The effect of sampling variability on overall performance and individual speakers’ behaviour in likelihood ratio-based forensic voice comparison

Xiao Wang

PhD

University of York
Language and Linguistic Science
October 2021

Abstract

In the past years, there is increasing awareness and acceptance among forensic speech scientists of using Bayesian reasoning and likelihood ratio (LR) framework for forensic voice comparison (FVC) and expressing expert conclusions. Numerous studies have explored overall performance using numerical LRs. Given that the data used for validation is a sample coming from an unknown distribution, little attention has been paid to the effect of sampling variability or individuals' behaviour. This thesis investigates these issues using linguistic-phonetic variables. First, it investigates how different configurations of training, test and reference speakers affect overall performance. The results show that variability in overall performance is mostly caused by varying the test speakers, while less variability is caused by sampling variability in the reference and training speakers. Second, this thesis explores the effect of sampling variability on overall performance and individuals' behaviour in relation to the use of linguistic-phonetic features. Results show that sampling variability affects overall performance to different extents using different features, while combining more features does not always improve overall performance. Sampling variability has limited effects on individuals in same-speaker comparisons, and most speakers are less affected by sampling variability in different-speaker comparisons when four or more features are used. Third, this thesis explores the effect of sampling variability on overall performance in relation to score distributions. Results reveal that system validity and reliability are more affected by different-speaker score skewness, and less affected by same-speaker score skewness. Using different calibration methods reduces the effect of sampling variability to different extents. The results in this thesis have implications for both FVC using numerical LRs and FVC in general, as experts need to make pragmatic decisions whether numerical LR is used or not, and every decision made has implication to final evaluation results. Further, the results on score skewness and different calibration methods have potential contribution for improving FVC performance using automatic systems.

Contents

Abstract......2

Contents......3

List of Tables......7

List of Figures......9

Acknowledgements......14

Declaration......16

Chapter 1 Introduction......18

1.1 Analytic Methods in FVC......18

1.2 Conclusion frameworks in FVC......20

1.3 Validation and regulation......24

1.4 Established procedures for validation......26

1.5 Research aims and questions......27

1.6 Thesis outline......29

Chapter 2 Research Context......31

2.1 Bayes' Theorem and the likelihood ratio (LR) framework......31

2.2 The LR approach in FVC......33

2.2.1 System testing......35

2.2.1.1 System Validity......36

2.2.1.2 System reliability......40

2.2.2 Factors affecting system validity and reliability......45

2.2.3 Current study......49

Chapter 3 General Methodology......51

3.1 Corpora 51
3.1.1 IARPA Babel Cantonese Language Pack 51
3.1.2 Dynamic Variability in Speech (DyViS) corpus 52
3.2 Segmental features 54
3.2.1 Cantonese segmental features 54
3.2.2 English filled pause *um* 55
3.2.3 Segmentation and feature extraction 55
3.2.4 Feature parameterisation 63
3.3 LR computation 66
3.3.1 Feature-to-score 67
3.3.2 Score-to-LR 69
3.3.2.1 Logistic regression 69
3.3.2.2 Regularised logistic regression 71
3.3.2.3 Bayesian model 72
3.3.2.4 Empirical lower and upper bound (ELUB) 74
3.4 Evaluation 77
3.4.1 System validity and reliability 78
3.4.1.1 Log LR cost function (C_{llr}) 78
3.4.1.2 Equal error rate (EER) 80
3.4.2 Individual behaviour 80
3.4.2.1 Root-mean-square-deviation (RMSD) 80
3.4.2.2 Zoo plot 81
3.4.3 Verbal LR scale 83
3.5 Chapter Summary 84
Chapter 4 Overall performance as a function of sampling variability 87
4.1 Introduction 87
4.2 Method 88
4.2.1 Input data and LR computation 88
4.3 Experiments 88
4.4 Results 91
4.4.1 Experiment 1: varying all speakers. 91

4.4.2 Experiment 2: varying test speakers. 94

4.4.3 Experiment 3: varying reference speakers 97

4.4.4 Experiment 4: varying training speakers 100

4.5 Discussion 102

4.5.1 The performance reliability of segmental variables 103

4.5.2 Test speakers *vs*. training speakers *vs*. reference speakers 105

4.5.3 System vs individual 106

4.6 Chapter summary 109

Chapter 5 Overall performance and individual speakers' behaviour as functions of choice of linguistic-phonetic features and configurations of training and reference speakers 110

5.1 Introduction 110

5.2 Method 111

5.2.1 Data 111

5.2.2 Experimental procedure 113

5.3 Results 114

5.3.1 Overall performance 114

5.3.2 Individual behaviour 119

5.3.2.1 Individual validity 119

5.3.2.2 Individual reliability 131

5.3.2.2.1 Single-feature systems 135

5.3.2.2.2 Two-feature systems 138

5.3.2.2.3 Three-feature systems 141

5.3.2.2.4 Four/five-feature systems 143

5.3.3 What happens in a real case? 146

5.4 Discussion 150

5.5 Chapter summary 152

Chapter 6 Overall performance as a function of score skewness, sample sizes and calibration methods 153

6.1 Introduction 153

6.1.1 Polit study 154

6.2 Method 156

6.2.1 Score distribution 156

6.2.2 Score simulation and LR computation 157

6.3 Results 161

6.3.1 Varying the skewness of both SS and DS scores 161

6.3.2 Varying only SS score skewness 163

6.3.3 Varying only DS skewness 164

6.4 Discussion 165

6.5 Chapter summary 167

Chapter 7 Discussion and Conclusion 168

7.1 Sampling variability at *feature-to-score* stage 168

7.2 Sampling variability at *score-to-LR* stage 169

7.3 Direct Implications 171

7.4 Wider Implications 173

7.5 Future studies 174

7.6 Conclusion 175

Appendix 177

Appendix A - Simulating raw scores 177

Appendix B - Zoo plots of SS and DS LLRs of 25 test speakers in 31 systems, and strip texts on top of the plots indicate specific system. 186

Appendix C - Plots of mean SS and DS LLR of individual speakers across 31 systems. Strip texts on top indicate speaker ID. 202

List of abbreviations 209

Legal cases 211

Bibliography 212

List of Tables

Table 1.1 Classical probability verbal scales (adapted from Broeders, 1999, p.229).............21

Table 2.1 Assumed probability of each individual appearance characteristics (Table from Finkelstein & Fairley, 1970, p. 491). 33

Table 3.1 Numbers of recordings across five regions in Guangdong and Guangxi provinces. 51

Table 3.2 Categorical calculation of EER using the LLR as a discriminant function. 80

Table 3.3. LLR threshold for animal groups in current study. 83

Table 3.4 Numerical LRs in verbal expression Champod & Evett (2000). 84

Table 4.1 Average number of tokens and recording sessions of each linguistic-phonetic variable...88

Table 4.2 Number of speakers used in each data set for each variable in Experiment 1.........89

Table 4.3 Number of speakers used in each data set for each variable in Experiments 2 to 4. 90

Table 4.4. Statistical summary of /a/ and *um* in experiment 1...93

Table 4.5. Minimum, maximum, median, mean, first and third quartiles of C_{llr}s and EERs of /a/, /haia/ and *um* in Experiment 2. 97

Table 4.6. Statistic summary of C_{llr}s and EERs of /a/, /haia/ and *um* in sampling reference speakers...100

Table 4.7. Statistic summary of C_{llr}s of /a/, /haia/ and *um* in Experiment 4. 102

Table 4.8 Mean, IQR and range of C_{llr} values of /a/, /haia/ and *um* in Experiments 2, 3 and 4. 106

Table 5.1 31 systems tested and cross (X) indicates the feature(s) used. *DUR* stands for durations of the vocalic and nasal of the FP *um*. 112

Table 5.2 Statistical summary of C_{llr} in systems with one and two features.........................116

Table 5.3. Statistic summary of C_{llr} in systems with three, four and five features. 117

Table 5.4 Animal groups of speakers in single-feature systems..124

Table 5.5 Animal groups of speakers in two-feature systems. 126

Table 5.6 Animal groups of speakers in three-feature systems. 128

Table 5.7 Animal groups of speakers in four-/five-feature systems. 130

Table 5.8 Correlation between the validity and reliability of individual speakers. 134

Table 5.9 SS and DS RMSE values of speakers with the least and most variable LLRs. 147
Table 6.1 Distribution parameters of SS/DS log scores. 157
Table 6.2 Parameter values used for ST distribution simulation. 158

Table A.1 Summary statistics of SS and DS Ln and raw scores. 177
Table A.2 Parameter values of fitted Weibull and gamma distributions from MLE. 179
Table A.3 AD and BIC values of gamma and Weibull fitted to empirical SS and DS scores. 183
Table A.4 Three sets of scale and shape parameters for SS and DS score simulation. 185

List of Figures

Figure 1.1 A demonstration of similarity and typicality. 24

Figure 2.1 Conceptual demonstration of validity and reliability (Figure modified from Morrison, 2016, Figure 1, p.372) 36

Figure 3.1 An example of the map used for task 1 containing information of the story told by mock suspect (Nolan et al. 2009:42). 53

Figure 3.2 Segmented Cantonese variable /a/ (top) and /haia/ (bottom). 56

Figure 3.3 Unanalysable F3 (top) in Gold (2009, p.39) and unanalysable F3 of /a/ from IARPA (bottom).. 58

Figure 3.4: Praat formant setting window... 60

Figure 3.5 Formant extraction using different formant settings in Praat. 61

Figure 3.6 Quadratic curve fitting to F0, F1, F2 and F3 of five tokens from #114 in DyViS Task 1.. 63

Figure 3.7 Example of formant dynamic analysis. Figure 3.8 from (Hughes, 2014, p.82). Measurement is taken with a 10% interval of /aɪ/ in the word *skype* from DyViS sample 027-01-060425.wav 64

Figure 3.8 Quadratic and cubic curve fitting to /a/ (left panel) and /haia/ (right panel) respectively. 65

Figure 3.9 Plot of raw formant values of /haia/ fitted with cubic polynomial curves. 66

Figure 3.10 Distributions of SS (red) and DS (blue) scores (upper panel), logistic regression fitted to SS and DS scores (middle panel) in probability space and linear relationship between SS and DS scores derived from the logistic regression modelling in log-odds space (bottom panel). (Morrison 2013: 182 Figure 4).. 70

Figure 3.11 An example of regularised logistic regression using different κ^{ψ} values (Figure 2 in Morrison & Poh, 2018, p. 205). 72

Figure 3.12 SS (H_p) and DS (H_d) scores fitted with different density models (Figure 2 in Vergeer et al., 2016, p.483) 75

Figure 3.13 A Zoo plot with the relative location of the new and old classes of animals (Dunstone and Yager, 2009, p. 168, Figure 8.5).. 82

Figure 4.1 Boxplots show the C_{llr} ranges of /a/, /haia/ and *um* by varying speakers in all three datasets in each replication. Two hinges are the first and third quartiles, and the central line represents the median value. 92

Figure 4.2 Boxplots represent the EER ranges of /a/, /haia/ and *um* by varying speakers in all three datasets in each replication. Two hinges are the first and third quartiles, and the central line represents the median value. 93

Figure 4.3 Boxplots represent the C_{llr} ranges of /a/, /haia/ and *um* by varying test speakers in each replication. Two hinges are the first and third quartiles, and the central line represents the median value. 95

Figure 4.4 Boxplots represent the EER ranges of /a/, /haia/ and *um* by varying test speakers in each replication. Two hinges are the first and third quartiles, and the central line represents the median value. 96

Figure 4.5 Boxplots represent the C_{llr} ranges of /a/, /haia/ and *um* by varying reference speakers in each replication. Two hinges are the first and third quartiles, and the central line represents the median value. 98

Figure 4.6 Boxplots represent the EER ranges of /a/, /haia/ and *um* by varying reference speakers in each replication. Two hinges are the first and third quartiles, and the central line represents the median value. 99

Figure 4.7 Boxplots represent the C_{llr} ranges of /a/, /haia/ and *um* by varying training speakers in each replication. Two hinges are the first and third quartiles, and the central line represents the median value. 101

Figure 4.8 Boxplots of C_{llr} and EER of /a/, /haia/ and *um* in four experiments. Two hinges are the first and third quartiles, and the central line represents the median value. 103

Figure 4.9 SS and DS LLRs of test speakers using FP *um* in Experiment 3. 108

Figure 5.1 A schematic diagram of experiment procedure in LR computation.........113

Figure 5.2 C_{llr} variation across 31 systems (Top panel: systems with one or two features; bottom left: systems with three features; bottom middle: systems with four features; bottom right: system with five features). 115

Figure 5.3 C_{llr} mean (upper panel) and range (lower panel) across 31 systems.........118

Figure 5.4 C_{llr} mean and range across 31 systems. 119

Figure 5.5 Mean SS and DS LLR of speakers #51, #48 and #53 across 31 systems. Two triangles marked by 'F02' in lower panels indicate that speakers #48 and #53 can be well separated using a two-feature system (F02). 120

Figure 5.6 Mean SS and DS LLR of all 25 test speakers across 31 systems (see Appendix C for full size). 121

Figure 5.7 Zoo plot of SS and DS LLR of 25 test speakers in single-feature systems (see Appendix B for full size zoo plots). 123

Figure 5.8 Zoo plot of SS and DS LLR of 25 test speakers in two-feature systems (see Appendix B for full size zoo plots). 125

Figure 5.9 Zoo plot of SS and DS LLR of 25 test speakers in three-feature systems (see Appendix B for full size zoo plots). 127

Figure 5.10 Zoo plot of SS and DS LLR of 25 test speakers in four-/five-feature systems (see Appendix B for full size zoo plots). 129

Figure 5.11 RMSE values of individual speakers across 25 systems. 132

Figure 5.12 SS and DS RMSE ranges across different systems. 133

Figure 5.13 SS RMSE values of all 25 test speakers across single-feature systems. 136

Figure 5.14 DS RMSE values of all 25 test speakers across single-feature systems. 137

Figure 5.15 SS RMSE values of all 25 test speakers across two-feature systems. 139

Figure 5.16 DS RMSE values of all 25 test speakers across two-feature system......140

Figure 5.17 SS RMSE values of all 25 test speakers across three-feature systems. ..142

Figure 5.18 DS RMSE values of all 25 test speakers across three-feature systems. .143

Figure 5.19 SS RMSE values of all 25 test speakers across four-/five-feature systems. 144

Figure 5.20 DS RMSE values of all 25 test speakers across four-/five-feature systems. 145

Figure 5.21 SS LLRs of speaker #30 and #40 using different training and reference speakers in the F0123 system. 148

Figure 5.22 DS LLRs of speaker #20 and #51 using different training and reference speakers in the F0123 system. 149

Figure 6.1 Histogram and density distribution of SS and DS Log scores from FP *um*. 157

Figure 6.2 Distributions of simulated SS log scores by varying the skewness..........159

Figure 6.3 Simplified process of *score-to-LR* computation. 160

Figure 6.4 Mean and range of C_{llr} values varying the skewness in both SS and DS scores. 162

Figure 6.5 Mean and range of C_{llr} values only varying the SS score skewness. 164

Figure 6.6 Mean and range of C_{llr} values only varying the DS score skewness. 165

Figure A.1 CF graph of SS and DS empirical scores (plot produced using *fitdistrplus* package (Delignette-Muller and Dutang, 2015) in R (RStudio Team 2019)). 178

Figure A.2 Goodness-of-fit plots of gamma and Weibull distributions fitted to FP *um* empirical raw scores. 182

Figure A.3 Goodness-of-fit plots of gamma and Weibull distributions fitted to FP *um* empirical raw scores. 183

Figure A.4 Simulated score distributions with different scale parameters. 184

Figure B.1 Zoo plots of SS and DS LLRs of 25 test speakers in DURATION and F0 systems. 186

Figure B.2 Zoo plots of SS and DS LLRs of 25 test speakers in F1 and F2 systems. 187

Figure B.3 Zoo plots of SS and DS LLRs of 25 test speakers in F3 system. 188

Figure B.4 Zoo plots of SS and DS LLRs of 25 test speakers in F01 and F02 systems. 189

Figure B.5 Zoo plots of SS and DS LLRs of 25 test speakers in F03 and F0DUR systems. 190

Figure B.6 Zoo plots of SS and DS LLRs of 25 test speakers in F12 and F13 systems. 191

Figure B.7 Zoo plots of SS and DS LLRs of 25 test speakers in F1DUR and F23 systems. 192

Figure B.8 Zoo plots of SS and DS LLRs of 25 test speakers in F2DURand F3DUR systems. 193

Figure B.9 Zoo plots of SS and DS LLRs of 25 test speakers in F03DUR and F13DUR systems. 194

Figure B.10 Zoo plots of SS and DS LLRs of 25 test speakers in F023 and F02DUR systems. 195

Figure B.11 Zoo plots of SS and DS LLRs of 25 test speakers in F12DUR and F23DUR systems. 196

Figure B.12 Zoo plots of SS and DS LLRs of 25 test speakers in F01DUR and F123 systems. 197

Figure B.13 Zoo plots of SS and DS LLRs of 25 test speakers in F012 and F013 systems. 198

Figure B.14 Zoo plots of SS and DS LLRs of 25 test speakers in F0123 and F012DUR systems. 199

Figure B.15 Zoo plots of SS and DS LLRs of 25 test speakers in F123DUR and F013DUR systems. 200

Figure B.16 Zoo plots of SS and DS LLRs of 25 test speakers in F023DUR and F0123DUR systems. 201

Figure C.1 Mean SS and DS LLR of speakers 8, 13, 17 and 20 across 31 systems. .202

Figure C.2 Mean SS and DS LLR of speakers 21, 26, 27 and 30 across 31 systems. 203

Figure C.3 Mean SS and DS LLR of speakers 36, 40, 46 and 47 across 31 systems. 204

Figure C.4 Mean SS and DS LLR of speakers 48, 51, 53 and 54 systems. 205

Figure C.5 Mean SS and DS LLR of speakers 56, 72, 77 and 79 across 31 systems. 206

Figure C.6 Mean SS and DS LLR of speakers 90, 94, 114 and 118 across 31 systems. 207

Figure C.7 Mean SS and DS LLR of speaker 120 across 31 systems. 208

Acknowledgements

I cannot express how grateful I am to my first supervisor Professor Paul Foulkes. Thank you for teaching me forensic phonetics since my master’s study at York. Thank you for your patience and help before my PhD journey back in 2015. Then, we met at IAFPA and had a brief discussion about my PhD plan in 2016 and I started my PhD in 2017. Thank you for your patience and help during my PhD where I shifted the focus of my PhD from sociophonetics (Cantonese) to forensic phonetics. Thank you for your patience and help at the end of my PhD, for bridging me with Professor Ghada Khattab so I could teach phonetics at Newcastle University. Thank you for your help along the journey and I really could not ask for more from you.

I would like to express my sincere gratitude to my second supervisor Dr Vincent Hughes. Thank you for the countless LR and scripting chat. Thank you for your guidance that helped me in understanding the differences between different statistical models. Thank you for your advice and guidance in the concept of validation in forensics as well as score simulations for system testing, and I would not get a better understanding of these without your help. Thank for the encouragement and being a big brother during all the casual conversations.

I would like to thank my Thesis Advisory Panel member, Dr George Bailey. Thank you for introducing me to *FAST TRACK*. I wish I could have known this earlier. Thank you for bringing up the idea of degree of objectivity in doing research and that really helped me with the Chapter 5 in this thesis. I would also like to thank you, as well as Professor Carmen Llamas and Dr Claire Childs, for your patience and help during my GTA work. Thanks to Dr George Brown for teaching me *Forced Alignment*. Thanks to Dr Phil Harrison for pointing me to the right literature in score normalisation in ASR. Thanks to Dr Michael Jessen for the discussion and your insights into the number of Gaussians in ASR systems.

I would like to express my deep gratitude to Dr Ricky Chan, who offered me the opportunity to work as an RA at the University of Hong Kong. Thank you for your guidance and patience throughout my days in Hong Kong and when I made mistakes. Thank you for taking care of me and for all the night chats in the lab when I was down. Thank you for all the fried chicken, sashimi and ramen noodles as well as the academic trips to Seoul and Chongqing. Thanks to Dave, Adas, Elva, Kelly and Fei Gor for spending time with me in HK.

I would also like to thank my PhD-peer Justin. Thank you for sharing the R package for LR analysis with me and your help with my R script, which speeded up my experiments to a whole different level. Thank you, Justin, for all the conversations. Thanks to Kati for your encouragement and constantly checking on me, just to make sure I am doing OK. Thanks to my housemates, Roky, Tom and Ray. Thank you for your companionship and for being patient with me when I nag a lot.

Finally, I want to thank my mum for letting me doing what I want, but sorry mum that I could not give you a grandchild yet (not ready). Thank you, Miss Chan, for sweeping me off my feet back in Portsmouth nearly 10 years ago. Thank you for your cuddles, smiles, words and support throughout my PhD. Thank you for keeping me off the ground ever since.

Declaration

Elements of Chapter 4 have previously been published:

- Wang, B. X., Hughes, V. & Foulkes, P. (2018). A preliminary investigation of speaker randomisation in likelihood-ratio based forensic voice comparison. Poster presentation at the 27th Annual conference of *International Association for Forensic Phonetics and Acoustics*. University of Huddersfield, UK. 29 July-1 August 2018.
- Wang, B. X., Hughes, V. & Foulkes, P. (2019). The effect of speaker sampling in likelihood ratio based forensic voice comparison. *International Journal of Speech, Language and the Law*, 26(1), 97 - 120. https://doi.org/10.1558/ijsll.38046

Elements of Chapter 5 have previously been published:

- Wang, B. X., Hughes, V. and Foulkes, P. (2021) *System performance and speaker individuality in LR-based forensic voice comparison.* Oral presentation at Association Italiana Scienze della Voce (AISV). University of Zurich, Switzerland. 4-5 February 2021.
- Wang, B. X., Hughes, V. & Foulkes, P. (under review). The effect of sampling variability on systems and individual speakers in likelihood ratio-based forensic voice comparison. *Speech Communication.*

Elements of Chapter 6 have previously been published:

- Wang, B. X., Hughes, V., & Foulkes, P. (2019). Effect of score sampling on system stability in Likelihood Ratio based forensic voice comparison. *Proceedings of the 19th International Congress of Phonetic Sciences,* Melbourne, Australia (pp. 3065 - 3069). Canberra, Australia: Australasian Speech Science and Technology Association Inc.
- Wang, B. X. and Hughes, V. (2021) System performance as a function of score skewness, calibration methods and sample size in likelihood ratio-based forensic speech comparison. Presentation at *International Association for Forensic Phonetics and Acoustics,* University of Marburg, Germany. 22 – 25 August 2021.

- Wang, B. X., Hughes, V. (2021) System Performance as a Function of Calibration Methods, Sample Size and Sampling Variability in Likelihood Ratio-Based Forensic Voice Comparison. Proc. *Interspeech* 2021, 381-385, doi: 10.21437/Interspeech.2021-267

Bruce Xiao Wang

October 2021

Chapter 1 Introduction

Forensic speech science (FSS) is the study and application of phonetics, acoustics, signal processing and logic to legal cases (Jessen, 2008; Nolan, 2001; Rose, 2002). Depending on the specific case conditions, experts are faced with different tasks when speech evidence is involved. For example, speaker profiling is requested when an offender's speech sample is available, but no suspect has been found. The job of forensic speech scientists is then to ascertain information about the speaker's regional, social and ethnic background based on the offender speech sample to narrow down the population of suspect(s) (Foulkes & French, 2001). Voice line-ups/parades are conducted when the voice of the offender is heard by the victim or witness at the scene of a crime. The ear-witness is then asked to pick out the voice of the suspect from a set of foils with a similar accent. However, many factors can affect the validity of voice parade evidence, e.g., length of utterance (Blatchford & Foulkes, 2006), obstacles between the speaker and listener (Fecher & Watt, 2013), memory decay of the victim (McGehee, 1937) etc. Sometimes, forensic speech scientists are faced with recordings containing questioned utterances, where the speech samples could be mis-transcribed or an alternative interpretation can be suggested. Expert analysis is often sought under those circumstances (see French & Harrison, 2006 for a case example). However, by far the most common type of casework carried out by forensic speech scientists is forensic voice comparison (FVC), which accounts for ca. 70% of cases (Foulkes & French, 2012). Such cases involve a comparison of speech samples, one of an unknown offender, and the other of a known suspect typically recorded during the police interview, e.g., in the UK (Home Office, 2003), or through wiretaps, e.g., in Germany and China (Liu, 2006). The job of the expert is to examine the similarity and typicality between these speech samples to estimate the extent to which the evidence supports the competing propositions of the prosecution and defence, which in turn assists the trier-of-fact in making a decision about the innocence or guilt of the accused.

1.1 Analytic Methods in FVC

Among forensic speech scientists, approaches used for FVC differ between different labs across different countries, and there is no universal consensus about how FVC analysis should be carried out. Two surveys were conducted by Gold & French (2011, 2019) reaching 36 participants across 13 countries in 2011 and 39 participants across 23 countries in 2019, which aimed to investigate current practices in FVC. The participants were professionals mainly from the European Network of Forensic Science Institutions (ENFSI), Forensic Speech and Audio Analysis Working Group (FSAAWG) and the International Association for Forensic Phonetics and Acoustics (IAFPA). Similarly, Morrison et al. (2016) conducted a survey that was circulated to law enforcement agencies in the 190 member countries of Interpol. 91 responses were received from 69 countries, and 44 out of 91 responses stated that they have the ability to carry out speech analysis. Although there is a mismatch between the targeting participants in the surveys conducted by Gold and French (2011, 2019) and Morrison et al. (2016), the methods employed in FVC across different individual practitioners, police labs or institutes can be classified into five categories, i.e., auditory phonetic analysis only (AuPA), acoustic phonetic analysis only (AcPA), auditory phonetic cum acoustic phonetic analysis (AuPA + AcPA), human-assisted automatic speaker recognition (HASR) and fully automatic speaker recognition (ASR).

In the AuPA method, forensic speech scientists make qualitative judgements using categorical-phonetic transcriptions and description following the conventions of International Phonetic Alphabet (IPA) (Jessen, 2008). Detailed analysis can be carried out upon segmental features (e.g., vowels and/or consonants) and/or suprasegment features (e.g., overall fundamental frequency, voice quality). According to Gold & French (2011), 94% of the respondents evaluate the auditory quality of vowels and 88% of the respondents evaluate the auditory quality of consonants when carrying out FVC tasks. Meanwhile, 94% of the respondents evaluate fundamental frequency and voice quality using AuPA analysis as part of the overall procedure. In Morrison et al. (2016), approximately 82% of the respondents reported that the AuPA method was employed in speech analysis; however, details about specific segmental or suprasegmental linguistic-phonetic features used in the analysis were not provided.

In the AcPA method, the analysis is carried out making acoustic measurements of segmental and suprasegmental linguistic-phonetic features. For example, the value of the first three formants of the same vowel from different speech samples are often extracted and compared, which allows experts to make quantitative judgements about the speech samples under analysis. Furthermore, the mean fundamental frequency is also often measured. The AuPA and AcPA method is then the combination of auditory and acoustic analysis and this is generally the most widely used method across the world. According to Gold & French (2011), 77% of the respondents (10 out of 13 countries) carry out speech analysis using the combination of auditory and acoustic analysis. The percentage is a little lower in Morrison et al. (2016), where 57% respondents (25 out of 44 countries) reported that both auditory and acoustic analysis were involved.

Both Gold & French (2011, 2019) and Morrison et al. (2016) surveyed the use of ASR and HASR in their questionnaires. In the ASR method, the speech signal is analysed holistically, i.e., speech is not as analysed with respect to specific linguistic-phonetic units. In ASR systems, typically the speech active portion is first identified and acoustic features (e.g., Mel-frequency cepstral coefficients, MFCCs) are extracted using a fixed-length window shifted at fixed intervals across the sample. By contrast, HASR differs in terms of the degree of human supervision in their operation. In both ASR and HASR, acoustic features are modelled to produce a speaker model, which in turn is used to generate a score capturing the similarity and typicality between a pair of recordings. State-of-the-art systems (xVectors; Snyder et al., 2018) now integrate machine learning at various stages within an ASR system's processing. The ASR systems seem to have grown in acceptance among the community of forensic speech scientists in recent years. Gold & French (2019) have shown that only 17% of respondents reported using an ASR system in their survey in 2011, while 41% of respondents stated they used an ASR system in the 2019 survey. However, it was not indicated explicitly how they used ASR and how the use of ASR was in relation to AuPA and AcPA. Similarly, in Morrison et al. (2016), a total of 63% of the respondents reported that ASR (18%) and HASR (45%) were used in their analysis. The advantage of using an ASR system is that it is less labour intensive; however, the drawback is that the ASR system is often described as a black-box and still struggles with the range of conditions that might be faced in forensic cases in the UK.

1.2 Conclusion frameworks in FVC

There is also diversity in the conclusion frameworks used by labs and institutions across different countries. Different terms were used in the surveys in Gold & French (2011, 2019) and Morrison et al. (2016); however, these conclusion frameworks are grouped into four types for simplicity in the current thesis, i.e., binary decision, classical probability scales (numerical or verbal), UK Position Statement (UKPS; French & Harrison, 2007) and likelihood ratio framework (LR; numerical or verbal). As the name suggests, binary decision simply gives the conclusion in terms of the voices in the suspect sample and offender sample match or mismatch.

The binary decision is the least commonly employed conclusion framework among the practitioners surveyed as only 5% of the participants reported using this framework in Gold & French (2011, 2019); the use of binary decision is not surveyed in Morrison et al. (2016).

Classical probability scales were introduced to allow practitioners to give conclusions in terms of the gradient probability scales. Table 1.1 shows an example of the classical probability scales. A comparison between Gold & French (2011, 2019) and Morrison et al. (2016) shows that this conclusion framework is gradually less preferred among forensic speech science communities. 31% of the participants reported using the classical probability scales in Gold & French (2011), while the percentage dropped to 16% and 13% in Morrison et al. (2016) and Gold & French (2019) respectively. This is probably due to the fact that the classical probability scales put forensic speech scientists in a position to evaluate the probability of proposition given evidence (posterior probability; see Chapter 2.1), which is neither logical nor legal (Rose & Morrison, 2009, p.8). The job of forensic speech scientists is to evaluate the strength of evidence (i.e., the probability of evidence given proposition), rather than making a decision about the innocence or guilt of the accused (i.e., the probability of proposition given evidence).

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

Table 1.1 Classical probability verbal scales (adapted from Broeders, 1999, p.229). Left column shows verbal scales for positive identification and right column shows verbal scales for negative identification.

Compared to the probability scales, the later developed UKPS (French & Harrison, 2007) allowed practitioners to give three conclusions, namely consistent (identification), inconsistent (exclusion) and no decision (inconclusiveness). Figure 1.1 shows a schematic flow chart of UKPS conclusion framework. Using the UKPS conclusion framework, suspect and offender samples are compared and the exclusion decision (different speakers) is given if they are deemed inconsistent. An inconclusive decision is given if no decision can be made by the practitioners. If the suspect and offender samples are judged consistent, then five levels of distinctiveness can be given, ranking from the least distinctive to the most. According to Gold & French (2011, 2019), the UKPS framework was employed by 31% of the participants surveyed in 2011; however, the figure dropped to 13% in 2019. In Morrison et al. (2016), 57% (25 out of 44) of the participants reported using a conclusion framework (i.e., identification/exclusion/inconclusion) that is similar to the UKPS framework. It is noted that the UKPS and identification/ exclusion/inconclusion frameworks were treated as two different frameworks in Morrison et al. (2016); however, they are grouped into the same framework in current thesis due to their similarity. Although the UKPS framework acknowledged that the experts' job should solely be examining the probability of evidence given the competing propositions, there are a few violations in the framework which in effect evaluates the probability of proposition given evidence. For example, Rose and Morrison (2009) pointed out that regarding to the *not-consistent* decision, the UKPS framework stated that "Where the samples are not consistent we see no logical flaw in making the statement that the samples are spoken by different speakers" (French & Harrison, 2007, p.141). A statement indicating the speech samples come from the same or different speakers due to consistency or inconsistency expresses the posterior probability, i.e., $\mathrm{p(H|E)}$; however, the ultimate role of forensic phoneticians is to evaluate the probability of evidence given proposition ($\mathrm{p(E|H)}$), i.e., what is the probability that the two speech samples are consistent given they are produced by same speaker and what is the probability that the two speech samples are inconsistent given they are produced by different speakers? (see Rose & Morrison, 2009 for detailed discussion).

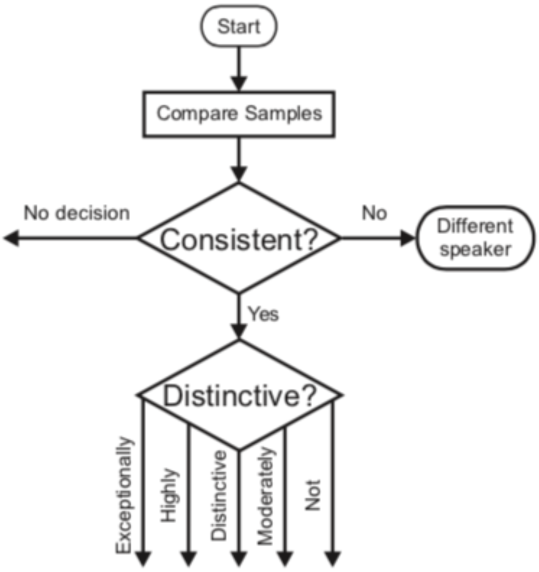


Figure 1.1 Flow chart representation of the UKPS (Figure 1 from; Rose & Morrison, 2009, p.3)

By contrast, the LR conclusion framework has come to be employed by more forensic speech scientists over the past decades. According to Gold and French (2011, 2019), 20% of the respondents stated that they express their evaluation conclusion using the LR framework (verbal or numerical) in 2011, while the figure almost doubled (39.5%) in 2019. Meanwhile, 43% of the participants from law enforcement agencies reported using the LR conclusion framework in Morrison et al. (2016). Expressing conclusions using the LR framework means that the experts need to explicitly assess both the similarity and typicality between two speech samples. Figure 1.2 gives a conceptualised demonstration of similarity and typicality in FVC (or any type of biometric/forensic evidence evaluation). The x- and y-axes represent hypothesised feature dimensions, and "off", "sus" and "ref" stand for the offender, suspect and reference speaker data. Both the left and right panels show a higher similarity between the suspect and offender samples; however, the left panel shows a lower typicality, while the right panel shows a higher typicality. As a result, the magnitude of the strength of evidence is higher in the left panel than that in the right panel (see Chapter 2 for more discussion in Bayesian reasoning and LR framework in FVC). It is important to note that the LR framework does not have to be numerical. The LR itself is a logical conceptual framework where the experts can assess the typicality and similarity between speech samples based on one's expertise or literature and internal calibration can be conducted by the expert.

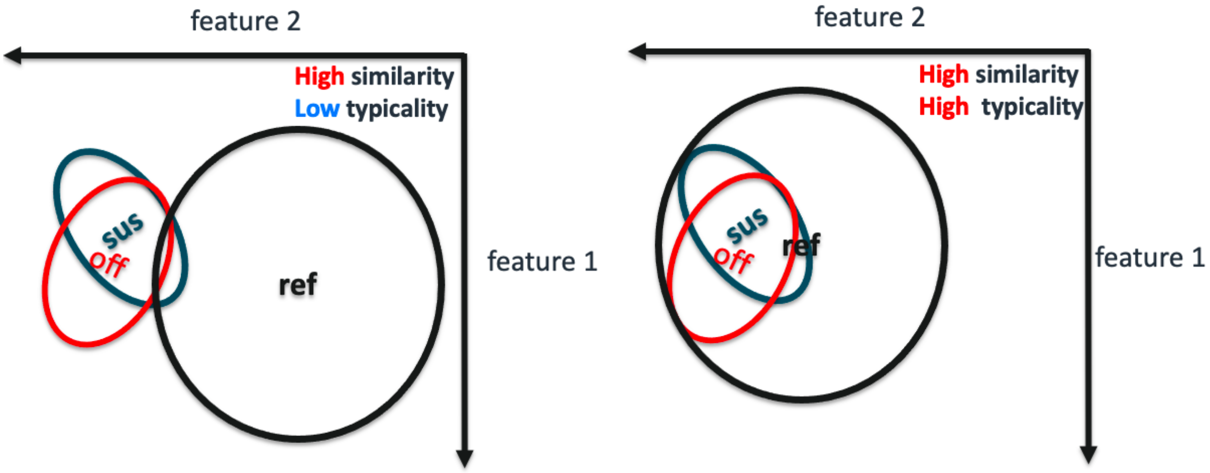


Figure 1.2 A demonstration of similarity and typicality. The left and right panels show different magnitude of the strength of evidence due to different degree of similarity and typicality.

1.3 Validation and regulation

Forensic experts and laboratories are now under increasing national and international regulatory pressure to empirically validate their methods (qualitative or quantitative) as well as demonstrate validity and reliability of the evaluation results to the trier of fact. The application of valid and reproducible methods is required for forensic evidence evaluation across many jurisdictions, e.g., the USA (Daubert ruling [1993]), the England & Wales Criminal Practice Directions 19A (CPD, 2015) and UK Crown Prosecution Service (CPS, 2019). The Daubert ruling [1993] sets forth a non-exclusive checklist for the test of admissibility of experts' testimony in courts:

> ***(1)*** *whether the expert's technique or theory can be or has been tested…;*
> ***(2)*** *whether the technique or theory has been subject to peer review and publication;*
> ***(3)*** *the known or potential rate of error of the technique or theory when applied;*
> ***(4)*** *the existence and maintenance of standards and controls;*
> ***(5)*** *whether the technique or theory has been generally accepted in the scientific community.*

Similar guidance is provided in CPD 19A (2015) that for expert evidence to be admitted the court is expected to consider:

> *(c)…whether the opinion takes proper account of matters, such as the degree of precision or margin of uncertainty, affecting the accuracy or reliability of those results;*
> *(d)…the extent to which any material upon which the expert's opinion is based has been reviewed by others with relevant expertise (for instance, in peer-reviewed publications);*
> *(h)…whether the expert's methods followed established practice in the field and, if they did not, whether the reason for the divergence has been properly explained.*

Based on the Daubert ruling and CPD 19A, validation is the process of testing how good or bad the expert and method achieve what they are claimed to achieve. Depending on the specific analytic methods used in FVC (Section 1.1), types of validation could be different. If AuPA and AcPA are employed, as individual expert plays the key role in AuPA and AcPA, the validation process in this context would centre around testing expert's competency (Kirchhübel & Brown, 2021). Meanwhile, validation could also concern the issues of repeatability and reproducibility, i.e., whether same evaluation result can be achieved if the speech samples are analysed by the same expert again (repeatability) and whether same evaluation result can be achieved if the speech samples are analysed again by different experts (reproducibility)? The expert-centred validation would be very labour intensive; meanwhile, there are practical issues, e.g., do we have the data that is relevant to the case for validation and whether the data can be transferred between experts?

On the other hand, there are established procedures to empirically validate methods using a data-driven LR-based framework. In doing so, normally three independent datasets (i.e., training, test and reference) and two stages are involved (see 2.2.1). Note that four datasets might be required in automatic systems using a scoring method, i.e., suspected speaker control database and suspected speaker reference database are used to generate same-speaker (SS) similarity scores and questioned speaker data and reference data are used to generate difference-speaker (DS) similarity scores (Alexander & Drygajlo, 2004). From a statistical point of view, any observed data is a sample that comes from a population where its distribution is unknow. Because the datasets used for validation are sampled from a much larger population where the true distribution is unknown, this validation method is subject to sampling variability, i.e., will the system have the same performance if different sets of samples are used for validation? This is the question that forms the basis of the current thesis.

1.4 Established procedures for validation

Section 1.2 briefly outlined the conclusion frameworks used in FVC chronologically, showing the gradual shifting in employing the LR framework in FVC. The increasing acceptance of using the LR framework in FVC is in line with other areas of forensic evidence evaluation (e.g., fingerprint, bite marks, tire marks, handwriting) in what is described as part of the so called *paradigm shift* (Saks & Koehler, 2005). DNA typing employs population data to empirically evaluate "the matches between suspects and crime scene DNA evidence in terms of the probability of random matches across different reference populations" (Saks & Koehler, 2005, p. 893). In the FVC context, that is to demonstrate how good or bad the system is at separating SS and DS pairs in relation to the relevant population. Over the years, there are well-established procedures for method and system validation using numerical LRs, and they have been employed in numerous studies (see Section 2.2.2). In such work analysts validate their systems empirically via a two-stage process using data where the ground truth is known, to present the results of validation tests to the end user. In stage one (the *feature-to-score* stage), (typically acoustic) data extracted from pairs of SS and DS recordings taken from the test and training datasets are compared to produce training and test scores which indicate the similarity between the SS and DS samples and assessing typicality with respect to a reference data. In stage two (the *score-to-LR* stage or calibration), the training scores are used to generate calibration coefficients which are then applied to the test scores to convert them to interpretable LRs. System validity and reliability metrics are generated from the calibrated LRs for the test set.

There are broadly two different contexts for validation, testing under specific case conditions (case-specific) and generic testing. Under both contexts, the training, test and reference speakers need to be sampled from the relevant population, i.e., "…a particular class of persons on the basis of age, sex, occupation…" (Coleman & Walls, 1974, p.276), based on assumptions about the speaker in the offender sample(s). Case-specific testing aims to test system validity and reliability by taking case specific conditions into consideration, e.g., speaking style, recording condition (see Enzinger & Morrison, 2017 for an example based on simulated conditions). Those tests should be conducted on a case-by-case basis as no two cases are identical and the system evaluation obtained from testing under conditions from one case might give little insight to another case. However, in practical terms there is a problem with this approach: as the identity of the offender is unknown under real case conditions, so too is the relevant population drawn based on the assumption of the offender's origin (Hughes & Foulkes, 2015a). Thus, the forensic speech scientist must make pragmatic decisions about how to delimit that population, based on observations that can be made about the speech sample. Usually this amounts only to a broad statement about the sex of the speaker and their language or major regional accent (e.g., Australian English; Hughes, 2014; Rose, 2004) or other demographic factors. Given this paradox, generic testing is often conducted to investigate overall performance when assumptions do and do not match the ground truth. For example, using a nominal suspect and offender with an Australian English accent but reference speakers with a different accent enables forensic speech scientists to investigate the effect of accent mismatch in the reference speakers on overall performance. Due to the fact that we will never have data which are exactly the same as the case conditions, ultimately it is not a dichotomy between case-specific and generic testing, but a continuum of how closely the data we use for testing match the case data and whether the difference matters.

