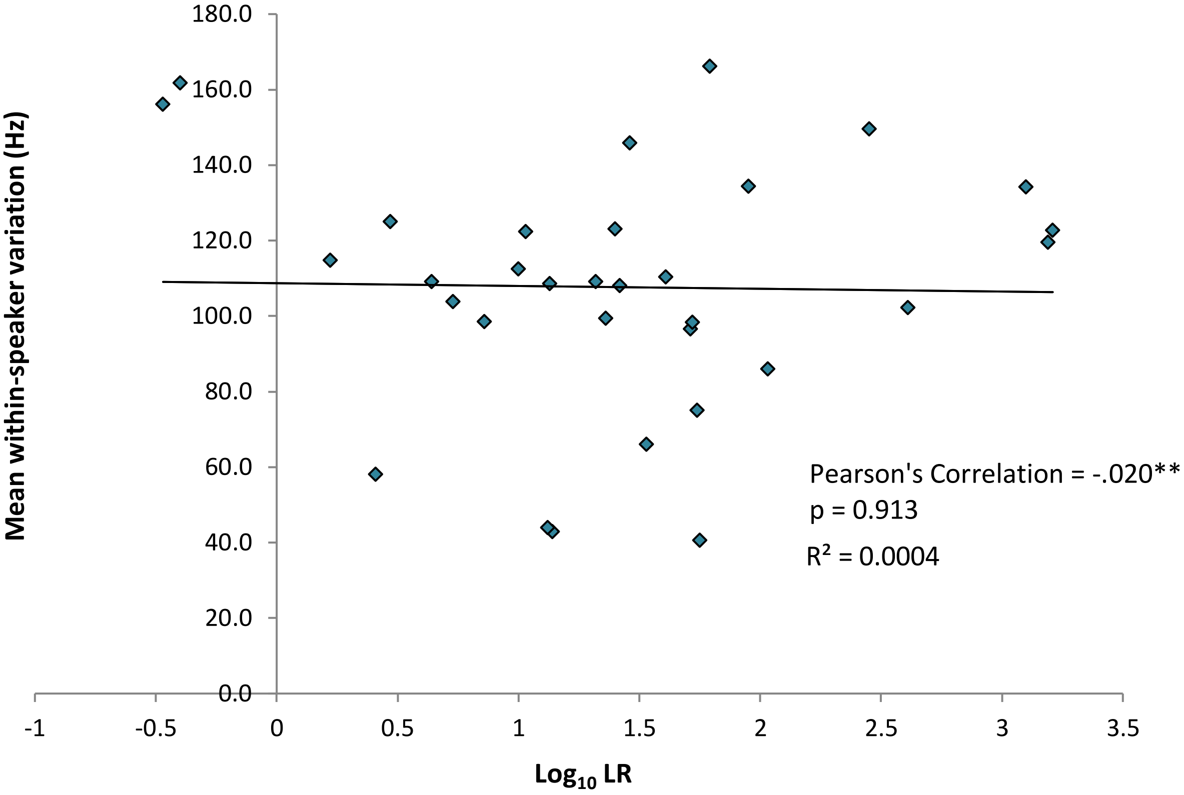

Figure 21 – *Scattergram of within-speaker variability (Hz) plotted against log$_{10}$ LRs based on quadratic polynomials for each of the 32 same-speaker LR comparisons performed (four test sets x eight speakers per set). The linear trend line indicates no significant correlation between the variables (x-y)*

Therefore, it is assumed that the LR differences between test sets are determined by the typicality of samples relative to the reference data. In order to assess this, ONZE is considered the 'control' group against which the magnitude of LRs for the other three sets is assessed, since it matches the reference data for regional variety.

Table 7 shows that for quadratic and cubic input, mean strength-of-evidence is around three times higher for the groups which do not match the reference data for accent. In all cases the divergence between ONZE and the other groups is equivalent to the difference between limited and moderate support for H$_p$.

Table 7 – *Mean log$_{10}$ LRs for same-speaker pairs from each of the test sets for quadratic and cubic systems. The magnitude of difference between the three sets from Manchester, Newcastle and York compared with ONZE is also indicated*

	Quadratic, F1-F2		**Cubic, F1-F2**	
	Mean Log$_{10}$ LR	Times > ONZE	Mean Log$_{10}$ LR	Times > ONZE
Manchester	1.77	3.06	1.64	3.32
Newcastle	1.49	2.58	1.52	3.09
York	1.82	3.15	1.73	3.51
ONZE	0.56		0.49	

Considering these results on a speaker-by-speaker basis, Figure 22 reveals a significant, positive correlation between log$_{10}$ LRs for the 32 SS pairs and their proximity to the reference data (§3.4). The relationship between these two variables was also significant in the cubic-based system (Pearson's correlation: -.831**, p<0.01).

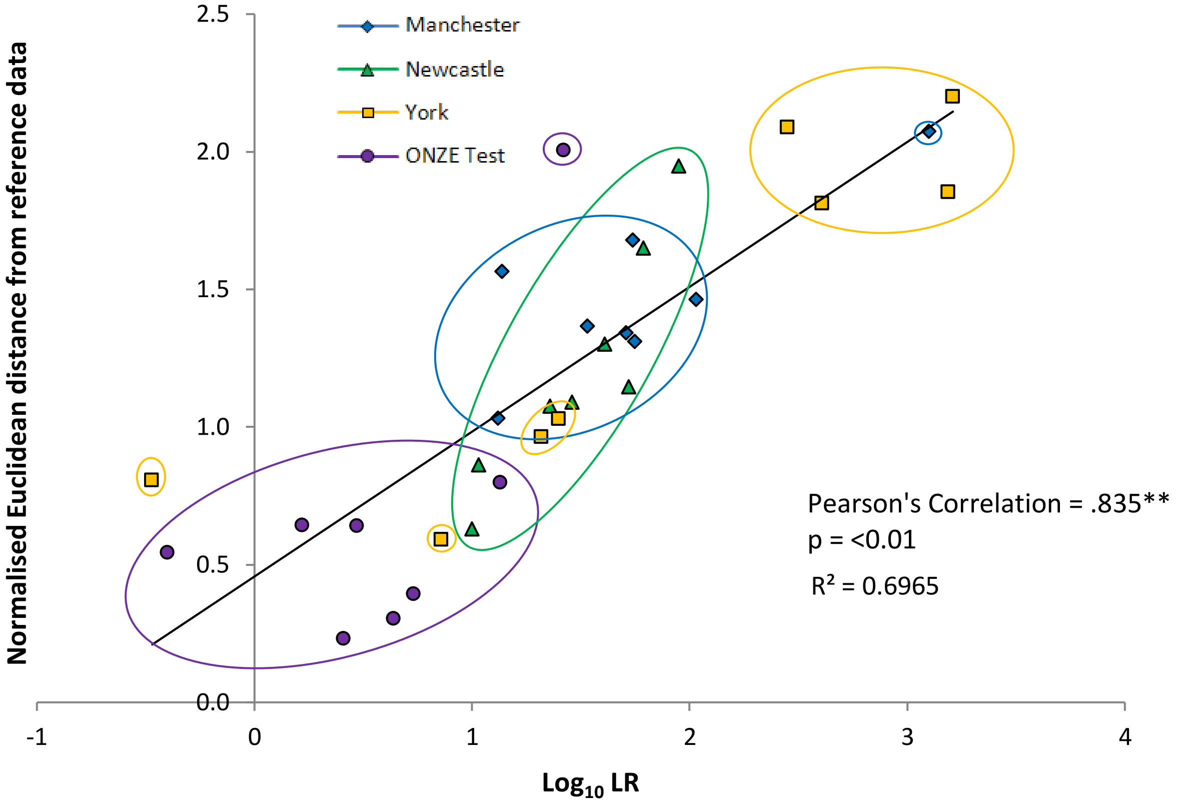

Figure 22 – *Scattergram of normalised Euclidean distance from the reference data of each individual speaker in each of the test sets (used to quantify 'proximity') plotted against the log$_{10}$ LR outcome of the SS comparison based on quadratic polynomials of F1-F2 for that individual. The correlation between variables (x-y) is indicated with a linear trend line*

Those speakers which exhibit greater Euclidean distance from the reference data generally achieved stronger strength-of-evidence than individuals positioned closer. Compared with the ONZE baseline mean log$_{10}$ LR of 0.56 (limited support), the speakers from other accent groups achieved log LRs equivalent to moderate, moderately strong or even strong support for H$_p$. Given that within-speaker variation is not found to be a significant factor, the results in Figure 22 are interpreted as over-estimation of the evidence due to regional variety.

4.2.1.2 Different-speaker pairs

More complicated patterns of variation are displayed across DS results. Tippett plots for quadratic (Figure 23) and cubic (Figure 24) polynomials show that the magnitude of DS LRs is greatest for York. The weakest support for H_d is displayed in the Newcastle results.