It is worth noting that the fully data-driven LR approach is employed in the current thesis to demonstrate the effect of sampling variability only, it does not mean that this approach is the gold standard in FVC and employed by the majorities in the real world. It was also indicated in Gold and French (2011) and Morrison (2016) that over half of the practitioners carry out speech analysis using the combination of auditory and acoustic analysis. The major limitations for practitioners to conduct fully data-driven LR approach are probably due to data limit and the selection of relevant population.

1.5 Research aims and questions

The ENFSI (2015) guidelines state that the LR framework "measures the strength of support the findings provide to discriminate between propositions of interest" (2.4, p. 6) which is scientifically accepted, and "this framework for evaluative reporting applies to all forensic science disciplines." (2.4, p. 6). It has also been shown that one of the major differences among the three surveys (Gold & French 2011, 2019; Morrison et al., 2016) is the increasing awareness and acceptance of Bayesian reasoning and likelihood ratio (LR) framework among forensic speech scientists. The LR framework allows forensic speech scientists to focus on the evaluation of evidence given a proposition; meanwhile, experts are less likely to fall into the prosecutor's fallacy or attorney's fallacy (see Sections 2.1 and 2.2) which could lead to misinterpretation of evidence and miscarriage of justice.

As mentioned in Sections 1.3 and 1.4, there are now established procedures for method validation using numerical LRs, i.e., using data where the ground truth is known to test how good or bad the system performs the task that it is designed to do. Numerous studies have investigated the factors that might affect the results of LR-based FVC (e.g., sample size, accent mismatch, statistical models) (see Chapter 2 for detailed discussion) focusing on the evaluation of *overall* performance. However, a further question is how representative those samples are (i.e., training, test and reference data) in relation to the larger relevant population that is assumed based on the offender sample? Given that training, test and reference speakers are sampled from a relevant population with matching demographic factors to the offender, the real size of the relevant population, although it depends on the specific defence proposition, is likely much larger than the number of speakers sampled for validation. Then, one crucial question for forensic speech scientists is that whether overall performance is consistent if we use different samples of speakers from a single relevant population (e.g., speakers with similar social class, regional accent, education etc.) and replicate experiments multiple times. This thesis aims to explore the effect of sampling variability on overall performance and individual speakers' behaviour at *feature-to-score* and *score-to-LR* levels. First, to investigate the effect of sampling variability on overall performance and individual speakers' behaviour in relation to configurations of training, test and reference speakers as well as choice of linguistic-phonetic features. Second, to investigate the effect of sampling variability on overall performance in relation to score skewness, calibration methods and sample size. The specific research questions (RQ) are as follows:

1. **What is the effect of sampling variability on overall performance?**

    a. To what extent does overall performance (validity and reliability) vary if different configurations of training, test and reference speakers (from the same relevant population) are used?
    b. Is the variability in the overall performance primarily caused by different configurations of training speakers, test speakers or reference speakers?

2. **What is the effect of sampling variability on individual speakers' behaviour?**

    a. Do certain combinations of linguistic-phonetic features outperform the others and are they less susceptible to sampling variability?
    b. How is individual speakers' LR output affected when different configurations of training and reference speakers are used?
    c. How is individual speaker' LR output affected when different linguistic-phonetic features are used?

3. **What is the effect of sampling variability on overall performance in relation to score skewness?**

    a. To what extent is overall performance affected by skewed scores?
    b. Are certain calibration methods more susceptible to score skewness than others?
    c. Would overall performance be improved with larger sample sizes when scores are skewed?

1.6 Thesis outline

Chapter 2 introduces the research context starting with a brief introduction to Bayes' Theorem and the LR framework. Established procedures for method and system validation using numerical LRs are explained. Following that, previous FVC studies using numerical LRs are discussed from the perspectives of system validity and reliability testing. Lastly, the rationale of the current study is explained.

Chapter 3 gives the general methodology employed in the current thesis, i.e., corpora and linguistic-phonetic features used, methods for feature segmentation, data extraction and parameterisation, and statistical methods used for LR computation and system evaluation.

Chapters 4, 5 and 6 are experimental chapters investigating the effect of sampling variability on LR-based FVC systems from different perspectives. Chapter 4 explores the effect of sampling variability on *overall* system validity and reliability focusing on the use of different configurations of training, test and reference speakers. RQs 1a. and 1b. are addressed in

Chapter 4. Chapter 5 looks at the effect of sampling variability with more focus on individual speakers and the use of different combinations linguistic-phonetic features. The possible range of individual speakers' LR output under a real case scenario is also demonstrated and discussed, and RQs 2a., 2b. and 2c. are addressed in Chapter 5. Chapter 6 uses simulated scores to investigate the effect of sampling variability on the score level aiming to address RQs 3a., 3b. and 3c. Scores were simulated based on empirical data from Chapter 4. Score distribution skewness, sample sizes and calibration methods were varied, and the overall performance was tested under those different conditions. In each experimental chapter, the chapter-specific methodologies and results are explained and discussed.

Finally, Chapter 7 discusses the effect of sampling variability at *feature-to-score* and *score-to-LR* stages in relation to the findings of Chapters 4, 5, and 6. The implications of this thesis for casework and future research are also discussed in the conclusion (Chapter 7).

Chapter 2 Research Context

Chapter 2 first gives a brief introduction to Bayes' Theorem and likelihood ratio as well as the logic behind them in evidence evaluation (Chapter 2.1). Following that, the development and application of the likelihood ratio framework in FVC are discussed in relation to system validity testing (Chapter 2.2.1.1) and system reliability testing (Chapter 2.2.1.2) respectively. Further, factors that affect overall performance are reviewed in relation to previous studies (Chapter 2.2.2). The overall rationale of the current thesis is given at last (Chapter 2.2.3).

2.1 Bayes' Theorem and the likelihood ratio (LR) framework

Bayes' Theorem explains how to update one's belief in a hypothesis when a new piece of evidence is incorporated can be expressed using a probability form in terms of hypothesis and evidence below,

$$p(H|E) = p(H) \times p(E|H)$$

Equation (2.1)

(Robertson et al., 2016)

where $p(H)$ is commonly known as the prior odds, i.e., our initial knowledge about the hypothesis (Robertson et al., 2016, p.15), before any evidence is taken into consideration. For example, if five suspects were arrested at a crime scene, and only one of them is the offender, then the prior probability of each suspect being the offender is 20% (the prior odds are 1: 4) before any evidence is considered. The prior probability needs to be multiplied by $p(E|H)$, i.e., the likelihood ratio (LR), to give the posterior probability. The LR estimates the strength of evidence under the two competing propositions of the prosecution and defence (Robertson et al., 2016), i.e., what is the probability of observing the evidence had it come from the same person or different people? The LR equation is given in Equation 2.2:

$$LR = \frac{p(E|H_p,I)}{p(E|H_{d,}I)}$$

Equation (2.2)

where $p(E|H_p)$ indicates the probability of observing the difference between the suspect and offender speech samples given the prosecution proposition, i.e., the speech sample comes from the suspect; $p(E|H_d)$ represents the probability of observing the difference between the suspect and offender speech samples given the defence proposition, i.e., the speech sample does not come from the suspect but someone else from the relevant population; I stands for background information about the case. Essentially, the numerator of the LR is an estimation of the similarity between the suspect and offender speech samples, while the denominator is an estimation of their typicality compared to the relevant population, i.e., the probability of observing the offender sample had it not come from the suspect but some other randomly selected member of the relevant population. The $p(H|E)$, the posterior odds, indicates the probability of the hypothesis given the evidence, i.e., the final estimate of the chance that a certain incidence happened. The posterior odds are governed and can be modified by both the prior odds (the "prior" assessment of the probability of certain incidence) and the LR (the statistical analysis of the strength of evidence given hypothesis) (Finkelstein & Fairley, 1970).

The Bayesian approach started to gain favour in courtroom evidence evaluation after the US case of *People v. Collins* [1968], where an elderly woman was robbed and assaulted by two people in Los Angeles. The victim mentioned that she saw a young woman with dark blond hair and a ponytail running away after robbing her. Another witness mentioned that the young woman with dark blond hair and a ponytail ran into a yellow car driven by a black man with a beard and moustache. Later, a couple with similar appearance characteristics were arrested by the police, and an instructor of mathematics was called to establish the probability of the arrested defendants being the offenders who robbed the elderly woman based on the similar appearance characteristics.

The witness was instructed by the mathematics instructor to give assumptions of individual probabilities of the relevant appearance characteristics (i.e., yellow car, male with moustache, female with ponytail, female with dark blond hair, black male with beard and interracial couple in car). The specific probabilities of each individual appearance characteristics are:

Individual appearance characteristics	**Probability**
Yellow car	1/10
Male with moustache	1/4
Female with ponytail	1/10
Female with blond hair	1/3
Black man with beard	1/10
Interracial couple in car	1/1000

Table 2.1 Assumed probability of each individual appearance characteristics (Table from Finkelstein & Fairley, 1970, p. 491).

Applying the "product rule", i.e., the probability of the relevant appearance characteristics occurring jointly equals the product of the individual probabilities of each of the relevant appearance characteristics (Finkelstein & Fairley, 1970, p. 491), the mathematics instructor then concluded that a couple selected at random having the same incriminating characteristics would be one out of twelve million. The court rejected the instructor's testimony based on several grounds. The two most important ones are as follows. First, the individual probabilities of each appearance characteristics (Table 2.1) assumed by the witness were not supported by any evidence. Second, there was no evidence to demonstrate the independence of the assumed individual probabilities of the appearance characteristics (Table 2.1). However, even if the individual probabilities were supported by evidence and they could be shown to be independent, the logic in calculating the probability in the *People v. Collins* [1968] case was flawed, also known as prosecutor's fallacy (Thompson & Schumann, 1987), i.e., using the probability of proposition for the probability of evidence. In *People v. Collins* [1968], the mathematics instructor inappropriately reported the probability that was conditional on the evidence (appearance characteristics) rather than on the propositions (the interracial couple arrested is guilty). A more justifiable probability estimate could be established if the Bayesian approach were adapted, i.e., the experts should focus on the evaluation of evidence given proposition, rather than the other way round.

2.2 The LR approach in FVC

It has been shown in Section 2.1 that Bayes' theorem and LR framework started to gain favour in forensic evidence evaluation after the case of *People v. Collins* [1968]. Researchers from other areas of forensic evidence evaluation later on started to adapt to the LR framework (see examples of fibre and DNA in Aitken & Taroni, 2004), also known as the *paradigm shift* (Chapter 1.4) (Saks & Koehler, 2005). However, it was not introduced and discussed explicitly in FVC until the late 1990s by Champod & Meuwly (1998, 2000), where Bayes' theorem was explained and the evaluation of speech evidence using the LR framework was proposed, as well as the discussion of the prosecutor's fallacy (i.e. using posterior probability as the LR) and the attorney's fallacy (i.e., using a larger number of possible suspects with no relation to the reality of the case) (Thompson & Schumann, 1987). Champod & Meuwly argued that forensic scientists should be "concerned solely with the LR" (2000, p. 201) not the posterior odds. This is because the posterior odds is the product of the prior odds and the LR, and the calculation of the posterior odds needs case-related background information (e.g., potential population of the suspects) where forensic experts normally do not have access to (Champod & Meuwly, 2000, p. 200). Meanwhile, Broeders (1999) also stated that "the question whether the suspect was the perpetrator is outside the province of the expert" (p.239), and reporting the probability of a proposition given the evidence is logically incorrect (p. 228). Further, he specified that forensic speech scientists should lay stress on two questions:

1. How likely is the questioned speech sample to sound the way it does under the hypothesis that it was produced by the suspect?

2. How likely is the questioned sample to sound like this on the hypothesis that it was produced by somebody other than the suspect? (Broeders, 1999, p. 238).

Ultimately, Broeders (1999) and Champod & Meuwly (1998, 2000) have all explained the logic behind the conceptual LR framework and suggested that it is essential for forensic speech scientists to shift from evaluating the probability of proposition given evidence to evaluating the probability of evidence given proposition. However, they did not demonstrate the application of the quantitative LR-based approach using linguistic-phonetic features in FVC explicitly. Kinoshita (2001) was one of the first to investigate the feasibility of employing quantitative LR-based approach in FVC, where she tested the speaker-discriminatory power of linguistic-phonetic features. Specifically, midpoint F3 values of /o m s/ from the Japanese telephone opening phrase *moshimoshi* were used, as well as midpoint F2 values of /i/ and F2 and F3 of /e/ from other target words. 10 male Japanese speakers were used for the testing. Aitken's formula (1995) was used for cross-validated comparisons to calculate the LRs, and discrimination was then carried out based on the posterior odds. The overall classification rates achieved were 90% for same-speaker (SS) comparisons and 97% for different-speaker (DS) comparisons.

Thereafter, numerous tests have been carried out using different linguistic-phonetic features. The majority of such studies focus on speaker discrimination performance. For example, Rose et al. (2004) explored the speaker-discriminatory power of three Japanese variables, /ŋ/, /ɕ/ and /ɔː/ using the midpoint data of F1 to F5, and Zhang et al. (2008) looked into two Chinese vowels (i.e. /i/ and /y/) using mean formant frequencies of F1, F2 and F3. Further, Morrison (2009) explored the speaker-discriminatory power using dynamic vowel formant data. Meanwhile, some other studies examined the speaker-discriminatory power using suprasegmental features, e.g., long-term F0 distribution (Kinoshita et al., 2009), lexical tones (Rose & Wang, 2016), speech tempo (Lennon et al., 2019) and voice quality (Enzinger et al., 2012; Hughes et al., 2019). Apart from testing different linguistic-phonetic features, many other studies have investigated the effect of non-linguistic factors on LR-based FVC systems, e.g., sample size (Hughes, 2017; Ishihara & Kinoshita, 2008), statistical models (Kinoshita & Wagner, 2014; Morrison, 2011a), calibration methods (Morrison & Poh, 2018), sampling variability (Ali et al., 2015), channel mismatch (Hughes, Harrison, et al., 2019), reference population mismatch (Watt et al., 2020). Ultimately, previous studies used speech data where the ground truth is known to investigate two major questions, i.e., whether the system does what it is designed to do (validity) and whether the system would yield the same result if the analysis were repeated (reliability). Section 2.2.1 lays out details of previous studies where system validity and reliability were explored using linguistic-phonetic features, and following Section 2.2.2 discusses factors that affect the overall performances of LR-based FVC. Lastly, Section 2.2.3 gives the rationale of current study.

2.2.1 System testing

In order to generate a LR (be that numerical or verbal (see Section 3.4.3); or indeed any form of conclusion in a FVC case), the expert employs a *system*. This is defined broadly as the particular courses of action that are used to compare the suspect and offender samples (Morrison, 2013), e.g., the data used to represent the relevant population, the linguistic-phonetic variables chosen for analysis, the methods of analysing those variables, the statistical models used for score generation, and calibration methods used for *score-to-LR* conversion. For an end user (e.g., jury and/or the court) to be able to interpret the conclusion provided by the expert appropriately, it is essential to understand the validity and reliability of the system used to generate that conclusion. Validity measures how well the system performs the task that it is designed to do, while reliability answers the question of whether the system would yield the same result if the analysis were repeated (Hughes and Kinoshita, under revision).

Figure 2.1 shows a simplified demonstration of the difference between validity and reliability. Systems with both good validity and reliability would have narrow clusters of arrows and gather around the centre of the target (bottom left), while systems with bad validity and reliability would have wide arrows off the target (top right).

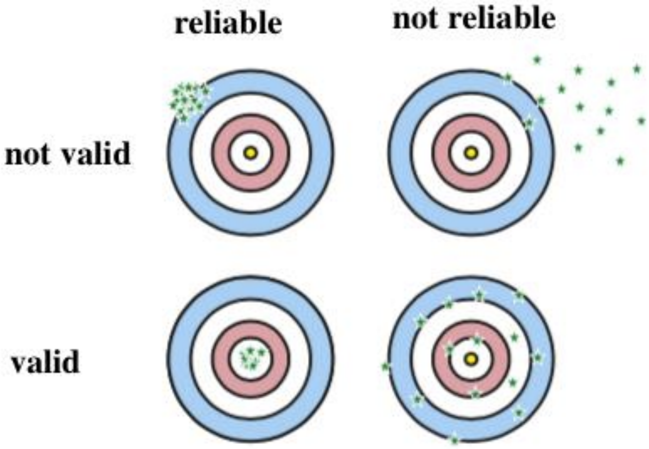


Figure 2.1 Conceptual demonstration of validity and reliability (Figure modified from Morrison, 2016, Figure 1, p.372).

Numerous studies have attempted to test system validity and reliability in line with the procedures for data-driven LR analysis outlined in Morrison et al., (2021). Sections 2.2.1.1 and 2.2.1.2 discuss previous studies into system validity and reliability testing.

2.2.1.1 System Validity

During early research in LR-based FVC studies, a major focus was placed on system validity, i.e., testing how well the system performs the task that it is designed to do using certain linguistic-phonetic variables and statistical models. Very often, overall performance is evaluated using Log LR cost function (C_{llr}; Brümmer & du Preez, 2006) and/or equal error rate (EER). C_{llr} evaluates overall performance based on the magnitude of evidence, i.e., what is the strength of the evidence given proposition, while EER makes categorial decisions where false hit equal to right miss. For both evaluation metrics, the lower the values the better the overall performance (see Section 3.4 for detailed discussion).

Rose et al. (2004) explored the speaker-discriminatory power of three Japanese variables, namely syllable-coda nasal /ŋ/, voiceless alveopalatal fricative /ɕ/ and long back mid-rounded vowel /ɔ:/. Non-contemporaneous telephone recordings from 60 Japanese speakers were used. The mid-point data of F1 to F5 and 12th order LPC cepstral coefficients were extracted for each token and the extracted data was modelled using multivariate kernel density (MVKD; Aitken & Lucy, 2004) for LR computation. They showed that the system yielded EERs of ca. 7.5% and ca. 13.5% using cepstral coefficients and mid-point formant data respectively. The results indicated that using cepstral coefficients provide stronger speaker-discriminatory power than using mid-point format data. However, this might be due to the fact that cepstral coefficients have more dimensions than mid-point formant data, which could possibly lead to a better speaker-discriminatory performance. Moreover, it was not clear how Rose et al. (2004) allocated the 60 speakers (e.g., 20 speakers used for each of the training, test and reference sets) or if the system was calibrated using a different set of data. Similarly, Zhang et al. (2008) explored the speaker-discriminatory power using the mean formant frequencies of F1, F2 and F3 from two Chinese vowels (i.e. close front unrounded vowel /i/ and close front rounded vowel /y/). Non-contemporaneous telephone recordings from 64 male speakers were used. Raw formant data was entered into the MVKD (Aitken & Lucy, 2004) to produce LRs. The best system validity was achieved using F1, F2 and F3 of /y/ (EER = 25.1%) and F2 and F3 of /i/ (EER = 26.1%). Later, Morrison (2009) explored the speaker-discriminatory power using dynamic vowel formant data (contrary to static formant data). Five diphthongs (i.e., /aɪ/, /eɪ/, /oʊ/, /aʊ/, /ɔɪ/) from Australian English were tested and non-contemporaneous speech samples from 27 male speakers were involved. The first to third formants of each diphthong were parametrised using polynomial curves and discrete cosine transform (DCT). Different degrees of polynomial curves and DCT were attempted, and each diphthong was tested under two conditions, i.e., use all three formants and only use F2 and F3. The coefficients were used as the input for MVKD (Aitken and Lucy, 2004) for LR computation and the system was calibrated using cross-validation. In terms of individual diphthongs, the lowest C_{llr} (0.095) was achieved using the 0th to 2nd DCT coefficients of F2 and F3 formant trajectories of /eɪ/.

However, the C_{llr} further lowered to 0.056 using all three formants fusing all five diphthongs. Morrison (2009b) showed that fusing multiple linguistic variables improves the system validity, i.e., lower C_{llr}; however, given the number of speakers (27 male speakers) used in his study, it is likely that the overall performance is overestimated. Similarly, Zhang et al. (2011) looked into the speaker-discriminatory power of formant trajectories using a Chinese triphthong /iau/. Several other studies have been carried out investigating the speaker-discriminatory power of vowel trajectories using Cantonese (Chen & Rose, 2012; Li & Rose, 2012; Pang & Rose, 2012; Rose & Wang, 2016).

Apart from segmental variables, previous studies have also investigated system validity using suprasegmental variables (Kinoshita, 2005; Kinoshita & Ishihara, 2010). For example, Kinoshita et al. (2009) explored the potential discriminatory power using long-term F0 (LTF0) distribution parameters. Non-contemporaneous speech samples from 201 male Japanese speakers were used and the average duration of the speech samples varied between 10 and 25 minutes. The raw F0 data was extracted at every 5 milliseconds from the entire duration of each speech sample. Six LTF0 distribution parameters (i.e. mean, standard deviation, kurtosis, skewness, modal F0, modal density) were calculated and used as the input of MVKD (Aitken & Lucy, 2004) for LR computation. An EER of 10.7% was obtained using all six LTF0 distribution parameters. Kinoshita et al. (2009) further explored whether the system validity is affected by the amount of speech used. They varied the duration of input speech from 5 to 180 seconds, showing that the EER lowers from 23% to 16% when the speech duration increased from 5 to 30 seconds. There were some improvements in EER using longer speech input; however, the rate was slower after the first 30 seconds. They emphasized that the focus of their study was to explore the system that produces the strongest LR, whether supporting prosecution or defence propositions; however, the use of EER as a metric for system evaluation in their study seems counter-intuitive. This is because EER only treats LRs categorially and it does not take the magnitude of evidence (i.e., how strong or weak the evidence is) into consideration.

Further, the speaker-discriminatory power of other suprasegmental variables have also been studied, e.g., Lennon et al. (2019) on speech tempo, Rose & Wang (2016) on tonal F0, Enzinger et al. (2012) and Hughes et al. (2019) on voice quality.

Beyond testing system validity using different segmental and suprasegmental variables, previous studies have also attempted to assess system validity as a function of using different statistical models. For example, Morrison (2011) tested the MVKD (Aitken & Lucy, 2004) and Gaussian mixture model-universal background model (GMM-UBM; Reynolds et al., 2000) using the same set of acoustic-phonetic data from Morrison (2009), i.e. the formant trajectories of five Australian diphthongs /aɪ/, /eɪ/, /oʊ/, /aʊ/ and /ɔɪ/. However, only F2 was used in his study. This is because recording samples in real cases are often recorded or transmitted via telephone or mobiles, and F2 is less susceptible to the telephone or mobile effects. Specifically, F1 is likely to be affected by landline telephone bandpass, i.e. the estimates of F1 are artificially raised by telephone transmission, an average of 14% for landlines (Künzel, 2001) and 29% for mobiles (Byrne & Foulkes, 2004). Similarly, F3 is likely to be affected by the codecs of GSM mobile network (Guillemin & Watson, 2009). In Morrison (2011), non-contemporaneous speech samples from 27 Australian male speakers were used and the F2 trajectories were fitted using DCT and polynomial curves (see Morrison, 2009, 2011 for detailed parameterisation procedure). The DCT and polynomial coefficients were used as the input for LR computation. The results show that the systems yield similar validity uing MVKD and GMM-UBM models when only individual diphthongs are involved. However, the GMM-UBM outperforms MVKD fusing all five diphthongs, with the C_{llr} values equal to 0.218 and 0.035 for MVKD and GMM-UBM respectively. Later, Kinoshita & Wagner (2014) showed a different pattern when testing system validity using MVKD and GMM-UBM but from a more case-realistic perspective. They used data from 27 male Australian English speakers, with a focus on the long monophthing /i:/. The midpoint measurements of the first three formants were used as the input data for LR computation. Due to the fact that speech samples available in real cases are often limited, they tested robustness of MVKD and GMM-UBM models using very small numbers of tokens (i.e. 2, 4 and 6 tokens per speaker). Meanwhile, different numbers of Gaussians used for GMM-UMB were also tested. They showed that, MVKD consistently outperformed GMM-UBM with different numbers of tokens. For MVKD, the lowest C_{llr} (0.451) was obtained using 6 tokens per speaker, while for GMM-UBM, the lowest C_{llr} (0.5) was achieved uisng 6 tokens per speaker with 4 Gaussians.

Apart from Morrison (2011a), the previous studies mentioned above mainly focused on system validity, i.e. aiming for lower C_{llr} or/and EER values. However, given what really matters in a real case is the voice of the speaker under analysis, the question is how representative or reliable is the specific LR in a given FVC case or would we get the same LR if the experiment were repeated? The next section (Section 2.2.1.2) discusses previous studies where system reliability is involved.

2.2.1.2 System reliability

System validity testing accounts for only part of the evidence evaluation process. System reliability is equally or even more important because it answers the question of whether one will get the same result if the evaluation is repeated. Taking a real case scenario for example, if two speech samples (one suspect and one offender) need to be compared, and we have collected the test, training and reference data from the relevant population based on our assumptions of the offender sample. The system is well trained and tested and is ready for the comparison of the suspect and offender samples. However, the size of the training, test and reference data is likely to be smaller than the actual size of the relevant population, the question would be how representative is the sampled data to the relevant population? Would the system give consistent results if the comparison is repeated with different samples (from the same relevant population) or different numbers of speakers in each on the training, test and reference data? Ishihara & Kinoshita (2008) seems to be one of the first to measure system reliability in relation to sample size using empirical speech data. In particular, they investigated the sample size of reference speakers as a function of LR reliability, i.e., they addressed the question of how variable the LRs are if different numbers of reference speakers are used. Non-contemporaneous speech samples (two sessions) from 241 male Japanese speakers were involved and the LTF0 data was used as the input for MVKD (Aitken & Lucy, 2004) and LR computation. They varied the sizes of the reference data from 10 to 120 with 10-speaker increments. Pairs of the 241 speakers were compared against each of the speakers in the reference data, resulting in 241 SS comparisons and 115,680 DS comparisons. The experiments were replicated three times, and 6 Log$_{10}$ LRs (LLR) were produced for each comparison session (i.e. one LLR per comparison session per reference data * 2 reference data per comparison * 3 replications). The difference between LLRs was used as a measure for system reliability. Ideally, more reliable systems would yield smaller LLR ranges. The results from Ishihara and Kinoshita's study suggested that, for SS comparisons, the number of reference speakers does have a substantial effect on the reliability of LR output (the difference between LLR varies from 0 to over 8) when the sample size is small (i.e., 10 and 20 speakers), and the LLR starts to stabilise when the sample size reaches 30 and onwards (LLR varies between ca. 2.5 and ca. 1). For DS comparisons, the LR output is much more variable, and the difference varies between ca. LLR 0 and ca. LLR 8 even when the sample size reaches 120. Ishihara & Kinoshita (2008) thus show that the size of the reference speaker sample does indeed have an effect on the reliability of LR output.

Kinoshita & Ishihara (2014) later replicated Ishihara & Kinoshita (2008) using the same dataset and experimental procedures but with logistic regression calibration (Brümmer et al., 2007) applied. Instead of using the subtraction of two LLRs, variability scores were used for reliability testing. The variability score is calculated using the subtraction of two converted LRs using the function below,

$$Converted\ LR = f(x) = \begin{Bmatrix} x, x \geq 1 \\ 0 - \frac{1}{x}, 0 < x < 1 \end{Bmatrix} \qquad \text{(Equation 2.3)}$$

Kinoshita & Ishihara (2014, p.208)

According to Kinoshita & Ishihara (2014), the variability scores in systems with equal numbers of reference speakers captures the fluctuation in LR outputs between different comparison sessions. Across different numbers of reference speakers (i.e., 10 and 120), the variability scores varied between ca. -0.4 and ca. 1 (SS comparisons) and ca. -2.5 and 6 (DS comparisons) using 10 reference speakers, while they varied between ca. -0.8 and 1.5 (SS comparisons) and ca. -2.8 and 8 (DS comparisons) using 120 reference speakers. Kinoshita & Ishihara (2014) claimed that, with calibration applied, varying the size of reference speakers does not have much effect on system reliability. However, Kinoshita & Ishihara (2014) were flawed as they did not give explanation for the interpretation of variability scores. Although they stated that lower variability score indicates more stable overall performance, variability scores were calculated from converted LRs, which made the variability score not as interpretable as LLRs. For example, the absolute differences of DS variability scores equal to 8.5 (|-2.5 - 6|) and 10.8 (|-2.8 - 8|) when 10 and 120 reference speakers are used respectively. The difference between the absolute differences is then 5.6 (|10.8-5.2|), which is not interpretable and cannot directly be compared to LLRs.

Further, Hughes (2017) used simulated data to explore the reliability of LR output to variation in the size of not only just reference data, but also training and test data. The mid-point values of the first three formants of the filled pause (FP) *um* were extracted from a sociolinguistically homogeneous set of 100 male speakers of Standard Southern British English (SSBE) from the DyViS corpus (Nolan et al., 2009). Then, the formant values were simulated for training, test and reference data respectively, and the sample size was varied from 10 to 100 speakers in each of the three datasets. The formant data was modelled using MVKD (Aitken & Lucy, 2004), and logistic regression (Brümmer et al., 2007) was used to calibrate scores to LRs. The result showed that "relatively precise LR output" (Hughes, 2017, p.28) can be achieved when training and test data reach 30 to 40 speakers, and the reference data reaches 15 speakers.

Apart from using the subtraction of LLRs and variability scores, Morrison (2011) and Morrison et al. (2010) adapted credible intervals (CI) from DNA evidence evaluation (Curran, 2005; Curran et al., 2002) to test the reliability of LR outputs in FVC. The CI is the Bayesian version of the confidence interval; however, unlike the confidence interval, the CI treats the boundaries (two intervals) as fixed variables while the estimated LR is treated as a random variable (note that under LR framework, any LR obtained is an estimate of the true unknown LR). A 95% CI is often used in FVC to measure the reliability of LR output (Morrison, 2011b; Morrison et al., 2011), and the wider the CI the less reliable the LR estimate. The CI is normally averaged across speakers, and the measurement is an indication of overall system reliability (i.e., does not indicate how each individual speaker behave). In Morrison et al. (2011), 95% CI was calculated using empirical data to estimate the reliability of the LR output in a FVC system (see Morrison et al. 2011 for discussion of validity as it is not the focus of this section). Non-contemporaneous recordings (two sessions) from 61 male Northern Mandarin speakers were involved. Three Mandarin monophthongs (/i/, /e/ and /a/), and the mean frequency of the first three formants of /e/ and /a/ and F2 and F3 of /i/ over the vocalic portion were used as the input for LR computation in MVKD (Aitken & Lucy, 2004). In LR computation, a cross-validation procedure was applied, i.e. all speakers were being used as the background data except for the speakers that were being compared. The calibration coefficients were then calculated separately for each of the monophthongs using logistic regression (Brümmer et al., 2007; Morrison, 2013) and cross-validated applied, i.e. all LRs obtained from MVKD were used to calculate the calibration coefficients except for the ones being calibrated. After that, logistic regression (Brümmer et al., 2007) was applied to fuse the three sets of calibrated LLRs that were obtained from three monophthongs into one single set. Due to the practical limitation on real-world data collection, the 95% CI was only calculated for each of the DS LLRs. To the best of the author's knowledge, the only available statistics for real-world data collection is found in Rose (2013b) where he used 35 male speakers for a telephone fraud case in Australia; meanwhile in studies for research purposes, the number of speakers used varies from 60 to 90 (which seems to be the reasonable limits of sample sizes in the real-world) across different studies (e.g., Morrison et al., 2011; Rose & Cuiling, 2018). For CI calculation, the within comparison DS LLR mean was first calculated using Equation 2.4:

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \qquad \text{(Equation 2.4)}$$

Morrison et al (2010, p.65)

where:
n_i is the number of DS LLRs per pair
i is the specific DS pair
x_{ij} is the j^{th} DS LLR for the comparison pair i
The deviation-from-the-mean (y_{ij}) of each DS LLR was the calculated using:

$$y_{ij} = x_{ij} - \bar{x}_i \qquad \text{(Equation 2.5)}$$

The 95% CI was then calculated using local linear regression with the k nearest neighbours based on the mean and deviation-from-the-mean values of each pair of DS LLRs (See Morrison et al., 2011 for detailed procedures). The overall system reliability can then be evaluated using the mean of the 95% CIs, i.e., the average positive and negative difference between the upper and lower intervals of the CI and the mean value for a given comparison. For example, if a DS LLR for a given comparison is -0.3 and the 95% CI is ± 1, then the estimated 95% LLR CI would range between -1.3 and 0.7 (i.e., an LR of 19.95 ($10^{1.3}$) in favour of DS proposition and 5.01 ($10^{0.7}$) in favour of SS proposition). A generalised statement using 95% CI was given by Morrison et al. (2011, p.64),

> "…, I have calculated that one would be *X* times more likely to obtain the acoustic properties of the voice samples if the questioned-voice sample had been produced by someone other than the accused than if it had been produced by the accused. … I am 95% certain that the true values will be within the range of interval$_{\text{lower}}$ and interval$_{\text{higher}}$. "

In general, the lower the 95% CI, the better the system reliability; however, it was not stated explicitly in Morrison et al. (2011) how low is a reasonable 95% CI for FVC.

Meanwhile, Kinoshita & Ishihara (2014) further suggested some other obstacles for using 95% CI. First, given most triers of fact are not familiar with LRs, it would be extremely difficult to convey the complex concepts of 95% CI and leave triers of fact themselves to interpret the information appropriately in courts. Second, in order to estimate a valid 95% CI for SS comparisons, each speaker needs to be recorded on at least three separate occasions. Given the limitation on data collection in real-world FVC cases, it is often extremely difficult to have offender samples recorded on different occasions, especially when linguistic features are involved. Although one could use different portions of the offender sample to serve as samples recorded on different occasions (given the offender sample is long enough), the system reliability is likely to be overestimated given the effect of using contemporaneous data.

Estimating the reliability of LR has been debated over the years. Some hold the position that reporting uncertainty in the LR is the best practice (Curran, 2016; Morrison & Enzinger, 2016). This is because LR is “the function of parameters which have a distribution” (Curran, 2016, p.381); LR should also have a distribution because LR is dependent on those parameters. It is true that CI is often used to measure the uncertainty in LR. However, others (Ommen, Saunders and Neumann, 2016) argue that since LR is our primary concern, nuisance parameters (i.e., parameters which are not of our immediate interest) should be integrated out using fully Bayesian approach which would give us Bayes Factor (BF). Since BF incorporates the uncertainty associated with the nuisance parameters, it would be “redundant to include an interval estimate for Bayes Factor” (p.386). Moreover, Ommen et al. points out that if CI is used for reporting uncertainty, it would not be clear to the decision makers (e.g., court, jury) about how to apply CI in the decision-making process. For example, should the decision makers use the entire interval, the lower endpoint or the upper endpoint of the CI? Different choices would lead to incoherent decision resulting in bias in favour of the prosecution or the defence. From the author’s point of view, whether reporting uncertainty in the LR or incorporating uncertainty into the LR itself remains open for discussion. Different calibration methods were used this thesis as an approach to test their susceptibility against sampling variability and sample size.

2.2.2 Factors affecting system validity and reliability

There are many factors that might affect the LR output of a FVC system. These include linguistic factors (i.e., relating to the voice of the speaker(s) under analysis; e.g., regional accent, speech tempo, voice quality), non-linguistic factors (e.g., sample size, statistical model, channel mismatch). Moreover, researchers' degrees of freedom, i.e., any methodological decisions that the researcher makes may affect the final outcome (e.g., which phonetic features to be analysed and which statistical model and calibration method to be used). For linguistic factors, Watt et al. (2020) investigated the effect of using accent-mismatched reference data on the overall performance. 100 male SSBE speakers from the DyViS corpus (Nolan et al., 2009) and 60 male Northern east English speakers from the The Use and Utility of Localised Speech Forms in Determining Identity (TUULS; Watt et al., 2018) corpus (20 speakers from Middlesbrough, Newcastle, and Sunderland respectively) were involved. The MFCCs of 120-second speech material were extracted and divided into two halves to serve as the suspect and offender samples respectively. The Nuance Forensics v.11.1 (an iVector system) was used to generate comparison scores that are converted into LLRs using match and mismatch reference data using linear discriminant analysis. Not surprisingly, when using SS and DS comparison scores from the TUULs speakers, matched reference data yielded the best system validity (C_{llr} = 0.145; using TUULS as the reference data) and mismatched reference data yielded the worst (C_{llr} = 0.427, using DyViS as the reference data). Interestingly, using combined reference data (i.e. TUULS and DyViS) yielded different system validity depending on the number of DyViS speakers involved. A C_{llr} of 0.218 was obtained when using all TUULS (60) and DyViS (100) speakers as the reference data, while the C_{llr} reduced to 0.198 when the number of DyViS reduced to 60 (i.e., 60 TUULS and 60 DyViS speakers respectively).

For non-linguistic factors, Hughes et al. (2019) investigated the effect of channel mismatch on vowel-based FVC systems. Note that this reference (i.e., Hughes et al.) is being used because they tested linguistic features. The first three formants of 60 seconds of vowel-only material were extracted from 97 SSBE speakers across four channels, i.e., studio quality, landline telephone, GSM mobile with high bit-rate (12.2kb/s) and GSM mobile with low bit-rate (4.75kb/s). The LRs were computed using GMM-UBM (Reynolds et al., 2000) (8 Gaussians) with maximum a posteriori (MAP) adaptation and calibrated using logistic regression (Brümmer et al., 2007). Overall, systems with matched conditions (e.g., studio quality vs. studio quality; landline telephone vs. landline telephone) outperform those with mismatched conditions. However, the greatest variability (least reliable) was observed in the studio quality vs. studio quality system with EERs ranging from 7.9% to 32.8% and C_{llr}s ranging from 0.28 to 0.84. As discussed in Section 2.2.1.2, Ishihara & Kinoshita (2008) showed that SS and DS LLRs both varied to a substantial degree when small number of reference speakers (i.e., 10 to 20) were used, while SS and DS LLRs started to stabilise when more than 30 reference speakers were used. Hughes (2017) also showed that reliable LLRs can be achieved when 30 to 40 speakers were used for training and test data, 15 speakers for reference data. However, these studies did not build a formal relationship between sample size and sampling variability, e.g., how does the system perform when different sets of training, test and reference speakers are sampled? It should be noted that, from a subjective Bayesian point of view, any observed data at hand is a sample that comes from an unknown distribution, and thus the calculated LR is an estimate of the true unknown LR (Morrison & Poh, 2018, p.200). In principle, the larger the sample size, the more reliable the output. Therefore, when the sampled data size is large, the sample distribution is likely to be a reasonable approximation of the true population distribution resulting in more probable estimate of the true LR. On the other hand, when the sampled data size is small, sampling variability may cause the shape of the sample distribution shifts away from that of the true distribution and resulting in the calculated LR a poor estimate of the true LR (Morrison & Poh, 2018). Therefore, sampling variability is another major factor that affects system validity and reliability.

It can be seen from previous studies that sample size is a key component in using numerical LRs for validation. The LRs are generated using statistical or probabilistic models with different sample sizes, which lead to the fact that the value of the LR is a function of statistical or probabilistic models and sample sizes. While it is possible for researchers to use different models, the choice of sample size is likely to be outside researchers' control in real world scenarios (although the researchers can decide the number of speakers assigned to each dataset given sample size). Some argue that the amount of uncertainty in relation to sample size should be embedded in the calculation of an LR, i.e., having conservative LRs (e.g., close to 1) when the uncertainty is high (i.e., small sample size) (Brümmer, 2013). Since analysts often deal with small sample sizes in linguistic casework, sampling variability can be introduced at not only the *feature-to-score* stage, but also the *score-to-LR* stage (calibration), i.e., when the sample size is small, and the theoretical density estimation is not well-supported by the data leading to extrapolation at the tails of the score distributions (see Section 3.4.2 for detailed discussion).

Many calibration methods have been developed for system improvement and optimisation and the performance has been compared, e.g., linear discriminant analysis, Morrison (2013); logistic regression, Brümmer et al. (2007); pool adjacent violators, Zadrozny & Elkan (2002); and Bayesian models, Brümmer & Swart (2014), although not all of them aimed at dealing with small sample size issues. Ali et al. (2015) investigated the sampling variability on overall performance at the *score-to-LR* stage. Different sample sizes and calibration methods, as well as different score distributions, were tested. They simulated 5000 sets of training scores with three different sample sizes, i.e., 20, 200, 2000 for SS training scores and 1000, 10000, 100000 for DS training scores from different probability density functions (PDF), i.e., normal PDFs, reversed Weibull PDFs, Uniform PDFs and Beta PDFs. Meanwhile, three calibration methods, kernel density estimation (KDE) (Parzen, 1962), logistic regression (Brümmer et al., 2007), and pool adjacent violators (PAV) (Zadrozny & Elkan, 2002a) were tested. The results show that sampling variability has the least effect on system perform using logistic regression when the training sample size is large (2000 SS and 100000 DS scores). When the sample size is small (20 SS scores and 1000 DS scores), the sampling variability has the least effect on system reliability using KDE calibration. However, the size of the test data was not taken into consideration and only three sets of sample size of the training data were considered. Their study did not provide explicit knowledge about the relationship between sampling variability and sample size.

Similarly, Morrison & Poh (2018) used simulated scores to explore the effectiveness of different calibration methods in shrinking LR output and tested the generalizability using data from real cases. SS and DS scores with different sample sizes (e.g., 10, 100) were generated from Gaussian distributions and the score generation was repeated 1000 times. Four calibration methods were tested, i.e., linear discriminant analysis (Morrison, 2013), regularised logistic regression (Morrison & Poh, 2018), Bayesian model (Brümmer & Swart, 2014) and empirical lower and upper bound (ELUB; Vergeer et al., 2016). The results show that the ELUB and regularised logistic regression are less affected by sampling variability across all sample sizes. However, they only compared the effectiveness of different calibration methods on simulated scores that follow Gaussian distributions with equal variance, while the score distributions might not follow Gaussian distributions with equal variance in reality.

Researchers' degrees of freedom seem to be a less studied factor that could affect overall performance in FVC and in linguistics in general. Whether a qualitative or quantitative approach is employed, researchers have the degrees of freedom to conduct analysis and make decisions. For example, Roettger (2019) discussed possible stages where researchers' degrees of freedom could be exploited. He used simulated data to demonstrate how researchers' degrees of freedom affects experimental results in quantitative-based phonetic studies. Taking studies in word stress as an example, he pointed out that researchers have the degrees of freedom of choosing different phonetic features to be measured (e.g., measuring duration, intensity, or F0 as the phonetic correlates of word stress); operationalizing the chosen features (e.g., using the mean or the onset or the maximum of F0); and taking different domains (e.g., over the entire syllable, the rhyme, or the coda) (Roettger, 2019, p. 5 - 7). All the decisions made at each stage have implications on the analysis, and could potentially affect the final evaluation results.

Similarly, in employing numerical LR approach, as well as auditory and acoustic analysis, experts need to decide which acoustic features to extract, how to extract them, and how to analyse them. Figure 2.2 shows six steps where researchers' degrees of freedom can be exploited in data-driven LR-based FVC using linguistic-phonetic features. It is acknowledged that Figure 2.2 is a simplification of the real-world scenario where the exploitation of researchers' degrees of freedom could be much complex and diverse. In step 1, training, test and reference data are sampled from the relevant population. Given previous studies have shown that it requires no fewer than 30 to 40 training and test speakers and 15 reference speakers to generate reliable LRs, it is the researcher's decision on how to assign speakers into each of the training, test and reference data. In step 2, different linguistic-phonetic features can be discarded or included for analysis (e.g., F0 and harmonics-to-noise ratios, Hughes et al., 2019; parametric cepstral distance Kinoshita et al., 2018; LTF0, Rose & Zhang, 2018; formants, San Segundo & Yang, 2019). In step 3, the different features selected will result in different feature extraction methods, e.g., polynomial curves, MFCC and DCT. In steps 4 and 5, different statistical methods (e.g., MVKD, Aitken & Lucy, 2004; GMM-UBM; Reynolds et al., 2000) can be used to generate speaker models. In step 6, researchers can choose different calibration methods to compute LRs.

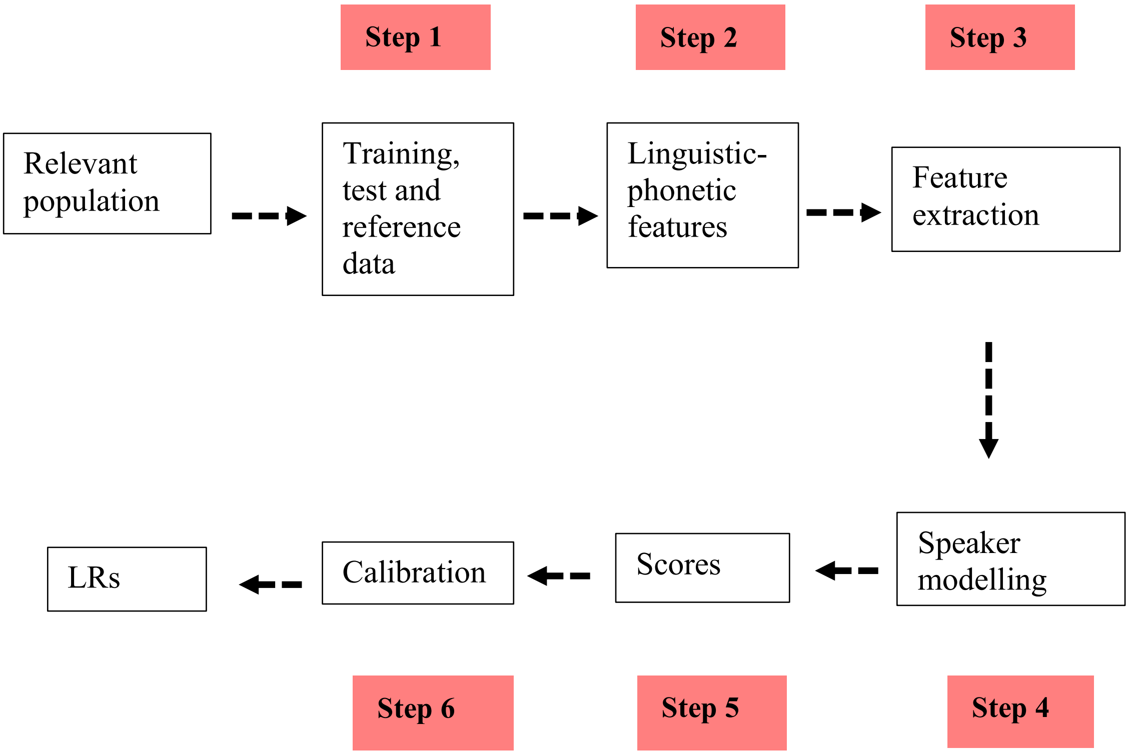


Figure 2.2 A schematic process of numerical LR computation and researchers' degrees of freedom involved in each step. This is a simplification of the real-world scenario and the order of steps differ among individual researchers.

In the current study, Chapters 4, 5 and 6 use numerical LRs to show how different configurations of training, test and reference speakers, different combination of linguistic-phonetic features and different calibration methods would affect system validity and reliability and individual speakers' behaviour.

2.2.3 Current study

Despite the fact that previous studies have explored potential factors that affect system validity and reliability, the validity and reliability have generally been evaluated on the basis of pooled LRs, i.e. the LR output across the overall system. What previous studies have not done is to explore factors such as sampling variability and accent and channel mismatch in relation to individual speakers. It has been shown, e.g., by Hughes (2017), that a set of 30 to 40 training and test speakers and 15 reference speakers sampled from the relevant population is sufficient to estimate relatively reliable LR output. However, given what really matters in a real case is the specific voice of the speaker(s) under analysis and the samples used for experiments are likely to be smaller than that of the relevant population, this raises issues about how representative the testing is in respect of a specific a case, and whether *overall* performance provides adequate insight into the validity and reliability of an LR in a specific case of individual speakers.

Therefore, it is important to investigate the representativeness of an LR output to different conditions. The current study therefore investigates the effect of sampling variability on both overall performance and individual behaviour. It further considers the issue of how representative the system is under different conditions from the perspectives of the configurations of training, test and reference speakers, choice of linguistic features, score distributions and choice of calibration methods. As discussed earlier, since the real size of the relevant population is likely to be much larger than the speakers sampled, Chapter 4 explores the questions about whether the system validity and reliability would be affected if different configurations of training, test and reference speakers (from a relevant population) are used. If so, is the variability in the overall performance primarily caused by different configurations of training speakers, test speakers or reference speakers? Chapter 5 then explores the relationship between sampling variability (i.e., the choice of speakers used for training and testing systems, rather than sample size) and the choice of linguistic features used with a focus on both overall performance and individual behaviour. While Chapters 4 and 5 focus more on the effect of sampling variability at the *feature-to-score* stage, Chapter 6 looks into the effect of sampling variability when score distributions do not follow normality and how different calibration methods can reduce the LR variability.

Chapter 3 General Methodology

This chapter explains the methods employed throughout this thesis. It discusses the corpora and linguistic-phonetic segmental features, LR computation process and evaluation metrics for both overall performance and individual speakers' behaviour. Meanwhile, chapter specific methods are given in each of the experimental Chapters (Chapter 4, Chapter 5, Chapter 6) as well.

3.1 Corpora

Two corpora were used in this thesis, namely the Intelligence Advanced Research Projects Activity (IARPA) Babel Cantonese Language Pack (Andrus et al., 2016) and the Dynamic Variability in Speech corpus (DyViS; Nolan et al., 2009). This Chapter outlines the general structure of the two corpora, while the specific features used are introduced in the corresponding experimental chapters.

3.1.1 IARPA Babel Cantonese Language Pack

The IARPA Babel Cantonese Language Pack (Andrus et al., 2016) is a Cantonese corpus designed by Appen for the Babel program for speech recognition purposes. Cantonese speakers recruited for the corpus were from Guangdong and Guangxi provinces across five dialectal groups in China (i.e., not including Hong Kong, Macau or overseas Cantonese speakers). A detailed breakdown of speaker number and their origin is given in Table 3.1 below:

	Total	**Male**	**Female**
Central Guangdong	262	138	124
Northern Guangdong	190	99	91
Northern Pearl River Delta group	221	102	119
Southern Pearl River Delta group	212	77	135
Guangxi and Western Guangdong	201	98	103
Total number of recordings	1086	514	572

Table 3.1 Numbers of recordings across five regions in Guangdong and Guangxi provinces.

The corpus contains 1086 recordings (approx. 215 hours) of natural Cantonese conversational speech, together with transcriptions in simplified Chinese characters and Romanised Chinese (Huang & Kok, 1999). Sociolinguistic factors are not controlled in this corpus, beyond language (Cantonese), regional accent, and biological sex (male/female). Age is the only social variable that is recorded, with participants aged between 16 and 67 years old at the time of the recording. All conversations were recorded through mobile phone calls in daily life environments with unexpected settings, such as in the office, street, karaoke bar, and inside vehicles. The conversations were recorded from the caller's end and saved into two separate audio files which contains the caller and receiver speech respectively. All the audio files were sampled at a rate of 8000Hz, meaning that only information up to 4000Hz was available for analysis.

Only recordings from male speakers were used. Recordings from Guangxi and Western Guangdong areas were excluded due to their distinctive accent from the rest of the groups, i.e., Central Guangdong, Northern Guangdong, Southern Pearl River Delta and Northern Pearl River Delta regions. This led to a total number of 416 recordings available. Further, approximately 27% of these recordings were excluded due to extreme background noise.

A major limitation in using the IARPA corpus is that it only has one recording session per speaker, which is not an ideal corpus for research studies in FVC. As such it does not fit with typical forensic conditions involving two samples recorded with a certain time interval between them. Using contemporaneous speech data of this kind is expected to overestimate the validity and reliability of the overall FVC system relative to real casework (Enzinger & Morrison, 2012). However, this corpus is forensically realistic in terms of recording channels since the conversations were recorded using telephones in different environments (e.g., indoor, outdoor, recording/transmission-channel mismatch) and large number of speakers. In principle, forensic recordings could be made in any situation by any recording device.

3.1.2 Dynamic Variability in Speech (DyViS) corpus

The DyViS corpus (Nolan et al., 2009), on the other hand, is a much more controlled corpus that was developed by the University of Cambridge and designed for forensic phonetic studies in British English. The corpus contains 100 male speakers of Standard Southern British English (SSBE) aged between 18 and 25. Each participant was recorded under both studio and telephone conditions, and 20 of them participated in a second recording session for non-contemporaneous variation analysis. Four tasks were involved in the DyViS corpus, namely, mock police interview, telephone conversation with an 'accomplice', reading passage and reading sentences. For the current thesis, only the first two tasks were used.

In task 1, subjects were asked to attend a mock police interview. However, the subjects were instructed to answer questions based on the information given in a map (Figure 3.1). Meanwhile, the subjects were told they could only answer questions containing the information provided in black type on a prompt sheet, while avoiding mentioning information provided in red type. This process was designed to create "a situation of 'cognitive conflict,' where speakers were made to lie" (Nolan et al. 2009: 41). The task 1 recordings were digitised and sampled at 44.1 kHz and a 16-bit depth, and last approximately 20 to 30 minutes in duration.

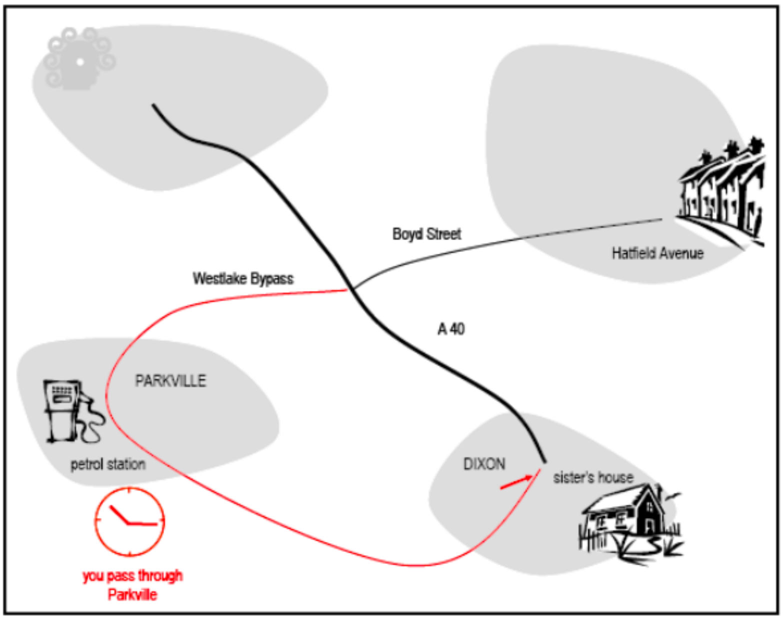


Figure 3.1 An example of the map used for task 1 containing information of the story told by mock suspect (Nolan et al. 2009:42). Participants should only answer questions containing the information provided in black type, while avoiding information provided in red type.

In task 2, subjects were instructed to undertake a telephone conversation with a mock accomplice. This task aimed to create a relaxed atmosphere so the subjects could use "a reasonably relaxed speaking style…, such as they might use when talking to a friend" (Nolan et al., 2009:43). The telephone conversation involves discussing a situation that happened in the police station. The conversation was structured via on-screen prompt cards containing required information. Task 2 was recorded both directly (high-quality; using 44.1 kHz sampling rate and 16-bit depth) and at the end of the telephone line for telephone effect (landline band-pass filter 300 to 3400 Hz; Byrne & Foulkes, 2004); however, only the high-quality recordings (i.e. 44.1 kHz sample rate, 16-bit depth) were involved in the current thesis.

3.2 Segmental features

Two Cantonese and one English segmental features were extracted and analysed in the current thesis: (1) Cantonese sentence final particle (SFP) /a/ '啊' (gloss: ah), (2) Cantonese disyllabic word /haia/ '係啊' (gloss: yes/yeah), and (3) British English Filled Pause (FP) *um.*

3.2.1 Cantonese segmental features

Numerous previous FVC studies have focused on English, and only limited work has been done in Cantonese (e.g., Chen & Rose, 2012; Rose & Wang, 2016). Given the large population of native Cantonese speakers (over 80 million), the current study in general contributes to FVC in Cantonese for research and practical purposes. The Cantonese sentence final particle (SFP) /a/ '啊 ah' and disyllabic word /haia/ '係啊 yes/yeah' were chosen as target variables. Cantonese SFPs are bound forms attached at sentence final position (Law, 2002). Functionally, they are often said to be the equivalent of intonation in English (Wakefield, 2011). The SFPs are potentially good variables for FVC, because they have a high frequency of occurrence in daily use (Leung, 2009). However, there is no empirical study showing the exact frequency of occurrence of each SFPs, which is probably due to the high variability of different types of SFPs (e.g., "ah, lah, wor, ga"). The number of different types of SFPs in Cantonese ranges from 30 (Kwok, 1984) to 95 (Leung, 1992) depending on how one counts them. The reason for selecting SFP /a/ '啊 ah' in the current experiment is that /a/ is one of the most common SFPs in Cantonese (Sybesma & Li, 2007) and occurred with high frequency in the current data set. Moreover, SFP /a/ occurs in sentence final positions, and thus it usually incurs syllable final lengthening which leads to longer duration and more canonical formant values than syllable onset /a/ (Lindblom, 1963). As a result, SFPs are easier to segment than syllable onset /a/.

The second variable /haia/ ‘係啊 yes/yeah’ is a disyllabic word consisting of two Chinese characters. The first part /hai/ ‘係’ means ‘yeah/yes’ by itself, and the second part is SFP /a/ which is attached to the end of the word for pragmatic purposes among Cantonese speakers. The words /hai/ and /haia/ have the same semantic meaning, but differ at morphemic level, i.e., whether the SFP /a/ is attached or not.

3.2.2 English filled pause *um*

Filled pauses (FP) are articulations between utterances used to ‘hold the floor’ while the speaker is thinking or hesitating (Clark, 2002). In English, FPs are generally assumed to be produced as a central vowel with or without a final bilabial nasal, i.e. [əː] or [əːm]. The FP *um* is similar to the Cantonese SPFs in the following aspects. Firstly, FPs have a high frequency of occurrence among speakers in most styles of spontaneous speech (Tschäpe et al., 2005). It is therefore highly likely that FPs would be available in sufficient numbers in most questioned samples (QS) and known samples (KS) containing spontaneous speech that are at least a few minutes long (Hughes et al., 2016). Secondly, FPs usually also have longer duration than lexical vowels, which gives longer and more stable formant trajectories assuming a normal recording circumstance. This makes segmentation and acoustic measurement easier to conduct. Thirdly, neither FPs nor Cantonese SFPs have a semantic meaning by themselves. Apart from these similarities, FPs are normally somewhat isolated from other syllables, in that they are often flanked by a pause on at least one side, and are thus less influenced by coarticulation. Therefore, it can be hypothesized that FPs should show low intra-speaker variation, because there is no other vowel or consonants at the onset or coda of FPs (Hughes et al., 2016). Moreover, FPs have been shown to be useful FVC variables in previous studies (e.g., Foulkes et al., 2004; Jessen, 2008). Only *um* is used in the current study because it was found to outperform its counterpart *uh* and has shown promising performance in separating SS and DS speaker pairs using the same data as the experiments in this thesis (Hughes et al., 2016).

3.2.3 Segmentation and feature extraction

This section explains the detailed data segmentation procedures for Cantonese segments. The target Cantonese variables were manually segmented and labelled on the same interval tier using a *TextGrid* in Praat (Praat version 6.0.36; Boersma and Weenink, 2017).

Figure 3.2 shows an example of segmented tokens. The boundaries were placed at the onset and offset of the full vocalic portion of each token. The onset of SFP /a/ was marked by the start of regular periodicity of the full vocalic portion, and the offset was marked by the last periodic wave of the full vocalic portion. The onset of /haia/ was marked by the start of the voiceless glottal fricative and the offset was marked by the last periodicity of the vocalic portion of /a/. A stable portion of the SFP /a/ was segmented when there is no clear-cut between the offset of preceding vowel/consonant and the onset of target variables.

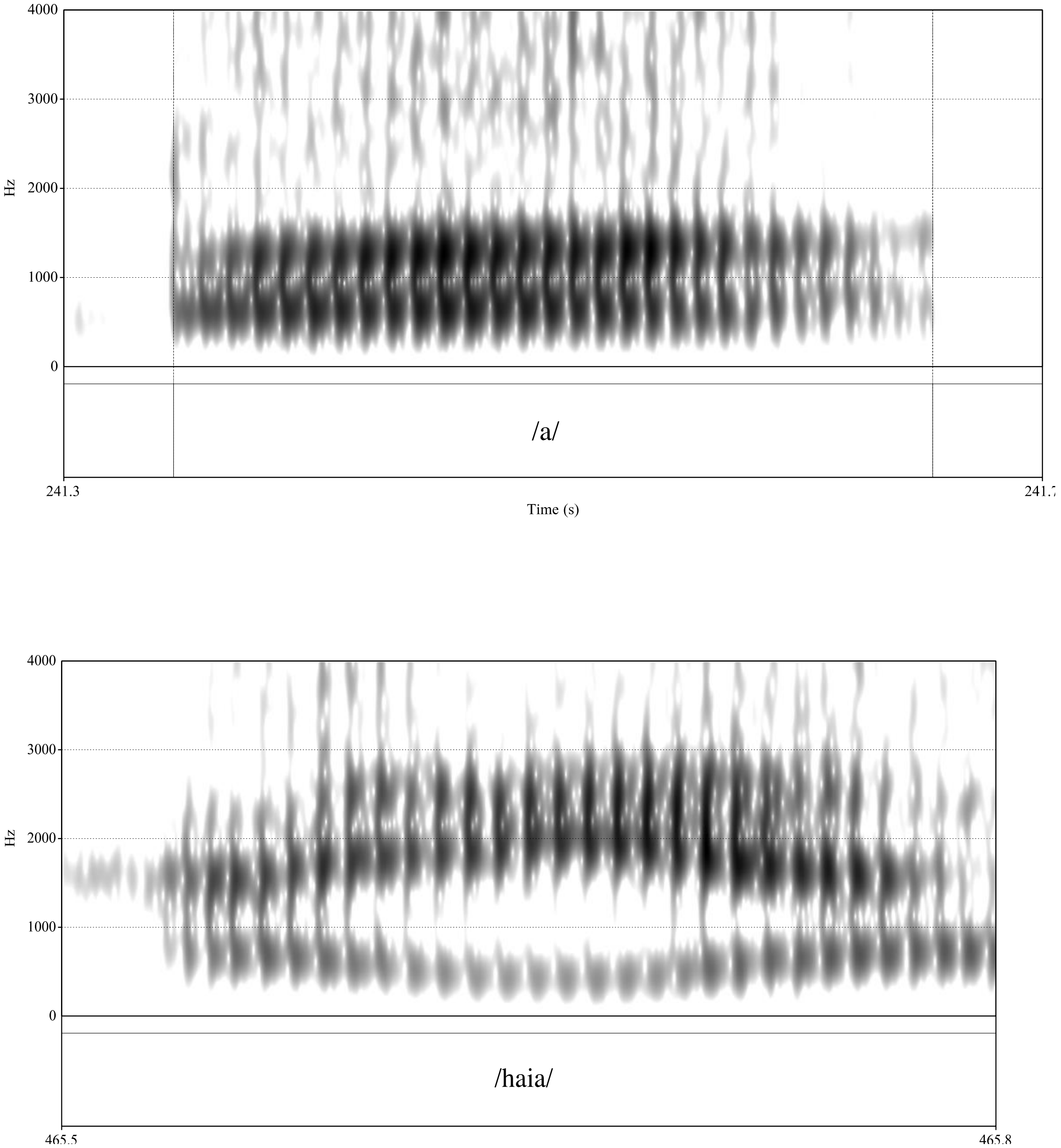


Figure 3.2 Segmented Cantonese variable /a/ (top panel) and /haia/ (bottom panel).

Due to poor transmission, there were large numbers of tokens where the F3s were extremely difficult to measure. A similar pattern was also reported in Gold (2009), where F3 measurement was extremely difficult to make in cellular phone recordings because the acoustic information was removed by “the high frequency cut-off of the codec” leading to the reduction in acoustic energy in the high frequency range. Although recordings in Gold (2009) were made internally using a voice recorder application rather than them being transmitted. Figure 3.3 (top panel) is an example of weak/unanalysable F3 from Gold’s study (2009, p.39) and Figure 3.3 (bottom panel) is an example from the Cantonese IARPA corpus, and the F3 is not traceable at all. As a result, only the first two formants were extracted from the Cantonese corpus IARPA in the current study.

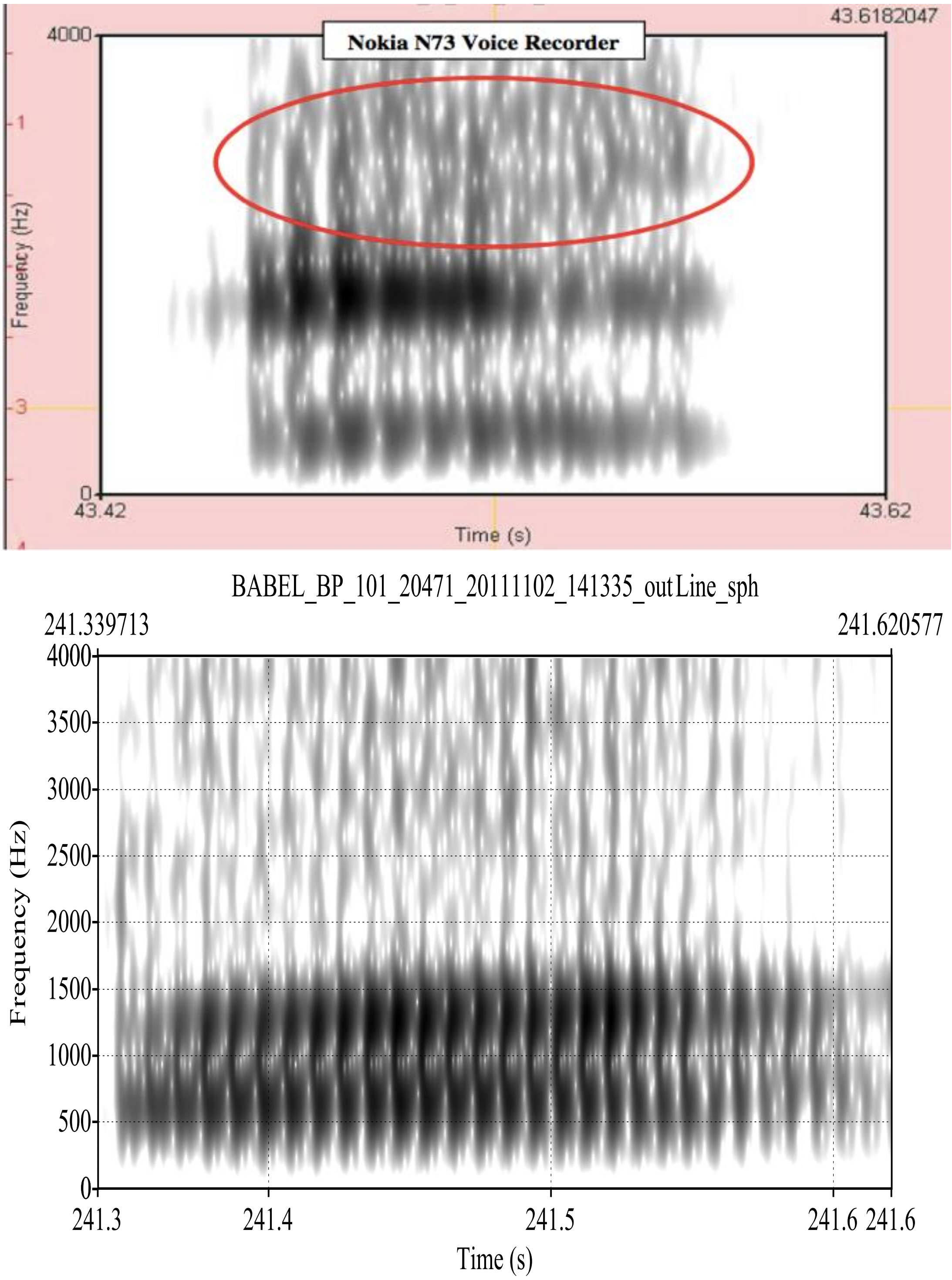
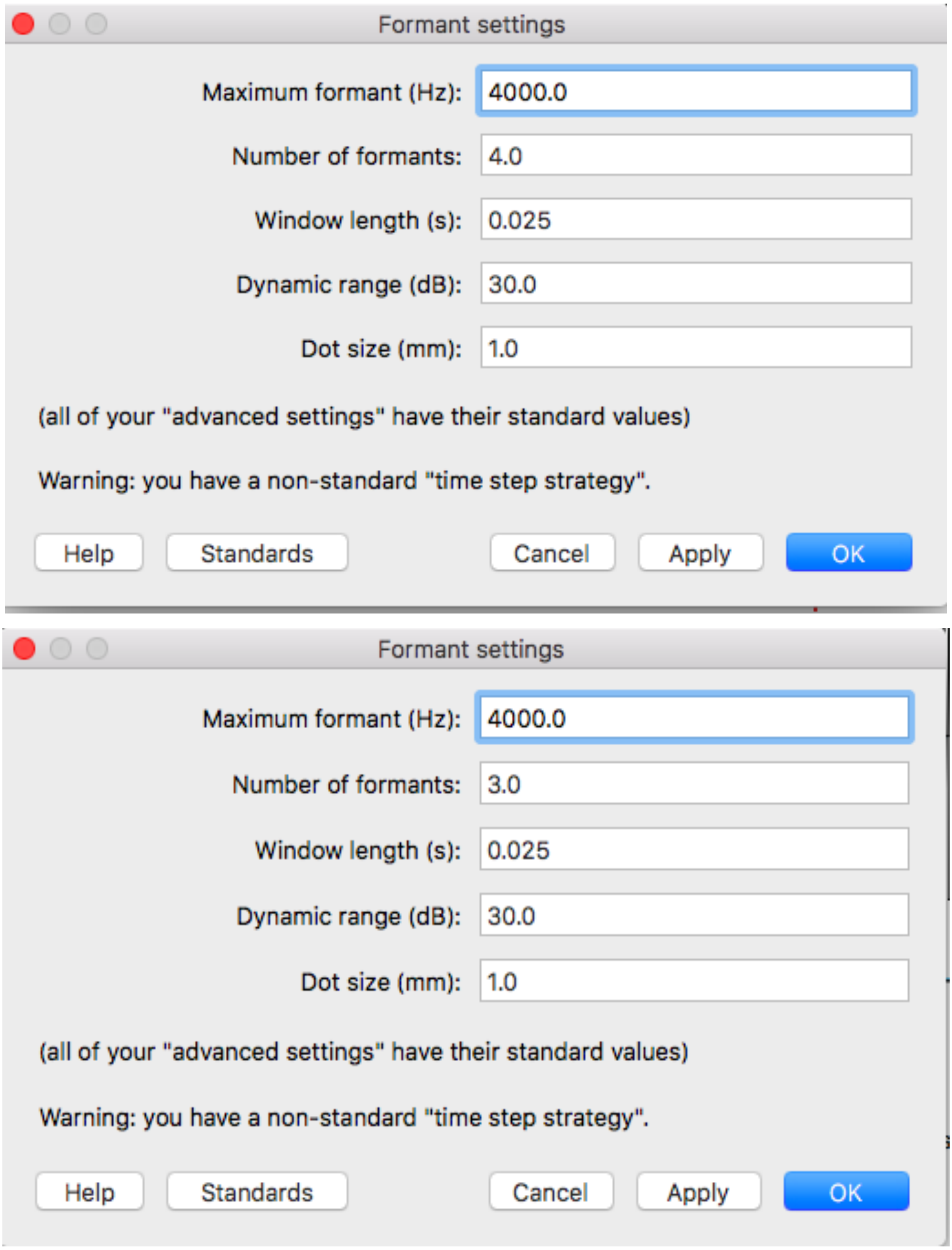


Figure 3.3 Top panel: unanalysable F3 of the word ‘bit’ from the voice recorder of the Nokia N73 (top) (Gold, 2009, Figure. 12, p.39). Bottom panel: unanalysable F3 of /a/ from IARPA, speaker_141335_outLine.

Two Praat scripts (Lennes, 2002, 2003) were then used to extract the first two formants of all the segmented tokens. The first script trimmed all the segmented tokens into a separate sound file and saved it to a specified directory. The reason for the first stage was to avoid running

Praat scripts with large sound files; extracting the tokens to individual sound files makes the process much more efficient. The second script took all the individual sound files and extracted the first three formants at 10% steps across the full vocalic portion. The data was then exported to text file.

The automatic formant extraction method makes experiments in acoustic-phonetic studies much more efficient. However, human-assisted examination is still an essential element. This is because Praat or any software might give inaccurate formant measurements under the default formant settings, and not all formants from all the speakers can be extracted accurately under the same format settings. Figure 3.4 (top panel) shows the default formant settings (this is the default settings used in current thesis, not Praat default settings) for formant extraction (i.e., maximum formant: 4000 Hz; number of formants: 4). Correspondingly, Figure 3.5 (top panel) shows the output of a linear predictive coding analysis (red dots) using the default formant settings. When the second script is being implemented, Praat takes measurements of formant values of those red points. However, there are several obvious errors in the top panel in Figure 3.5. This is because Praat would treat the lowest red dots as F1, the second lowest as F2 and the third lowest as F3, leading to inaccurate overall measurements. The bottom panel (Figure 3.5) shows the same token with a different formant setting, i.e., maximum formant: 4000 Hz, number of formants: 3, which would yield a more accurate measurement.



Formant settings

Maximum formant (Hz): 4000.0

Number of formants: 4.0

Window length (s): 0.025

Dynamic range (dB): 30.0

Dot size (mm): 1.0

(all of your "advanced settings" have their standard values)

Warning: you have a non-standard "time step strategy".

Formant settings

Maximum formant (Hz): 4000.0

Number of formants: 3.0

Window length (s): 0.025

Dynamic range (dB): 30.0

Dot size (mm): 1.0

(all of your "advanced settings" have their standard values)

Warning: you have a non-standard "time step strategy".

Figure 3.4: Praat formant setting window. Number of formants changed from 4 (top panel) to 3 (bottom panel).

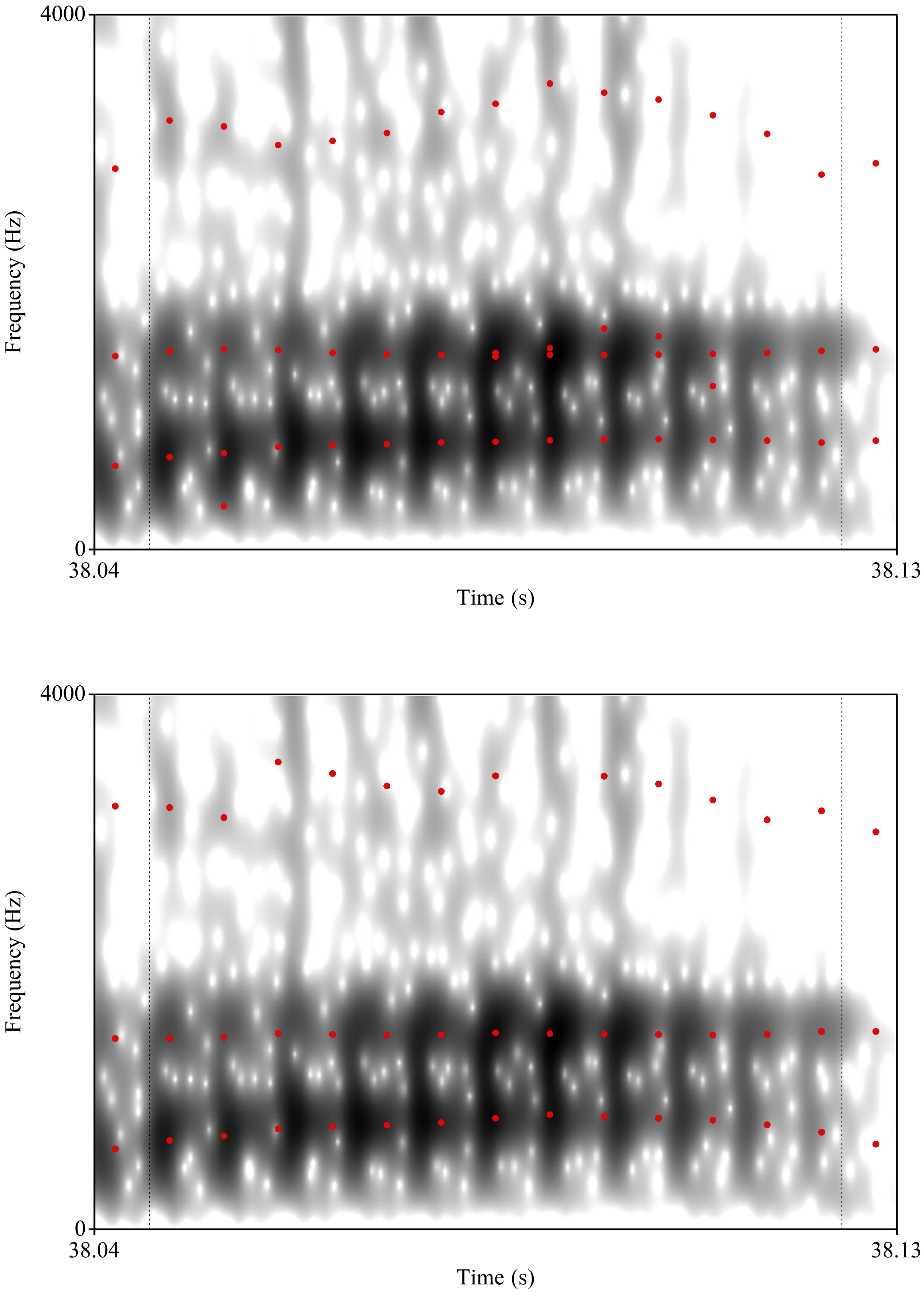


Figure 3.5 Formant extraction using different formant settings in Praat (Top panel: 4000 Hz, 4 formants; bottom panel: 4000 Hz, 3 formants). The top panel shows errors, i.e., the lowest red dot at the beginning and a few around F3 towards the end of the segment.

Therefore, the raw formant values were extracted twice under two different formant settings (Figure 3.4) and the measurements were saved in two separate text files. Formant values in two excel files were compared and filtered using acceptable measurement ranges adapted from Hughes & Foulkes (2015b), i.e., 200 Hz to 900 Hz for F1, and 1000 Hz to 2000 Hz for F2. Where tokens had values outside the range for one configuration of settings in one excel file they were replaced by a more reasonable measurement from the other configuration of settings from another excel file. If not, the formant values were removed when they were outside the range in both files. It was presumed that measurements would cause data points to deviate from the group mean. In order to delete the most extreme outliers in the raw formant values, the pooled mean across all speakers was calculated and a z-score of +/- 3.29 was applied to each set of 10% formant measurements. Measurements that were 3.29 standard deviations greater/less than the mean were removed. In order to preserve as many tokens as possible, missing values were replaced by the mean of two adjacent measurements. However, the whole token was removed where the first and/or last measurements were missing.

Hand-corrected acoustic data for the FPs was already available. The FP *um* data was an extension of that used by Hughes et al. (2016), containing the FP *um* data of 90 SSBE speakers. The data contained the raw values of the first three formants and F0 as well as nasal and vocalic duration (see Hughes et al. 2016 for detailed data extraction procedure). There were on average 35 FP *um* tokens per speaker per task. The first three formants and F0 of each token were fitted using quadratic polynomial curves, as no single formant or F0 was expected or observed to have more than one turning point. Figure 3.6 shows an example of quadratic fitting to five tokens from speaker #114 in Task 1 (I refer throughout the thesis to DyViS speaker codes in the form #xxx, and those speaker numbers are the original DyViS speaker numbers). The quadratic coefficients of the first three formants and F0 as well as the vocalic and nasal duration were then used as the input features for computing LRs.

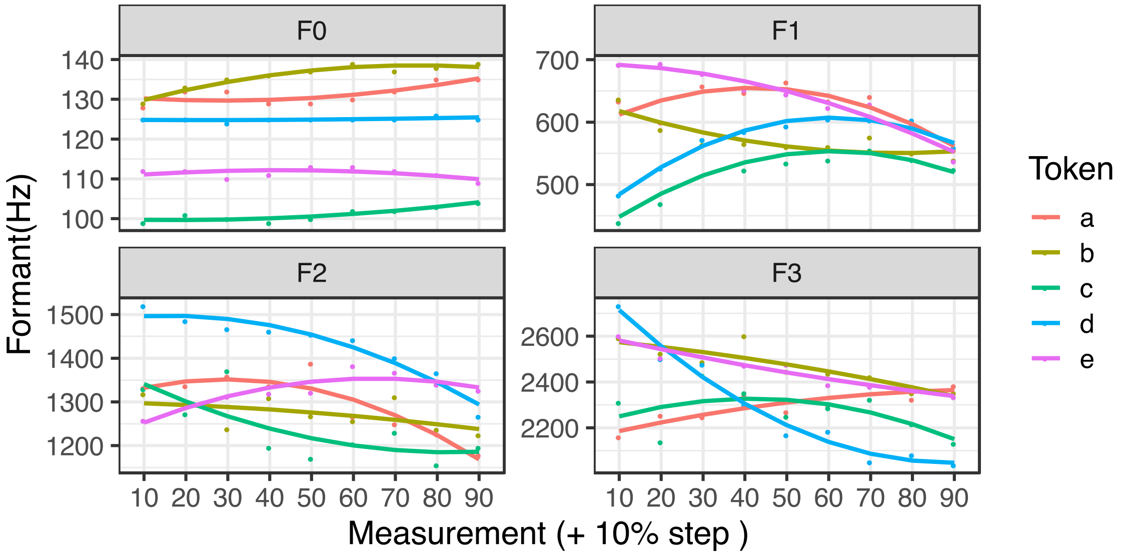


Figure 3.6 Quadratic curve fitting to F0, F1, F2 and F3 of five *um* tokens from #114 in DyViS Task 1.

3.2.4 Feature parameterisation

In FVC systems using linguistic-phonetic features, feature extraction is carried out over segmentals/suprasegmentals such as vowels or disyllabic words. In FVC studies, multiple measurements are often taken over the entire formant trajectories, and then trajectories are fitted with polynomial curves to capture the shapes. This process is known as the dynamic approach (McDougall, 2004). Numerous empirical studies have shown that using the dynamic approach gives promising speaker discrimination results and works better than static measurement (Greisbach et al., 1995; McDougall, 2004; Morrison, 2009; Rodman et al., 2002).

Figure 3.7 gives an example of the formant dynamic approach in extracting the first three formants over a diphthong /aɪ/. It shows that 9 measurements were taken throughout the duration of the segment, i.e., a 10% interval between each measurement.