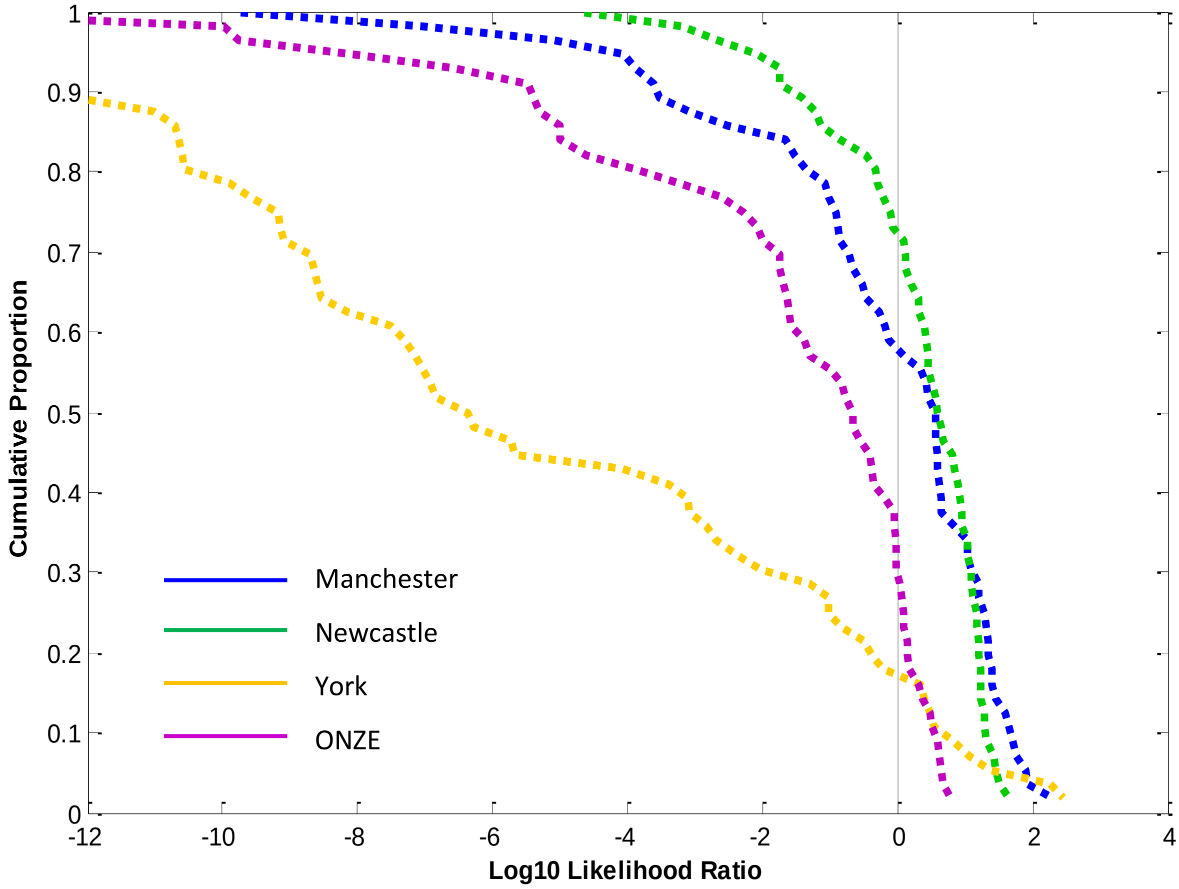

Figure 23 – *Tippett plot of different-speaker comparisons based on quadratic polynomials of F1 and F2 for each of the test sets*

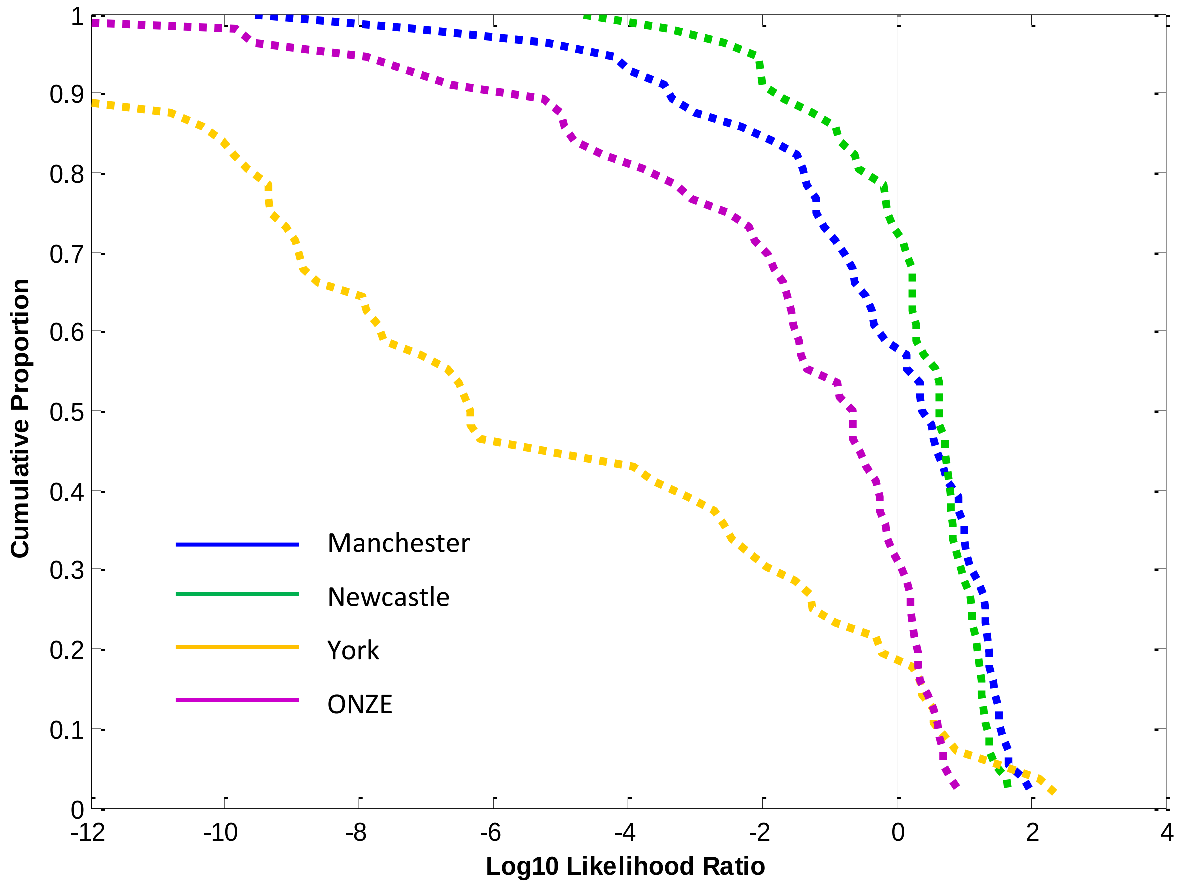

Figure 24 – *Tippett plot of different-speaker comparisons based on quadratic polynomials of F1 and F2 for each of the test sets*

The magnitude of DS LRs seems primarily determined by acoustic differences between the mock suspect and offender samples. Those speakers displaying the greatest Euclidean distance from each other achieve the strongest strength-of-evidence in support of H_d. DS pairs which are closer within the speaker-space display weaker strength-of-evidence. This relationship was significant in both quadratic- (Figure 25) and cubic-based systems (Pearson's correlation: -.256**, p<0.01).

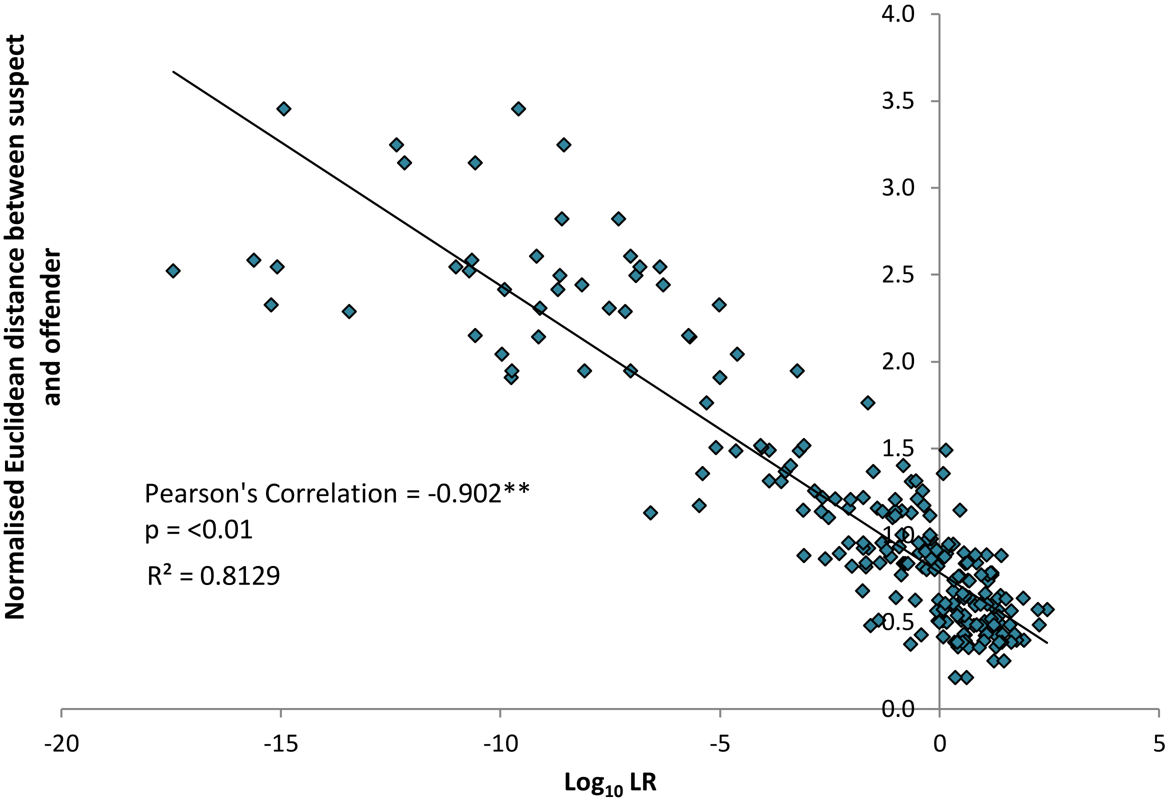

Figure 25 – *Normalised Euclidean distance between mock different-speaker suspect and offender samples plotted against the log_{10} LR output based on quadratic polynomials of F1-F2 for that DS comparison for all speakers in each of the four test sets*

To a lesser extent the proximity of two speakers to the reference population also seems relevant to LR output (Figure 26). However, any strong evidence of overestimation due to regional background is overshadowed by differences between sets in between-speaker variation.