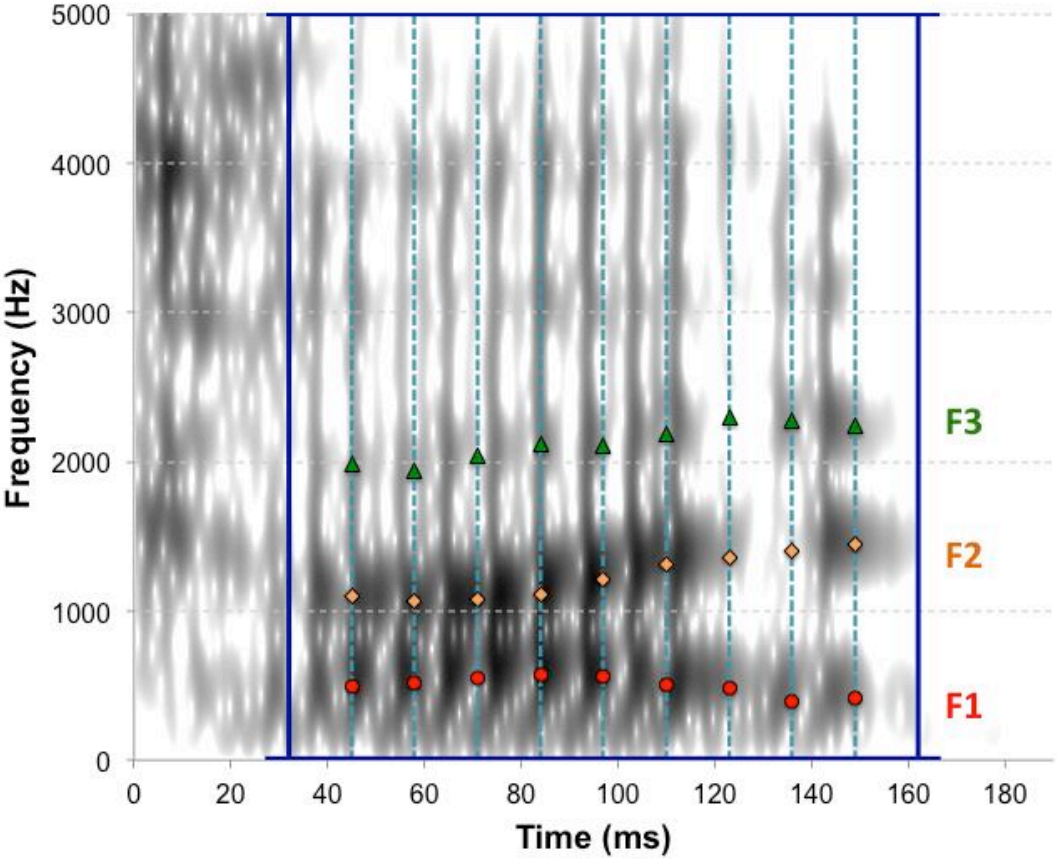


Figure 3.7 Example of formant dynamic analysis. Figure 3.8 from Hughes (2014, p.82). Measurement is taken with a 10% interval of /aɪ/ in the word *skype* from DyViS sample 027-01-060425.wav

In the current thesis, quadratic and cubic polynomial curves were fitted over the nine measurements of F1 and F2 over the vocalic portion of /a/ and /haia/ respectively. The polynomial fitting was conducted in R (R Core Team, 2018). The polynomial coefficients model the whole trajectory of each formant, which reduces the dimensionality of the dataset and improves the discrimination performance (Hughes et al., 2016; McDougall, 2006). The quadratic and cubic polynomial formulas are shown below,

Quadratic polynomial regression:

$$y = f(x) = ax^2 + bx + c \qquad \text{(Equation 3.1)}$$

Cubic polynomial regression:

$$y = f(x) = ax^3 + bx^2 + cx + d \qquad \text{(Equation 3.2)}$$

Figure 3.8 shows examples of quadratic and cubic fitting for /a/ (left panel) and /haia/ (right panel). The blue and red dots are the raw F1 and F2 values, while the lines are polynomial curves fitted to raw F1 and F2 values. The reason for using quadratic curves for SFP /a/ is because the trajectory of SFP /a/ is basically linear with no more than one turning point. Therefore, a quadratic curve seems to be enough to capture the trajectory information. By contrast, /haia/ has more complex trajectories. It is expected to have a rising-falling F2 and falling-rising F1 in general, as the vowel sequence contains an open vowel followed by a close front vowel, and then a second open vowel. The formant trajectories are therefore expected to have two turning points. As a result, cubic polynomial curves should suffice to capture the formant complexity in /haia/.

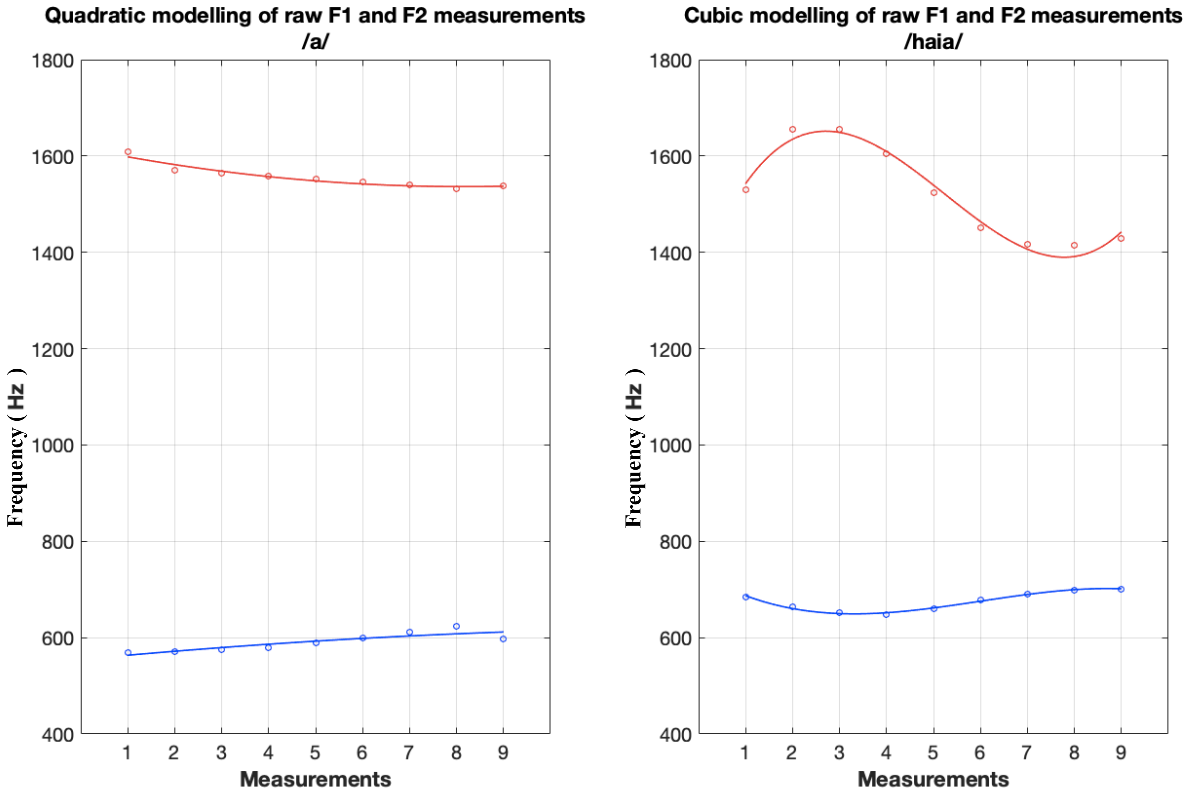


Figure 3.8 Quadratic and cubic curve fitting to /a/ (left panel) and /haia/ (right panel) respectively.

Each token was plotted and fitted with corresponding polynomial curves. Obvious errors in the raw formant data were again removed, and polynomial curves were re-fitted. There were cases where the raw data points had a large deviation from the polynomial curves. Figure 3.9 gives an example of the same seven /haia/ tokens fitted with cubic polynomial curves before (left panel) and after (right panel) hand-correction. Outliers (highest values of the second and third measures of F2) in the left panel in Figure 3.9 were located in the original recording, the data points were measured manually, and the tokens were fitted again by using cubic polynomial curves (Figure 3.9 right panel). The final dataset contains 155 speakers for SPF /a/ and 64 speakers for /haia/, with on average 13 to 14 tokens per speaker.

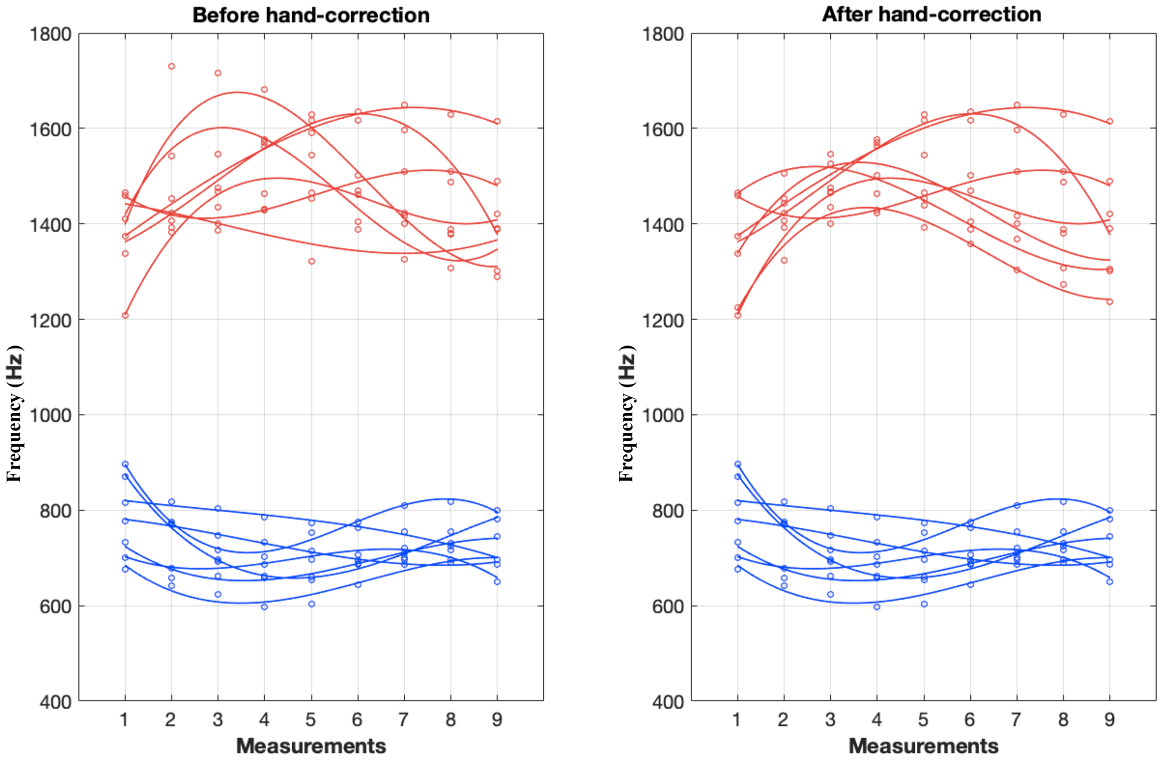


Figure 3.9 Plot of raw formant values of /haia/ fitted with cubic polynomial curves. An example of showing error measurements in the F2 (left panel) and hand-corrected version (right panel). The F1 values were sensible and did not need hand-correction.

3.3 LR computation

In LR-based FVC, two stages (i.e., *feature-to-score* and *score-to-LR*) and three datasets (i.e., training, test and reference data) are normally involved (Section 1.4). This section explains the methods involved in the two stages in relation to functions of training, test and reference data.

3.3.1 Feature-to-score

After the speech features are extracted from the speech data, speaker models are built using mathematical models, i.e. *feature-to-score* conversion (Morrison, 2013). Suspect and offender models are built from training and test data respectively to assess the similarity between two speech samples; meanwhile, typicality between suspect and offender samples are evaluated using a model based on data from the reference speakers. Throughout this thesis, the multivariate kernel density model (MVKD; Aitken & Lucy, 2004) was used for the *feature-to-score* stage. In the MVKD procedure, the within-speaker variation is modelled with a normal distribution, while the between-speaker variation is modelled using kernel density estimation (Aitken & Lucy, 2004 115:117; Rose, 2013a).

The numerator of the MVKD formula is given in Equation 3.3, evaluating the probability of observing the difference between the suspect and the offender given they come from the same speaker,

Numerator of MVKD:

$$f_0(\bar{y}_1, \bar{y}_2 \,|U, C) \; = (2\pi)^{-p}|D_1|^{-\frac{1}{2}}|D_2|^{-\frac{1}{2}}\,|C|^{-\frac{1}{2}}$$

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

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

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

where $y^* = \; (D_1^{-1} + D_2^{-1})^{-1}\,(D_1^{-1}\bar{y}_1 + D_2^{-1}\bar{y}_2)$

(Equation 3.3)

The denominator of the MVKD formula (Equation 3.4) evaluates the probability of observing the difference between the suspect and offender samples given that those samples are produced by different speakers that are from the relevant population,

Denominator of MVKD:

$$f_1(\overline{y}_1 + \overline{y}_2 \,|U, C) = (2\pi)^{-P}|C|^{-1}(mh^p)^{-2}$$

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

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

(Equation 3.4)

Aitken & Lucy (2004: 116-117)

where U, C = intra-, inter-speaker variance/covariance matrices
n_1, n_2 = number of replicates per speaker
m = number of speakers in reference population
p = number of assumed correlated variables per speaker
D_1=D_1, D_2 = offender, suspect variance/variance/covariance matrices = $n_1^{-1}U$, $n_2^{-1}U$
h = optimal smoothing parameter for kernel density = $(4/(2p+1))^{1/(p+4)_{m}-1/(p+4)}$
$\overline{y}_1$ = $\overline{y}_1$, $\overline{y}_2$ = offender, suspect mean vectors
$\overline{x}_i$ = intra-speaker means of reference population.

The results of the MVKD formula are LR-like *scores* taking both similarity and typicality into consideration. The higher the score, the greater the support for prosecution proposition, while a lower score gives more support to the defence proposition (Brümmer & du Preez, 2006). However, scores are purely non-negative values that cannot be interpreted directly; therefore, the *score-to-LR* stage is needed to transform scores into interpretable LRs and to optimise overall performance.

3.3.2 Score-to-LR

In the second stage, calibration is conducted to convert test scores into LRs. This is done using calibration models trained on scores computed for a set of training data. Four calibration methods were used in the current thesis, namely logistic regression (Brümmer et al., 2007), regularised logistic regression (Morrison & Poh, 2018), Bayesian model (Brümmer & Swart, 2014) and empirical lower and upper bound (ELUB; Vergeer et al., 2016). In Chapters 4 and 5, only logistic regression was used, while all four calibration methods were used in Chapter 6. The following sections give a brief introduction of the rationales behind these calibration methods. Apart from logistic regression, the other three calibration methods incorporate uncertainty into the LR itself, such that LRs will be closer to 1 when uncertainty is high (i.e., when sample size is low). In this thesis, the logistic regression calibration is carried out using an R package `fvclrr` (Lo, 2018) and the other three are implemented using a Matlab script (Morrison, 2018).

3.3.2.1 Logistic regression

Logistic regression (Brümmer et al., 2007) is one of the most widely used calibration methods in data-driven LR-based FVC. A set of training scores is normally required to train the logistic regression model where the coefficients from the model are then applied to the test scores to generate calibrated LRs. Logistic regression, and calibration methods in general, optimises overall performance, minimises C_{llr} and serves to “ameliorate what would otherwise be very misleading results” (Grigoras et al., 2013:620). A visualisation of logistic regression modelling using training scores is given in Figure 3.10,

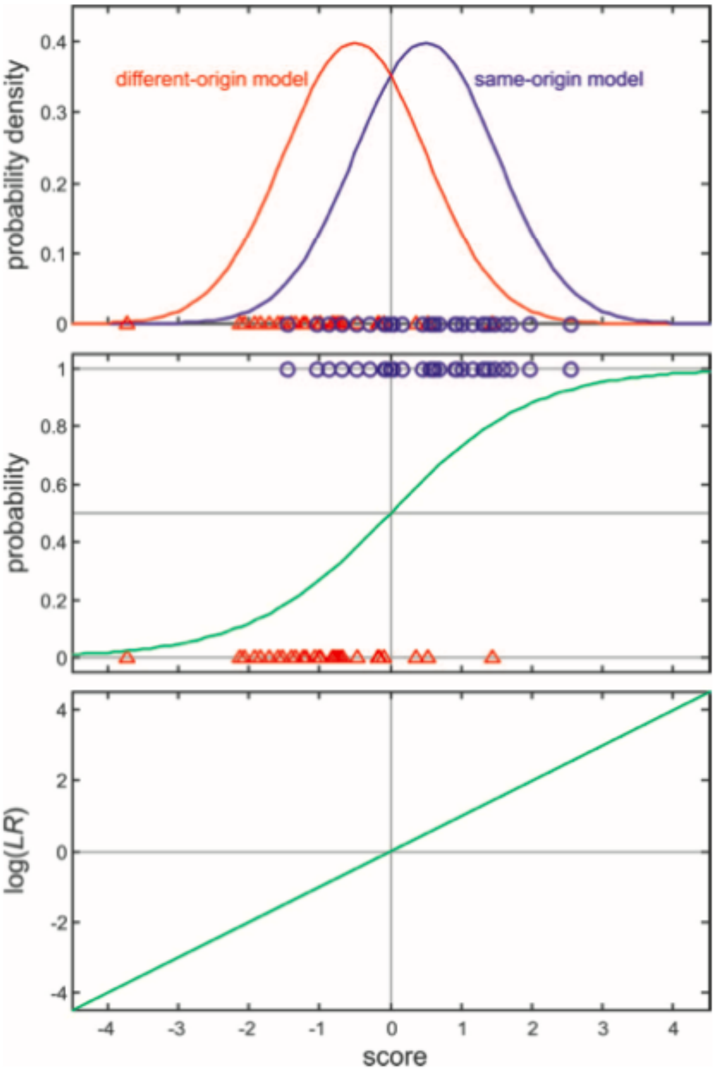


Figure 3.10 Distributions of SS (red) and DS (blue) scores (upper panel), logistic regression fitted to SS and DS scores (middle panel) in probability space and linear relationship between SS and DS scores derived from the logistic regression modelling in log-odds space (bottom panel). (Morrison 2013: 182 Figure 4).

Ramos-Castro (2007:120) and Morrison (2013:184) explain that zero and one are assigned to the distributions of SS and DS training scores (Figure 3.10, upper panel) and are modelled using logistic regression where the sigmoidal curve is fitted to SS and DS training scores using maximum likelihood function in probability space (Figure 3.10, middle panel). Logistic regression computes the best fit of the sigmoidal curve by making the probability of hypothesis given evidence (i.e., score) as close as possible to zero for DS comparisons and as close as possible to one for SS comparisons. The sigmoidal curve is then transformed from probability space to log-odds space to generate the linear relationship between SS and DS log scores

(Figure 3.10, bottom panel). Once the linear relationship is obtained, the conversion from scores to LLRs can be achieved using the form of linear regression (Equation 3.5),

$$LLR = \alpha + \beta s \qquad \text{(Equation 3.5)}$$

where α and β are the regression coefficients, s is the score from test data and LLR is the calibrated LR. The intercept α and coefficient β obtained using logistic regression and training scores are the shift and scale values. The shift and scale values are added and multiplied to the test scores respectively to generate calibrated LLRs.

3.3.2.2 Regularised logistic regression

The rationale behind regularised logistic regression is similar to logistic regression, where the shift and scale values are obtained by fitting a logistic regression model to the training data. The shift and scale values are then applied to test data to generate interpretable LRs. However, the regularisation process is implemented by adding sets of pseudodata (Menard, 2010; Morrison & Poh, 2018) with uninformative uniform distribution to avoid numerical problems or shrink extreme LR outputs. Note that pseudodata can have other distributions, but the one used in the current thesis is uninformative uniform distribution. The pseudodata is then weighted using w^{ψ} which is calculated using a κ^{ψ} value divided by two times the sum of the number of SS and DS comparisons in the training data. In this way, the effect of regularisation (i.e., the w^{ψ} value) decreases as the sample size in the training data increases (Morrison & Poh, 2018). Once the pseudodata is weighted, a regularised logistic regression model can be fitted to the weighted pseudodata with reduced fitted slope to shrink the LR output. One disadvantage of regularised logistic regression is that the κ^{ψ} value needs to be specified, and the choice of κ^{ψ} values is arbitrary to some extent. Depending on the purpose of calibration, small prior values ($\kappa^{\psi} \leq 0.1$) deal with the complete separation issue (i.e., SS and DS speakers are completely separated and no overlap between SS and DS scores before calibration) and larger prior values ($\kappa^{\psi} \geq 1$) deal with extreme LR outputs (namely shrink the range of LR output) (Morrison & Poh, 2018).

Figure 3.11 shows an example of applying different κ^{ψ} values using regularised logistic regression (ibid). In Figure 3.11 (panel (a)), logistic regression was fitted to training data with complete separation between SS (circles coded as 1) and DS (triangles coded as 0) scores. As a result, the likelihood is maximised and extreme LRs are produced when the intercept lies between the highest DS training score and lowest SS training score. Figure 3.11 panel (b) and panel (c) shows an example of regularised logistic regression fitted with extra copy of SS and DS training data using κ^{ψ} equals to 0.1 and 5 respectively. In the current thesis, the focus is placed on shrinking the range of LR output, and therefore I follow Morrison and Poh (2018) in adopting κ^{ψ} equals to 5 in using regularised logistic regression for calibration.

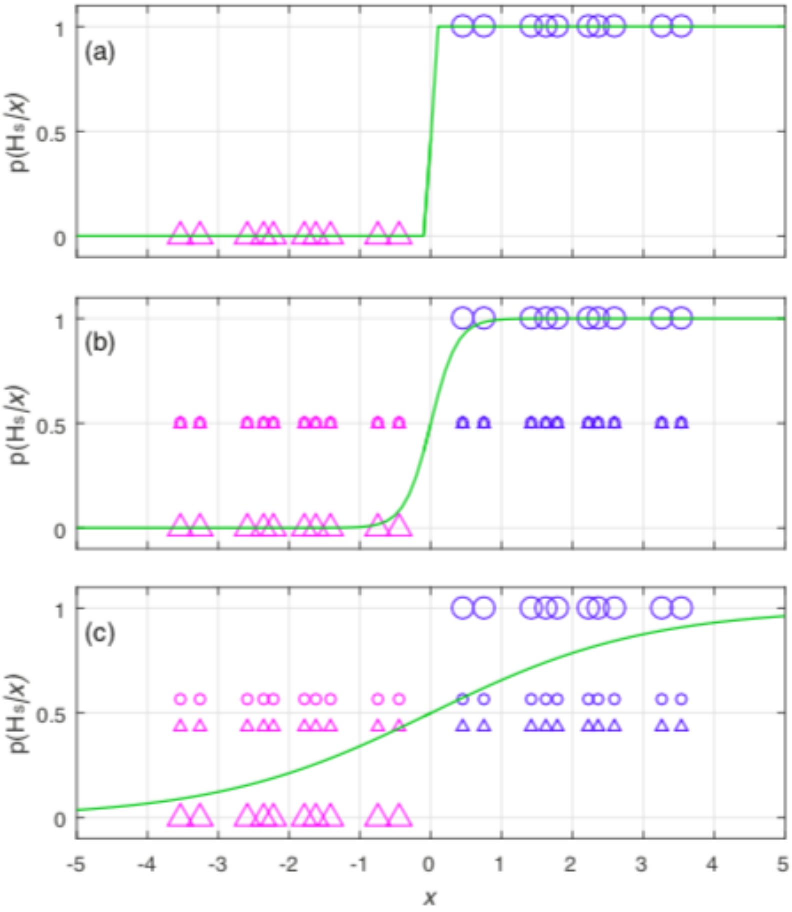


Figure 3.11 An example of regularised logistic regression using different κ^{ψ} values (Figure 2 in Morrison & Poh, 2018, p. 205). Panel (a) shows an example of logistic regression fitted without regularisation. Panels (b) and (c) show logistic regression fitted with small and large regularisation to avoid complete separation and to induce shrinkage respectively. Large symbols are the sampled data, and the small symbols are the weighted pseudo data with an uninformative uniform distribution.

3.3.2.3 Bayesian model

The fully Bayesian approach involves the use of priors (i.e. hyperparameters) to reduce the magnitude of the LRs when uncertainty is high (Brümmer & Swart, 2014; Morrison & Poh, 2018). The fully Bayesian calibration models need to be estimated using SS and DS training scores respectively. The likelihood of the Bayesian models is then evaluated using test scores (Brümmer & Swart, 2014). Meanwhile, the prior belief and the strength of the belief for the mean and variance of the training scores need to be specified. However, due to the nature of FVC, the ground truth is impossible to know and it has been shown that uninformative priors yield more constrained Bayes factors (BF, the Bayesian counterpart of the frequentist LR) than informative priors (Morrison et al., 2014). In this thesis, following Morrison & Poh (2018), the Jeffreys reference uninformative priors were used. The formula for Bayesian model estimation can then be simplified as shown in Equation 3.6.

Bayesian model using Jeffreys reference:

$$\lambda^B = t_{n-1}(x|\hat{\mu}, \frac{n+1}{n-1}\hat{\sigma}^2) \qquad \text{(Equation 3.6)}$$

Where t is a t distribution, n is the sample size, x is the test score, $\hat{\mu}$ and $\hat{\sigma}^2$ are the sample mean and variance of the training score. The calculation of BF is then the ratio between the likelihood of the Bayesian models evaluated using test scores is shown in [3.7].

$$\log\text{(BF)} = \log\left(t_{n_{ss}-1}\left((x \mid \widehat{\mu_{ss}}, \frac{n_{ss}+1}{n_{ss}-1}\hat{\sigma}_{ss}^2)\right)\right) - \log\left(t_{n_{ds}-1}\left((x \mid \widehat{\mu_{ds}}, \frac{n_{ds}+1}{n_{ds}-1}\hat{\sigma}_{ds}^2)\right)\right)$$

(Equation 3.7)

However, Morrison and Poh (2018) explained that monotonicity is not guaranteed in [3.7] as *t* distribution is used for both numerator and denominator. Therefore, certain constraint needs to be imposed to reduce the extent of non-monotonicity. The BF is then calculated using Equation 3.8 where the pooled sample variance ($\hat{\sigma}^2$) is used; meanwhile, the degrees of freedom (n_{ss}+$n_{ds} - 2$) are adjusted to take the pooled variance calculation into consideration and the $\bar{n}$ is the sum of SS and DS samples divided by 2.

$$\log\text{(BF)} = \log\left(t_{n_{ss}+n_{ds}-2}\left((x \mid \widehat{\mu_{ss}}, \frac{\bar{n}+1}{\bar{n}-1}\widehat{\sigma}^2)\right)\right) - \log\left(t_{n_{ss}+n_{ds}-2}\left((x \mid \widehat{\mu_{ds}}, \frac{\bar{n}+1}{\bar{n}-1}\widehat{\sigma}^2)\right)\right)$$

(Equation 3.8)

Morrison and Poh (2018: 203)

3.3.2.4 Empirical lower and upper bound (ELUB)

The ELUB (Vergeer et al., 2016) method uses empirical data to set maximum and minimum values to the LRs that a given system can output based on the training set, then all other LRs produced by the test data are limited within that range. The advantage in using the ELUB calibration methods lies in that it avoids extreme LRs caused by data extrapolation at the distribution tails. Figure 3.12 shows an example of data extrapolation using different density models. The right panel shows the SS (H_p) and DS (H_d) scores fitted with different density models and the left panel gives the enlarged figure at the distribution tails. The distribution estimation at the tail of SS scores (H_p) is not well supported by the observed data, and it is more likely to obtain extreme LRs using exponential density models than kernel density estimation.

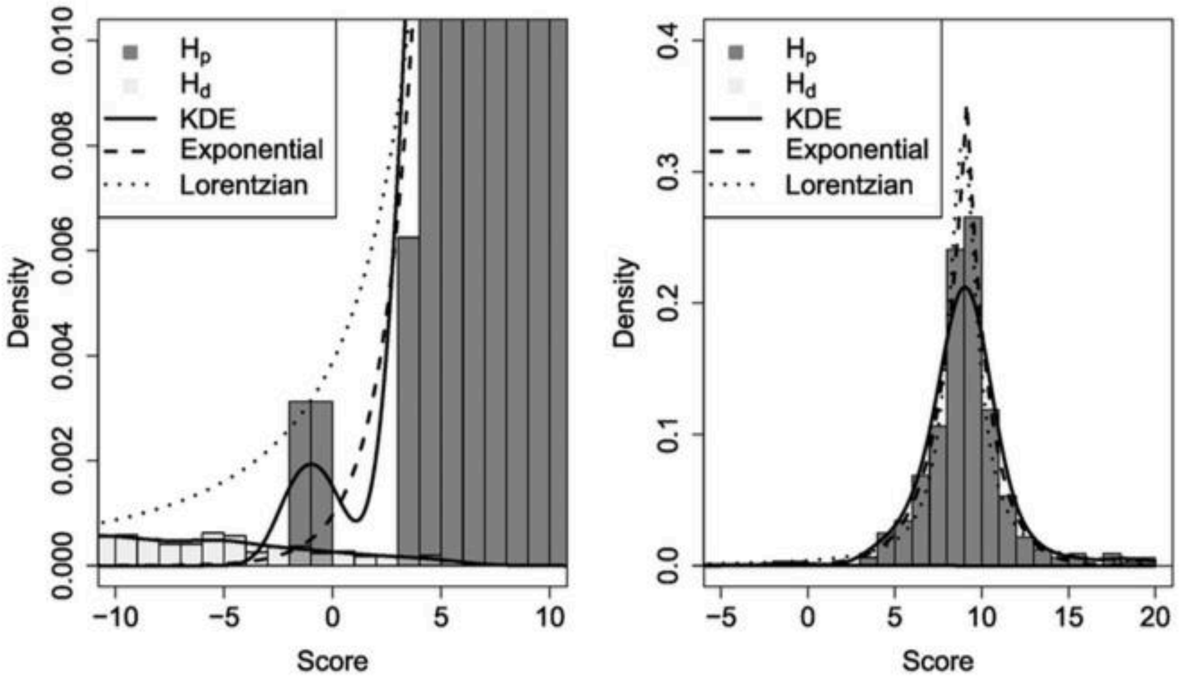


Figure 3.12 SS (H_p) and DS (H_d) scores fitted with different density models (Figure 2 in Vergeer et al., 2016, p.483). The left panel shows enlarged plot at the distribution tail showing different degrees of extrapolation using KDE, Exponential and Lorentzian modelling for SS scores.

As mentioned above, ELUB shrink extreme LR outputs by setting a maximum and minimum value to the LR output by the system based on the sizes of the training data, and the rationale behind the ELUB calibration method lies in a rule of thumb that the LRs should be smaller than the sample size of the training data for H_d, or smaller than 1 divided by the size of training data for H_p (Vergeer et al., 2016). For example, if there are one disputed speech and 4 suspects, and it is known that one of suspects produced that disputed speech. The probability of each suspect produced the disputed speech is 25% (1/4) before any evidence is taken into consideration. Therefore, a conservative estimate of the LR should be no larger than 1/4 when H_p is true or smaller than 4 when H_d is true.

The implementation of ELUB is carried out using the expected utility (EU) ratio. Ultimately, the EU serves as a reference for decisions when the uncertainty is high. Vergeer et al. explained that the utility function, to be associated with the court decisions, consists of four possible binary decisions, i.e., U_{cp} (conviction of a perpetrator), U_{ci} (conviction of an innocent), U_{ap} (acquittal of a perpetrator), U_{ai} (acquittal of an innocent). It is assumed that the trier of fact should aim to maximize EU rather than decisions, and EU can be expressed using equation [3.9].

$$\text{EU} = \max\{\ \text{U}_{\text{cp}} \times P(p) + \text{U}_{\text{ci}} \times (1 - P(p))\ ;\ \text{U}_{\text{ap}} \times P(p) + \text{U}_{\text{ai}} \times (1 - P(p))\ \}$$

Equation 3.9

Vergeer et al. (2016, p. 485)

Where $P(p)$ stands for the probability that the suspect is the perpetrator, and maximizing EU leads to a Bayes decision rule (Brümmer, 2010),

Decision "convict":

$$\frac{P(p)}{(1 - P(p))} > \frac{(U_{ai} - U_{ci})}{(U_{cp} - U_{ap})}$$

Equation 3.10

Decision "acquit":

$$\frac{P(p)}{(1 - P(p))} < \frac{(U_{ai} - U_{ci})}{(U_{cp} - U_{ap})}$$

Equation 3.11

(Vergeer et al., 2016, p.486)

The left-hand sides of Equations 3.10 and 3.11 are the posterior odds in a court decision, i.e., $P(p) = P(H_p)$ and $1 - P(p) = P(H_d)$. The $\frac{(U_{ai} - U_{ci})}{(U_{cp} - U_{ap})}$ then can be regarded as a threshold odds (*Odds_{th}*) for the decision. Based on Bayes rule, the decision rule can be rearranged into:

Decision "convict":

$$\frac{P(E|H_p)}{P(E|H_d)} > Odds_{th} \times \frac{P(H_d)}{P(H_p)}$$

Equation 3.12

Decision "acquit":

$$\frac{P(E|H_p)}{P(E|H_d)} < Odds_{th} \times \frac{P(H_d)}{P(H_p)}$$

Equation 3.13

Vergeer et al. (2016, p.486, Equations 3(a) and 3(b))

Because the left hand side of Equations 3.12 and 31.3 is in fact the LR, the right hand side can then be used as a threshold LR (LR_{th}). The LR_{th} can then be used to calculate the empirical uppder and lower bounds using Equation [3.14].

Calculating EUratio:

$$\text{EUratio(ELUB)} = \frac{\begin{cases} 1 & if\ LR_{th} > 1 \\ LR_{th} & if\ LR_{th} \leq 1 \end{cases}}{\frac{nLR_s \leq LR_{th} + 1}{nLR_s + 1} + LR_{th} \times \frac{nLR_d > LR_{th} + 1}{nLR_d + 1}}$$

(Equation 3.14)

(Morrison & Poh, 2018)

The numerator is the neutral system LR that is no larger than 1 and the exact value depends on the LR_{th}. For the denominator, nLR_s and nLR_d are the number of the SS and DS LRs in the training data respectively. The $nLR_s \leq LR_{th}$ represents the number of SS LRs that is no larger than LR_{th} and $nLR_d > LR_{th}$ represents the number of DS LRs that is higher than LR_{th}. The upper and lower boundaries obtained from EUratio can then be applied to the test data to shrink the LR output.

3.4 Evaluation

In the current thesis, evaluation was carried out for both overall performance and individual speakers' behaviour. The main metrics used for overall performance were Log LR cost function (C_{llr}; Brümmer & du Preez, 2006) and Equal error rate (EER), while zoo plots (Doddington et al., 1998; Dunstone & Yager, 2009) and root-mean-square-deviation (RMSD) were used for the evaluation of individual speakers.

3.4.1 System validity and reliability

3.4.1.1 Log LR cost function (C_{llr})

C_{llr} (Brümmer & du Preez, 2006) is used as the main metric for system validity evaluation in the current thesis. The C_{llr} function (Equation 3.9) evaluates overall performance based on the magnitude of evidence (i.e., LR), specifically the contrary-to-fact LRs. This means that not all errors are problematic, i.e., contrary-to-fact LRs with lower magnitude (e.g., closer to 0) are less problematic for overall performance than contrary-to-fact LRs with higher magnitude (e.g., 100).

Log LR cost function (C_{llr}):

$$C_{llr} = \frac{1}{2}\left[\frac{1}{N_{ss}}\sum_{i=1}^{N_{ss}} log2\left(1+\frac{1}{LR_{ss_i}}\right)+\frac{1}{N_{ds}}\sum_{i=1}^{N_{ds}} log2\left(1+LR_{ds_i}\right)\right]$$

(Equation 3.15)

Gonzalez-Rodriguez et al. (2007)

where:

N_{ss} = the number of comparisons using speech samples produced by the same speaker
LR_{ss} = LRs of speech samples produced by the same speaker
N_{ds} = the number of comparisons using speech samples produced by different speakers
LR_{ds} = LRs of speech samples produced by different speakers

The left and right part within the square brackets of the equation (either side of the + sign) is the mean of the output of a function applied to all the LRs obtained from same speaker (SS) and different speaker (DS) comparisons respectively (Morrison, 2011b), and C_{llr} is the sum of the mean divided by two. Therefore, each part of the equation contributes to the final value of C_{llr}. Since C_{llr} evaluates overall performance based on the magnitude of LRs (not based on system decisions directly), it is not easy to interpret *per se*. Therefore, C_{llr} is often used to compare performance between systems. A C_{llr} between 0 and 1 indicates that the system is capturing some useful information, and the closer to 0 the better the system validity is. A C_{llr} of 1 is equivalent to a system that consistently produces LLRs of 0 (LRs = 1) irrespective of whether they came from SS or DS comparisons, and a LLR equals to 0 gives no useful information for FVC purpose. As such, a C_{llr} of higher than 1 indicates that the system is not capturing any useful information. Following the consensus on validation of FVC outlined in (Morrison et al., 2021), 1 is considered as the only relevant threshold for C_{llr} in the current thesis.

In each of the experimental chapters (i.e., Chapters 4, 5 and 6), the experiments were replicated 100 times and the system validity was assessed using mean C_{llr} values, i.e., the lower the mean C_{llr} the better the system validity. Chapter 2 has discussed system validity testing using mean C_{llr} as well as system reliability testing using different metrics, e.g., 95% credible interval, variability scores and C_{llr} range, in previous studies. Since there is no consensus about which measure should be used for system reliability testing in FVC, the range between C_{llr} values was used for assessing system reliability for the sake of simplicity (i.e., subtraction of the highest and lowest C_{llr} across replications). A system with good performance should have both low mean C_{llr} values and small C_{llr} ranges. Due to the limit of sample size, it is acknowledged that the training, test and reference speakers across 100 replications are not completely independent of each other. This is likely to underestimate the variability in system output, compared with using low quality recordings and truly independent samples of speakers. However, given the limitations on the availability of data in the real world, it is likely that any type of replication study of this sort would use samples which are not entirely independent. Therefore, the results should in some ways be treated as the 'base case scenario' for variability in system performance as a function of sampling, and that even wider variability in system performance would be expected where poor quality recordings are used and independent samples are drawn from a much larger database. Meanwhile, the 100 times replication, similar to previous studies (Ali et al., 2015; Morrison & Poh, 2018), is an arbitrary choice in the current thesis.

3.4.1.2 Equal error rate (EER)

EER measures the system validity using the LLR as a discriminant function, which indicates the threshold values when the *false rejection* rate equals the *false hit* rate (Table 3.2). EER was calculated with an adaptation of Ketabdar's (2004) MATLAB function in R (Lo, 2018). 5000 thresholds were used across the entire range of LLRs.

	SS comparison	DS comparison
LLR > 0	*Hit*	*False hit*
LLR < 0	*False rejection*	*Correct rejection*

Table 3.2 Categorical calculation of EER using the LLR as a discriminant function. *Hit* and *Correct rejection* are when speaker pairs are classified correctly, and *false rejection and false hit* are when speaker pairs are wrongly classified.

The major limitation in using EER for LR-based FVC system evaluation is that EER only treats LLRs categorially and it does not take the magnitude of evidence into consideration, i.e., what is considered an "error" is not being judged on a threshold of LLR = 0; therefore, a system that consistently yields high contrary-to-fact LLRs could have the same system validity to the one that produces low contrary-to-fact LLRs.

3.4.2 Individual behaviour

3.4.2.1 Root-mean-square-deviation (RMSD)

The RMSD is frequently used to measure the differences between predicted values and observed values. However, RMSD is used here to capture the mean distance between a speaker's individual LLR and mean LLR. A low RMSD value indicates that the specific speaker is more stable, while a high RMSD value indicates that the speaker is less stable. Essentially, the RMSD value captures the stability of each individual speaker's behaviour throughout the experiments. The RMSD values of SS and DS comparisons were calculated for each speaker in each system, measuring how far each comparison LLR is from the mean for that speaker. The individual speakers' RMSDs were calculated using LLRs. The formula of RMSD is defined below in [3.10]:

Root-mean-square-deviation:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x_i - y_i)^2} \qquad \text{(Equation 3.16)}$$

Where n is the total number of SS or DS comparisons,
x_i is the individual SS/DS LLR in each comparison,
y_i is the mean of SS/DS LLRs of each individual speaker in each system.

3.4.2.2 Zoo plot

The zoo plot uses animal names to categorise individuals (in this case, speakers) for identifying the trends of overall performance and problematic speakers. Initially, four animal names were included in the biometric menagerie (Doddington et al., 1998; Dunstone & Yager, 2009), i.e. *sheep, goats, lambs* and *wolves* (Figure 3.13), and these four animals are defined based on user scores in biometric systems (same-speaker (SS) or different-speaker (DS) scores in FVC). In the context of FVC and other biometric systems, *sheep* contains the majority of the speakers, who tend to match well against themselves and well against others, having high SS and low DS scores; *goats* are difficult to match when they are compared against themselves but easy to match with other speakers, often characterised by low SS and low DS scores; *lambs* and *wolves* are speakers who are both likely to yield good performance when they are compared against themselves (high SS scores). However, the difference is that *lambs* are more likely to be impersonated by other speakers and *wolves* “are successful at impersonation” (Yager & Dunstone, p.161, 2010).

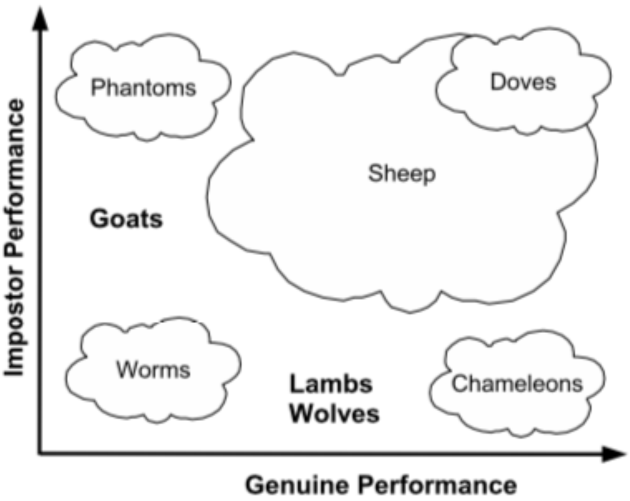


Figure 3.13 A Zoo plot with the relative location of the new and old classes of animals (Dunstone and Yager, 2009, p. 168, Figure 8.5). New class of animals: phantoms, doves, worms and chameleons; old class of animals: goats, sheep, lambs and wolves.

Dunstone and Yager (2009) later proposed a new class of creatures, i.e., *chameleons, phantoms, doves,* and *worms* (Figure 3.13). *Chameleons* are speakers who are always obtaining high SS and low DS scores when being compared against themselves and other speakers, while *phantoms* are rarely similar to any other speakers and likely to yield low SS and DS scores. *Doves* are speakers who have especially good performance when being compared with themselves and can be easily separated from other speakers. *Worms* are speakers who always give bad performance when being compared against themselves and other speakers, and they are characterised by high DS and low SS comparison scores.

By convention, lower and upper quartiles are applied to speakers' SS and DS comparison scores to define the corresponding locations of each animal group, i.e. *phantoms* are speakers with the lowest 25% SS and DS comparison scores (top-left corner in Figure 3.13); *doves* are speakers with highest 25% SS and lowest 25% DS comparison scores (top-right corner in Figure 3.13); *worms* are speakers who are among the lowest 25% SS and highest 25% DS comparisons scores (bottom-left corner in Figure 3.13), and *chameleons* are speakers with the highest 25% SS and 25% DS comparison scores (bottom-right corner in Figure 3.13). However, it is worth noting that one speaker can be in different animal groups depending on the specific configurations of the system, e.g., speech feature, statistical models and calibration methods used, meaning that speakers' animal groups are system specific.

In this thesis, the new class of animal labels (i.e., *chameleon, phantom, dove, worm*) are used because they are "defined in terms of a relationship" between SS and DS scores among a group of speakers, while the old animal labels are defined in terms of only score distributions (Dunstone & Yager, 2009, p.161). Further, instead of using upper and lower quartiles, the thresholds for defining different classes of animals are adjusted using LLRs. This is because LLRs are comparable between speakers and across systems, which are different from comparison scores used in ASR systems. The LLR thresholds for animal groups are given in Table 3.3 below, where LLR equals 0 indicates equal support for prosecution and defence proposition. The thresholds between 0 and 1 and 0 and -1, similar to the 25% threshold used in automatic systems, are rather arbitrary; meanwhile, different thresholds can be selected depending on the specific research aim and system configuration.

Animal group	SS LLR	DS LLR
Phantoms	≤ 0	≤ -1
Worms	≤ 0	≥ -1
Doves	≥ 1	≤ -1
Chameleons	≥ 1	≥ 0

Table 3.3. LLR thresholds for animal groups in the current study.

3.4.3 Verbal LR scale

Although the LR itself serves as an indication of the strength of evidence, the extent to which end users (e.g., jury and courts) are capable of comprehending numerical LRs is very much a concern for the courts and forensic analysts (Kinoshita & Ishihara, 2014; Marquis et al., 2016). As a result, Champod & Evett (2000) proposed a conversion from numerical LRs to verbal expressions. Table 3.4 shows the conversion between LLRs and verbal expression.

It is acknowledged that there are certain drawbacks in using the verbal scale, e.g., the subjective interpretation could vary between and within groups (Rose, 2002). Moreover, cliff edge effects that are imposed at categorical boundaries could also be problematic (Morrison & Enzinger, 2016), e.g., the difference between two LRs of 99 and 100 is equivalent to the difference between *moderate* and *moderately strong* support for the prosecution. Nevertheless, both the forensic scientists and the courts are still facing difficulties in assessing the weight of evidence in terms of numerical LRs (Marquis et al., 2016). The verbal scale is adopted as a reference for discussion and contextualising numerical LRs in the thesis.

LLR	**Verbal Expression**
4 : 5	Very strong support for same-origin
3 : 4	Strong support for same-origin
2 : 3	Moderately strong support for same-origin
1 : 2	Moderate support for same-origin
0 : 1	Limited support for same-origin
0 : -1	Limited support for different-origin
-1 : -2	Moderate support for different-origin
-2 : -3	Moderately strong support for different-origin
-3 : -4	Strong support for different-origin
-4 : -5	Very strong support for different-origin

Table 3.4 Numerical LRs in verbal expression Champod & Evett (2000).

3.5 Chapter Summary

This Chapter explained general methods, i.e., corpora, linguistic-phonetic variables, feature parameterisation method, statistical models for *feature-to-score* and *score-to-LR,* evaluation metrics, used in the current thesis. A summary using bullet points are given below.

Corpora used for this thesis:

- Intelligence Advanced Research Projects Activity (IARPA) Babel Cantonese Language Pack (Andrus et al., 2016)
- Dynamic Variability in Speech corpus (DyViS; Nolan et al., 2009).

Linguistic-phonetic variables used in this thesis:

- Cantonese SFP /a/
- Cantonese disyllabic word /haia/
- English FP *um*

Feature parameterisation method:

- Polynomial curves
    - quadratic curves for Cantonese SFP /a/ and English FP *um*
    - cubic curves Cantonese disyllabic word /haia/

Statistical models used for *feature-to-score*:

- MVKD

Calibration methods for *score-to-LR:*

- Logistic regression
- Regularised logistic regression
- Bayesian model
- Empirical lower and upper bound

Evaluation metrics:

- C_{llr}
- EER
- RMSD

Chapters 4, 5 and 6 outline a series of experiments investigating the effect of sampling variability in LR-based FVC, and the RQs listed in Section 1.3 are addressed. Detailed experimental procedures, specific linguistic-phonetic variables, input features as well as calibration methods used for speech comparison in Chapters 4, 5 and 6 are explained in each chapter accordingly.

Chapter 4 Overall performance as a function of sampling variability

This chapter explores the effects of sampling variability on overall performance with regard to different configurations of training, test and reference speakers. The Cantonese SFP /a/ and /haia/ as well as English FP *um* were used for comparison. Overall performance was evaluated using C_{llr} and EER.

4.1 Introduction

In the evaluation of LR-based FVC, it is known that training, test and reference speakers need to be sampled from a relevant population that reflects the conditions of the voice(s) under analysis in a real case. Typically, a group of speakers, often around 60 (e.g., Zhang et al., 2013) to 90 (e.g., Rose & Cuiling, 2018) in total, is selected, split equally into training, test and reference speakers (e.g., 20-20-20/30-30-30) before running the analysis. As discussed in Section 1.3, the size of the relevant population is likely to be larger than the speakers sampled for the expriment. It is then essential to investigate whether a system yields stable performance when different groups of speakers (randomly sampled from the same relevant population) are used. The current chapter explores this issue in relation to the specific choice of speakers (rather than the size of the data set) in the training, test and reference data sets, and RQs 1a. and 1b. are addressed.

RQs 1.

a. To what extent does overall performance (validity and reliability) vary if different configurations of training, test and reference speakers (from the same relevant population) are used?
b. Is the variability in the overall performance primarily caused by different configurations of training speakers, test speakers or reference speakers?

4.2 Method

4.2.1 Input data and LR computation

The Cantonese SFP /a/, disyllabic word /haia/ and English FP *um* were used as the linguistic variables. In total, 155 and 64 speakers were used for /a/ and /haia/ respectively, with an average of 14 tokens per speaker (Table 4.1). As mentioned in Section 3.2.4, only the F1 and F2 were used for /a/ and /haia/, and the polynomial coefficients (quadratic for /a/ and cubic for /haia/) were used for LR computation. As the Cantonese corpus only contains one recording session per speaker, data for each speaker was divided in half with the first half acting as the suspect sample and the second half acting as the offender sample. For FP *um,* 90 speakers were used with an average of 35 tokens per speaker per Task (i.e., the separate recording contexts in the DyViS corpus). The quadratic coefficients of the first three formants as well as nasal and vocalic duration were used for LR computation. Task 1 data was used as the suspect sample and Task 2 data as the offender sample. The training and test data were first computed using MVKD (Aitken & Lucy, 2004) to generate training and test scores, and calibration was carried out using logistic regression (Brümmer et al., 2007).

Linguistic-phonetic variable	Average number of tokens per task	Numbers of Speakers	Recording session/Task
/a/	14	155	1
/haia/	14	64	1
um	35	90	2

Table 4.1 Average number of tokens and recording sessions of each linguistic-phonetic variable.

4.3 Experiments

Four experiments are described in this Chapter (see below Experiments 1 to 4). The number of speakers available for each linguistic-phonetic variable was different; thus, the number of speakers used in each data set (i.e., training, test and reference) was adjusted in four experiments accordingly to enable speaker sampling. Each experiment was replicated 100 times and the procedure was explained below.

- Experiment 1: varying all speakers
    - Speakers were randomly sampled into each of the three data sets in each replication.

Experiment 1 aimed to explore if the system would give consistent performance if the experiment was replicated multiple times by arranging speakers differently in the training, test and reference sets, and Table 4.2 shows the number of speakers used in each data set for each linguistic-phonetic variable. This is the equivalent of replicating validation studies for the same relevant population but using different speakers. Meanwhile, this is also where researchers have got the degrees of freedom on which speakers are assigned to which dataset. For example, if there are 90 speakers (numbered 1 to 90) available for system testing and 30 speakers are assigned to each of the training, test and reference data. The question is should speakers 1-31 or speakers 31-60 be assigned to the training, test or reference data and will it make a difference to system validity and reliability?

	Linguistic-phonetic variable		
	/a/	/haia/	*um*
Number of training speakers	30	15	30
Number of test speakers	30	15	30
Number of reference speakers	30	15	30

Table 4.2 Number of speakers used in each data set for each variable in Experiment 1.

- Experiment 2: varying test speakers
    - Experiment 2 was replicated but only varying the configurations of test speakers. Training and reference speakers were fixed.

Experiment 2 aimed to explore how overall performance is affected using different configurations of test speakers. The results can then be compared with Experiments 3 and 4 to investigate which dataset is more susceptible to sampling variability that leads to worse overall performance (i.e., validity and reliability). Table 4.3 shows the number of speakers used in each data set for each linguistic-phonetic variable.

	Linguistic-phonetic variable		
	/a/	/haia/	*um*
Number of training speakers	30	15	25
Number of test speakers	30	15	25
Number of reference speakers	30	15	25

Table 4.3 Number of speakers used in each data set for each variable in Experiments 2 to 4.

- Experiment 3: varying reference speakers
    - Experiment 3 was replicated but only varying the configurations of reference speakers. Test and training speakers were fixed.

Experiment 3 aimed to explore whether a random selection of speakers from the relevant population (e.g., a matched dialectal group) adequately represents the population, and how sensitive the system reliability was to the reference data. Table 4.3 shows the number of speakers used in each data set for each linguistic-phonetic variable.

- Experiment 4: varying training speakers
    - Experiment 4 was replicated but only varying the configurations of training speakers. Test and reference speakers were fixed.

Experiment 4 aimed to assess the sensitivity of the overall performance to different sets of training data given that speakers are all chosen from the relevant population. The EER was not reported in detail in Experiment 4, because the calibration coefficients derived from the training data only affect the C_{llr}. Thus, the EER will be the same across all replications where the test and reference speakers are fixed. Table 4.3 shows the number of speakers used in each data set for each linguistic-phonetic variable.

The implementation of the above four experiments were conducted in R (R Core Team, 2018) using the R package (*fvclrr;* Lo, 2018). The R script randomly samples speakers, runs the speech comparison and saves the results into a list (R Core Team, 2018). Each experiment was replicated 100 times with different configurations of training, test and reference speakers, as explained above. Details of the results are discussed below.

4.4 Results

4.4.1 Experiment 1: varying all speakers.

Figure 4.1 shows the overall performance sampling speakers in all three datasets. The boxplot shows the range of the C_{llr} values in the 100 replications. It is shown that varying test, training and reference speakers causes the overall performance to vary to different extents for different variables. Over the 100 replications, all the C_{llr} values of /a/ are lower than 1, which indicates that the system is giving some useful information in each replication. The C_{llr} of /a/ ranges from 0.60 to 0.97, and the median and interquartile range (IQR) are 0.74 and 0.07. The variation in C_{llr} does not indicate a stable overall performance, given that the logical threshold of C_{llr} is 1 (Morrison et al., 2021).

Turning now to the overall results for /haia/, Figure 4.1 shows that /haia/ has a wider C_{llr} range than /a/. The C_{llr} range of /haia/ varies from 0.29 to 1.15 and the median and IQR are 0.46 and 0.16. Despite this wide range, 75% of the C_{llr} values of /haia/ are lower than 0.55, and 50% of the C_{llr}s fall between 0.40 and 0.55. These results indicate that the system for /haia/ is giving a comparatively less stable performance than that for /a/. It is worth noting that the lowest C_{llr} for /haia/ is 0.29, which is fairly good given that only F1 and F2 were used. However, a large range of C_{llr} from 0.29 to 1.15 indicates an unstable overall performance.

As for the FP *um*, the C_{llr} values vary from 0.13 to 1.22 and the median and IQR are 0.31 and 0.11. In this data set, 75% of the C_{llr}s are lower than 0.38, and 50% of the C_{llr}s fall between 0.26 and 0.38, thus indicating a more stable performance than both /a/ and /haia/. However, there are nine outliers among the results from *um* and two of them are larger than 1, which means that the system is not giving any useful information in those two replications. Three out of the remaining seven outliers had a C_{llr} larger than 0.8, which also shows a fairly poor performance.

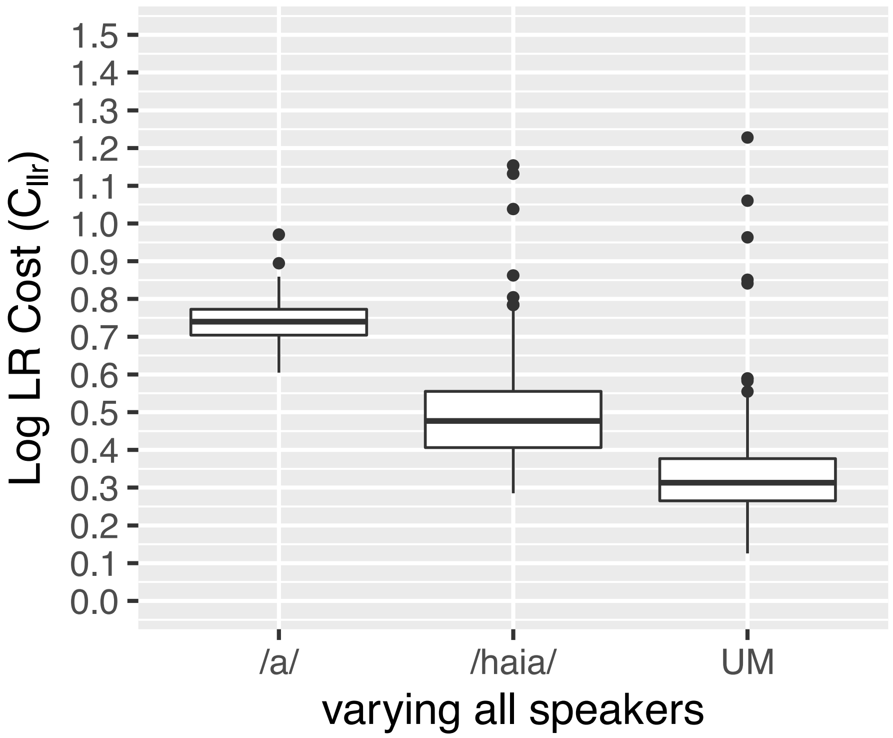


Figure 4.1 Boxplots show the C_{llr} ranges of /a/, /haia/ and *um* by varying speakers in all three datasets in each replication. Two whiskers are the first and third quartiles, and the central line represents the median value.

Overall, /a/ had the best system reliability in terms of IQR (0.07) and C_{llr} range, and /haia/ and *um* yielded lower mean C_{llr} than /a/.

Figure 4.2 shows the distributions of the EERs for /a/, /haia/ and *um* across the 100 replications, and Table 4.4 shows the statistics of C_{llr} and EER values across 100 replications. The EER of /a/ varies from 19.77% to 33.56% and the median and IQR are 26.26% and 3.50%. An IQR of 3.50% indicates that half of the EERs are concentrated between 23.28% and 26.78%. The EER of /haia/ ranges from 5.24% to 30.00% and the median and IQR are 13.33% and 6.66%. Half of the replications had an EER between 12.86% and 19.52%, indicating that /haia/ has better performance than /a/ in terms of EER, while /a/ seems to have a better reliability by having a lower IQR (3.50%). It is worth noting that 75% of the EERs from /haia/ are lower than all the EERs from /a/. However, a total EER range between 5.24% and 30.00% makes the system reliability of /haia/ no better than /a/. The FP *um* obtained an EER range varying from 2.70% to 13.33%, while the median and IQR are 6.70% and 2.66%. This indicates that *um* has a more stable performance than /a/ and /haia/ in terms of EER. All the EER values of FP *um* are lower than those of /a/. However, a total EER range between 2.70% and 13.33% is still considerably high.

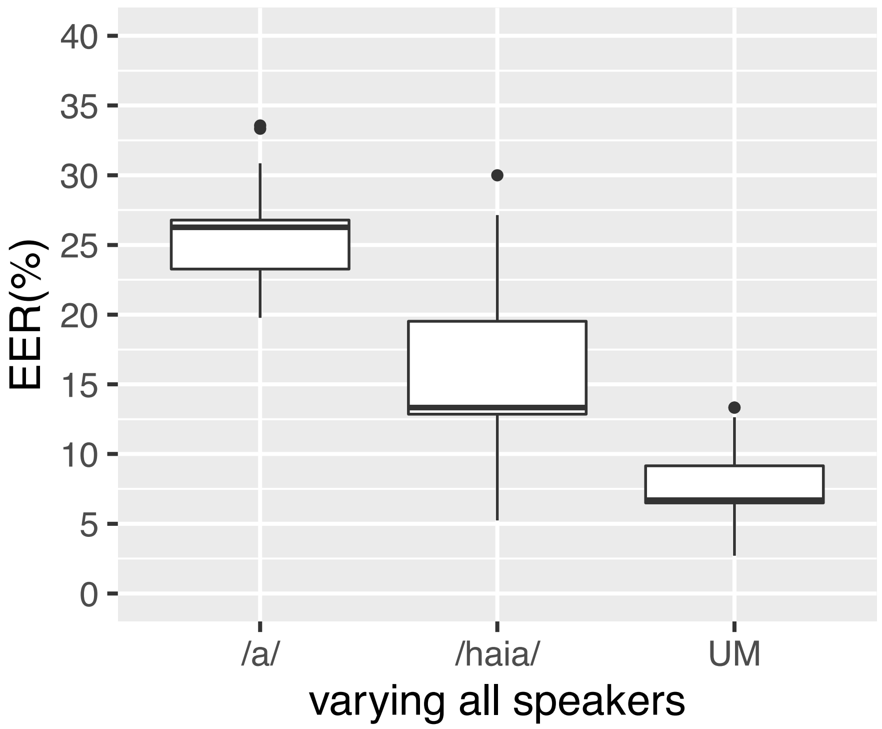


Figure 4.2 Boxplots represent the EER ranges of /a/, /haia/ and *um* by varying speakers in all three datasets in each replication. Two whiskers cover the first and third quartiles, and the central line represents the median value.

	C_{llr}			EER(%)		
	/a/	**/haia/**	***um***	**/a/**	**/haia/**	***um***
Min	0.60	0.29	0.13	19.77	5.24	2.70
1st qu.	0.70	0.40	0.26	23.28	12.86	6.49
Median	0.74	0.46	0.31	26.26	13.33	6.70
Mean	0.74	0.51	0.35	25.61	14.88	7.14
3rd qu.	0.77	0.56	0.38	26.78	19.52	9.15
Max.	0.97	1.15	1.22	33.56	30.00	13.33
IQR	0.07	0.16	0.11	3.50	6.66	2.66
Range	0.37	0.86	1.10	13.79	24.76	10.63

Table 4.4. Summary statistics of C_{llr} and EER of /a/, /haia/ and *um* in Experiment 1 (varying all speakers).

4.4.2 Experiment 2: varying test speakers.

In Experiment 2, training and reference speakers were fixed, while test speakers were varied in each replication. Figure 4.3 shows that using different test speakers in each replication causes the C_{llr}s to vary to different extents, and Table 4.5 shows the statistics of C_{llr} and EER values across 100 replications. However, the boxplots show that all the C_{llr} values are lower than 1 for all three linguistic-phonetic variables. The C_{llr} values of /a/ vary from 0.58 to 0.86 in the 100 replications. The median and IQR are 0.74 and 0.08, which indicates that 50% of the C_{llr} values vary within a small range. Turning to /haia/, the C_{llr} varies from 0.31 to 0.64, and the median and IQR are 0.49 and 0.08. In general, /haia/ outperformed /a/ in terms of mean C_{llr} (0.49). The overall C_{llr} range of /haia/ (0.33) is slightly higher than that of /a/ (0.28), but 75% of the C_{llr} values of /haia/ are lower than all the C_{llr} values of /a/. This indicates that using more acoustic input could potentially increase the system validity without increasing the variability. For the FP *um,* the C_{llr} values vary from 0.24 to 0.48, which shows a lower overall range than those produced by /a/ and /haia/. Moreover, FP *um* has a lower median (0.37), IQR (0.06), and mean (0.37) than SFP /a/ and /haia/, indicating that FP *um* has a better overall performance in terms of both validity and reliability.

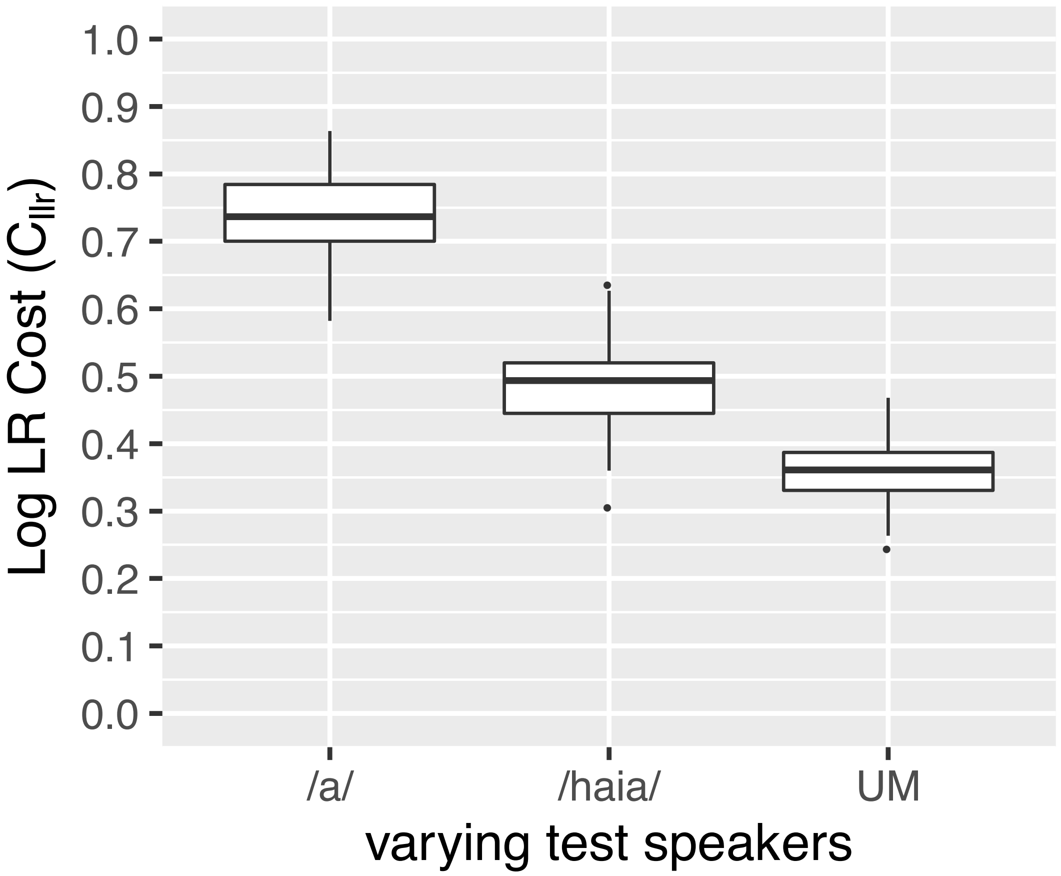


Figure 4.3 Boxplots represent the C_{llr} ranges of /a/, /haia/ and *um* by varying test speakers in each replication. Two whiskers cover the first and third quartiles, and the central line represents the median value.

Figure 4.4 shows the boxplots of EER of SFP /a/, /haia/ and FP *um* in Experiment 2. Similar to Experiment 1, the EERs of the three linguistic-phonetic variables vary to different extents. However, due to the limited number of speakers used, the EERs have more consistent patterns than C_{llr}. The EER of /a/ ranges from 16.84% to 36.72%, and the median and IQR are 27.01% and 6.38%, which is thus the least stable system among the three linguistic-phonetic variables. /haia/ on average yielded a lower EER than /a/. However, the EER of /haia/ ranges from 5.71% to 25.71% and a 20.00% EER range in the 100 replications makes the overall reliability of /haia/ no better than /a/. On the other hand, the median and IQR of /haia/ are 13.33% and 4.88%, which are lower than those of /a/. For FP *um*, the EER values range from 4% to 12%, which are the lowest among the three variables. The EER IQR of *um* (0.67%) is also lower than those of SFP /a/ and disyllabic word /haia/, indicating a much more stable overall performance.

There are only a discrete number of EERs possible given the number of speakers used. The smaller the number of speakers used, the smaller the number of possible EERs. It might also be due to the small number of speakers that led to some overlap in the configurations of speakers in each data group.

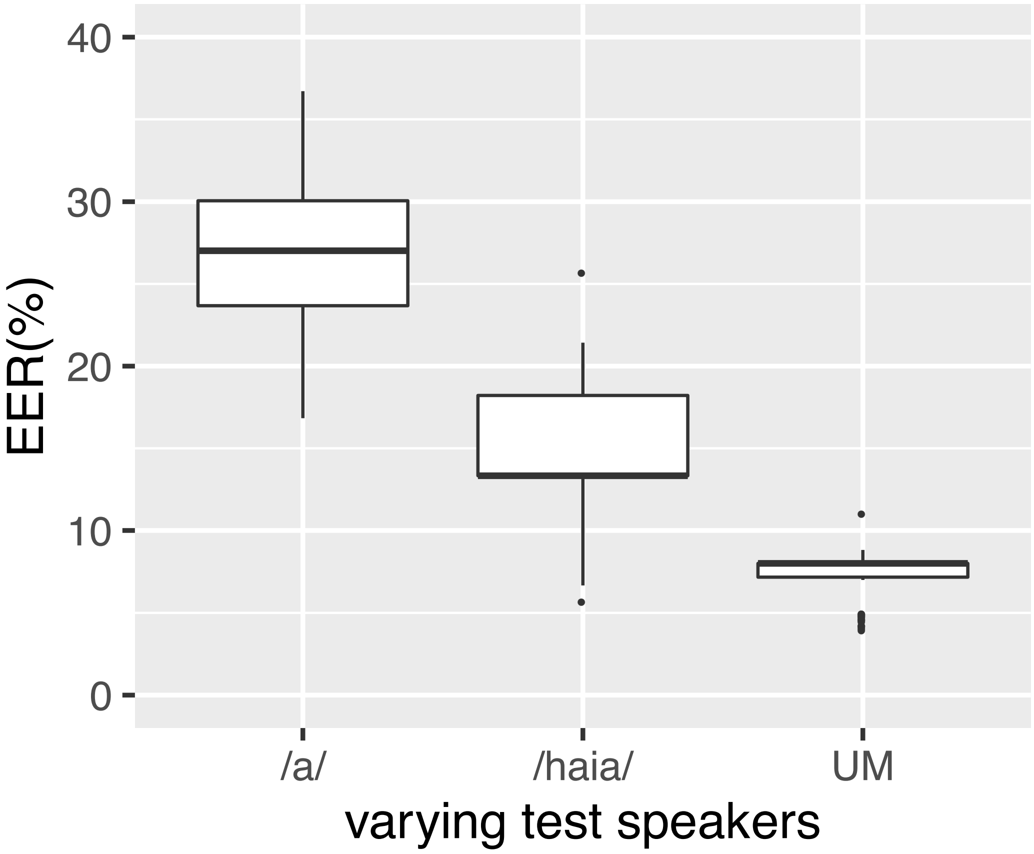


Figure 4.4 Boxplots represent the EER ranges of /a/, /haia/ and *um* by varying test speakers in each replication. Two whiskers cover the first and third quartiles, and the central line represents the median value.

	C_{llr}			**EER (%)**		
	/a/	**/haia/**	***um***	**/a/**	**/haia/**	***um***
Min	0.58	0.31	0.24	16.84	5.71	4.00
1st qu.	0.70	0.44	0.34	23.68	13.33	7.92
Median	0.73	0.49	0.37	27.01	13.33	8.00
Mean	0.74	0.49	0.37	27.53	14.55	8.34
3rd qu.	0.78	0.52	0.40	30.06	18.21	8.58
Max.	0.86	0.64	0.48	36.72	25.71	12.00
IQR	0.08	0.08	0.06	6.38	4.88	0.67
Range	0.28	0.33	0.24	19.88	20.00	8.00

Table 4.5. Summary statistics of C_{llr} and EER of /a/, /haia/ and *um* in Experiment 2 (varyting test speakers).

4.4.3 Experiment 3: varying reference speakers

In Experiment 3, only reference speakers were varied, while the training and test speakers were fixed throughout 100 replications. Figure 4.5 shows the boxplots of C_{llr} values of /a/, /haia/ and *um* in 100 replications, and Table 4.6 shows the statistics of C_{llr} and EER values across 100 replications. /a/ yielded a very small C_{llr} range (0.11), varying between 0.66 and 0.77, and the IQR is as low as 0.03. /haia/, on the other hand, is more variable in terms of C_{llr} values. The C_{llr} values of /haia/ range from 0.40 to 0.95. The C_{llr} IQR of /haia/ is low at 0.08, but this is nearly three times higher than /a/. However, over 75% of the replications in /haia/ achieved a better validity than that in /a/, as over 75% C_{llr} values of /haia/ are lower than all those of /a/. The FP *um* yielded a better overall performance than /a/ and /haia/ in terms of validity and reliability. The C_{llr} values vary from 0.18 to 0.25, and the C_{llr} IQR is 0.02, which is the lowest among the three variables.

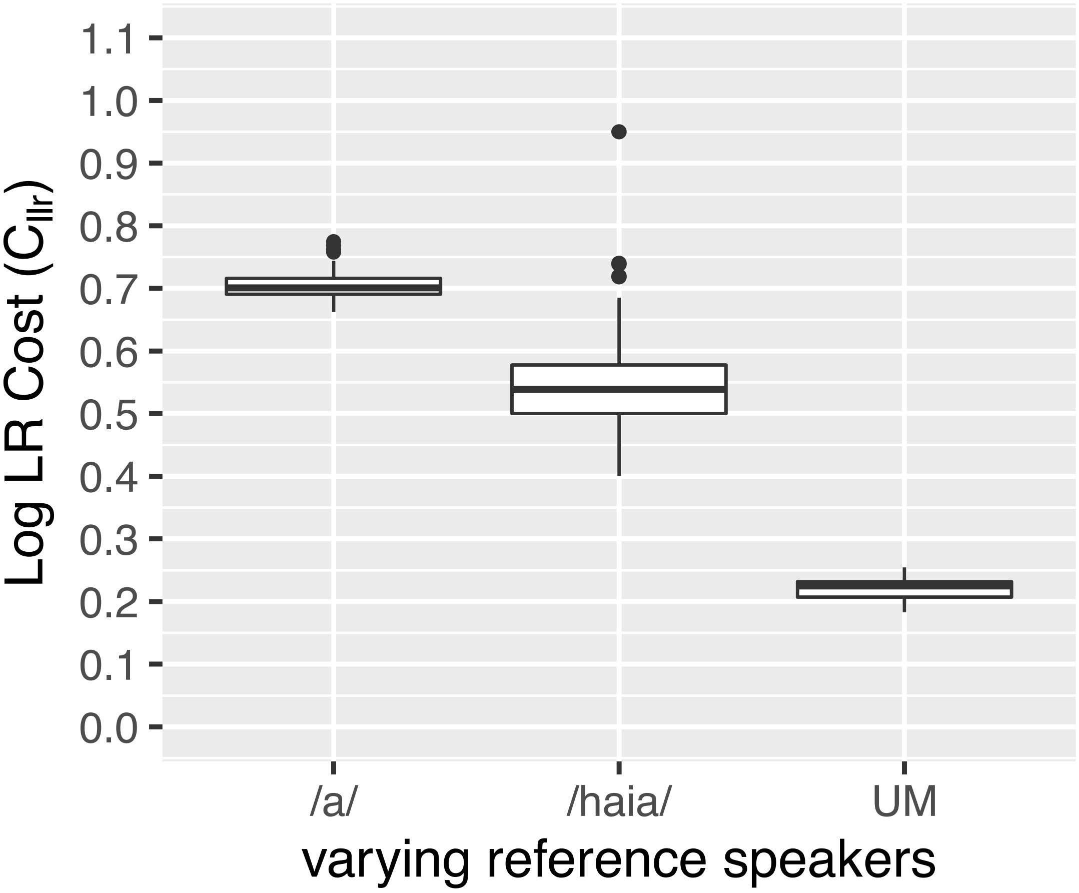


Figure 4.5 Boxplots represent the C_{llr} ranges of /a/, /haia/ and *um* by varying reference speakers in each replication. Two whiskers cover the first and third quartiles, and the central line represents the median value.

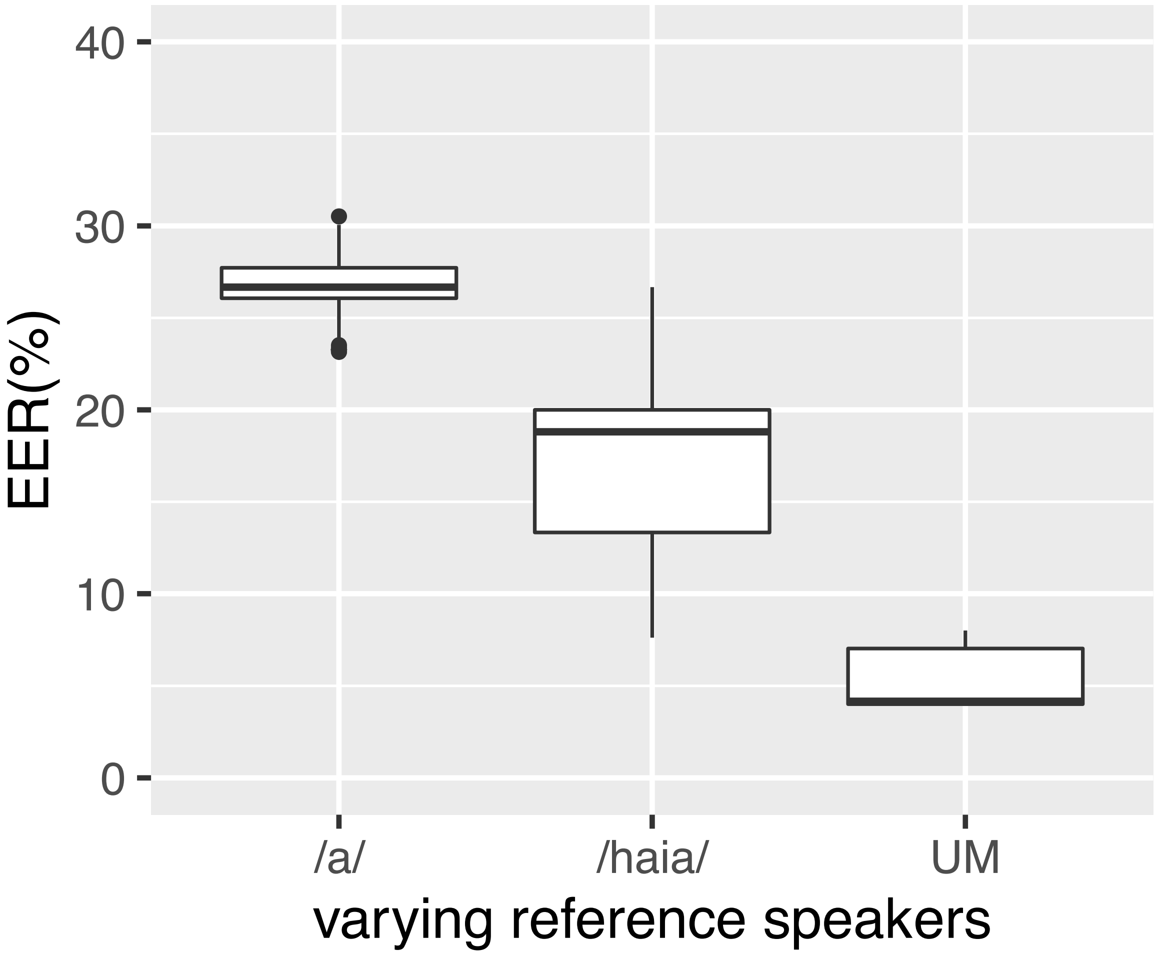


Figure 4.6 Boxplots represent the EER ranges of /a/, /haia/ and *um* by varying reference speakers in each replication. Two whiskers are the first and third quartiles, and the central line represents the median value.

Figure 4.6 shows the boxplots of EER values in the 100 replications using /a/, /haia/ and *um*. The EER values ranges from 23.16% to 30.52%, 7.62% to 26.67% and 3.92% to 8% for /a/, /haia/ and *um* respectively. FP *um* outperformed SFP /a/ and /haia/ in terms of the lowest EER value (3.92% vs 23.16% and 7.62%). Moreover, most EER values in FP *um* are lower than those in SFP /a/ and /haia/.

	C_{llr}			**EER (%)**		
	/a/	**/haia/**	***um***	**/a/**	**/haia/**	***um***
Min	0.66	0.40	0.18	23.16	7.62	3.92
1st qu.	0.69	0.50	0.21	26.06	13.33	4.00
Median	0.70	0.54	0.23	26.67	18.81	4.17
Mean	0.70	0.55	0.22	26.77	17.25	5.15
3rd qu.	0.72	0.58	0.23	27.72	20.00	7.02
Max.	0.77	0.95	0.25	30.52	26.67	8.00
IQR	0.03	0.08	0.02	1.66	6.67	3.02
Range	0.11	0.55	0.07	7.36	19.05	4.08

Table 4.6. Summary statistics of C_{llr}s and EERs of /a/, /haia/ and *um* in Experiment 3 (varying reference speakers).

Experiment 3 shows that using different reference speakers from a matched dialect group causes the system reliability to fluctuate to different extents. The FP *um* yields a better system reliability than SFP /a/, while /haia/ shows the importance of linguistically fine-grained relevant population as *um* is extracted from a much more controlled dataset than the Cantonese SFP. Moreover, the results from Experiment 3 also show that a narrower tailored relevant population not only contributes to system validity, but also reliability.

4.4.4 Experiment 4: varying training speakers

In Experiment 4, test and reference speakers were fixed throughout the 100 replications, while only training speakers were varied in each replication. Experiment 4 shows that sampling training speakers has a limited effect on system reliability for all three variables. Figure 4.7 and Table 4.7 show the C_{llr} values of /a/, /haia/ and FP *um* in 100 replications. All three variables yielded a very stable overall performance in terms of C_{llr} IQR values, which are 0.01, 0.03 and 0.005 for /a/, /haia/ and *um* respectively. The C_{llr} range of SFP /a/ is the lowest among the three variables, which is only 0.12. Meanwhile, /haia/ and the FP *um* yielded more outliers, making the C_{llr} range vary from 0.43 to 0.72 and from 0.27 to 0.53 respectively. Even /a/ yielded the best system reliability, all of the C_{llr} values from each replication of /haia/ and FP *um* are lower than those of SFP /a/. The EER of the three variables are not presented here as calibration has no effect on EER values.

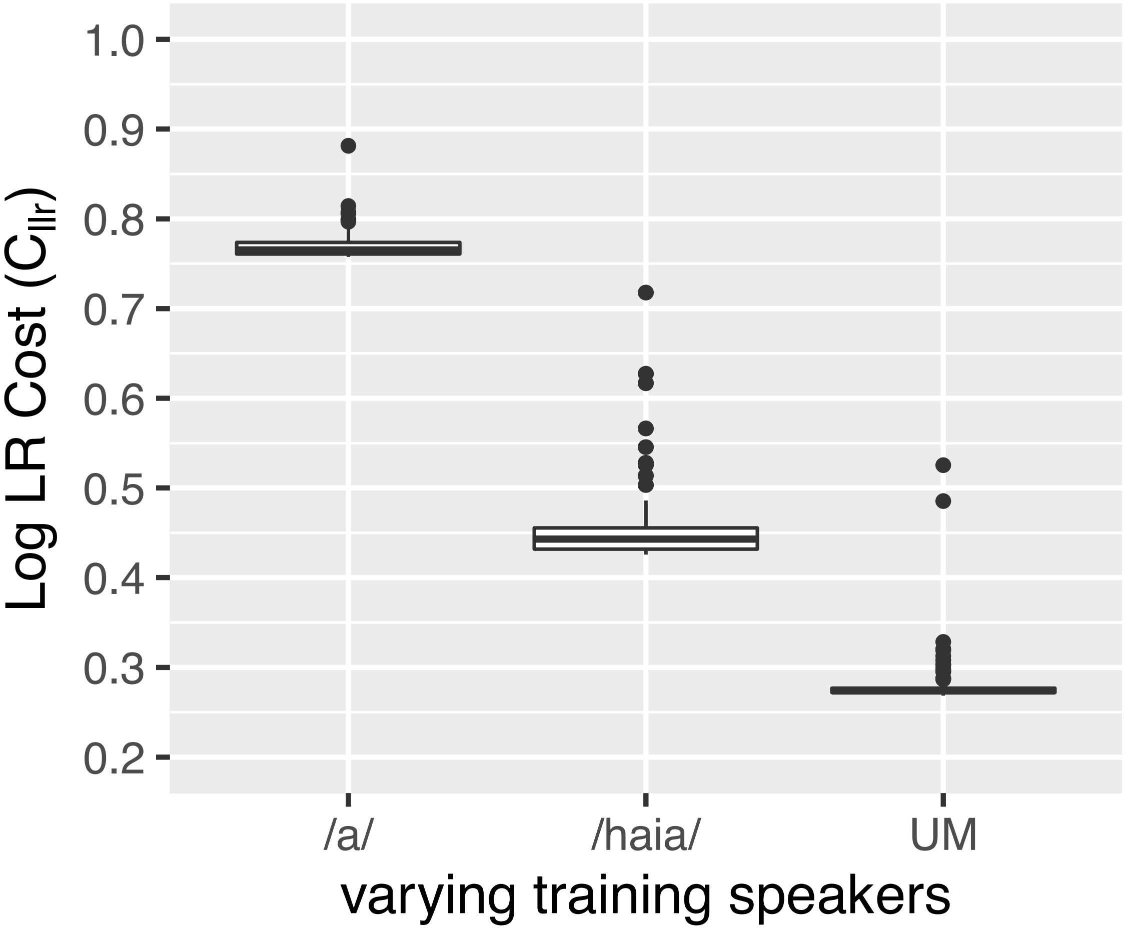


Figure 4.7 Boxplots represent the C_{llr} ranges of /a/, /haia/ and *um* by varying training speakers in each replication. Two whiskers are the first and third quartiles, and the central line represents the median value.

	C_{llr}		
	/a/	**/haia/**	***um***
Min	0.76	0.43	0.27
1st qu.	0.76	0.43	0.27
Median	0.77	0.44	0.27
Mean	0.77	0.46	0.28
3rd qu.	0.77	0.46	0.28
Max.	0.88	0.72	0.53
IQR	0.01	0.03	0.01
Range	0.12	0.29	0.26

Table 4.7. Summary statistics of C_{llr} of /a/, /haia/ and *um* in Experiment 4 (varying training speakers).