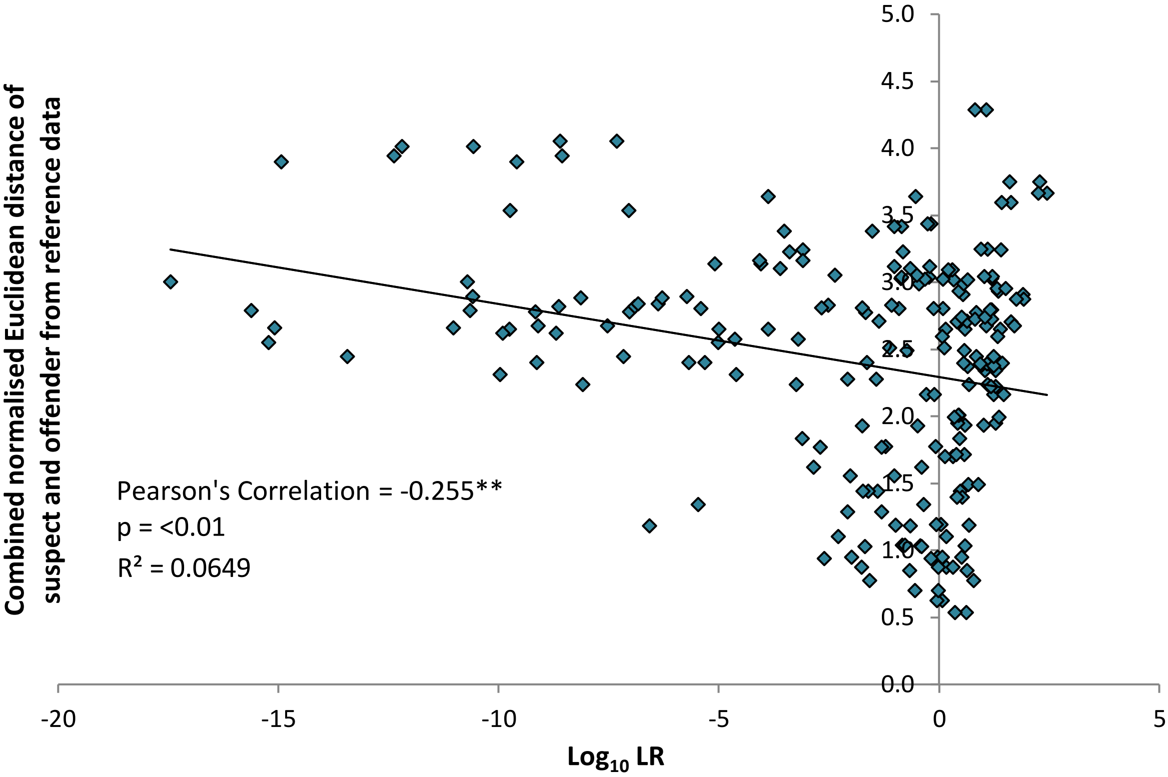

Figure 26 – *Combined normalised Euclidean distance of the different speakers acting as suspect and offender plotted against log_{10} LRs for those DS pairs calculated on the basis of quadratic polynomials*

A concerning outcome of the DS results is the high level of contrary-to-fact support for H_p. This is most prevalent in the Manchester and Newcastle sets, and is evidenced by the high cumulative proportion at which the results contour crosses zero on the x axis in DS tippett plots. The fact that two DS samples can offer support for H_p is a necessary facility of the LR framework. However, levels of between-speaker variation in this study do not predict the results in Table 8 and more than any other outcome, contrary-to-fact support for H_p must be reliable since it could contribute towards posterior odds in favour of guilt over innocence. The reliability of LR output is assessed at §4.2.3.

Table 8 – *Percentage of different-speakers pairs in the Manchester and Newcastle test sets with LR output which supported the same-speaker hypothesis (H_p)*

	Quadratic % DS pairs offering support for H_p	**Cubic** % DS pairs offering support for H_p
Manchester	57	57
Newcastle	71	71

4.2.2 F2

A significant issue for FVC involving formants is the bandwidth restrictions imposed by telephone transmission which have been shown to artificially alter F1 (Künzel 2001; Byrne and Foulkes 2004). As is common in LR studies based on formants (Rose 2006a; Rose 2007b), comparisons in the present study were also performed using F2 alone.

4.2.2.1 Same-speaker pairs

Similar trends are displayed in Figures 27 and 28 to those found for F1-F2. However, all test sets experience a leftwards shift in their results contour reflecting a weakening of strength-of-evidence in the F2-only condition. The most marked decrease in mean LR is displayed across the Manchester and Newcastle sets. Contrary to Rose et al (2006:334), the results suggest that F1 is making a considerable contribution to SS strength-of-evidence.

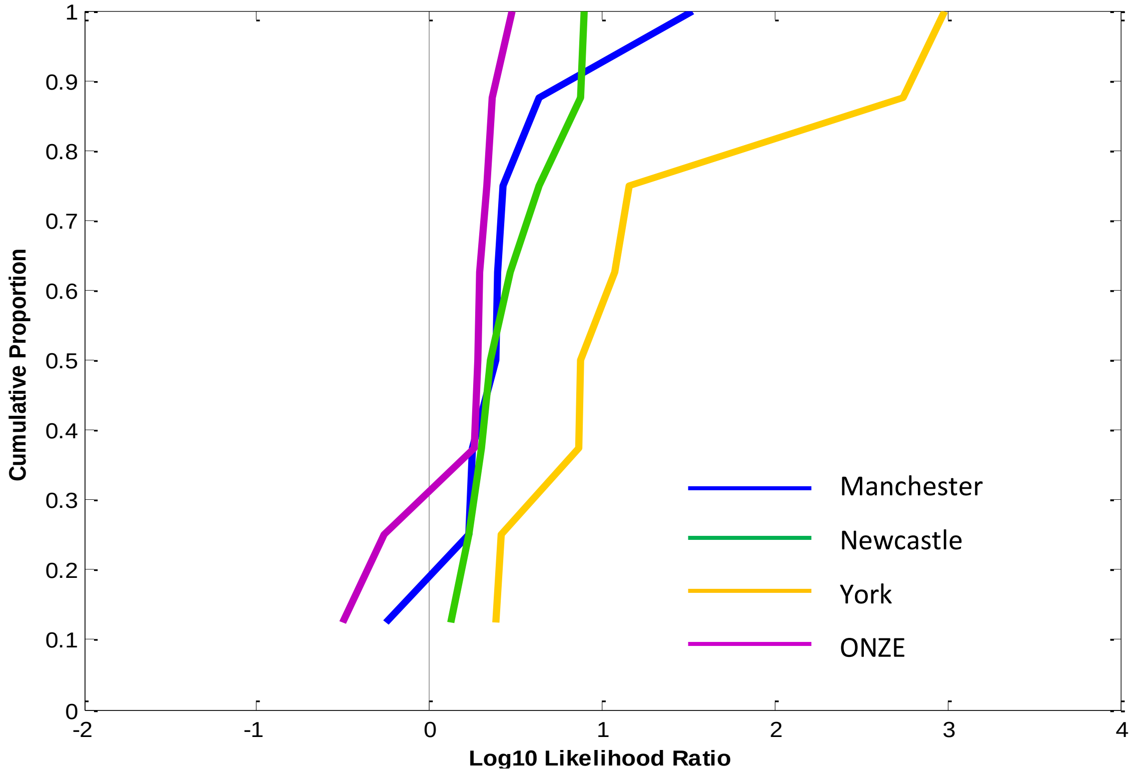

Figure 27 – *Tippett plot of same-speaker comparisons based on quadratic polynomials of F2 for each of the test sets*

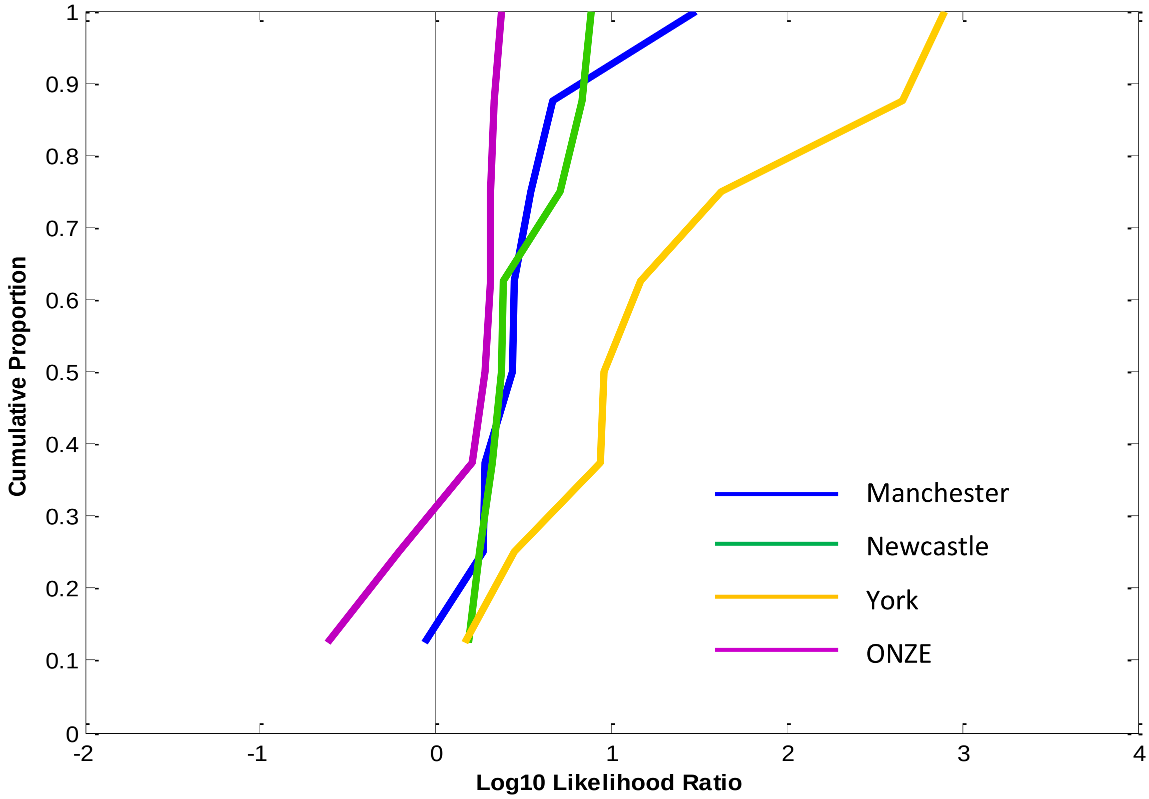

Figure 28 – *Tippett plot of same-speaker comparisons based on cubic polynomials of F2 for each of the test sets*

Euclidean distance was calculated on the basis of F2-only. A significant linear relationship was again discovered between positioning of individuals in the speaker-space and the outcome of SS LRs. However, the clustering of speakers is less clearly defined than for F1-F2. On the basis of Figure 29, it is evident that the gradient of the correlation is also skewed by three outlying speakers.