4.5 Discussion

There were some differences in the results for the variables examined in this chapter. Cantonese /a/ was less sensitive to speaker sampling than *um*, reflected in a narrower range of C_{llr} values across the four experiments. One reason for this may be the inherent speaker-discriminatory power of /a/ compared with *um*. The median C_{llr} value for *um* in Experiment 1 was 0.46, compared with 0.74 for Cantonese /a/. The lower end of the distribution of C_{llr} values for *um* also shows that it has the potential to produce very good performance with certain configurations of speakers. This shows that inherently better speaker discriminants will produce more variable overall performance because they have the potential to produce a wider range of results depending on which speakers are being used (principally, in the test set). However, poorer speaker discriminants, such as /a/, will produce poor overall performance irrespective of the speakers used. This relationship between speaker discriminatory power and sensitivity to speaker sampling replicates findings reported in Wang et al. (2019) based on simulated data.

The following sections discuss the results from the four experiments with regard to the research questions and compared to previous studies. Figure 4.8 shows the boxplots of C_{llr} and EER of three variables from all four experiments.

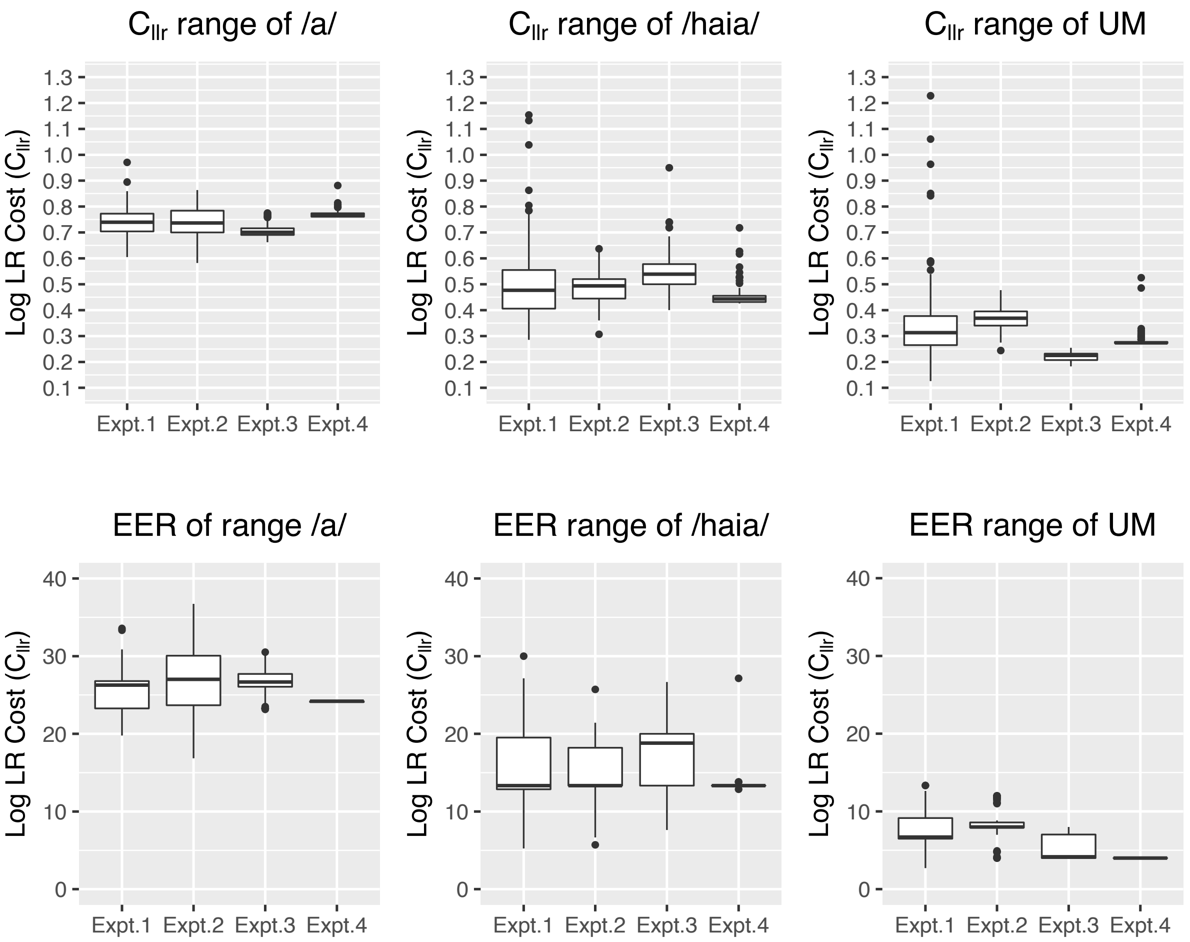


Figure 4.8 Boxplots of *C*llr and EER of /a/, /haia/ and *um* in four experiments. Two whiskers are the first and third quartiles, and the central line represents the median value.

4.5.1 The performance reliability of segmental variables

In previous data-driven LR-based FVC studies using linguistic-phonetic variables, most of the studies have placed large focus on obtaining better validity (i.e., a lower *C*llr) as well as aiming to compare the inherent speaker-discriminatory power of linguistic-phonetic variables, i.e., some linguistic-phonetic variables have better speaker-discriminatory power than others. For example, several Cantonese variables such as /ɔy/, /iau/, /ei/, /i/ and /daijat/ have been tested in previous studies (Chen & Rose, 2012; Li & Rose, 2012; Pang & Rose, 2012; Rose & Wang, 2016; Wang & Rose, 2012). The best system validity was achieved in Rose and Wang (2016) using the Cantonese disyllabic word /daijat/ 'first'. They used non-contemporaneous recordings from 23 male Hong Kong Cantonese speakers. The first three formants and tonal

F0 were fitted by using polynomial curves and the coefficients were used as the input of an MVKD model. A C_{llr} of 0.16 was achieved using the formant dynamics of /daijat/. The fusion of /daijat/ formant dynamics and tonal information reduced the C_{llr} to 0.1, and they fused /daijat/ with the Cantonese vowel /i/ which further reduced the C_{llr} to 0.03. Similarly, Zhang et al. (2011) compared the performance of the acoustic-phonetic system with the automatic system using a Mandarin triphthong /iau/. They used non-contemporaneous recordings of 60 northeast male Mandarin speakers. The first, second and last sets of 20 speakers were used as background, training and test speakers respectively. In the acoustic-phonetic system, the discrete cosine transform (DCT) was fitted to the first three formants of each token and the zeroth to fourth DCT coefficients of F2 and F3 were used as the input of the MVKD model. In the automatic system, the 16 mel-frequency-cepstral-coefficient (MFCC) values were taken every 10-milliseconds over the entire speech-active portion of the recordings, and a 20-millisecond wide Hamming window was used. A Gaussian mixture model-universal background model (GMM-UBM) was used for the background data and 1024 Gaussians were used for the training and test data. A C_{llr} of 0.349 and 0.029 was achieved by using the acoustic-phonetic and automatic system respectively, and it was further reduced to 0.009 by fusing the two systems. In terms of English variables, the FPs (*uh, um*) are assumed to be useful FVC variables by previous studies (e.g., Jessen 2008, Foulkes et al 2004) and the speaker discrimination power of FPs has been explored in several studies (e.g., Hughes et al., 2016; King, 2012; Wood et al., 2014). All of these three previous studies used contemporaneous speech data from DyViS (Nolan et al., 2009), but with different numbers of speakers involved. The C_{llr} varied from 0.506 in King (2012) using 20 speakers and static formant measurements to 0.12 using 60 speakers and formant dynamics in Hughes et al..

All previous studies mentioned in the previous paragraph aimed to obtain a lower C_{llr}; however, the current chapter shows that linguistic-phonetic variables yield promising system validity under one system setting might give poor validity under another system setting, i.e., depending on the configurations of speakers in training, test and reference data. Taking Experiment 1 for example, the best C_{llr} and EER achieved for /a/, /haia/ and *um* were 0.6 and 19.77%, 0.29 and 5.24% and 0.13 and 2.7%, while the worst were 0.97 and 33.56%, 1.15 and 30% and 1.22 and 13.33%. It is then questionable whether previous studies would achieve consistent or similar C_{llr} values were the experiments replicated. For example, would Rose and Wang (2016) obtain similar C_{llr} values were more speakers involved and the experiment replicated? Would Zhang et al. (2011) and Hughes et al. (2016) achieve consistent C_{llr} values if the speakers were rearranged (e.g., use the first 20 speakers as training, second 20 speakers as test and the last 20 speakers as reference speakers), and speaker sampling was carried out? Based on the variability in C_{llr} values reported in this chapter, it is suggested that researchers' aim in system testing using linguistic-phonetic variables should not be driven by obtaining better validity, but better reliability or reducing the uncertainty, i.e., a system producing reliable performance under various conditions (e.g., different number or/and configurations of speakers) should be preferred, other than a system that has the potential of obtaining a very low C_{llr} under one condition but high C_{llr} values under other conditions.

4.5.2 Test speakers *vs.* training speakers *vs.* reference speakers

Experiments 2 (varying test speakers), 3 (varying reference speakers) and 4 (varying training speakers) aimed to investigate which dataset is more susceptible to different speaker configurations (sampling variability) that leads to fluctuating overall performance. The EERs of the three experiments are not discussed here, because test and reference speakers are fixed in Experiment 4 and changing the training speakers does not affect the EERs. The results show that varying only training or reference speakers both have less effect on overall performance than varying only test speakers. However, the overall performance fluctuates among the three linguistic variables. On one hand, all three variables have the lowest C_{llr} IQR when only training speakers are varied (Table 4.8), indicating that overall performance is less sensitive to different configurations of training speakers. On the other hand, for SFP /a/ and FP *um*, varying only reference speakers in general gives lower C_{llr} mean and range than varying only training or test speakers, while it is the opposite for /haia/. This pattern probably indicates that overall performance is more susceptible to reference speakers when more acoustic information is involved (i.e., /haia/ is a disyllabic word and /a/ and *um* are monophthongs) and when the number of reference speakers used is low (only 15 speakers were used for /haia/, while 30 and 25 speakers were used for /a/ and *um* respectively). It is noted that FP *um* yields better performance than /a/ even though FP *um* has fewer reference speakers. This is probably due to the fact that the FP *um* reference data was extracted from a narrower tailored relevant population.

It is worth noting that although the results show that varying only training or reference speakers have less effect on overall performance than varying only test speakers, the results from Experiments 2, 3 and 4 are to some extent dependent on the speakers that are fixed in the test, reference, and training sets. That is to say, the results here should not be treated as any sort of generalisation, and different results could well be seen if different speakers were included in the fixed speaker groups or different number of speakers and different linguistic-phonetic features are used.

	/a/_C_{llr}		
	Varying test	Varying training	Varying reference
Mean	0.74	0.77	0.7
IQR	0.08	0.01	0.03
Range	0.28	0.12	0.11
	/haia/_ C_{llr}		
	Varying test	Varying training	Varying reference
Mean	0.49	0.46	0.55
IQR	0.08	0.03	0.08
Range	0.33	0.29	0.55
	***um*_C_{llr}**		
	Varying test	Varying training	Varying reference
Mean	0.37	0.28	0.22
IQR	0.06	0.01	0.02
Range	0.24	0.26	0.07

Table 4.8 Mean, IQR and range of C_{llr} values using /a/, /haia/ and *um* in Experiments 2 (varying test speakers), 3 (varying reference speakers) and 4 (varying training speakers).

4.5.3 System vs individual

The FP *um* yielded the lowest C_{llr} mean and range in Experiment 3 (only varying reference speakers); however, C_{llr} measures the overall performance and does not necessarily provide insights on the fluctuation in individual speakers' LLRs. Figure 4.9 shows the SS (upper panel) and DS (lower panel) LLRs of test speakers involved in Experiment 3. The x- and y-axis represent speaker numbers and LLRs respectively, and each boxplot indicates the C_{llr} range of individual speakers across 100 replications. The verbal expression scale (Champod & Evett, 2000, p.240) is used here for reference.

For SS comparisons, most of them generally vary within one LLR magnitude in terms of strength of evidence, i.e., SS LLR varies between 0 and 1 or 1 and 2. However, speaker 44 shows a different pattern. Among the 100 SS comparison replications, only one replication yielded the expected factual result (i.e., a positive LLR, in this case LLR = 0.00131), while the rest of the replications all yielded contrary-to-fact results indicating that speaker 44 is susceptible to different configurations of reference speakers. For DS comparisons, speakers seem to be more variable. Most speakers yielded contrary-to-fact results with *limited support* for SS origin (e.g., speaker 1, 9, 10) while a few speakers (e.g., 35, 42) yielded contrary-to-fact results that reached *moderate support* for SS origin (i.e., LLR between 1 and 2) given that it is DS comparison. Meanwhile, most speakers have factual DS LLRs that vary from 0 to over -5, which is a difference between *limited* to *strong* support in terms of strength of evidence. The patterns of individual speakers' LLRs show that individual speakers could be behaving differently in systems with different configurations of reference speakers.

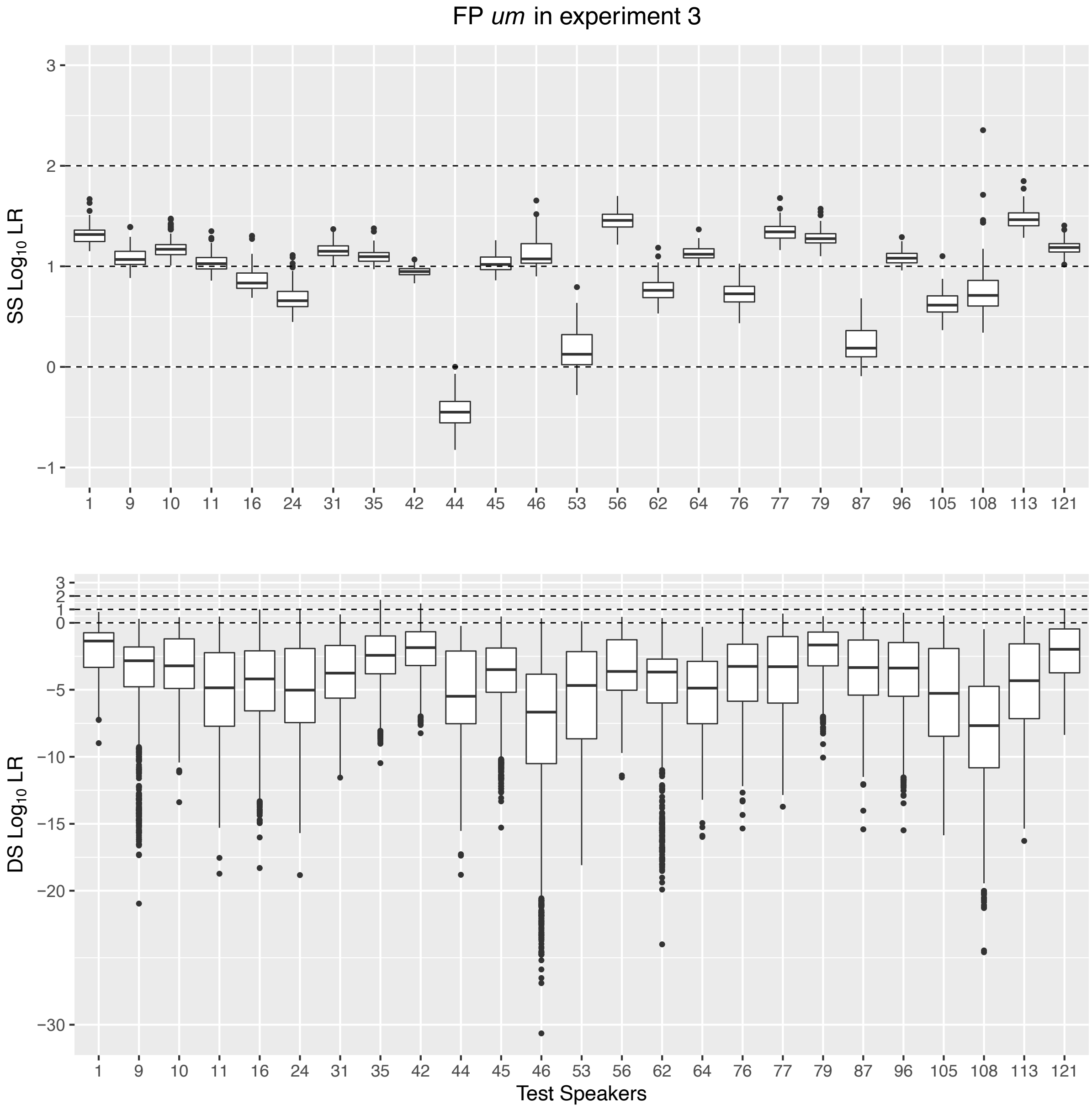


Figure 4.9 SS and DS LLRs of test speakers using FP *um* in Experiment 3 (varying reference speakers).

4.6 Chapter summary

Chapter 4 has investigated the effect of sampling variability on overall performance and which of the training, test and reference datasets is less susceptible to sampling variability. The results show that sampling variability indeed has a marked effect on overall performance and key results and patterns from Chapter 4 are given below.

- Overall performance is affected not only by the number of speakers, but also the specific choice of speakers used in each of the training, test and reference sets. Meanwhile, the overall performance also depends on the quality of the data extraction (i.e., no F3 for Cantonese variables) and the polynomial curve fitting, as well as the suitability of the statistical modelling in MVKD.
- The variability in overall performance is mainly due to different configurations of speakers in the test set.
- It is likely that different configurations of speakers in the training and reference sets have little effect on overall performance when 25 or more speakers are used; however, this pattern needs to be tested using a larger database.
- The inherent speaker-discriminatory power of linguistic-phonetic variables might affect the stability of results, e.g., FP *um* with better speaker-discriminatory power (than SFP /a/) has more variable overall performance. However, more tests should be conducted by altering the number of speakers more systematically.

Chapter 5 Overall performance and individual speakers' behaviour as functions of choice of linguistic-phonetic features and configurations of training and reference speakers

The previous chapter investigated the validity and reliability of LR output of *overall* performance and individual speakers' behaviour. Section 4.4 considers variation in the overall performance depending on *who* exactly is used in the training, test and reference data. Meanwhile, given a comparatively reliable system, the LR output of individual speakers can vary to different extents due to sampling variability. In the current Chapter, the effect of sampling variability on individual speakers' behaviour is investigated in a more systematic manner as well as the overall performance. Only the English FP *um* is used in the current chapter as it allows greater control over other potentially confounding variables, such as regional accents, age and social class before the sampling variability is introduced, and DyViS is a more controlled corpus in terms of both social and linguistic factors than the IARPA Babel Cantonese corpus.

5.1 Introduction

As discussed in Section 2.2.1, numerous empirical studies have explored the validity and reliability of the LR output focusing more on the *overall* system validity and reliability (Hughes et al., 2016; Morrison et al., 2011; Morrison & Enzinger, 2016; Taylor et al., 2016). By contrast, few have explored the behaviour of individual speakers within these systems (but see e.g., Lo, 2021). Given that what really matters in a real case is the specific voice of the speaker(s) under analysis, this raises issues about whether generic testing provides adequate insight into the validity and reliability of an LR system in a specific case, given the reasonable limits of sample size in the real world. The current chapter explores the relationship between overall performance and individual speaker behaviour in relation to sampling variability in the training and reference speakers and the choice of linguistic features.

There are two principal aims. First, to explore the effect of sampling variability on overall performance when different features are used, e.g., single-feature systems using only F1, F2 or F3; multi- feature systems using the combination of F1 and F2 or F2 and F3. Second, to explore the effect of sampling variability on individual speakers' behaviour, i.e., how individual speakers' LR output is affected when different configurations of training and reference speakers are used. The specific RQs addressed in this chapter are RQs 2a., 2b. and 2c.

RQs 2.

a. Do certain combinations of linguistic-phonetic features outperform the others and are they less susceptible to sampling variability?
b. How is individual speakers' LR output affected when different configurations of training and reference speakers are used?
c. How is individual speaker' LR output affected when different linguistic-phonetic features are used?

5.2 Method

5.2.1 Data

The FP *um* from 90 SSBE speakers was used as the variable for analysis in the current chapter. The quadratic coefficients of the first three formants and F0 as well as the duration of the vocalic and nasal portion were used as the input for LR computation (Section 3.2.3). In order to explore overall performance and individual speaker behaviour with respect to different conditions, experiments were carried out using all 31 possible combinations of features. i.e., 5 single-feature systems, 10 two-feature systems, 10 three-feature systems, 5 four-feature systems and 1 five-feature system (Table 5.1).

SYSTEM	F0	F1	F2	F3	DUR.	NUMBER OF FEATURES
F0	X					1
F1		X				1
F2			X			1
F3				X		1
DUR					X	1
F01	X	X				2
F02	X		X			2
F03	X			X		2
F0DUR	X				X	2
F12		X	X			2
F13		X		X		2
F1DUR		X			X	2
F23			X	X		2
F2DUR			X		X	2
F3DUR				X	X	2
F012	X	X	X			3
F013	X	X		X		3
F01DUR	X	X			X	3
F023	X		X	X		3
F02DUR	X		X		X	3
F03DUR	X			X	X	3
F13DUR		X		X	X	3
F123		X	X	X		3
F12DUR		X	X		X	3
F23DUR			X	X	X	3
F0123	X	X	X	X		4
F012DUR	X	X	X		X	4
F013DUR	X	X		X	X	4
F023DUR	X		X	X	X	4
F123DUR		X	X	X	X	4
F0123DUR	X	X	X	X	X	5

Table 5.1 31 systems tested and cross (X) indicates the feature(s) used. *DUR* stands for durations of the vocalic and nasal of the FP *um*.

5.2.2 Experimental procedure

25 speakers were randomly sampled to act as the test, training and reference data respectively. Task 1 and Task 2 were used as the suspect and offender samples respectively. The SS and DS pairs of test and training data were compared using the MVKD formula (Aitken & Lucy, 2004) to produce test and training scores, and calibration was then carried out using logistic regression (Brümmer et al., 2007; Morrison, 2011). Experiments were again replicated 100 times for each system using different configurations of training and reference speakers, but keeping the 25-speaker test set fixed. This gives access to assess the LR results for the same test speakers using different input features and different configurations of training and reference speakers. 50 out of the remaining 65 speakers were used in each replication to allow for different configurations of training and reference speakers. A schematic diagram of the experimental procedure is given in Figure 5.1.

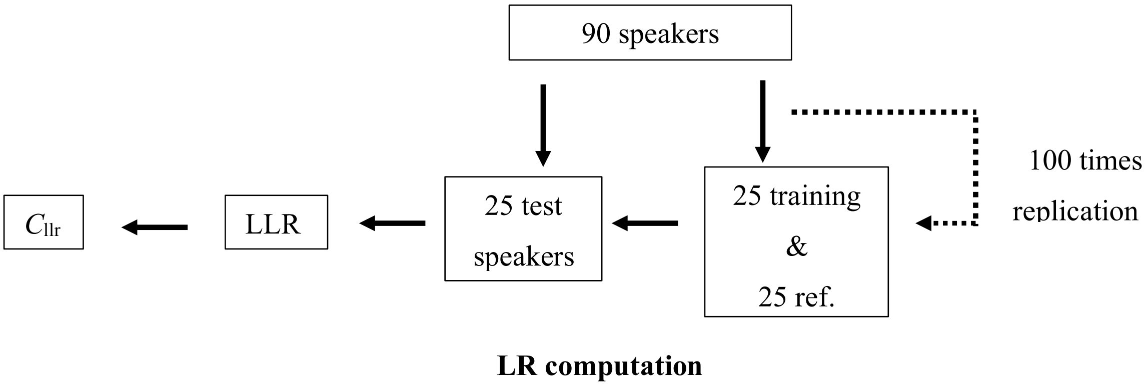


Figure 5.1 A schematic diagram of the experiment procedure for LR computation. Only training and reference speakers were sampled 100 times, while the 25 test speakers were sampled once. LR computation was conducted using different sets of training and reference speakers and the same set of test speakers.

Overall performance was evaluated using C_{llr} (Brümmer & du Preez, 2006), and the mean and range (i.e., difference between the maximum and minimum C_{llr} values across 100 replications) of C_{llr} values were used to assess the system validity and reliability respectively. The individual speakers' behaviour was assessed using mean LLR and RMSD values with reference to the LR verbal scale (Champod & Evett, 2000).

5.3 Results

The results of overall performance and individual speaker's behaviour are presented in this section. The validity and reliability of overall performance with single and different combinations of features are presented first, followed by the variability of LR outputs for individual speakers.

5.3.1 Overall performance

The boxplots in Figure 5.2 show the variation of C_{llr} values across the 31 systems, and Table 5.2 shows the statistical summary of C_{llr} in systems with one and two features. The x-axis shows the different systems. For example, 'F0' refers to the system where only F0 was used as input, while 'F01' indicates the combination of F0 and F1. The y-axis represents C_{llr} values. The top panel shows the C_{llr} range of systems with one and two features, and the bottom panel shows the C_{llr} range of systems with three, four and five features. For single feature systems, F2 yields the lowest C_{llr} mean (0.39) and range (0.04), which is consistent with previous studies (Hughes et al., 2016). The other four single feature systems yield similar overall performance, with the mean C_{llr} varying between 0.66 and 0.77. F1 yields the least stable overall performance with an overall C_{llr} range of 0.47, while F0, F3, and DUR systems yield a C_{llr} overall range between 0.12 and 0.18. The performance among single feature systems indicate that F2 is the least sensitive to different configurations of training and reference speakers.

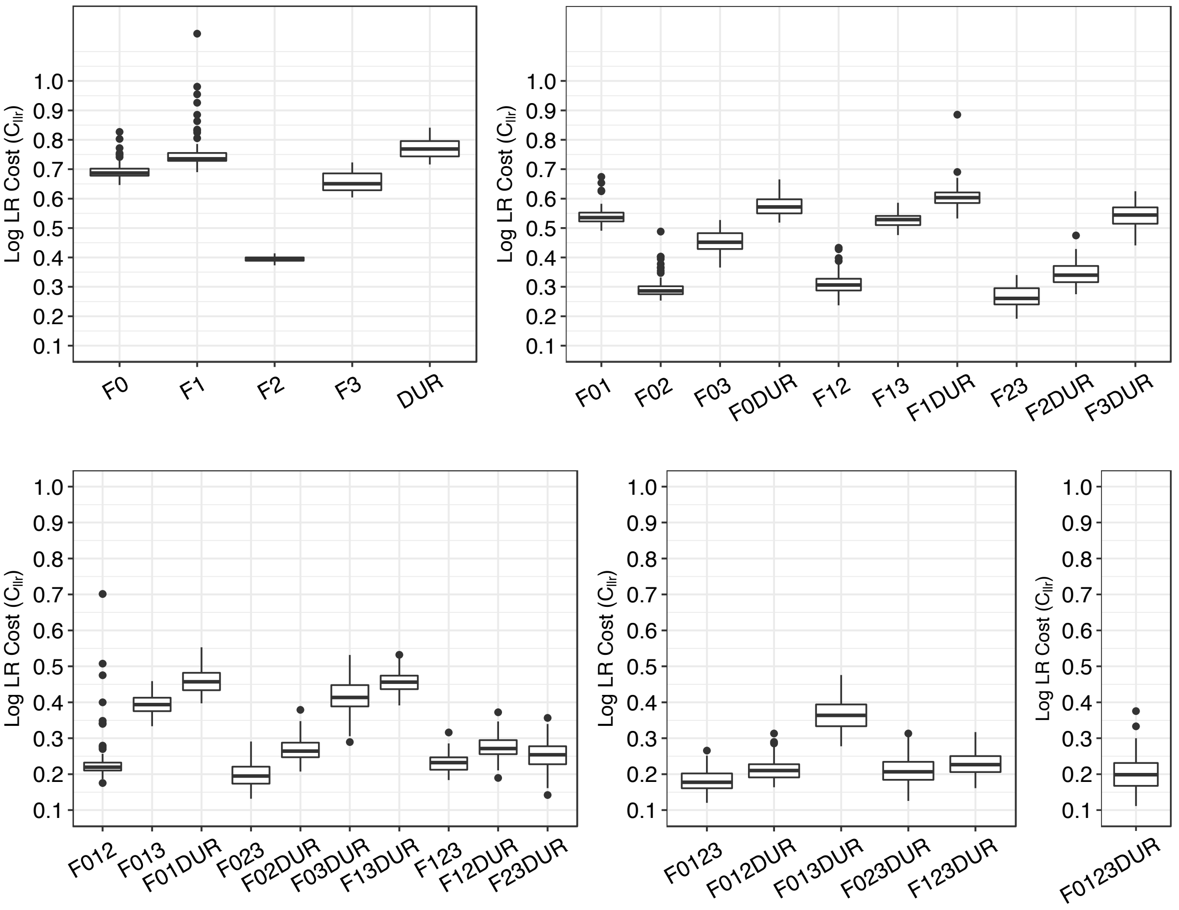


Figure 5.2 C_{llr} variation across 31 systems (Top panel: systems with one or two features; bottom left: systems with three features; bottom middle: systems with four features; bottom right: system with five features).

The C_{llr} patterns are more variable among systems with two features. The F23 system yields the lowest C_{llr} mean (0.27), while the F0DUR system gives the highest mean C_{llr} (0.64). In terms of system reliability, the F13 system yields the lowest C_{llr} range (0.11), and the F1DUR system yields the highest (0.35). A consistent pattern in two-feature systems is that systems with F2 involved outperform systems without F2 in terms of mean C_{llr}. In terms of the overall performance (i.e., lowest C_{llr} mean and range), the F23 system seems to be the best. It can be observed that the F13 system has a lower C_{llr} range than the F23 system; however, all of the C_{llr} values in the F23 system are lower than those in the F13 system and the C_{llr} range of the F23 system is only marginally higher than that of the F13 system.

System\\C_{llr}	mean	Max.	Min.	IQR	Range
F0	0.69	0.83	0.65	0.02	0.18
F1	0.75	1.16	0.69	0.03	0.47
F2	0.39	0.41	0.37	0.01	0.04
F3	0.66	0.72	0.60	0.06	0.12
DUR	0.77	0.84	0.72	0.05	0.13
F01	0.54	0.67	0.49	0.03	0.18
F02	0.29	0.49	0.25	0.03	0.23
F03	0.45	0.53	0.37	0.05	0.16
F0DUR	0.58	0.67	0.52	0.05	0.15
F12	0.31	0.43	0.24	0.04	0.20
F13	0.53	0.59	0.48	0.03	0.11
F1DUR	0.61	0.89	0.53	0.04	0.35
F23	0.27	0.34	0.19	0.06	0.15
F2DUR	0.34	0.47	0.27	0.06	0.20
F3DUR	0.54	0.62	0.44	0.06	0.18

Table 5.2 Summary statistics of C_{llr} in systems with one and two features.

Table 5.3 summarises the results for systems with three, four and five features. Among systems with three features, the F023 system yields the lowest mean C_{llr} (0.20), while the F01DUR and F13DUR systems yield the highest (0.46). Similar to systems with single and two features, systems with F2 involved again outperform those without F2 in terms of mean C_{llr}, e.g., the F012, F123, and F023 systems have lower mean C_{llr} values than the F013 and F01DUR systems. In terms of C_{llr} range, the F013 and F123 systems yield the lowest C_{llr} range (0.13), while the F012 system yields the highest due to extreme outliers (0.53). The systems with the duration feature, i.e., F01DUR, F12DUR, F23DUR, F02DUR, F03DUR and F13DUR yield similar C_{llr} range varying from 0.16 to 0.24. Overall, the F023 system has a marginally lower mean C_{llr} and higher C_{llr} range than the F123 system, and these two systems seem to have similar overall performance and are less sensitive to different configurations of training and reference speakers than other systems.

For systems with four and five features, the F013DUR system gives the highest mean C_{llr} (0.37), while other systems seem to have similar mean C_{llr} values. Combining all features does not improve the overall performance, as the mean C_{llr} (0.2) of the F0123DUR system is slightly higher than that of the F0123 (0.18) system, and the C_{llr} range of the F0123DUR system (0.26) is higher than all other systems with four features.

System\ C_{llr}	**mean**	**Max.**	**Min.**	**IQR**	**Range**
F012	0.24	0.70	0.18	0.02	0.53
F013	0.39	0.46	0.33	0.04	0.13
F01DUR	0.46	0.55	0.40	0.05	0.16
F123	0.23	0.32	0.18	0.03	0.13
F12DUR	0.27	0.37	0.19	0.04	0.18
F23DUR	0.25	0.36	0.14	0.05	0.21
F023	0.20	0.29	0.13	0.05	0.16
F02DUR	0.27	0.38	0.21	0.04	0.17
F03DUR	0.41	0.53	0.29	0.06	0.24
F13DUR	0.46	0.53	0.39	0.04	0.14
F0123	0.18	0.27	0.12	0.04	0.15
F012DUR	0.21	0.31	0.16	0.04	0.15
F013DUR	0.37	0.48	0.28	0.06	0.20
F023DUR	0.20	0.31	0.13	0.05	0.19
F123DUR	0.23	0.32	0.16	0.04	0.16
F0123DUR	0.20	0.38	0.11	0.06	0.26

Table 5.3. Summary statistics of C_{llr} in systems with three, four and five features.

Figure 5.3 shows the average system validity (C_{llr} mean; upper panel) and reliability (C_{llr} range; lower panel) across 31 systems. The x-axis shows the corresponding system and the y-axis indicates the C_{llr} mean and range. The systems were plotted from the highest (left end of the x-axis) to lowest (right end of the x-axis) based on C_{llr} mean and range. Figure 5.3 (upper panel) shows that the system validity improves when more features are involved, and it starts to stabilise when four or more features are used. However, exceptions can be found, for example, the F2 system yield a lower mean C_{llr} than other two- and three-feature systems (e.g., F01, F13DUR). Meanwhile, systems with F2 have a better validity than systems without F2.

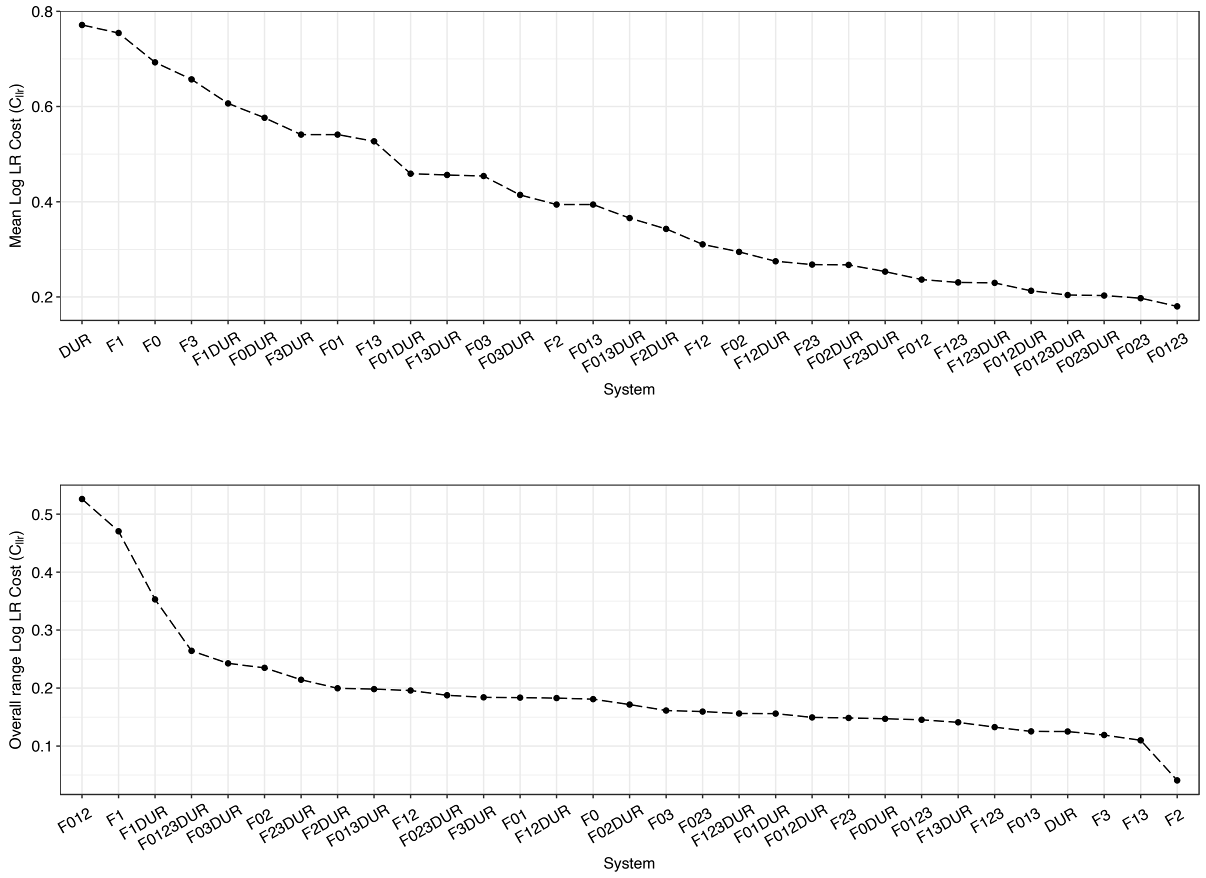


Figure 5.3 C_{llr} mean (upper panel) and range (lower panel) across 31 systems. X-axis shows different systems and y-axis shows C_{llr} range (i.e., the difference between maximum and minimum C_{llr} values across 100 replications in each system).

For system reliability (Figure 5.3 lower panel), there does not appear to be any general relationship between the system reliability and number of features used, i.e., single or different combinations of features are affected by sampling variability to different extents and systems with more features are not necessarily more sensitive to sampling variabilities. For example, systems with four features, e.g., the F0123, F012DUR, F123DUR, have lower overall C_{llr} ranges than some of the systems with single, two- or three-feature systems such as the F1, F1DUR and F012 systems. On the other hand, some other systems with fewer features, e.g., the DUR, F3 and F2 systems, yield lower overall C_{llr} ranges than systems with more features such as the F012, F1DUR and F0123DUR systems.

Figure 5.4 shows the relationship between C_{llr} mean (x-axis) and C_{llr} range (y-axis) of the 31 systems. In general, systems with more features have a tendency to move further to the bottom left corner (i.e., have better overall performance). However, this is not always the case; for example, the system using F2 alone outperforms two-feature (e.g., F03, F13) and three-feature systems (e.g., F03DUR, F13DUR) in terms of both validity and reliability; two-feature systems (F02, F12, F23) outperform three- and four-feature systems (e.g., F03DUR, F013DUR) in terms of validity (x-axis).

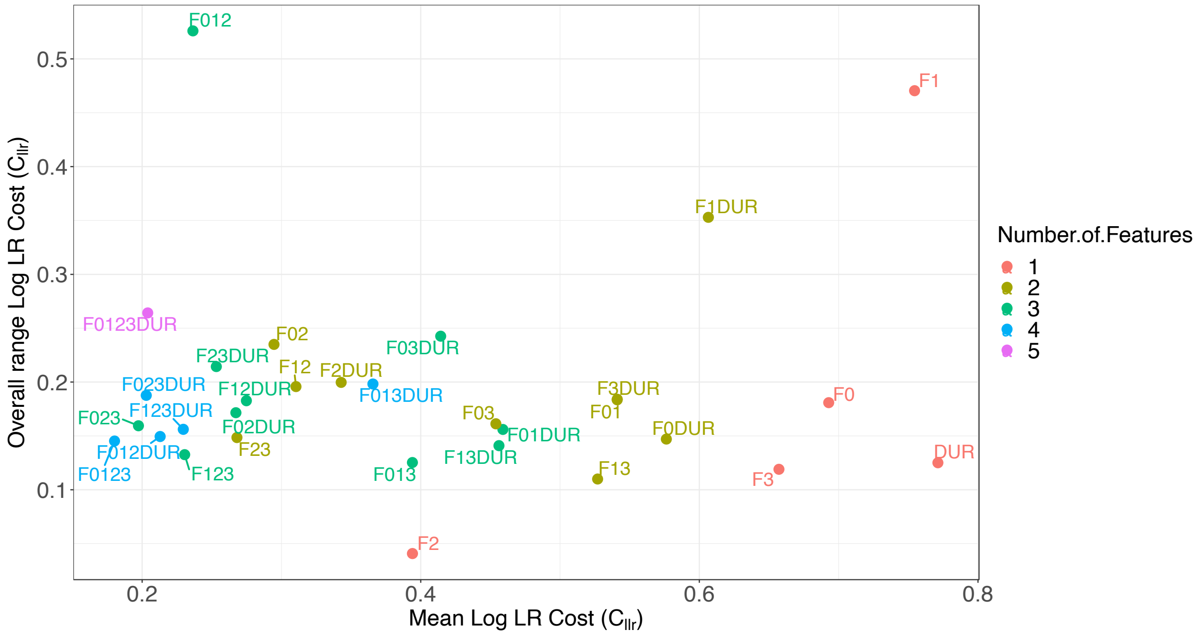


Figure 5.4 C_{llr} mean plotted against C_{llr} range of 31 systems.

5.3.2 Individual behaviour

5.3.2.1 Individual validity

Figure 5.6 shows the mean SS and DS LLR of all 25 test speakers across 31 systems and 100 replications. The x- and y-axis represent the SS and DS LLRs, and the legend indicates the number of features used in the system. The vertical and horizontal lines indicate SS/DS LLR equal to 0. Speakers that are easy to be separated from others and themselves should have crosses, circles and triangles clustering at top right, while speakers that are difficult to be separated from others and themselves should have those symbols clustering at bottom left. Most of the speakers yielded a similar pattern to speaker #51 (Figure 5.5), showing that speakers are more likely to yield more accurate behaviour in systems with more features, i.e., the crosses, black circles and black triangles tend to move to the top right corner as more features are included. Meanwhile, the majority of speakers produced consistent-with-fact results in systems with three or more features. However, exceptions can be found in speakers

#48 and #53 (Figure 5.5). Systems with four or five features do not seem to improve the mean validity of the LRs of speakers #48 and #53. Instead, #48 and #53 can be well separated using a two-feature system (i.e., F02, indicated by floating label F02 next to the triangle). The maximum magnitude of LLRs exceeds 2.3 for SS comparisons, equivalent to *moderately strong support* for H_p, while the DS LLR is over -20, equivalent to *very strong support* for H_d. Moreover, speaker #48 produced contrary-to-fact results in SS comparisons in the F013 (SS LLR = -0.38) and F13DUR (SS LLR = -0.02) systems and speaker #53 produced contrary-to-fact results in the F13DUR system (SS LLR = -0.42).

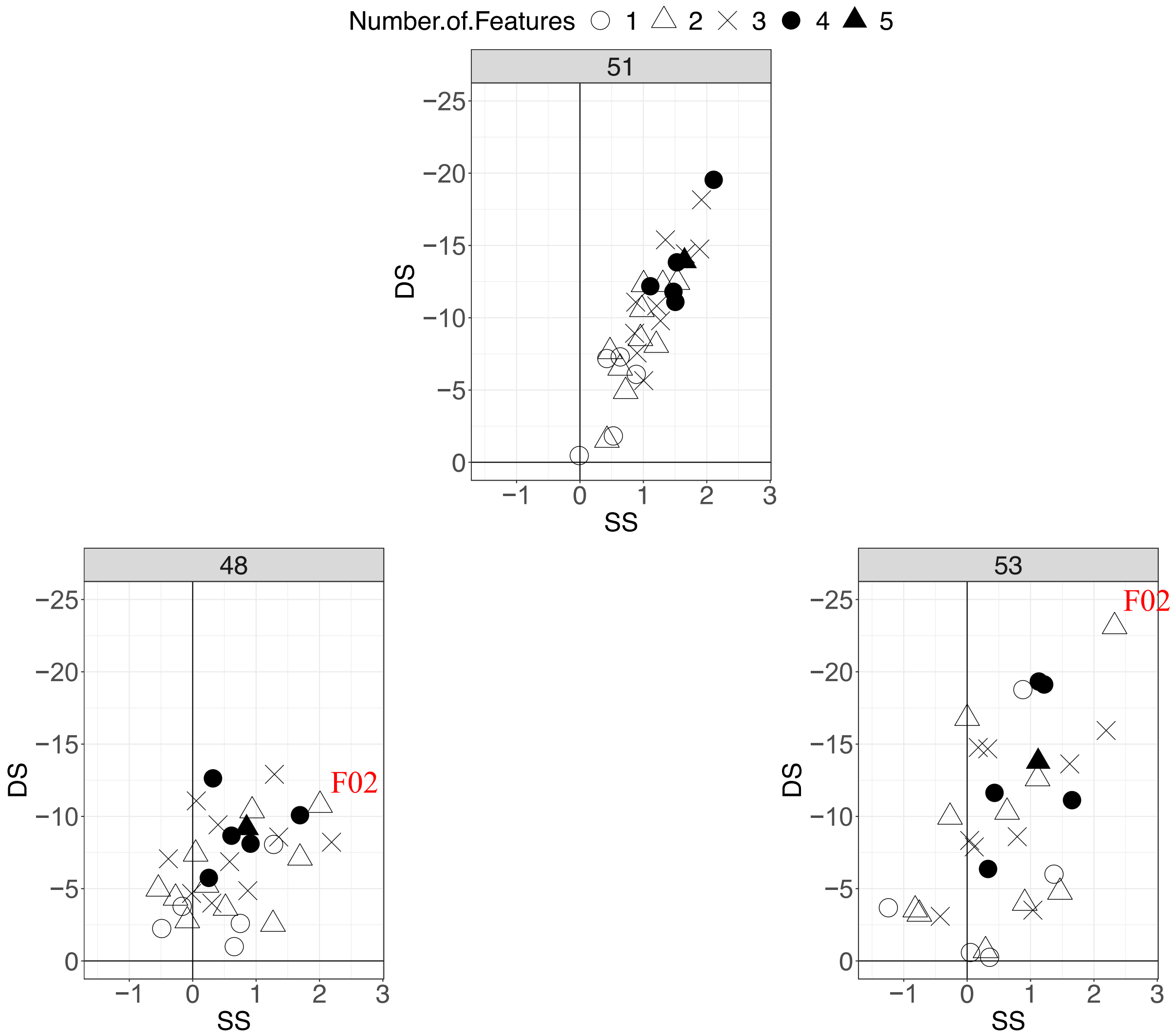


Figure 5.5 Mean SS and DS LLR of speakers #51, #48 and #53 across 31 systems. Two triangles marked by ‘F02’ in lower panels indicate that speakers #48 and #53 can be well separated using a two-feature system (F02).

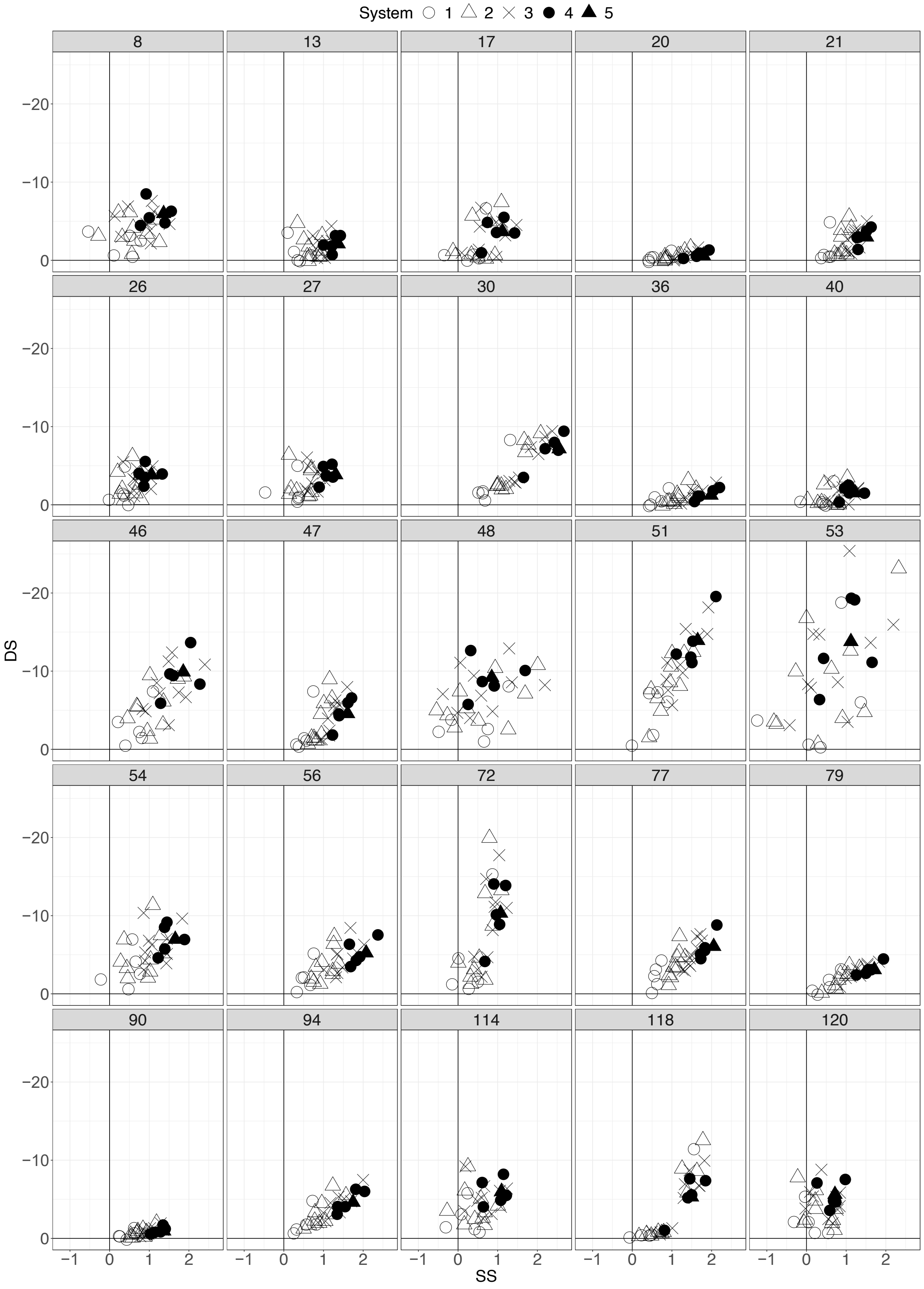


Figure 5.6 Mean SS and DS LLR of all 25 test speakers across 31 systems (see Appendix C for full size).

Figures 5.7 to 5.10 give a more detailed portrait of individual speakers' LLRs across different systems, showing the zoo plots of the 25 test speakers. Each black point represents the mean of SS and DS LLRs of each individual speaker in the 100 replications across different systems and the neighbouring numbers indicate speaker numbers. The animal group, i.e., *phantom, worm, dove* and *chameleon*, of individual test speakers is assigned on the basis of mean SS and DS LLRs, and the red dash lines indicate the adjusted thresholds for the zoo plots (Table 3.3).

Figure 5.7 shows that most of the 25 test speakers yielded consistent-with-fact results using single-feature systems, and Table 5.4 shows animal groups of speaker numbers in single-feature systems. In terms of speaker's animal groups, there are only five speakers (i.e., speakers #30, #46, #48, #53, #118) who fall into the *dove* group (i.e., speakers who have especially good performance when being compared with themselves and can be easily separated from other speakers) in the F2 system, while no speakers fall into the *dove* group in the other four single-feature systems. On the other hand, there are *phantom* speakers (i.e., speakers who are rarely similar to any other speakers and likely to yield low SS and DS scores) in all single-feature systems (Table 5.4), and there are no *chameleon* (i.e., speakers who are always obtaining high SS and low DS scores when being compared against themselves and other speakers) or *worm* (i.e., speakers who always have high DS and low SS comparison scores when being compared against themselves and other speakers) speakers. Figure 5.7 shows that individual speakers' behaviour varies depending on which feature is being used, e.g., speaker 48 falls into the *phantom* group in the F1 and F3 systems, while speaker #48 is in the *dove* group in the F2 system. Moreover, individuals seem to have more within- and between-speaker variation in the F2 system than others as the SS and DS LLRs are more spread out.

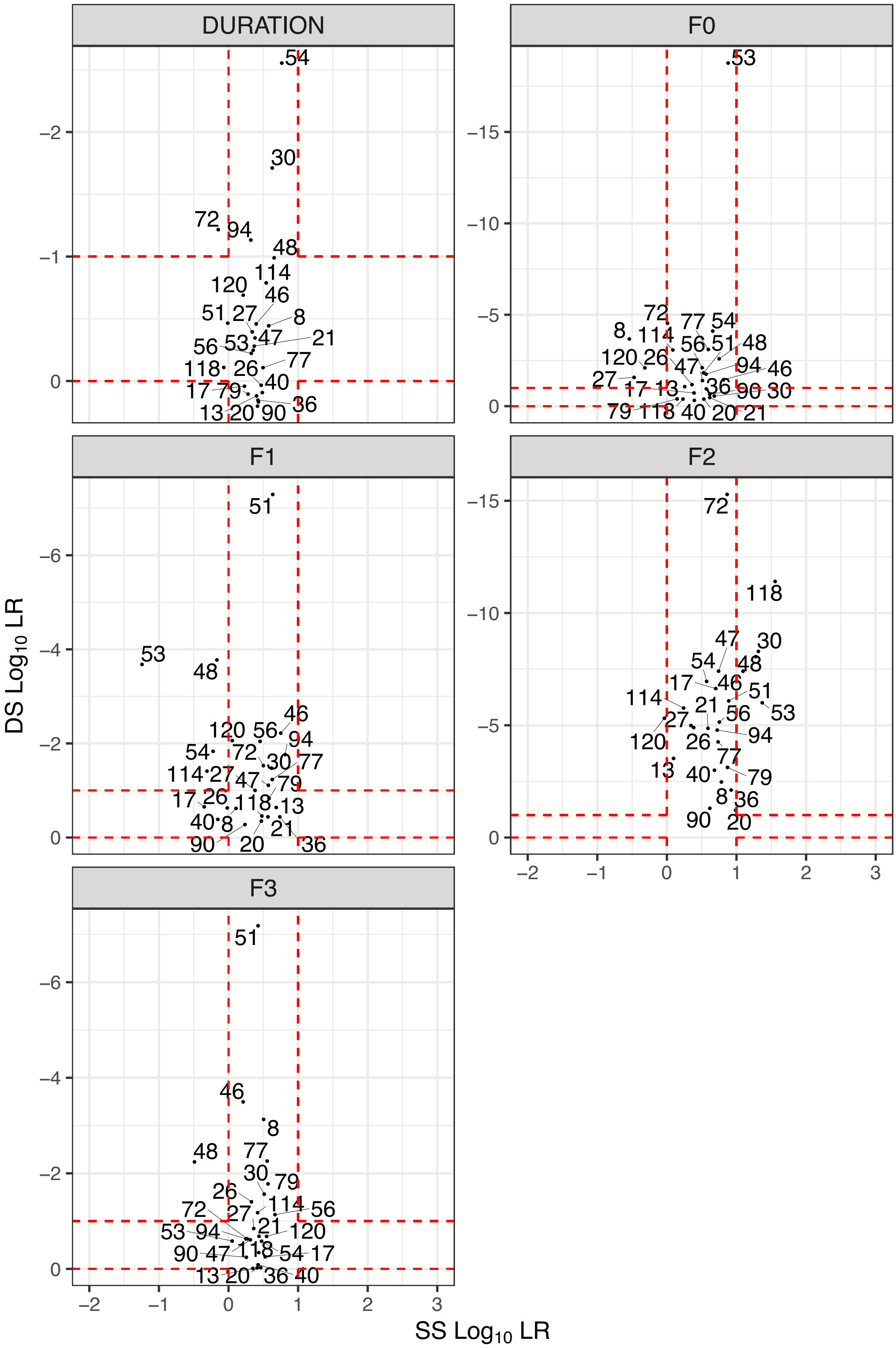


Figure 5.7 Zoo plot of SS and DS LLR of 25 test speakers in single-feature systems. Note that y-axes are different (see Appendix B for full size zoo plots).

Systems\Animal group	Phantoms	Doves	Worms	Chameleons
DURATION	72	N.A.	N.A.	N.A.
F0	8, 27, 120	N.A.	N.A.	N.A.
F1	48, 53, 54, 114	N.A.	N.A.	N.A.
F2	120	30, 46, 48, 53, 118	N.A.	N.A.
F3	48	N.A.	N.A.	N.A.

Table 5.4 Animal groups of speakers in single-feature systems. N.A. indicates that there is no speaker occur in that group.

In two-feature systems, as with single-feature systems, most speakers yielded consistent-with-fact results in SS and DS comparisons and speakers seem to shift between animal groups depending on different combinations of features (Figure 5.8). Table 5.5 shows animal groups of speaker numbers in two-feature systems. Individual speakers are more likely to fall into the *dove* group and less likely to fall into the *phantom* group in systems with F2 than systems without F2 as most speakers (18 out of 25) fall into the *dove* group in the F02 system, and there is no *dove* speaker in the F3DUR system. Meanwhile, apart from the F23 and F2DUR systems, all other systems have *phantom* speakers. There is no speaker who consistently falls into the same animal group across all systems. The nearest such speakers are speaker #30, who falls into the *dove* group in all systems but the F1DUR and F3DUR systems, and speakers #36 and #118, who only appear in the *dove* group in systems with F2, i.e., the F02, F12, F23 and F2DUR systems. In terms of the *phantoms,* there seem to be no systematic pattern either. Speaker #48 falls in the *phantom* group in the F03, F13 and F3DUR systems, but shifts to the *dove* group in the F02, F0DUR and F2DUR systems. Similarly, speakers #8, #17, #53 and #114 are in the *phantoms* in one system, but *doves* in another. However, speaker #120 seems to be the only speaker who occurs in the *phantom* group in the F02, F0DUR and F1DUR systems without shifting to other animal groups in other systems.

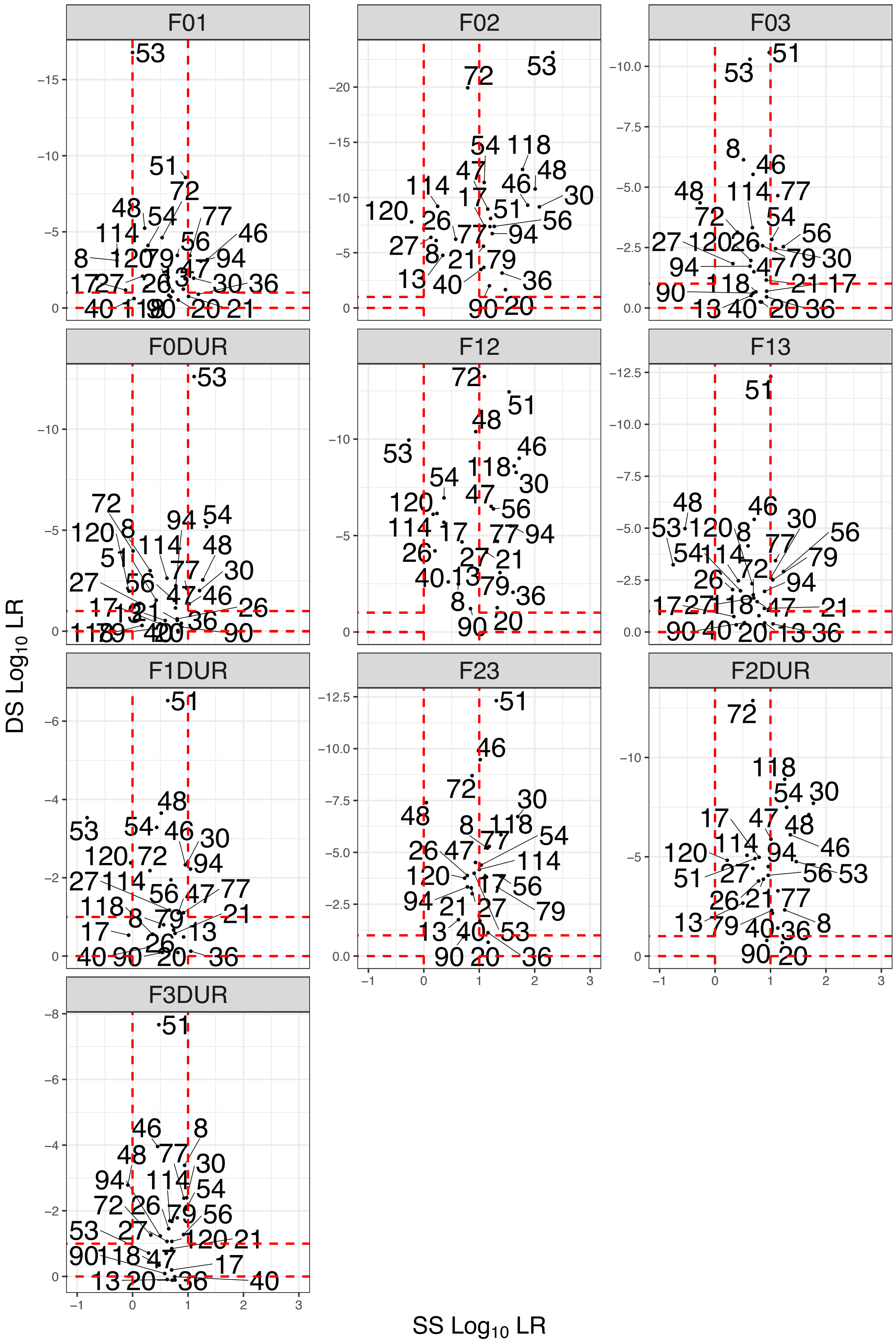


Figure 5.8 Zoo plot of SS and DS LLR of 25 test speakers in two-feature systems. Note that y-axes are different (see Appendix B for full size zoo plots).

Systems\Animal group	Phantoms	Doves	Worms	Chameleons
F01	8, 17, 53, 114	30, 46, 77, 94	N.A.	N.A.
F02	120	17, 20, 21, 30, 36, 40, 46, 47, 48, 51, 53, 54, 56, 77, 79, 90, 94, 118	N.A.	N.A.
F03	48	30, 54, 56, 77	N.A.	N.A.
F0DUR	120	30, 46, 48, 53, 54	N.A.	N.A.
F12	53	20, 21, 30, 36, 46, 47, 51, 56, 72, 77, 79, 94, 118	N.A.	N.A.
F13	48, 53	30, 51, 56, 77, 79	N.A.	N.A.
F1DUR	53, 120	94	N.A.	N.A.
F23	N.A.	8,17, 30, 36, 40, 46, 51, 54, 56, 77, 79, 114, 118	N.A.	N.A.
F2DUR	N.A.	8, 30, 36, 40, 46, 47, 48, 53, 54, 77, 79, 118	N.A.	N.A.
F3DUR	48	N.A.	N.A.	N.A.

Table 5.5 Animal groups of speakers in two-feature systems. N.A. indicates that there is no speaker occur in that group.

Figure 5.9 shows individuals' mean validity across systems with three features, and Table 5.6 shows animal groups of speaker numbers in three-feature systems. Most speakers shift to the *dove* group (top right corner) in three-feature systems comparing with single and two-feature systems. Despite that, *phantoms* speakers are observed in the following systems: F013 (speaker #48), F01DUR (speaker #120) and F13DUR (speakers #48, #53). Furthermore, a *chameleon* speaker (speaker #20) is observed in the F03DUR system. All other systems, i.e., the F012, F023, F02DUR, F123, F12DUR, F23DUR, yield factual LLRs. Most speakers fall into the *dove* group in the F123 system (19 speakers). Meanwhile, three speakers – #30, #56, #77 – consistently fall into the *dove* group in all three-feature systems. Similar to two-feature systems, speaker #48 shifts between *phantom* and *dove* in different three-feature systems, while speaker #120 does not fall into the *dove* group in any three-feature systems.

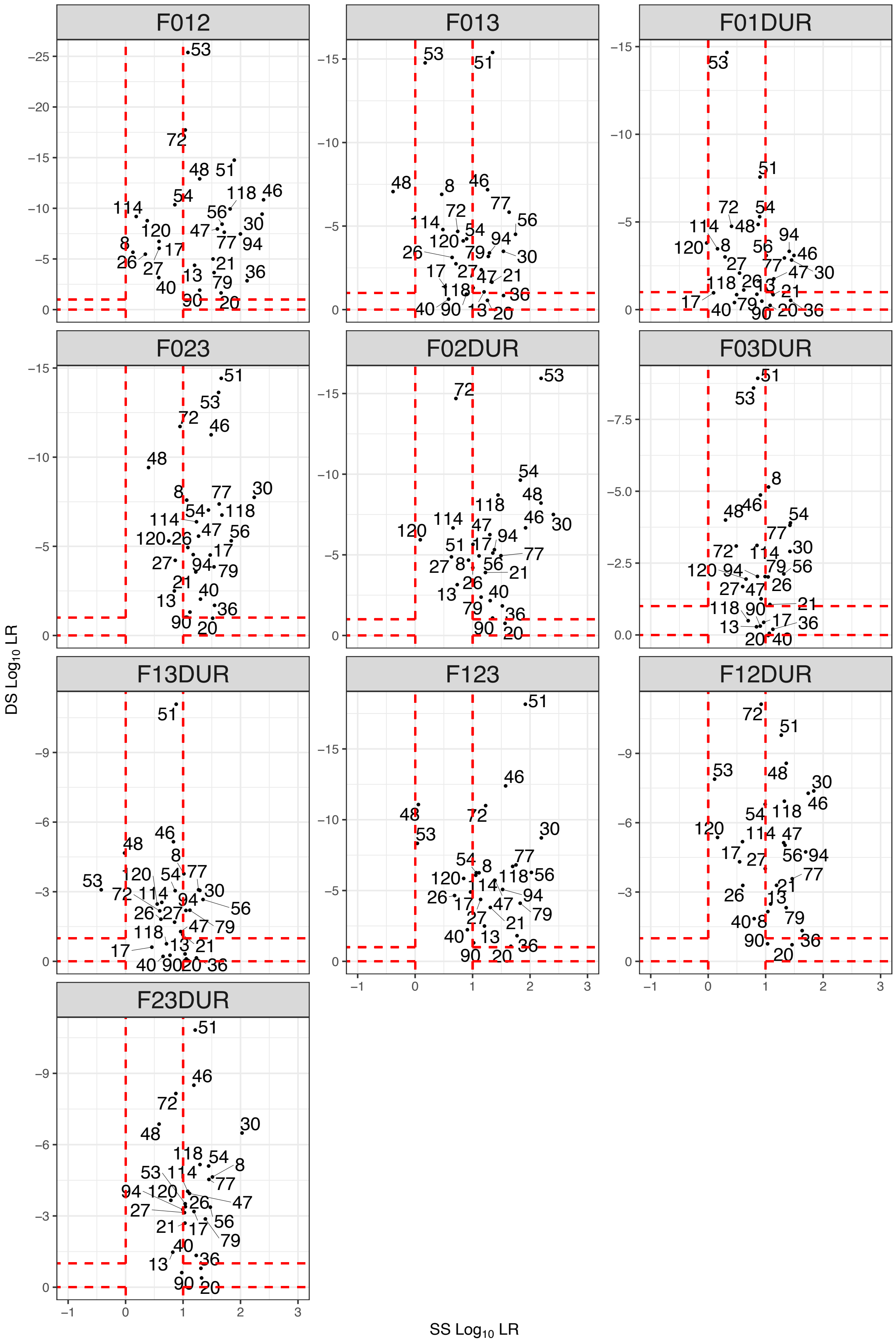


Figure 5.9 Zoo plot of SS and DS LLR of 25 test speakers in three-feature systems. Note that y-axes are different (see Appendix B for full size zoo plots).

Systems\Animal group	Phantoms	Doves	Worms	Chameleons
F012	N.A.	13, 20, 21, 30, 36, 46, 47, 48, 51, 53, 56, 72, 77, 79, 90, 94,118	N.A.	N.A.
F013	48	13, 21, 30, 46, 47, 51, 56, 77, 79, 94, 118	N.A.	N.A.
F01DUR	120	30, 46, 47, 56, 77, 94	N.A.	N.A.
F023	N.A.	8, 17, 21, 26, 30, 36, 40, 46, 47, 51, 53, 54, 56, 77, 79, 90, 94, 114, 118	N.A.	N.A.
F02DUR	N.A.	17, 21, 26, 30, 36, 40, 46, 47, 48, 51, 53, 54, 56, 77, 79, 90, 94, 118	N.A.	N.A.
F03DUR	N.A.	8, 21, 26, 30, 54, 56, 77	N.A.	20
F13DUR	48, 53	8, 30, 56, 77, 79, 94	N.A.	N.A.
F123	N.A.	13, 20, 21, 27, 30, 36, 46, 47, 51, 54, 56, 72, 77, 79, 8, 90, 94, 114, 118	N.A.	N.A.
F12DUR	N.A.	8, 13, 21, 30, 36, 46, 47, 48, 51, 56, 77, 79, 94, 118	N.A.	N.A.
F23DUR	N.A.	8, 17, 21, 26, 27, 30, 40, 46, 47, 51, 53, 54, 56, 77, 79, 114, 118	N.A.	N.A.

Table 5.6 Animal groups of speakers in three-feature systems. N.A. indicates that there is no speaker occur in that group.

Figure 5.10 shows individuals' mean validity across systems with four or five features, and Table 5.7 shows animal groups of speaker numbers in four-/five-feature systems. All speakers yielded consistent-with-fact results in SS and DS comparisons in systems with four or more features, as no speaker falls into the *phantom, worm* or *chameleon* group. The F0123 and

F0123DUR systems yield the same number of *dove* speakers (21 speakers), indicating that adding the duration feature does not necessarily increase the mean magnitude of the strength of evidence in individual speakers. Moreover, the F0123 system outperforms the other four-feature systems in terms of number of *dove* speakers, suggesting that individual speakers are in general more likely to yield stronger strength of evidence in systems without duration. However, exceptions can be found in speakers #48 and #120, as speaker #48 only falls into the *dove* group in the F012DUR system, and speaker #120 does not fall into the *dove* group in any systems. Despite the fact that some speakers, e.g., speakers #8, #17, #20, shift in and out of the *dove* group across different systems, 10 out of 25 speakers fall into the *dove* group consistently in all systems with four or more features, i.e., speakers #21, #30, #46, #47, #51, #54, #56, #77, #79 and #94. This shows that different combinations of features and configurations of training and reference speakers have less effect on an individual speaker's mean validity when four or more features are used.

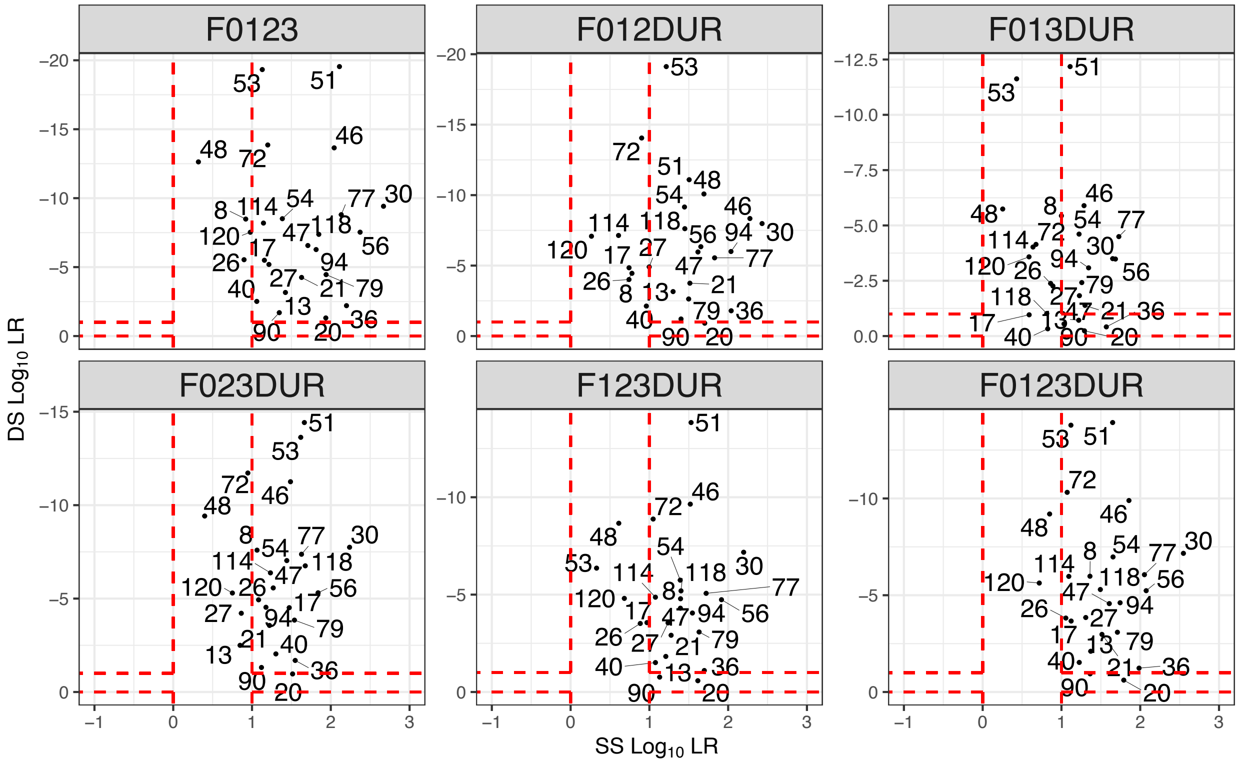


Figure 5.10 Zoo plot of SS and DS LLR of 25 test speakers in four-/five-feature systems. Note that y-axes are different (see Appendix B for full size zoo plots).

Systems\Animal group	Phantoms	Doves	Worms	Chameleons
F0123	N.A.	13, 17, 20, 21, 27, 30, 36, 40, 46, 47, 51, 53, 54, 56, 72, 77, 79, 90, 94, 114, 118	N.A.	N.A.
F012DUR	N.A.	13, 21, 30, 36, 46, 47, 48, 51, 53, 54, 56, 77, 79, 90, 94, 118	N.A.	N.A.
F013DUR	N.A.	21, 30, 46, 47, 51, 54, 56, 77, 79, 94	N.A.	N.A.
F023DUR	N.A.	8, 17, 21, 26, 30, 36, 40, 46, 47, 51, 53, 54, 56, 77, 79, 90, 94, 114, 118	N.A.	N.A.
F123DUR	N.A.	8, 13, 21, 27, 30, 36, 40, 46, 47, 51, 54, 56, 72, 77, 79,94, 114, 118	N.A.	N.A.
F0123DUR	N.A.	8, 13, 17, 21, 26, 27, 30, 36, 40, 46, 47, 51, 53, 54, 56, 72, 77, 79, 94, 114, 118	N.A.	N.A.

Table 5.7 Animal groups of speakers in four-/five-feature systems. N.A. indicates that there is no speaker occur in that group.

Across the 31 systems, there is only one speaker who falls into the *chameleon* group in the F03DUR group, while there is no *worm* speaker in any system. Among systems with the same number of features, individual speakers' mean validity varies depending on which feature, or features, are being used, i.e., the results are system specific. Across the 31 systems, the general trend shows that speakers are more likely to fall into the *dove* group (top-right, i.e., absolute LLR ≥ 1) when more features are used. Different combinations of features and configurations of training and reference speakers seem to have less effect on individual speakers' behaviour when four or more features are used.

5.3.2.2 Individual reliability

Figure 5.11 shows the RMSD values in SS (upper panel) and DS (lower panel) comparisons across the 31 systems based on different combinations of input features, indicating the reliability in individual speakers' LLRs within systems. The black dots represent the RMSD values of each individual speaker, and the coloured triangles are the mean RMSD values of all speakers in each system. The RMSD values were plotted from the highest (left end of the x-axis) to lowest (right end of the x-axis) in terms of the mean RMSD values. Higher RMSD values indicate that speakers have more fluctuating LLRs relative to their own mean LLR for that system, and are more sensitive to configurations of training and reference speakers. Overall, all speakers tend to be more fluctuating in DS comparisons than SS comparisons (note that the y-axis scales are different). This is likely due to the fact that speakers can only be so similar to themselves, but infinitely different from each other (Kinoshita et al., 2009). For SS comparisons, speakers fluctuate more when more features are included in the system. However, the fluctuation in individual speakers' SS LLR is minor, with all SS RMSD values smaller than 1 across the 31 systems.

For DS comparisons, increasing the number of features does not necessarily lower the reliability in individual speakers' mean LLR outputs, e.g., the mean DS RMSD value is higher in the F2 system (ca. 6.25) than that in the F0123DUR system (ca. 5). However, among systems with the same number of features, speakers fluctuate more when F2 is involved, i.e., speakers in systems at the left end of x-axis, e.g., the F012, F0123, F0123DUR systems, fluctuate more than those at the right end, e.g., the F1, F1DUR, F3DUR, F3 systems. It is also worth noting that most speakers have DS RMSD values fluctuating between 2.5 and 7.5 in most of the systems (e.g., F2, F12, F123, F0123), indicating that the LLRs of speakers in DS comparisons among 100 sampling replications could be 5 above or below the mean LLR. For example, if one speaker has a mean LLR of -2 in DS comparisons in the F2 system, the possible LLR outputs in the 100 replications could be between -7 and 3.

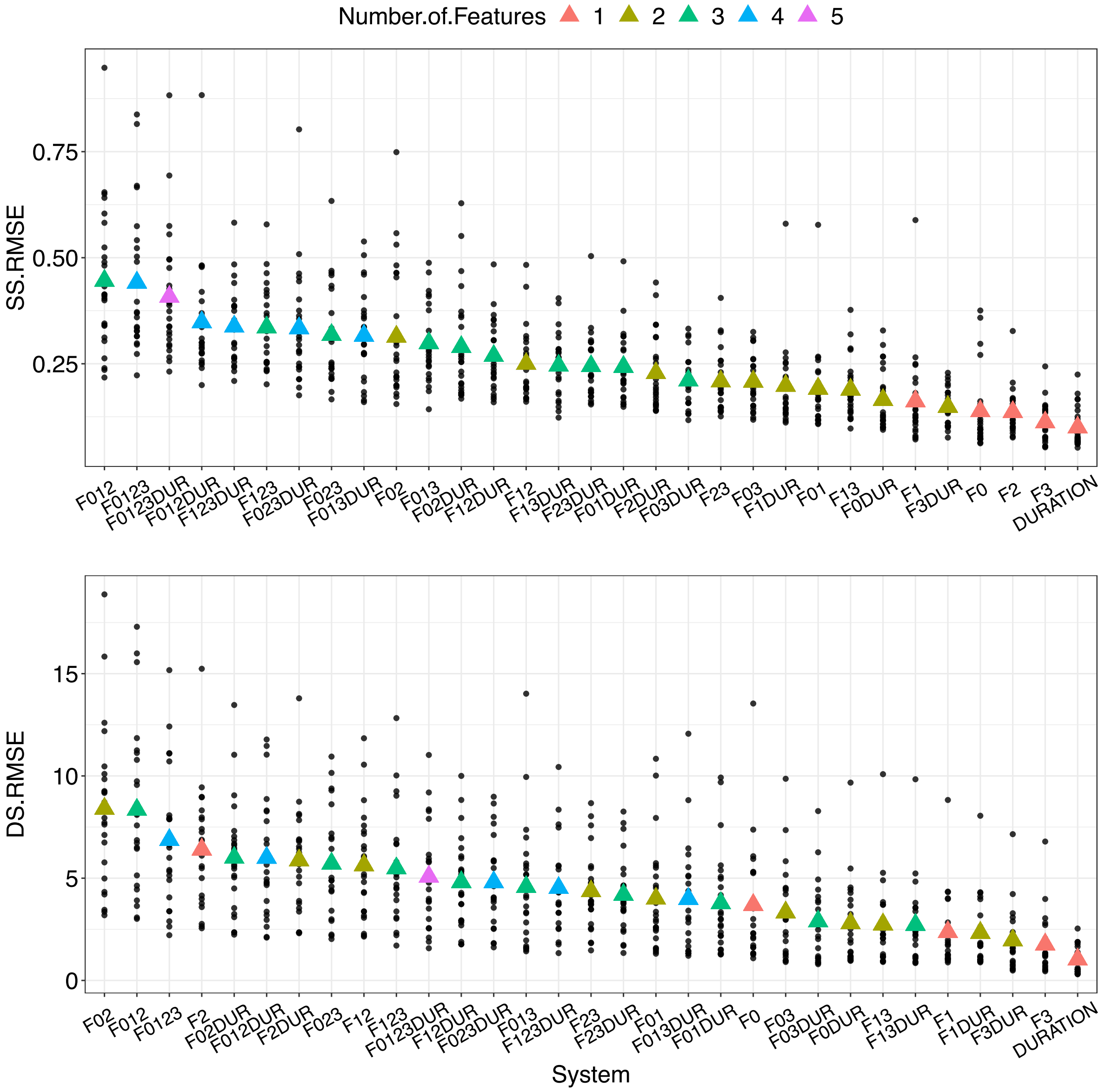


Figure 5.11 RMSD values of individual speakers across 31 systems. Black dots indicate the RMSD values of each individual speaker, and the coloured triangles are the mean RMSD values of all speakers in each system.

Figure 5.12 shows the reliability of individual speakers' LLR in systems with the same number of features (the five-feature system is not included as there is only one such system). Each dot represents the difference between the maximum and minimum RMSD values (i.e., the range) of each speaker across systems with equal number of features. In DS comparisons, the majority of speakers tend to be least stable in one- and two-feature systems and the most stable in systems with three and four features. Around half of the speakers yield DS RMSD values higher than 5 in one-feature systems, while most of the speakers have DS RMSD values lower than 5 in systems with four features. However, speakers show different patterns in SS comparisons. Some speakers start with low SS RMSD values (high reliability) in one-feature systems and end up with high SS RMSD values (low reliability) in four-feature systems, while some other speakers show an opposite pattern. The remaining speakers show similar patterns to those in DS comparisons where they yield the most fluctuating performance in systems with two or three features. All the speakers have SS RMSD values lower than 1, indicating that the fluctuation in SS comparisons caused by different combinations of features and configurations of training and reference speakers is lower than one LLR magnitude in terms of strength of evidence.

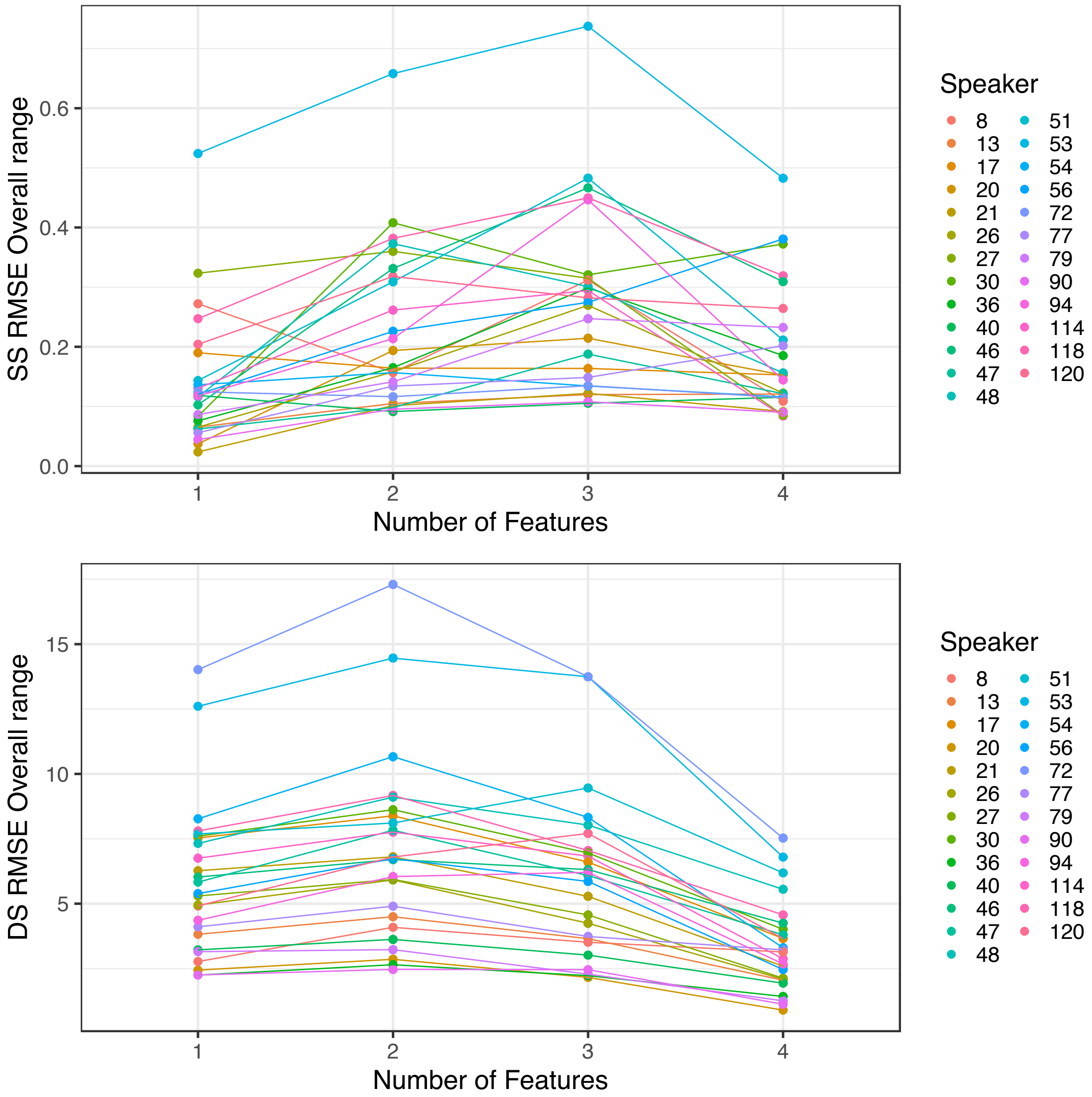


Figure 5.12 SS and DS RMSD ranges across systems with different numbers of features.