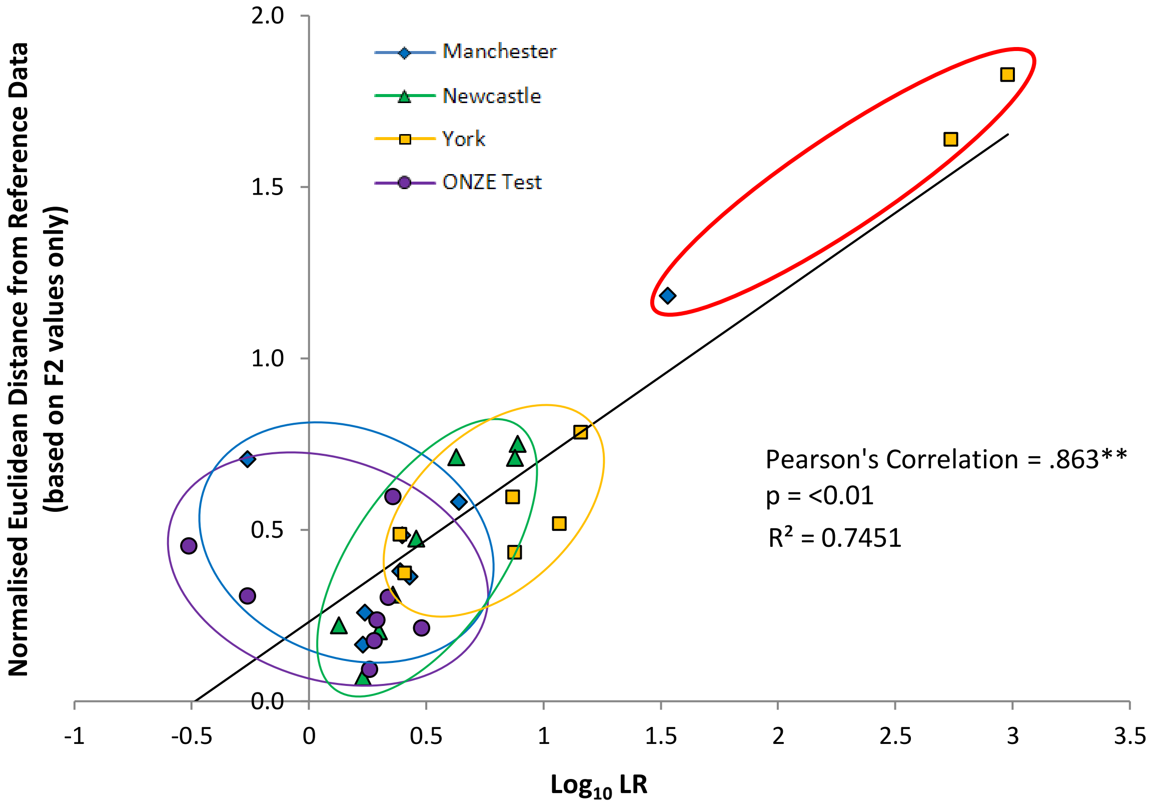

Figure 29 - *Scattergram of normalised Euclidean distance from the reference data based on quadratic polynomials of F2 for each individual speaker in each of the test sets plotted against the* log_{10} *LR outcome of the SS comparison for that individual with outlying values marked in red. The correlation between variables (x-y) is indicated with a linear trend line*

Given the lack of clear groupings based on F2-only input, it appears that F1 contains considerable regional-variety-defining information, the lack of which impacts on strength-of-evidence. Therefore, the magnitude of the effect of variety mismatch is minimised by removing F1. Consequently, mean LRs for Manchester and Newcastle relative to the ONZE baseline no longer differ by a verbal category; on average limited strength-of-evidence was achieved for these three groups. The difference between the ONZE and York sets remains limited to moderate.

4.2.2.2 Different-speaker pairs

Similarly, in F2-only DS comparisons there is an overall decrease in the magnitude of LR output for all test sets. DS pairs for York achieve the strongest strength-of-evidence, followed by Manchester, ONZE and finally Newcastle.

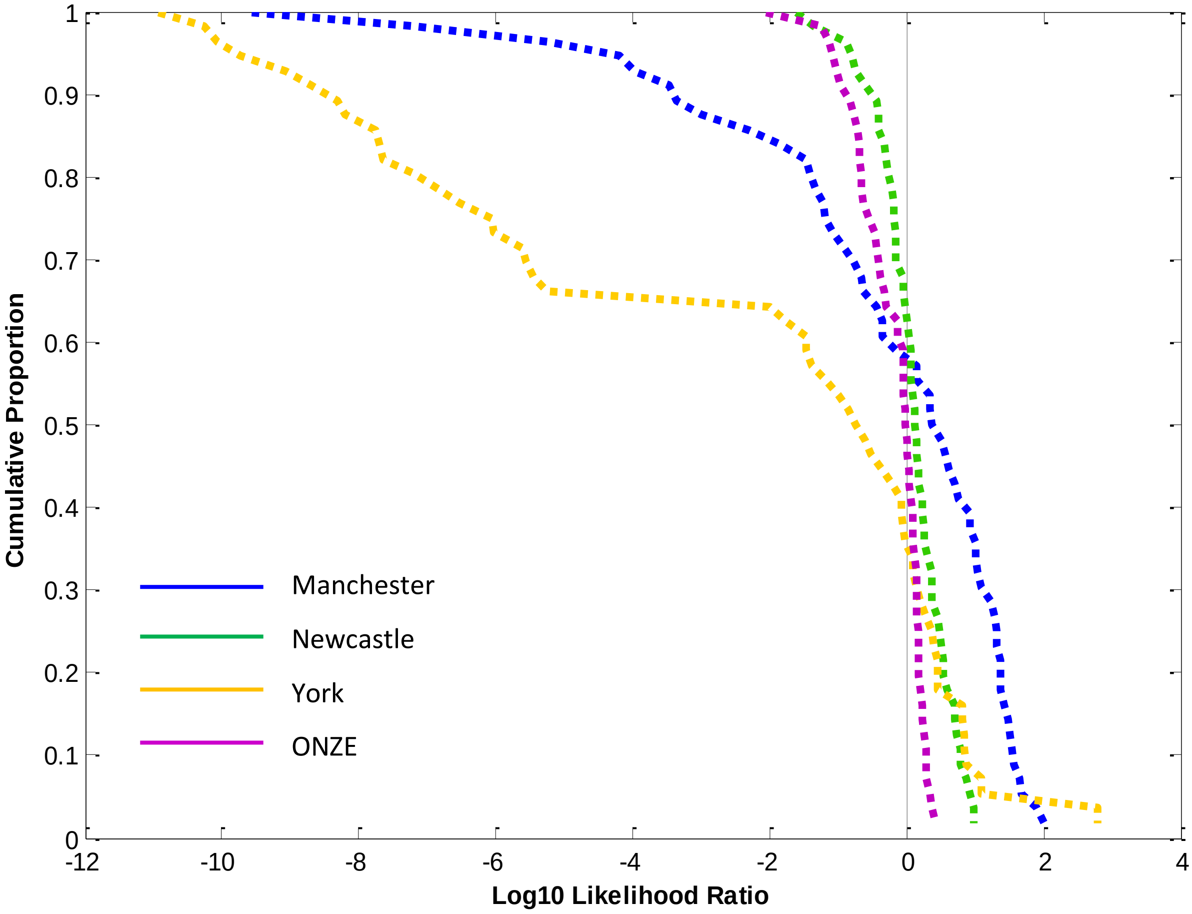

Figure 30 – *Tippett plot of different-speaker comparisons based on quadratic polynomials of F2 for each of the test sets*

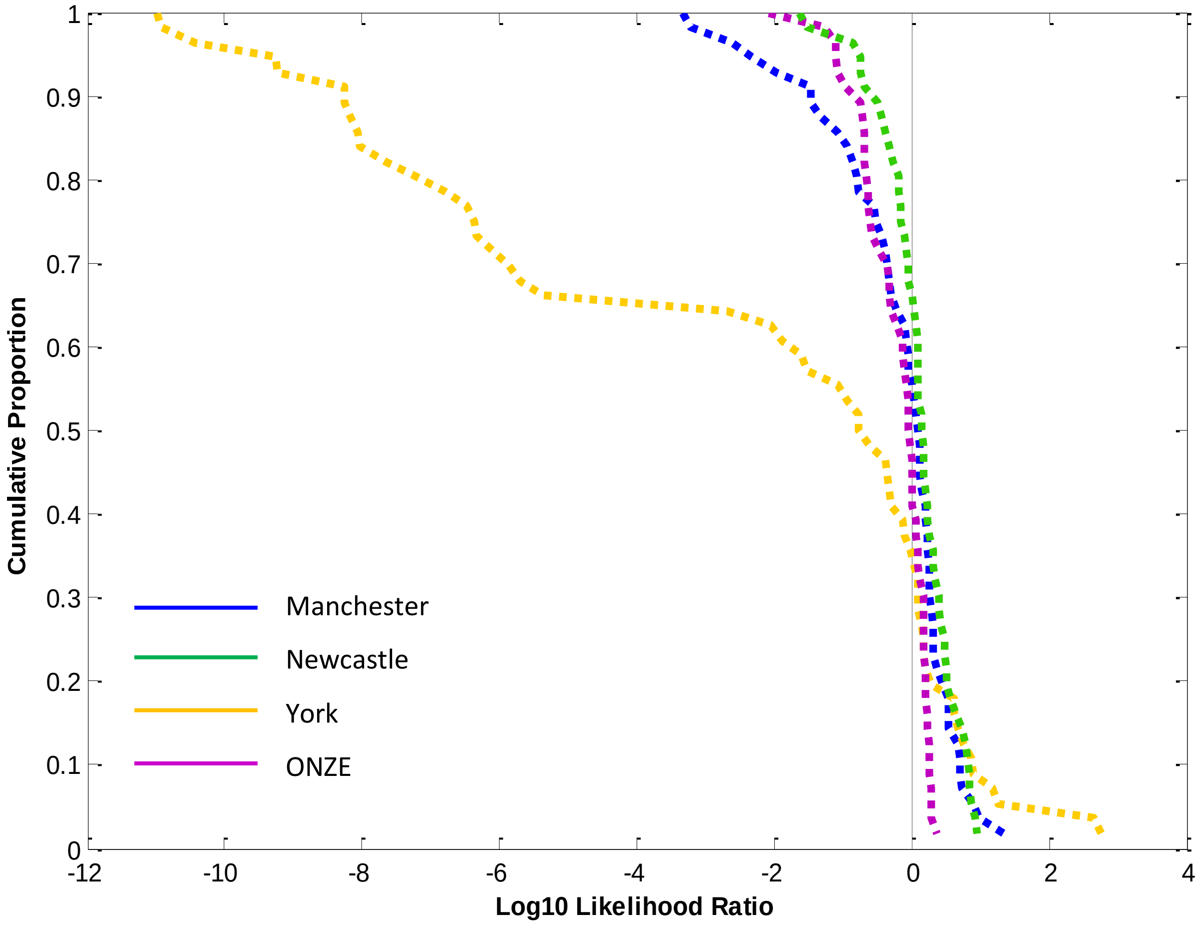

Figure 31 – *Tippett plot of different-speaker comparisons based on quadratic polynomials of F2 for each of the test sets*

High proportions of contrary-to-fact support for H_p are again displayed for Manchester and Newcastle. The proportion of DS pairs offering H_p support was also found to be much higher in the ONZE set with the removal of F1 (false-positive rate: quadratic=48.2%, cubic=46.4%).

4.2.3 System performance

Across all conditions error was greatest for the Manchester and Newcastle data. This is primarily a consequence of the high proportion of DS pairs offering H_p support. Using ONZE as the 'control' group, it is hypothesised that error is greater for DS pairs in regional-defined groups which are further away from the reference data in the speaker-space. Whilst Manchester and Newcastle display comparable between-speaker variation with ONZE, the high levels in the York data mean that between-sample differences are more marked. As a result, strongest support for H_d is displayed across DS pairs for York, thus minimising error.

For Manchester and Newcastle, F1-F2 performs consistently better (i.e. lower error) than F2-only, whilst the opposite is found for ONZE and York. Therefore, the removal of F1 has variable effects on LR reliability for different regionally-defined groups.

Although comparisons of mean C_{llr} indicate that error is marginally higher for the quadratic-based analyses, non-significant differences were found between the performances of the two systems (two-tailed t-test, p:0.575). Whilst it appears that over-fitting with cubic polynomials is not an issue, more accurate modelling of formant contours using higher order polynomials was also not found to improve the reliability of LR output to any considerable degree.

Improvement in the severity of error may be achieved with better calibration of the systems (Morrison 2009b; Morrison et al 2011). Following Morrison and Kinoshita (2008) it is possible that a global improvement in LR performance could also be achieved through the inclusion of more test and reference data.

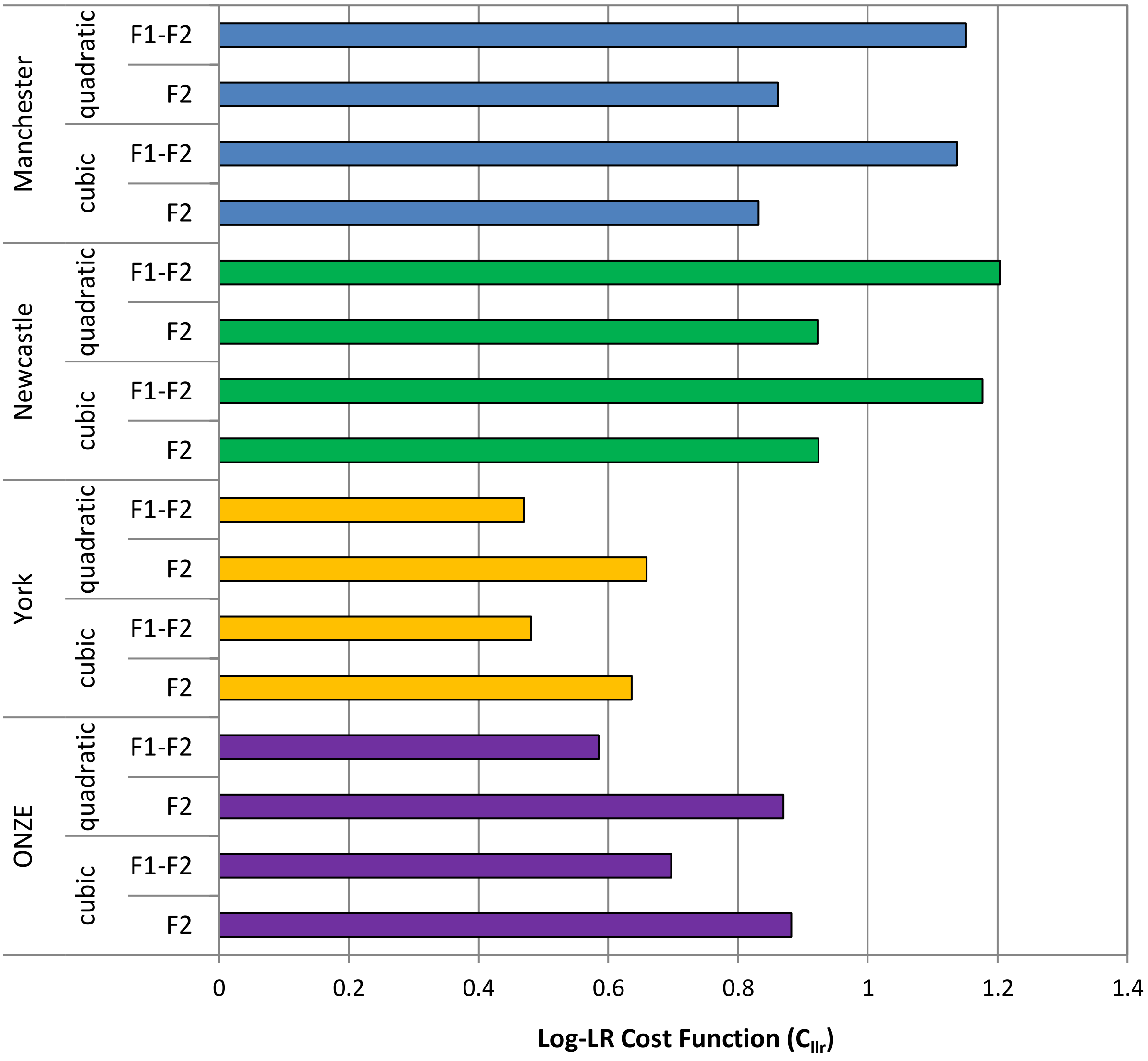

Figure 32 – *Log-LR cost for each of the test sets under each of the experimental conditions (polynomials-formants)*

5.0 DISCUSSION

5.1 Number of speakers

Results confirm Ishihara and Kinoshita's claim that LR reliability is "heavily compromised if the population data (is) limited to a small number of speakers" (2008:1941). Whilst mean SS LRs were relatively robust to varying population size, DS pairs were found to be more sensitive. Further, severity of error improved continually by adding more speakers to the reference set. It is predicted that error too would stabilise, although this is predicted only with greater than 120 speakers. Therefore, it should be assumed that the bigger the population, the more reliable the estimation of strength-of-evidence.

However in practical terms, the availability of reference data imposes restrictions on how many speakers are included. §4.1.1 suggests that population size should, in part, be dependent on the feature under investigation. The point at which mean DS strength-of-evidence was substantially affected by population size was 20 speakers higher for cubic polynomials compared with quadratic. Further, the cubic system displayed higher C_{llr} values than the quadratic system. Therefore, fewer speakers were required in the quadratic analysis to achieve the same LR reliability as that for cubic coefficients.

It may also be appropriate to delimit the number of speakers on the basis of the individual pair of evidential samples. The trend for an increase in SD as a function of decreasing population size highlights that pairs of samples are affected in different ways, with certain pairs more or less sensitive to such variation. This suggests that there is no universal optimum. Rather, the expert should minimally acknowledge the issues surrounding population size, and perhaps even conduct pre-testing to assess how sensitive evidential samples are to variability.

5.2 Number of tokens per speaker

Positively, the quadratic system displayed no change in categorical strength-of-evidence as the number of tokens was reduced, with moderate support achieved in each condition. Further, SS SD in the quadratic system was relatively stable across conditions. Mean SS LRs in the cubic system were more susceptible to change, displaying a jump from moderate support for H_p to limited support for H_d below three tokens. Further, C_{llr} decreased significantly as more tokens were included.

On the basis of the three metrics investigated (mean LRs, SD and C_{llr}) DS pairs displayed greater sensitivity to variation. No period of stability was displayed in mean LRs and the overall increase in strength-of-evidence as the number of tokens reduced represents the difference between moderately strong and very strong support in the cubic system. This would suggest that the greater the number of tokens, the more accurate the estimation of strength-of-evidence and that small numbers of reference tokens should be avoided. This is because input based on just two tokens per speaker is insufficient for capturing the within-speaker variability displayed by an individual.

Compared with the C_{llr} of 0.013 achieved using similarly small numbers of tokens in Rose (2011), the magnitude of error in the present study is rather severe. There are a number of potential explanations for this. Firstly, the number of tokens analysed across test and reference data was not comparable. In §4.1.2 there is a mismatch between test and reference data with regard to the number of tokens per speaker. It is considered probable that the effects displayed across mean LRs, SD and C_{llr} are consequences of such mismatch, since the atypicality of within-speaker distributions is likely to be overestimated compared with the level of within-speaker variation across a reference population with fewer tokens per speaker.

Phonological conditioning may also account for general trends in §4.1.2. It was not possible to control for phonological context in the reference data beyond the removal of ____l tokens. Therefore the process of removing tokens in a quasi-random fashion will result in variation in the proportion of tokens in each context. Given the predictable contextual effects on GOOSE, it is expected that LRs will have been affected by mismatch between test and reference speakers with regard to the number of tokens in specific contexts. This should be controlled in FVC casework.

5.3 Regional variety

The SS results for F1-F2 reveal that in the absence of significant differences in within-speaker variation, there is a positive linear correlation between proximity to the reference data and LR magnitude. This finding confirms earlier predictions based on the LRs assessment of typicality. However, this linear relationship has implications for the definition of the reference population. The positioning of individuals in the speaker-space is determined not only by idiosyncratic patterns, but also by regional variety. Given the regionally-defined groupings in Figure 5 (§3.4), this is true of GOOSE. Using the ONZE test data as a baseline, SS F1-F2 $\log_{10}$ LRs were found to be around three times greater for groups which do not match the reference population for accent. Considering raw LRs, the differences in mean SS strength-of-evidence are magnified: (Quadratic/F1-F2) Manchester=16.2x>ONZE, Newcastle=8.5x>ONZE, York=18.2x>ONZE.

Using mean formant values for Newcastle to indicate the positioning of a Newcastle reference set within the speaker-space, the effect of accent mismatch may be quantified on an individual basis. Speaker3 (S3) achieves a quadratic SS $\log_{10}$ LR of 1.0 (raw LR=10) and is 0.628 away from the ONZE reference data (a). The distance between S3 and Newcastle(mean) is 0.704 (b). Based on the linear correlation in Figure 34, an estimate of the difference in $\log_{10}$ LR for S3 may be calculated using the distances from these Euclidean distances. This is applied to the known LR to estimate the strength-of-evidence if the reference population were positioned at Newcastle(mean).

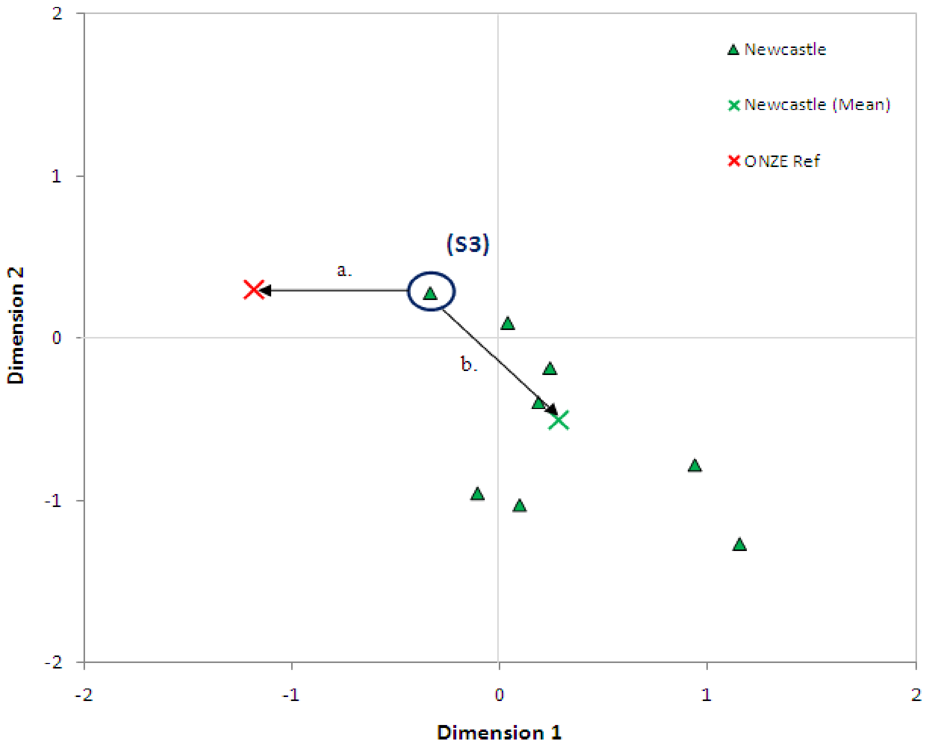

Figure 33 – *MDS plot of the Euclidean distances between the eight Newcastle test speakers, their combined means and the ONZE reference data, with the positing of speaker3 marked relative to the two reference data points*

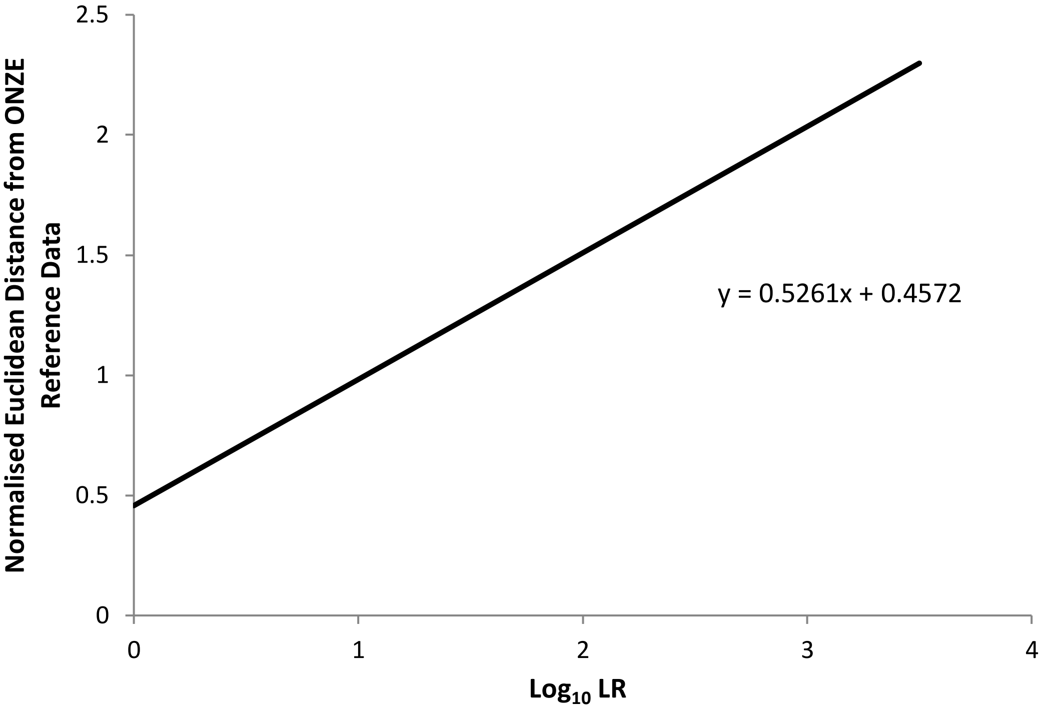

Figure 34 – *Linear trendline for the correlation between normalised Euclidean distance from the ONZE reference data and LR output based on quadratic polynomials of F1-F2 together with the equation used for determining the yfit value*

Based on Figure 35, the SS $\log_{10}$ LR for S3 is predicted to increase from 1.0 to 1.449 with a reference population which matches the test data for accent. This difference does not alter the verbal categorisation of limited support for H_p, since S3 is almost equidistant from ONZE Ref and Newcastle(mean), although precision is necessarily affected.

SPEAKER3 (S3) (SS $\log_{10}$ LR = 1.0/ raw LR = 10)

x = $\log_{10}$ LR

y = Normalised Euclidean distance

Linear equation: $y = 0.5272x + 0.4572$

$$\therefore \quad x = \frac{y - 0.4572}{0.5261}$$

Since y between S3 and ONZE reference/Newcastle mean is known, the linear correlation is used to generate LRs (x) based on these y values:

ONZE: $$\frac{0.628(y) - 0.4572}{0.5261} = 0.325 \ (\log_{10} \mathrm{LR})$$

Newcastle(mean): $$\frac{0.704(y) - 0.4572}{0.5261} = 0.471 \ (\log_{10} \mathrm{LR})$$

In order to assess this with regard to the SS $\log_{10}$ LR achieved by S3 the magnitude of the difference between the two predicted values above is calculated:

$$\frac{0.471}{0.325} = 1.449$$

The linear correlation therefore predicts that the $\log_{10}$ LR based on the Newcastle mean as the reference population will be 1.449 times the value achieved with the ONZE reference data (which was 1.0):

Estimated $\log_{10}$ LR = 1.0 x 1.449 = 1.449

Converted to raw LR = $10^{1.449}$ = 28.1

Figure 35 – *Calculation used to estimate the LR output for the SS comparison for Speaker3 relative to a set of reference data positioned at the mean for the Newcastle test data within the speaker-space*

However, for speakers who are further away from the ONZE data but closer to Newcastle(mean), the effect of accent mismatch will be greater. Speaker1 (S1) achieves a SS $\log_{10}$ LR of 1.79 (raw LR:61.7) in the quadratic F1-F2 system with the ONZE reference data.

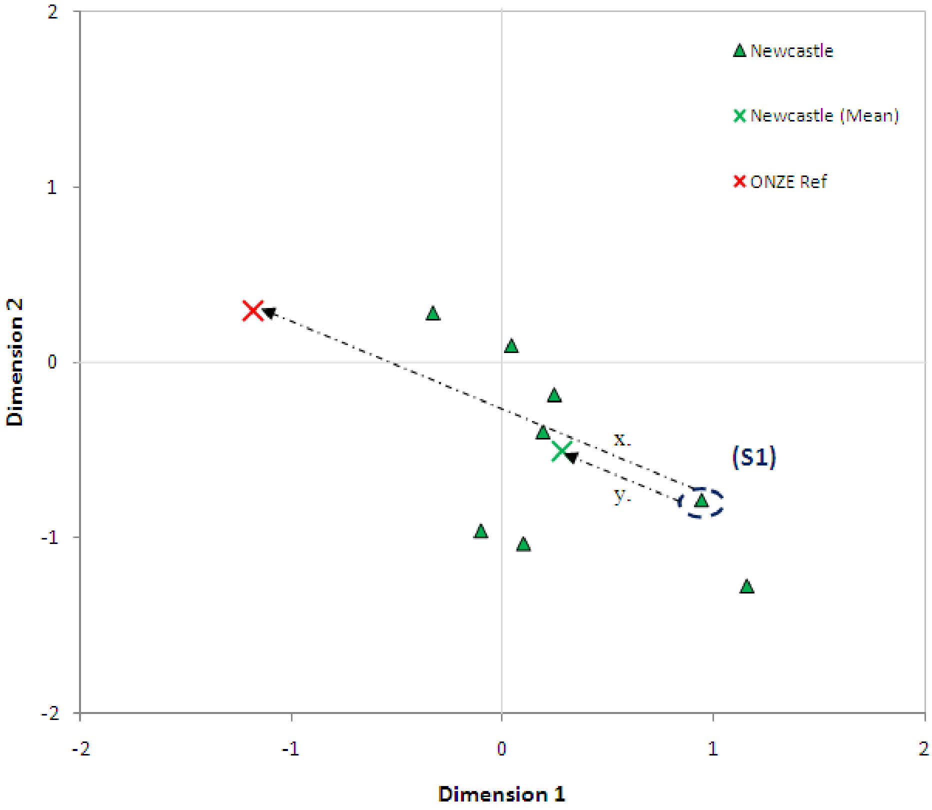

Figure 36 – *MDS plot of the Euclidean distances between the eight Newcastle test speakers, their combined means and the ONZE reference data, with the positing of speaker1 marked relative to the two reference data points*

Using the same procedure as Figure 35, the estimated SS $\log_{10}$ LR for S1 based on a Newcastle(mean) reference set is 0.06 (appendix 5). This is equivalent to the difference between moderate support based on the ONZE reference data and limited support based on 'accent-matching' reference material. The divergence is highlighted by the raw LR, which provides around 54 times greater strength-of-evidence based on ONZE than on Newcastle(mean). Thus for GOOSE, which is not a marked accent feature, LRs may be under- or over-estimated depending on the relative positioning of individuals in the speaker-space when test and reference data differ with regard to accent.

An unexpected outcome of this study is the high proportion of DS pairs offering contrary-to-fact support for H_p in the Manchester and Newcastle data. DS pairs in Manchester and Newcastle were more likely to record a 'false positive' than those for ONZE. More systematic testing is needed to assess whether this is a consequence of accent mismatch.

Similar patterns of variation to those for F1-F2 were displayed in the F2-only results. However, the magnitude of LR differences between ONZE and the other three groups was found to be lower based on F2-only. Therefore, it is assumed that F1 is offering variety-specific information, such that its removal creates greater group overlap in the speaker-space.

Differences in the magnitude of F1-F2 and F2-only LRs suggest that it may be preferable to define the *relevant* population on the regional salience of the variable under consideration. This is because results do not indicate that one reference data set will perform sufficiently well for all features analysed in FVC casework. In particular, tighter demographic restrictions on the reference data are required for phonetic-acoustic features which are stronger sociolinguistic markers, since the separation of regional varieties within the speaker-space will be greater. Even for features which display minimal regional variation (such as GOOSE) a more narrowly defined H_d than Rose's (2004) default assumption (§2.3) is considered essential.

However, for features which are not expected to display between-variety differences, a more general sampling of the population may be appropriate since the removal of accent-defining information leads to greater consistency in the LRs achieved across regional groups. Therefore, despite the considerable regional variation in British English it may be possible to assess typicality of regionally unmarked features against a 'general English' corpus, thus increasing the efficiency of the numerical LR approach without compromising reliability.

5.4 Implications and applications

These findings have broader implications on how theory and practise should be combined. Buckleton et al claim that "the population to be surveyed can be modified if there is some information which would cause us to reconsider our choice" (1991:464). Based on the principle of logical relevance (§2.3) Rose's (2004) 'default' assumption should be modified to include a consideration of the regional background of the perpetrator in FVC.

However, Loakes claims that “while it might be useful to control the reference sample in terms of sociolinguistic variables (...), it is impossible to determine what these sociolinguistic variables are for a particular case” (2006:206). This is because, theoretically, H_d is determined by the defence prior to the expert’s involvement. Perhaps it may be preferable for H_d to be informed by independent expert opinion to ensure the LR estimate is reliable. However, the issues of how narrowly to define regional variety and whether an accurate assessment of regional background may be extracted from an offender sample at all still remain. The implementation of the numerical LR approach is likely to be hindered if speaker profiling is required prior to each case.

A further issue which has emerged from the use of raw LRs, $\log_{10}$ LRs and verbal expressions is the interpretation of evidence by the trier-of-fact. On a $\log_{10}$ scale the difference between 1 and 2 is equal to the raw LR difference between 10 and 100 times more likely, all of which represent moderate evidence in verbal terms. However, Kahneman et al (1982) claim that “people are irrational in their use of numbers in the face of uncertainty” (in Evett 1991:18). Indeed, the results of Cudmore (2011) suggest that people are more likely to misinterpret raw LRs when their value is subjectively high. Therefore, it is predicted that under- or over-estimation of strength-of-evidence as a result of accent mismatch will be exacerbated if presented solely in the form of raw LRs.

6.0 CONCLUSION

At a time when the LR is misunderstood and mistrusted by the criminal justice system (following R-v-T [2010]) it is essential that the framework is shown to be scientifically and legally reliable. As a conceptual framework, this paper offers no challenge to the legitimacy of the LR as the correct way of assessing strength-of-evidence. However, as emphasised by Buckleton et al, the challenge lies in "applying these ideas to real situations" (1991:463). The present study highlights the need for a consideration of sociolinguistic factors beyond the default assumption in Rose (2004) when defining and delimiting the population acting as reference in FVC. Further, results suggest that the most reliable LRs may only be achieved when H_d is defined on a variable-by-variable basis.

Given that FVC cases performed by J P French Associates involve around 15 hours of analysis (French p.c.) and that the reference data here took two weeks of trial and error to ensure that a sufficient number of speakers with a sufficient number of reliable tokens could be included, it would seem "prohibitively difficult" (French et al 2010:147) for the expert to collect separate sets of tailored reference material for each variable. However, positively, results suggest that reliable strength-of-evidence may be achieved with moderate numbers of reference speakers and tokens per speaker, although error and variance may be continually improved with larger data sets. More significantly, the logical relevance of accent appears to be reduced as region-defining information is removed.

Therefore, it is essential that future research assesses the possibility of a more general H_d for variables which are not expected to differ by region. The logical relevance of other sociolinguistic factors such as age and class should also be investigated. Further, to address the limitations of this study it may be useful to compare LR performance based on data extracted from force aligned TextGrids with that of manually extracted data. The results of this study should also be tested under more forensically realistic conditions, notably using non-contemporaneous speech, telephone recordings and more systematic controls over adjacent phonological context. Finally, it would be useful to conduct similar studies with the inclusion of F3, since higher formants have been shown to display greater speaker-discriminatory potential (Simpson 2008).

At present, there remain a number of unanswered questions relating to the definition of H_d in FVC. However, it is hoped that this study will contribute towards a more comprehensive understanding of the factors which affect LRs, and which experts must consider when conducting casework.

References

Aitken, C. G. G. (1991) Populations and samples. In Aitken, C. G. G. and Stoney, D. A. (eds.) *The use of statistics in forensic science.* London: Ellis Horwood*.* 51-82.

Aitken, C. G. G. and Lucy, D. (2004) Evaluation of trace evidence in the form of multivariate data. *Applied Statistics 54*: 109-122*.*

Aitken, C. G. G. and Stoney, D. A. (1991) The use of statistics in forensic science. London: Ellis Horwood.

Aitken, C. G. G. and Taroni, F. (2004) Statistics and the evaluation of evidence for forensic scientists (2nd edition). Chichester: John Wiley.

Alderman, T. (2004) The Bernard data set as a reference distribution for Bayesian likelihood ratio-based forensic speaker identification using formants. *Proceedings of the 10th Australian International conference on Speech Science and Technology.* 8-10 December 2004, Sydney, Australia. 510-515.

Ash, S. (1996) Freedom of movement: /uw/-fronting in the Midwest. In Arnold, J., Blake, R., Davidson, B., Schwentner, S. T. and Soloman, J. (eds.) *Sociolinguistic Variation: Data, Theory and Analysis – Selected Papers from NWAV 23 at Stanford*. Stanford CA: Centre for the Study of Language and Information (CSLI) Publications, Stanford University. 3-23.

Baldwin, D. J. (2005) Weight of evidence for forensic DNA profiles. Chichester: John Wiley.

Baldwin, J. and French, P. (1990) Forensic phonetics. London: Pinter.

Bernard, J. R. (1967) Some measurements of some sounds of Australian English. PhD Dissertation, University of Sydney.

Bernard, J. R. (1970) Toward the acoustic specification of Australian English. *Zeitschrift für Phonetik* 23: 113-128.

Broeders, A. P. A. (1999) Some observations on the use of probability scales in forensic identification. *Forensic Linguistics* 6(2): 228-241.

Brümmer, N. and du Preez, J. (2006) Application-independent evaluation of speaker detection. *Computer Speech and Language* 20(2-3): 230-275.

Buckleton, J. S., Walsh, K. A. J. and Evett, I. W. (1991) Who is 'random man'? *Journal of Forensic Science Society* 31: 463-468.

Byrne, C. and Foulkes, P. (2004) The mobile phone effect on vowel formants. *International Journal of Speech, Language and the Law* 11, 83-102.

Champod, C. and Evett, I. W. (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 Meuwly, D. (2000) The inference of identity in forensic speaker recognition. *Speech Communication* 31: 193-203.

Clarke, S., Elms, F. & Youssef, A. (1995) The Third Dialect of English: Some Canadian evidence. *Language Variation and Change* 7(2): 209-228.

Coleman, R. F. and Walls, H. J. (1974) The evaluation of scientific evidence. *Criminal Law Review*: 276-287.

Cudmore, A. (2011) The interpretation of forensic speaker comparison evidence by potential jurors in the UK. Unpublished MSc Dissertation, University of York.

Easton, L. and Bauer, L. (2000) An acoustic study of the vowels of New Zealand English'. *Australian Journal of Linguistics* 20(2): 93-117.

Eriksson, E., Cepeda, L. F., Rodman, R. D., McAllister, D. F. and Bitzer, D. (2004) Cross-language speaker identification using spectral moments. *Proceedings of the XVIIth Swedish Phonetics Conference FONETIK.* 76-79*.*

Evett, I. W. (1991) Interpretation: a personal odyssey. In Aitken, C. G. G. and Stoney, D. A. (eds.) *The use of statistics in forensic science*. London: Ellis Horwood. 9-22.

Evett, I. W., Jackson, G., Lambert, J. A. and McCrossan, S. (2000) The impace of the principles of evidence interpretation on the structure and content of statements. *Science and Justice* 40(4): 233-239.

Ferragne, E. and Pellegrino, F. (2010) Formant frequencies of vowels in 13 accents of the British Isles. *Journal of the International Phonetic Association* 40(1): 1-34.

Flynn, N.E.J. (2011) GOOSE-Fronting: It's happening in Nottingham t[ʉː]. Paper presented at the 8th UK Conference on Language Variation and Change (UKLVC8). 12-14 September 2011, Edge Hill University.

Foulkes, P., Docherty, G., and Jones, M. (2010). Analyzing stops. In Di Paolo, M. and Yaeger-Dror, M. (eds.) *Best Practices in Sociophonetics.* London: Routledge.

French, J. P. and Harrison, P. (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., Nolan, F., Foulkes, P., Harrison, P. and McDougall, K. (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.

Fridland, V. (2008) Patterns of /uw/, /ʊ/, and /ow/ fronting in Reno, Nevada. *American Speech* 83(4): 432-454.

Gold, E. (2011) Forensic speaker comparison evidence: the international picture. Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference. 24-28 July 2011, Acoustics Research Institute Vienna, Austria.

Gonzalez-Rodriguez, J., Rose, P., Ramos, D., Toledano, D. T. and Ortega-Garcia, J. (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.

Gordon, E., Maclagan, M. and Hay, J. (2007) The ONZE corpus. In Beal, J. C., Corrigan, K. P. and Moisl, H. (eds.) *Models and Methods in the Handling of Unconventional Digital Corpra: Volume 2, Diachronic Copora.* London: Palgrave. 82-104.

Greisbach, R., Osser, E. and Weinstock, C. (1995) Speaker identification by formant contours. In Braun, A. and Köstner, J. (eds.) *Studies in Forensic Phonetics. Beiträge zur Phonetik und Linguistik 64.* Trier: Wissenschaftlicher Verlag Trier. 49-55.

Haddican, B. and Richards, H. (2008-2013) A comparative study of language change in northern Englishes. Economic and Social Research Council (ESRC) of Great Britain. RES-061-25-0033.

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 Hardcastle, W. and Laver, J. (eds.) *A Handbook of Phonetic Science.* Oxford: Blackwell. 81-129.

Harrison, P. and French, P. (2010) Evaluation of the BATVOX automatic speaker recognition system for use in UK based forensic speaker comparison casework Part II. Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference. 18-21 July 2010, University of Trier, Germany.

Hawkins, S. and Midgley, J. (2005) Formant Frequencies of RP monophthongs in four age-groups of speakers. *Journal of International Phonetic Association* 35(2): 183-199.

Hawkins, S. and Clermont, F. (2009) A new approach to evaluating the likelihood ratio test for forensic speaker comparison: sample size, confidence intervals and intrinsic dimension. Unpublished paper submitted for presentation at InterSpeech International Conference. Brighton, United Kingdom.

Hay, J., Maclagan, M. and Gordon, E. (2008) Dialects of English: New Zealand English. Edinburgh: Edinburgh University Press.

Hughes, V., McDougall, K. and Foulkes, P. (2009) Diphthong dynamics in unscripted speech. Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference. 2-5 August 2009, University of Cambridge, Cambridge.

Hughes, V., Foulkes, P., Haddican, B. and Richards, H. (2011) Vowel variation in Manchester: a dynamic approach. Paper presented at the 8th UK Conference on Language Variation Change (UKLVC8). 12-14 September 2011, Edge Hill University.

Ingram, J. C. L., Prandolini, R. and Ong, S. (1996) Formant trajectories as indices of phonetic variation for speaker identification. *Forensic Linguistics* 3(1): 129-145.

Ishihara, S. and Kinoshita, Y. (2008) How many do we need? Exploration of the Population Size Effect on the performance of forensic speaker classification. Paper presented at the 9th Annual Conference of the International Speech Communication Association (Interspeech). Brisbane, Australia. 1941-1944.

Jessen, M. (2011) Conclusions on voice comparison evidence in Germany and a challenging case. Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference. 18-21 July 2010, University of Trier, Germany.

Johnson, K. (2003). Acoustic and auditory phonetics (2nd edition). Oxford: Blackwell.

Jones, D. (1966) The pronunciation of English (4th Edition). Cambridge: Cambridge University Press.

Kahneman et al (1982)

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.

Kaye, D. H. (2008) DNA probabilities in *People v. Prince*: When are racial and ethnic statistics relevant? In Speed, T. And Nolan, D. (eds.) *Probability and Statistics: Essays in Honour of David A Freedman*. Beachwood, OH: Institute of Mathematical Statistics. 289-301.

Kinoshita, Y. (2001) Testing realistic forensic speaker identification in Japanese: a likelihood ratio-based approach using formants. PhD Dissertation, Australian National University.

Kinoshita, Y. (2002) Use of likelihood ratio and Bayesian approach in forensic speaker identification. *Proceedings of the 9th Australian International conference on Speech Science and Technology.* 2-5 December 2002, Melbourne, Australia. 297-302.

Kinoshita, Y. (2005) Does Lindley's LR estimation formula work for speech data? Investigation using long-term f0. *Journal of Speech, Language and the Law* 12(2): 235-254.

Künzel, H. J. (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.

Labov, W. (1971) The study of language in its social context. In Fishman, J. A. (ed.) *Advances in the Sociology of Language (vol. 1)*. The Hague: Mouton. 152-216.

Ladefoged, P. (2006) A course in phonetics (5th edition). Boston: Wadsworth Cengage Learning.

Lenth, R. V. (1986) On identification by probability. *Journal of the Forensic Science Society* 26: 197-213.

Lindau, M. (1978) Vowel features. *Language* 54(3): 541-563.

Lindley, D. V. (1977) A problem in forensic science. *Biometrika* 64: 207-213.

Loakes, D. (2006) A forensic phonetic investigation into the speech patterns of identical and non-identical twins. PhD Dissertation, University of Melbourne.

Lynch, M. and McNally, R. (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. and Gordon, E. (1999) Data for New Zealand social dialectology: the Canterbury Corpus. *New Zealand English Journal* 13: 50-58.

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.

McDougall, K. (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.

McDougall, K. and Nolan, F. (2007) Discrimination of speakers using the formant dynamics of /uː/ in British English. In Trouvain, J. and Barry, W. J. (eds.) *Proceedings of the 16th International Congress of Phonetic Sciences.* 6-10 August 2007, Saarbrücken: Universität des Saarlandes. 1825-1828.

Milroy, L., Milroy, J. and Docherty, G. (1994-1997) Phonological Variation and Change in Contemporary British English. Economic and Social Research Council (ESRC) of Great Britain. Ref: R000234892.

Morrison, G. S. (2007) MatLab implementation of Aitken and Lucy's (2004) forensic likelihood ratio software using multivariate-kernel-density estimation. Downloaded: 31st May 2011.

Morrison, G. S. (2008) Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/. *Journal of Speech, Language and the Law* 15(2): 249-266.

Morrison, G. S., (2009a). The place of forensic voice comparison in the ongoing paradigm shift. Written version of an invited presentation given at the 2nd International Conference on Evidence Law and Forensic Science. 25-26 July 2009, Beijing, China. 1-16.

Morrison, G. S. (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.

Morrison, G. S. (2011a) Measuring the validity and reliability of forensic likelihood-ratio systems. *Science and Justice*. doi:10.1016/j.scijus.2011.02.002.

Morrison, G. S. (2011b) 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: 242-256.

Morrison, G. S. and Kinoshita, Y. (2008) Automatic-type calibration of traditionally derived likelihood ratios: forensic analysis of Australian English /o/ formant trajectories. *Proceedings of Interspeech 2008 Incorporating SST 2008*, International Speech Communication Association. 1501-1504.

Morrison, G. S., Zhang, C. and Rose, P. (2011) An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic Science International* 208: 59-65.

Narayanan, S., Byrd, D. and Kaun, A. (1999) Geometry, kinematics, and acoustics of Tamil liquid consonants. *Journal of the Acoustical Society of America* 106(4): 1993-2007.

Nolan, F. (1997) Speaker recognition and forensic phonetics. In Hardcastle, W. J. and Laver, J. (eds.) *The Handbook of Phonetic Sciences.* Oxford: Blackwell. 744-767.

Nolan, F. (2001) Speaker identification evidence: its forms, limitations, and roles. *Proceedings of the 'Law and Language: Prospect and Retrospect' Conference.* 12-15 December 2001, Levi, Finland.

Proctor, M. I. (2009) Gestural characterisation of a phonological class: the liquids. PhD Dissertation, Yale University.

Ramos Castro, D. (2007) Forensic evaluation of the evidence using automatic speaker recognition systems. PhD Dissertation, Universidad Autónoma de Madrid.

Redmayne, M. (1998) Bayesianism and proof. In Freeman, M. and Reece, H. (eds.) *Science in Court*. Aldershot: Athenaeum Press. 61-82.

Reynolds, D. A., Quatieri, T. F. and Dunn, R. B. (2000) Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing* 10: 19-41.

Robertson, B. and Vignaux, G. A. (1995) Interpreting evidence: evaluating forensic science in the courtroom. Chichester: John Wiley.

Rodman, R., McAllister, D., Bitzer, D., Cepeda, L. and Abbitt, P. (2002) Forensic speaker identification based on spectral moments. *Forensic Linguistics* 9(1): 22-43.

Rose, P. (2002) Forensic Speaker Identification. London: Taylor and Francis.

Rose, P. (2004) Technical Forensic Speaker Identification from a Bayesian Linguist's Perspective. Keynote paper, *Forensic Speaker Recognition Workshop, Speaker Odyssey '04*. 31 May - 3 June 2004, Toledo, Spain. 3-10.

Rose, P. (2006a) The intrinsic forensic discriminatory power of diphthongs. *Proceedings of the 11th Australasian International Conference on Speech Science and Technology.* 6-8 December 2006, University of Aukland, New Zealand. 64-69.

Rose, P. (2006b) Accounting for correlation in linguistic-acoustic likelihood ratio-based forensic speaker discrimination. *Proceedings of the Odyssey Speaker and Language Recognition Workshop*. 1-8.

Rose, P. (2007a) Going and getting it - Forensic Speaker Recognition from the perspective of a traditional practitioner/ researcher. 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.

Rose, P. (2007b) Forensic speaker discrimination with Australian English vowel acoustics. In Trouvain, J. and Barry, W. J. (eds.) *Proceedings of the 16th International Congress of Phonetic Sciences.* 6-10 August 2007, Universität des Saarlandes, Saarbrücken. 1817-1820.

Rose, P. (2011) Forensic voice comparison with Japanese vowel acoustics – a likelihood ratio-based approach using segmental cepstra. *Proceedings of the 17th International Congress of Phonetic Sciences.* 17-21 August 2011, Hong Kong. 1718-1721.

Rose, P., Kinoshita, Y. and Alderman, T. (2006) Realistic extrinsic forensic speaker discrimination with the diphthong /aI/. *Proceedings of the 11th Australasian International Conference on Speech Science and Technology.* 6-8 December 2006, University of Aukland, New Zealand. 329-334.

Rose, P. and Morrison, G. S. (2009) A response to the UK Position Statement on forensic speaker comparison. *International Journal of Speech, Language and the Law* 16(1): 139-163.

Saks, M. J. and Koehler, J. J. (2005) The coming paradigm shift in forensic identification science. *Science* 309: 892–895.

Seber, G. A. F. and Wild, C. J. (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. *Procceedings of Fonetik*: 93–96.

Smith, R. L. and Charrow, R. P. (1975) Upper and lower bounds for the probability of guilt based on circumstantial evidence. *Journal of the American Statistical Association* 70: 555-560.

Tabachnick, B. G. and Fidell, L. S. (1996) Using Multivariate Statistics (5th Edition). New York: Harper Collins.

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.

Thompson, W. C. and Schumann, E. L. (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.

Torgersen, E. N. and Kerswill, P. (2004) Internal and external motivation in phonetic change: dialect levelling outcomes for an English vowel shift. *Journal of Sociolinguistics* 8(1): 23-53.

Van Leeuwen, D. A. and Brümmer, N. (2007) An introduction to application-independent evaluation of speaker recognition systems. In Müller, C. (ed.) *Speaker Classification I, LNAI 4343*. Berlin: Spinger-Verlag. 330-353.

Watt, D. (1998) *Variation and change in the vowel system of Tyneside English. PhD Dissertation, University of Newcastle.*

Wells, J. C. (1982) The accents of English (3 vols). Cambridge: Cambridge University Press.

Whittle, P. (1983) Prediction and regulation by linear least-squared methods (2nd edition). Minneapolis: University of Minnesota Press.

Young, F. W. (1985) Multidimensional Scaling. In Kotz, S. and Johnson, N. L. (eds.) *Encyclopaedia of Statistical Sciences* 5: 649-659.

Zhang, C., Morrison, G. S. and Thiruvaran, T. (2011) Forensic voice comparison using Chinese /iau/. *Proceedings of the 17th International Congress of Phonetic Sciences.* 17-21 August 2011, Hong Kong. 2280-2283.

Legal Rulings

Daubert-v-Merrell Dow Pharmaceuticals, 509 U.S. 579 [1993]

R-v-Doheny & Adams [1996] EWCA Crim 728

R-v-T [2010] EWCA Crim 2439