Pearson's r was calculated to explore the correlation between the LLR and RMSD values, i.e., the correlation between the validity and reliability of individual speakers' LLRs. Table 5.8 shows the correlation coefficients using individual speakers' SS/DS LLR and RMSD values from systems with the same number of features. Among SS comparisons, the SS LLR shows a positive correlation with RMSD in the five-feature system (*Pearson's r = 0.41*), indicating that speakers who yield high LLR (more likely to be separated in SS comparisons) are also more likely to yield unstable behaviour. In systems with one to four features, the SS LLR and RMSD values do not seem to have much correlation. For DS comparisons, the DS LLR and RMSD are negatively correlated in all systems, meaning that speakers who can be well separated from others in DS comparisons are also more likely to have unstable behaviour (DS LLR scale is reversed). The overall pattern in DS comparisons suggests that the validity and reliability of individual speakers' LLR are likely to be negatively correlated (i.e., speakers that are easier to be separated from others are more likely to have unstable behaviour).

	SS LLR vs. SS RMSD	**DS LLR vs. DS RMSD**
Number of features	*Pearson's r*	*Pearson's r*
1	0.16	-0.73
2	0.14	-0.52
3	0.14	-0.62
4	0.13	-0.58
5	0.41	-0.69

Table 5.8 Correlation between the validity and reliability of individual speakers.

Figures 5.13 to 5.20 give a more detailed portrait of the fluctuation in individual speakers' LLRs across different systems. The x-axis indicates different systems, and the y-axis indicates RMSD values. A flat line would indicate that different combinations of features and configurations of training and reference speakers have limited effect on individual speaker's LLR output.

5.3.2.2.1 Single-feature systems

Figures 5.13 and 5.14 show the LLR fluctuations for each speaker for each single-feature system. In SS comparisons, most speakers do not show much fluctuation across five single-feature systems, indicating that individual speakers' behaviour are comparatively stable regardless of the choice of features and training and reference speakers. However, there are some exceptions, e.g., speakers #8 and #27 have more fluctuating SS LLRs in the F0 system and speaker #53 is more fluctuating in the F1 system.

For DS comparisons, most speakers tend to have more fluctuating DS LLRs in the F2 system, e.g., speakers #17, #30, #48, #72, while other speakers, e.g., speakers #51 and #53, are fluctuating in the F1 and F0 systems respectively. Figures 5.13 and 5.14 also show high fluctuation in SS comparisons does not lead to high fluctuation in DS comparisons, e.g., speaker #53 is the most fluctuating in the F1 system for SS comparisons, but he is the most fluctuating in the F0 system for DS comparisons.

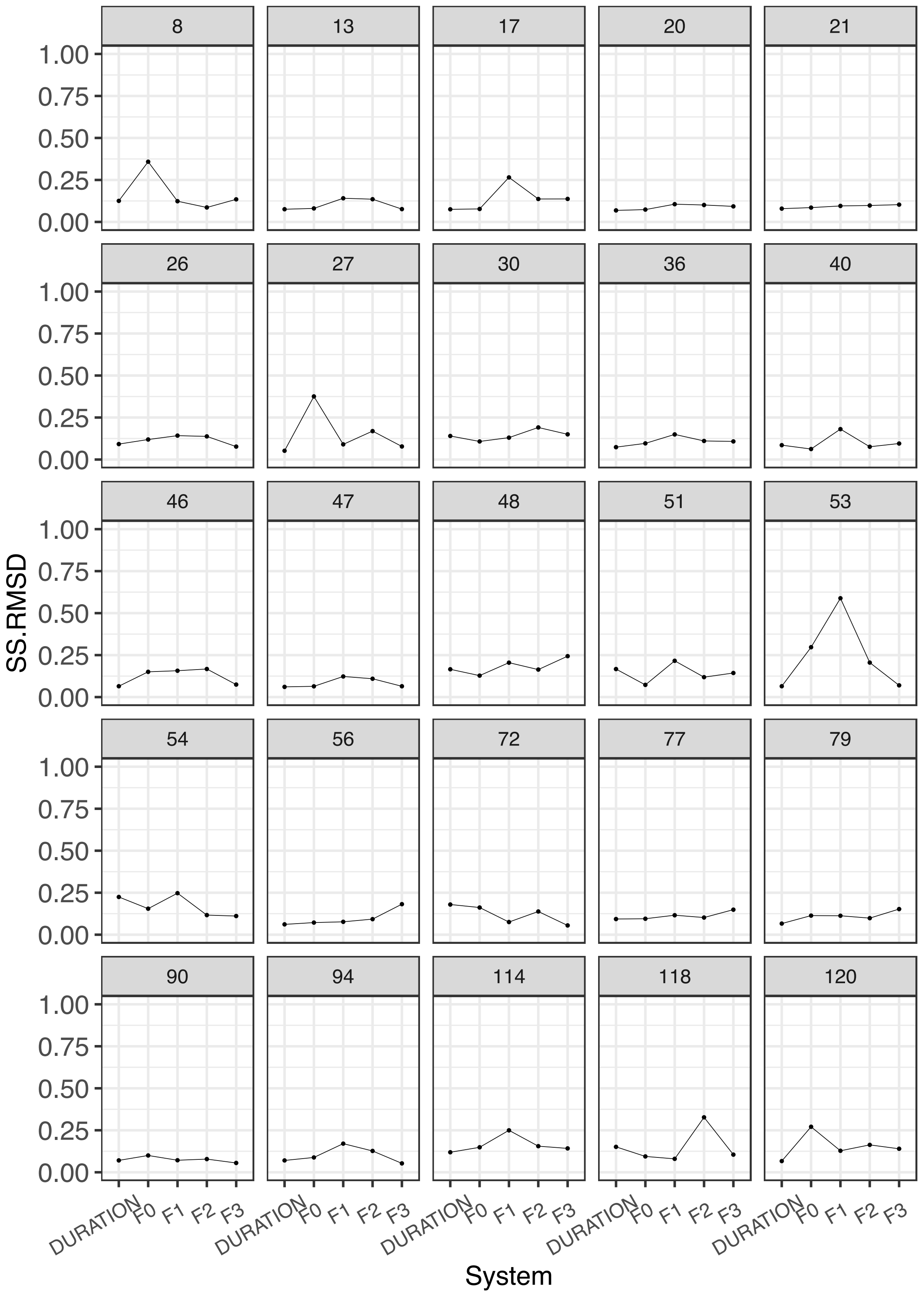


Figure 5.13 SS RMSD values of all 25 test speakers across single-feature systems. Numbers at the top of each panel indicate speaker numbers.

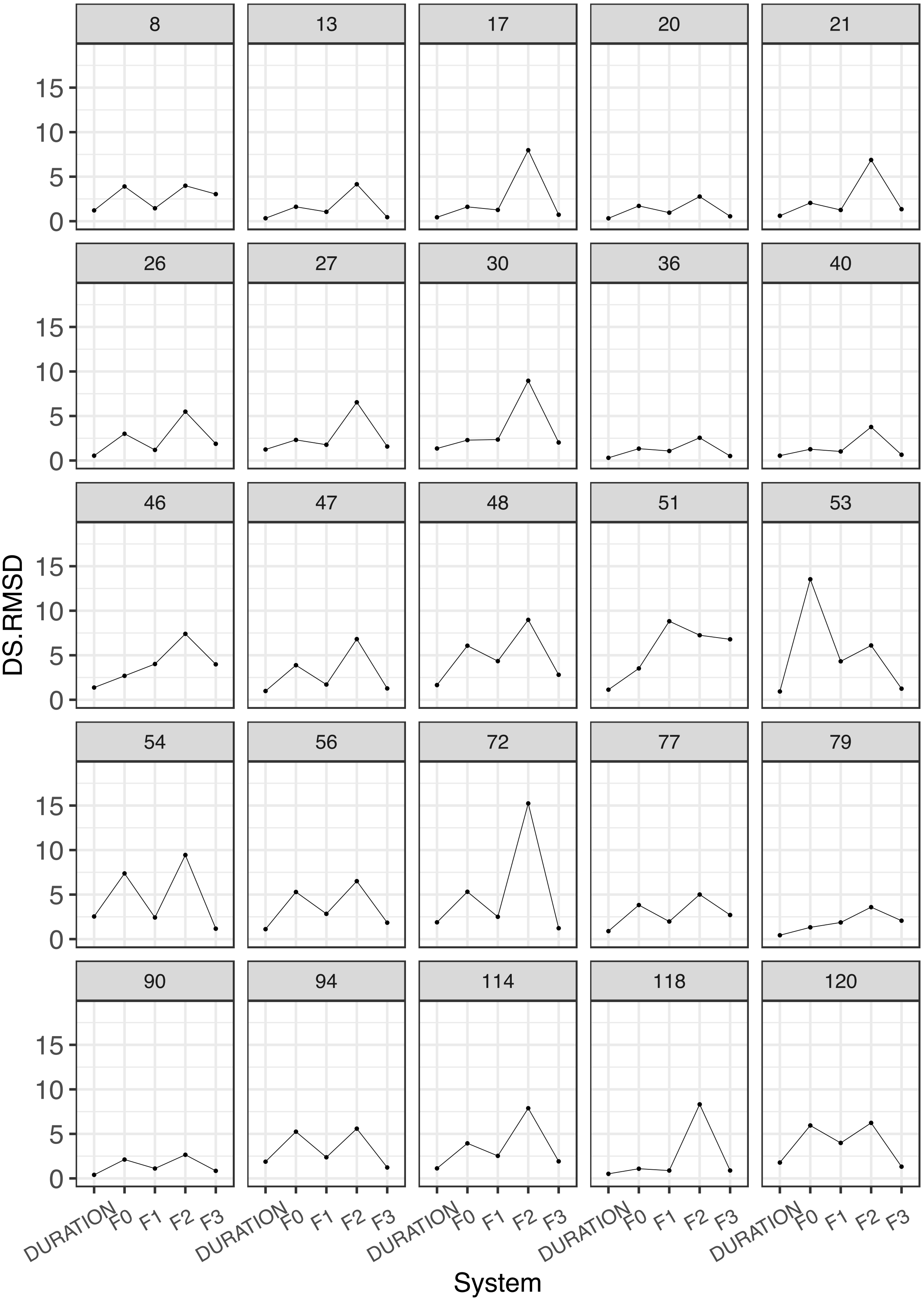


Figure 5.14 DS RMSD values of all 25 test speakers across single-feature systems. Numbers at the top of each panel indicate speaker numbers.

5.3.2.2.2 Two-feature systems

Figures 5.15 and 5.16 show the fluctuation in individual speakers across two-feature systems. Similar to single-feature systems, most speakers show limited fluctuation across different two-feature systems in SS comparisons. Some speakers, i.e., #20, #27, #30, #46, #48, #53, #114, #118 and #120, have a clear peak in the SS RMSD values in the F02 system, indicating that these speakers are more sensitive to different combinations of features and configurations of training and reference speakers in SS comparisons. In terms of DS comparisons, it seems that only a few speakers, i.e., #8, #13, #20, #36, #40, #90, are comparatively less affected by different combinations of features. Meanwhile, 19 out of 25 speakers, i.e., #13, #17, #20, #21, #26, #27, #30, #46, #47, #48, #53, #54, #56, #72, #77, #94, #114, #118, #120, have the most fluctuating DS LLRs in the F02 system.

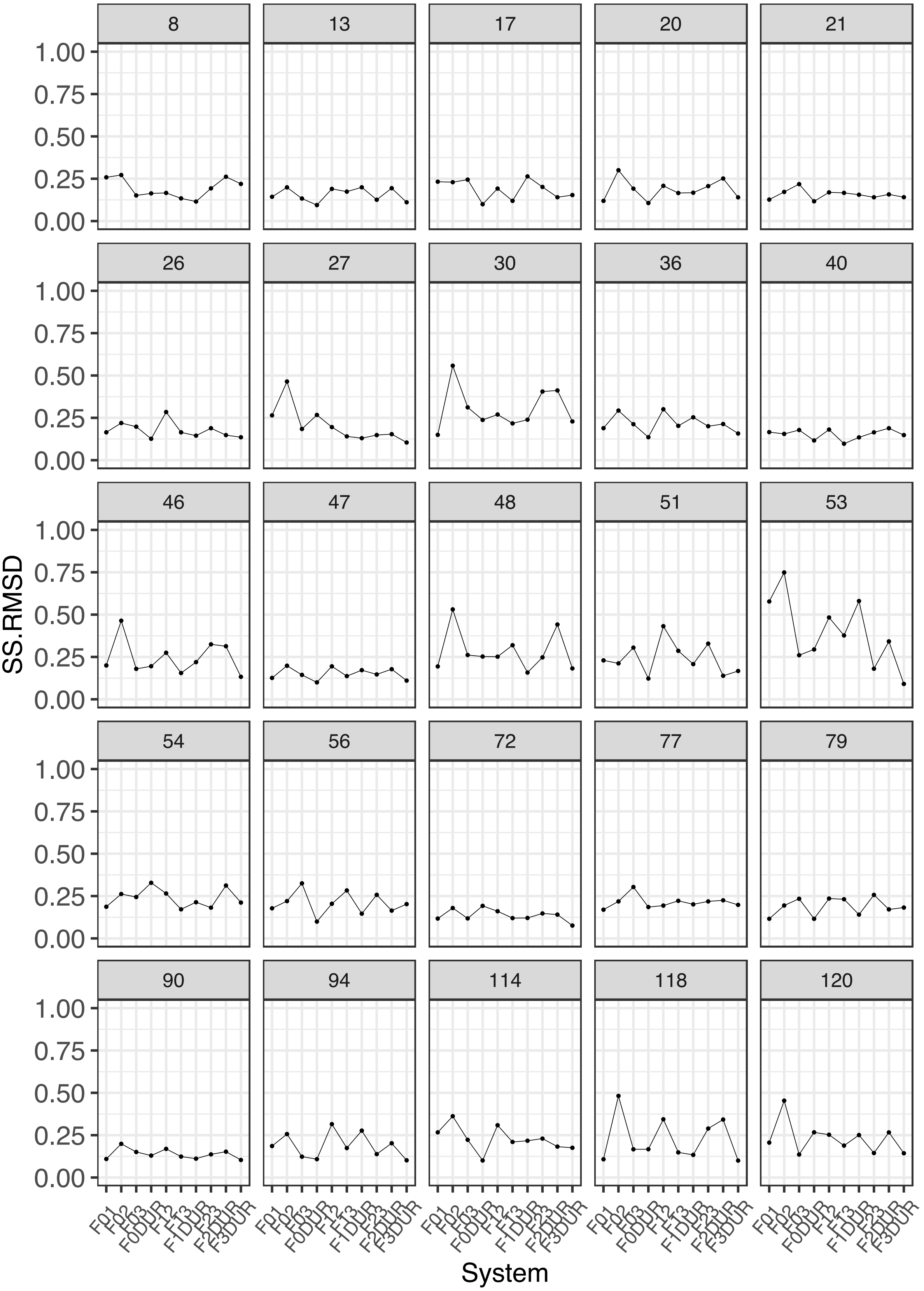


Figure 5.15 SS RMSD values of all 25 test speakers across two-feature systems. Numbers at the top of each panel indicate speaker numbers.

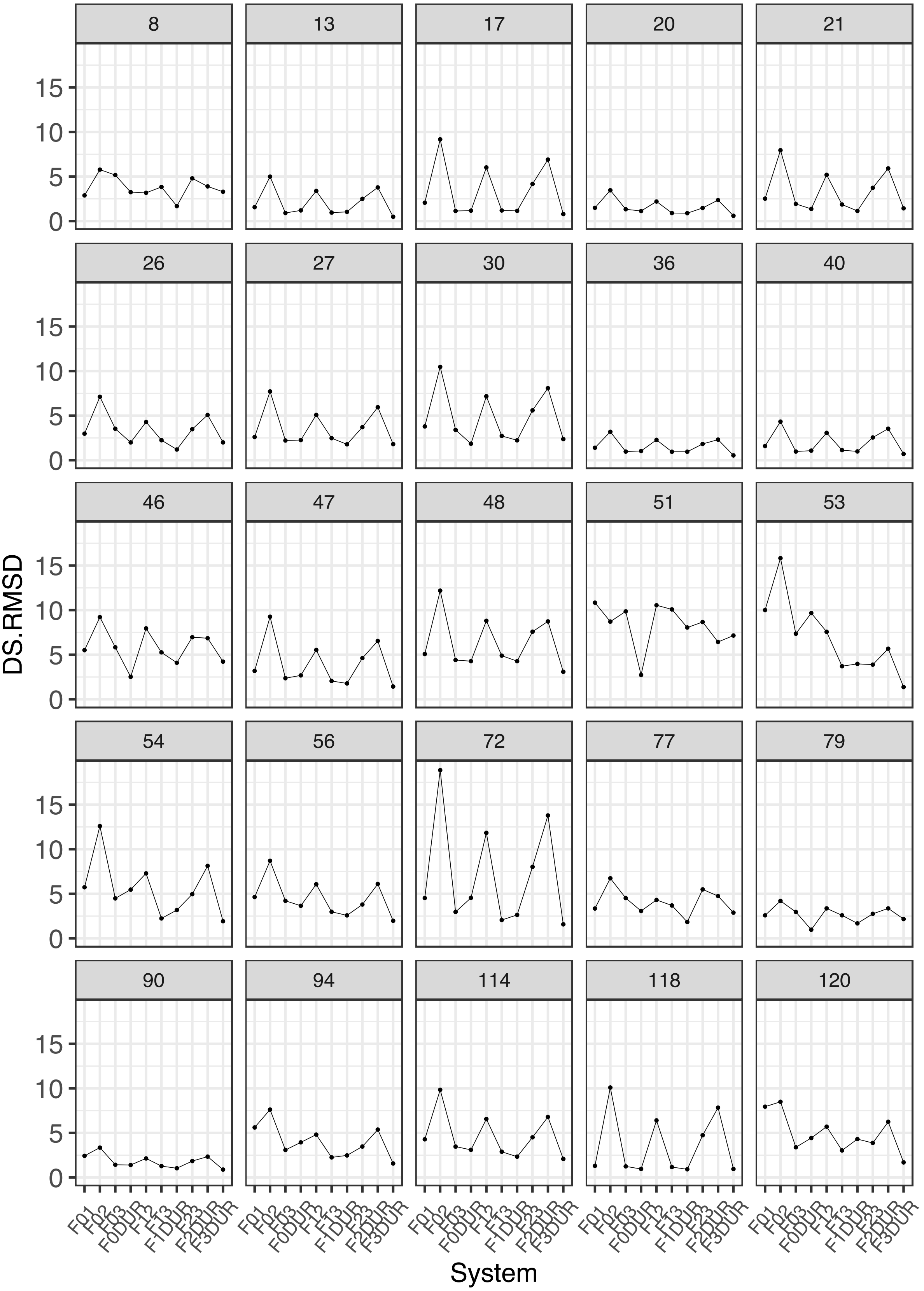


Figure 5.16 DS RMSD values of all 25 test speakers across two-feature system. Numbers at the top of each panel indicate speaker numbers.

5.3.2.2.3 Three-feature systems

Figures 5.17 and 5.18 show the fluctuation in individual speakers across three-feature systems. Similar to the single- and two-feature systems, i.e., more than half of the speakers have a stable performance across different systems. However, there are some exceptions, i.e., speakers #30, #46, #48, #51, #53, #56, #94, #118 and #120, fluctuate across different systems. In DS comparisons, speakers seem to be less fluctuating across different systems with three features than with two features. Most speakers have a fairly flat line across three-feature systems, while a few speakers such as #17, #30, #48, #53, #72, #118 and #120 have comparatively larger fluctuations.

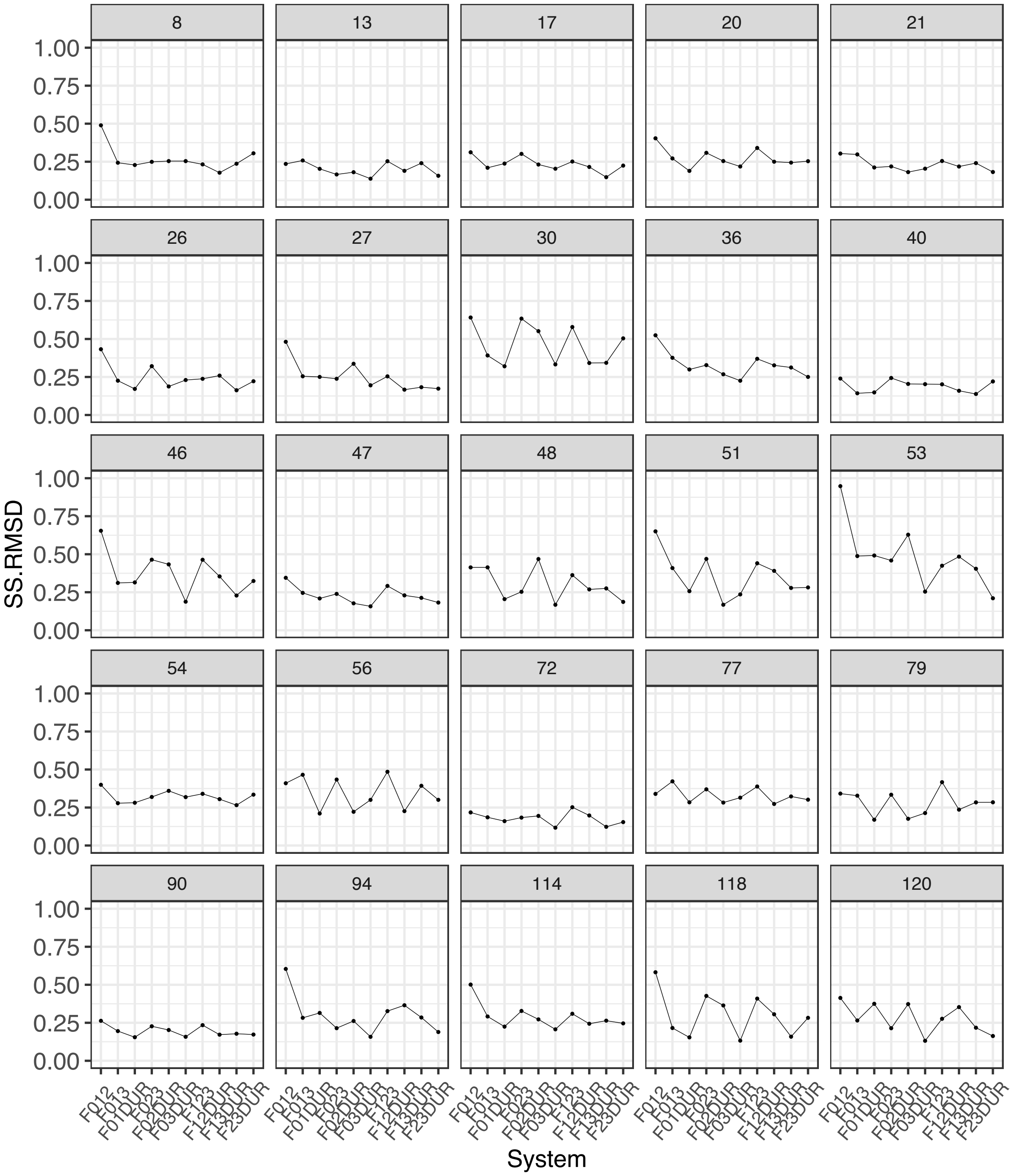


Figure 5.17 SS RMSD values of all 25 test speakers across three-feature systems. Numbers at the top of each panel indicate speaker numbers.

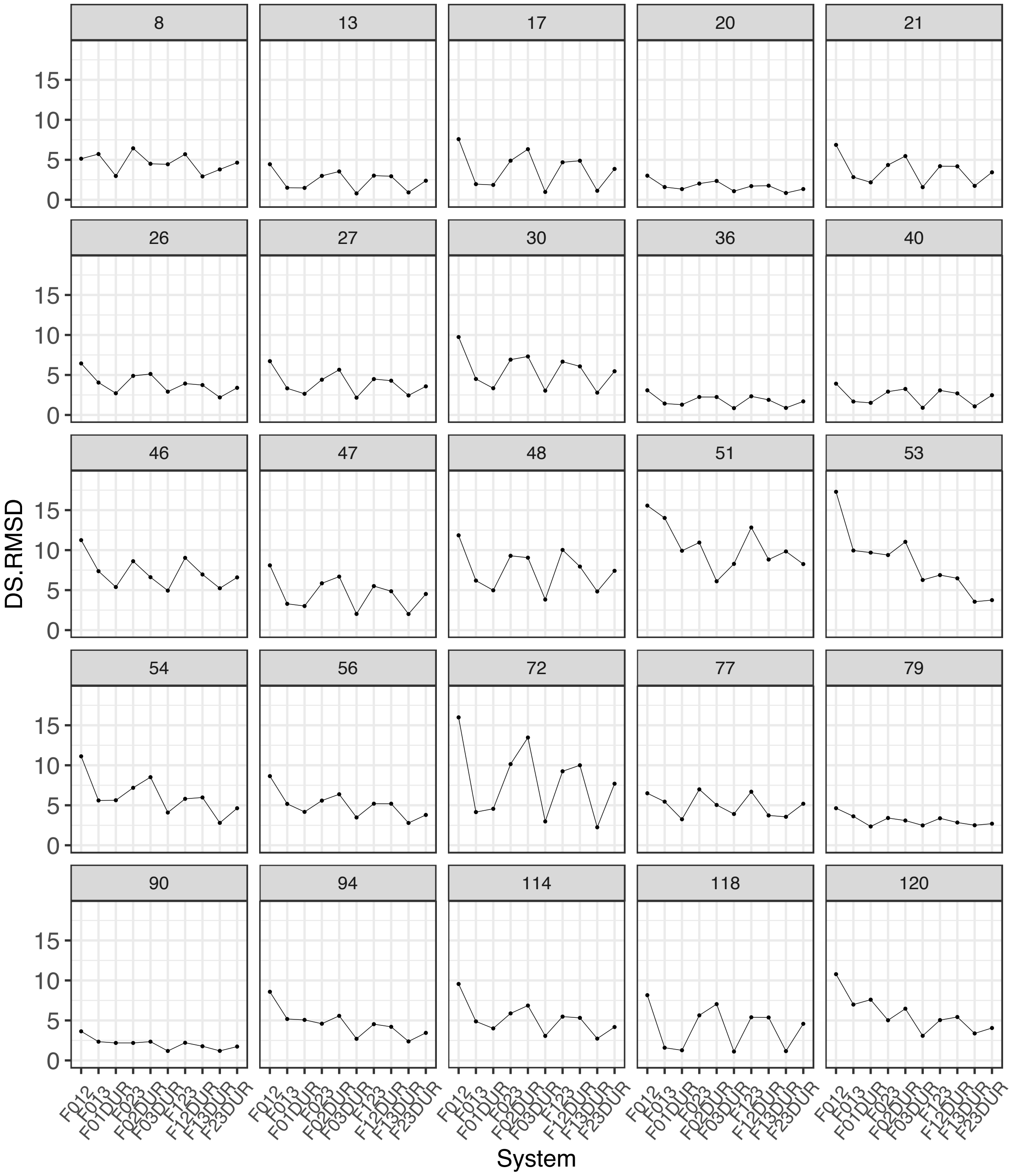


Figure 5.18 DS RMSD values of all 25 test speakers across three-feature systems. Numbers at the top of each panel indicate speaker numbers.

5.3.2.2.4 Four/five-feature systems

Figures 5.19 and 5.20 show individual speakers' fluctuation across systems with four or five features. For SS comparisons, the majority of the speakers have little LLR fluctuation across different four-/five-feature systems, which is similar to other systems with less than four features. Other speakers show some LLR fluctuation depending on which combinations of features are used, e.g., speaker #56 is the least fluctuating in the F012DUR system, while speaker #53 is the least fluctuating in the F123DUR system.

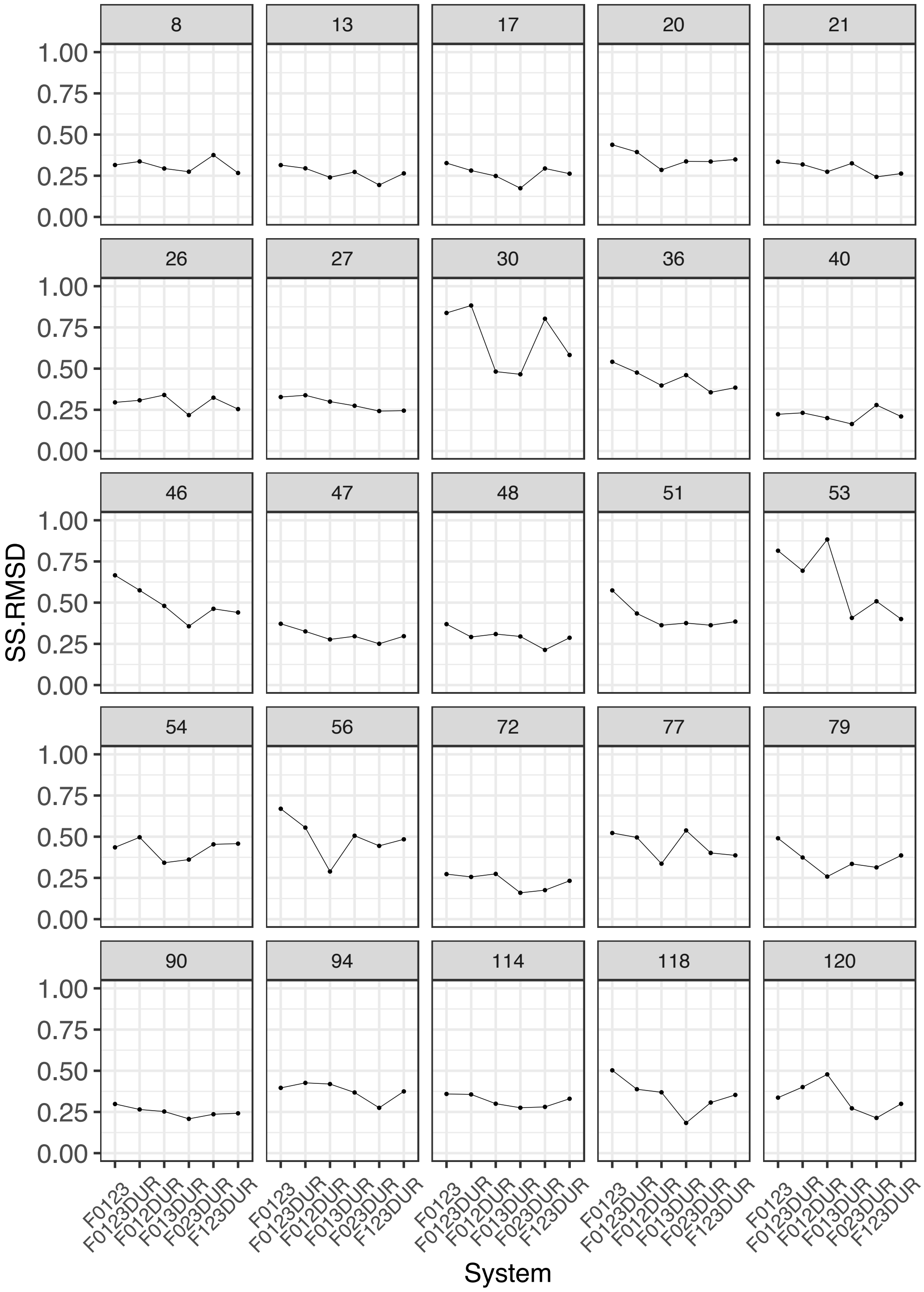


Figure 5.19 SS RMSD values of all 25 test speakers across four-/five-feature systems. Numbers at the top of each panel indicate speaker numbers.

For DS comparisons, most speakers seem to have limited fluctuation across systems with four or five features, suggesting that the majority speakers can potentially be less fluctuating across different system when four or more features are used. However, there are still a few exceptions, e.g., speakers #72 and #118 show a clear dip in DS LLR in the F013DUR system.

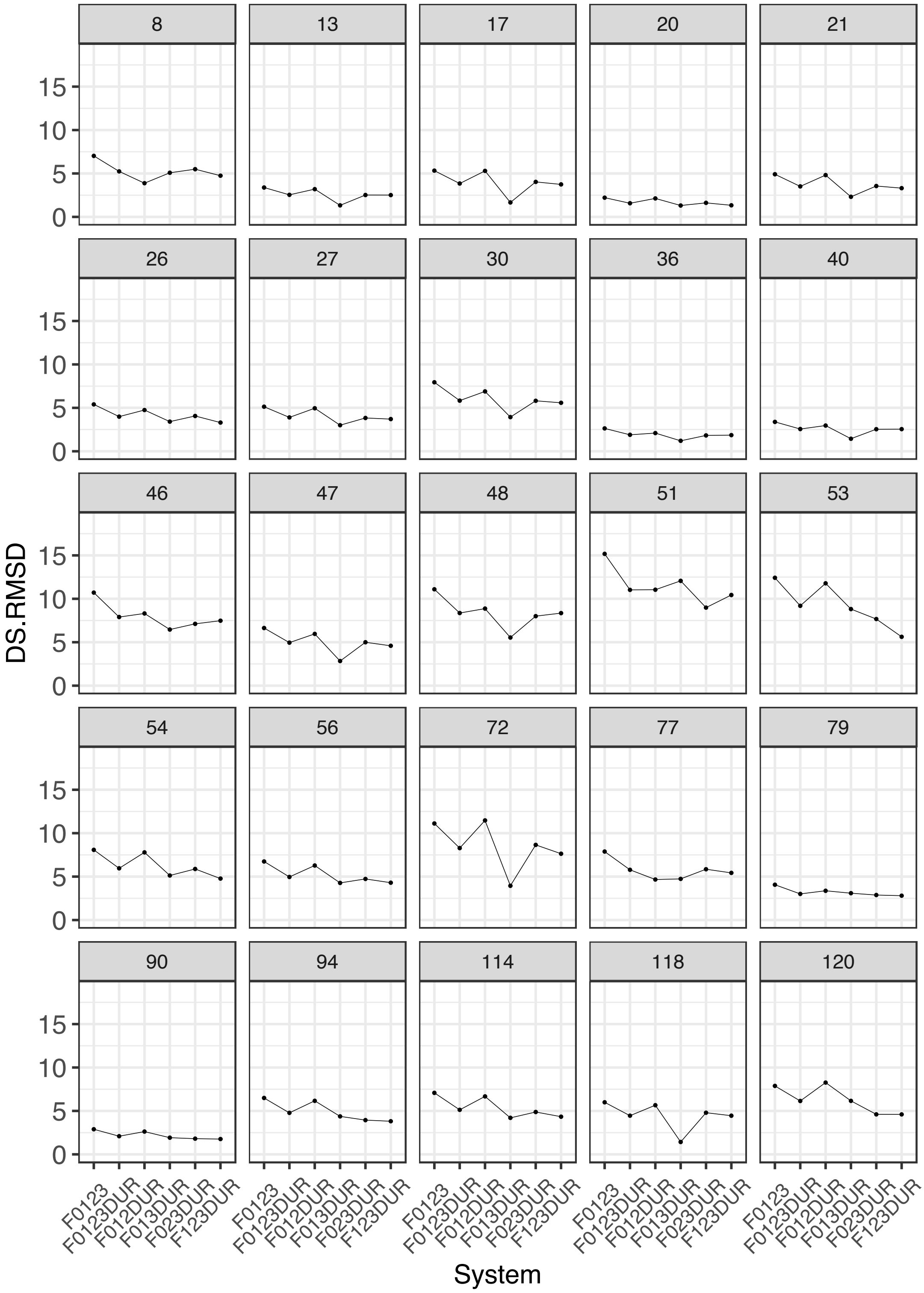


Figure 5.20 DS RMSD values of all 25 test speakers across four-/five-feature systems. Numbers at the top of each panel indicate speaker numbers.

Based on the patterns observed in the above sections (5.3.2.2.1 - 5.3.2.2.4), most speakers show limited variation in SS comparisons across systems with different combinations of features, indicating that speakers are less sensitive to different configurations of training and reference speakers regardless of whether a one-feature or five-feature system is used. However, there are always some speakers giving fluctuating behaviour depending on which features are being used, e.g., speakers #8 and #27 in the F0 system, speakers #27 and #118 in the F02 system. For DS comparisons, the majority of the speakers tend to have stable behaviour when three or more features are used; meanwhile, again some speakers (e.g., #53, #72, #118 and #120) have fluctuating behaviour depending on which features are being used in three-, four- and five-feature systems. The patterns in SS and DS comparisons using different combinations and different numbers of features indicate that individual speakers' behaviour is not only affected by sampling variability, but which exact features are being used. Comparatively, researchers and experts might have less control over the issue of sampling variability, while they have more control over which features included for analysis. It is then essential for researchers and experts to acknowledge that the same speaker might show different behaviour when different linguistic-phonetic features are used, especially in real case scenarios. The following section demonstrates potential LR ranges one might obtain in real case scenarios.

5.3.3 What happens in a real case?

The results above show that some speakers are more affected by the different combinations of features and configurations of training and reference speakers, while others are less affected. Under real case scenarios, 30 to 40 training and reference speakers are likely to be sampled from a relevant population (e.g., 35 speakers in Rose, 2013b), and the size of the relevant population itself is, in most cases, considerably larger than the number of training and reference speakers sampled. Although it is possible to sample more speakers from the relevant population, empirical studies (e.g., Ali et al., 2015; Morrison & Poh, 2018) and the current Chapter have shown that the effect of sampling variability on both overall performance and individual behaviour is inevitable. It is then a practical consideration for casework, i.e., would we obtain the same results for this particular pair of speakers if the experiment is replicated? It is then important to explore possible LLR ranges that one could obtain for a pair of speakers in a real case. The following section examines results from the best system in terms of C_{llr} mean and range, i.e., the F0123 system, to analyse in detail the speaker-specific effects, and thus consider the potential outcomes of this variability in the context of a real case. Since the RMSD values indicate how stable one specific test speaker behaves when different configurations of training and reference speakers are used, the test speakers who have the highest (i.e., having fluctuating behaviour when different configurations of training and reference speakers are used) and lowest RMSD values (i.e., having stable behaviour when different configurations of training and reference speakers are used) in SS and DS comparisons are selected for the demonstration of possible LLR ranges. Table 5.9 shows the RMSD values of the selected speakers from the F0123 system,

F0123 System			
Speaker	SS RMSD	Speaker	DS RMSD
#40	0.22	**#20**	2.21
#30	0.84	**#51**	15.17

Table 5.9 SS and DS RMSD values of speakers with the least and most variable LLRs.

Figure 5.21 shows the ranges of SS LLRs of speakers #30 (most fluctuating) and #40 (least fluctuating) across the 100 replications, i.e., #30 and #40 are compared with themselves 100 times with different configurations of training and reference speakers using F0, F1, F2 and F3. X-axis is the SS LLR values and y-axis indicate speaker IDs.

As Figure 5.21 shows, both speakers yielded positive SS LLRs, i.e., they all yielded consistent-with-fact results, in all 100 replications. Speaker #40 yielded the least variable LLRs, varying from ca. 0.6 to 1.9, which is a difference between *limited* and *moderate support* for SS origin in terms of verbal LLR expression. However, the SS LLR of speaker #30 varies between 1.3 and 6.6, which is a difference between *moderate* and *very strong support* for SS origin. Comparatively, speaker #40 seems to be less problematic, while speaker #30 would be more problematic to assess in a real case because the SS LLRs show a much wider variation depending on who is used in the training and reference data.

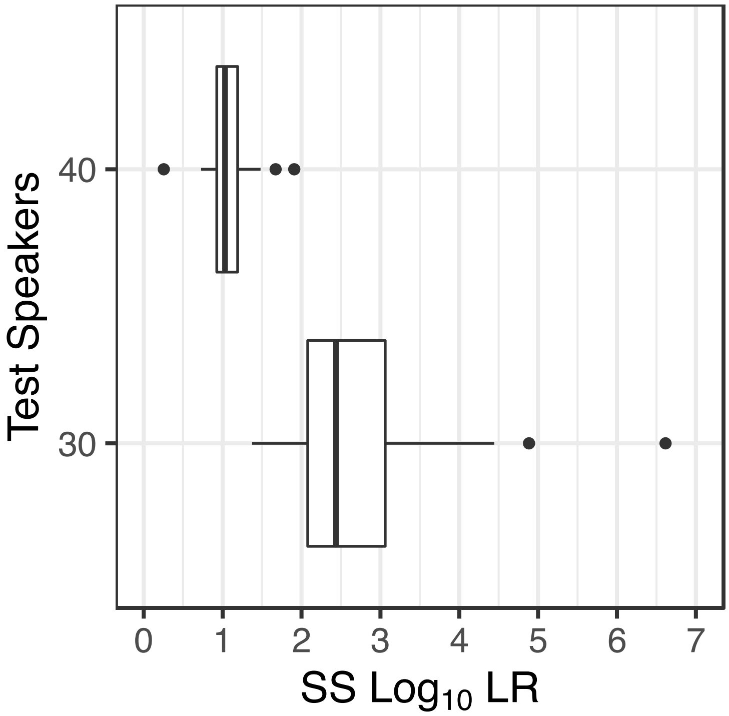


Figure 5.21 SS LLRs of speaker #30 and #40 using different training and reference speakers in the F0123 system.

Figure 5.22 shows the DS LLRs of the most and least variable speakers, i.e., speaker #20 and #51, in DS comparisons. Note that the x-axis scales are different for the top and bottom panels. The x-axis indicates the DS LLR values, while the y-axis shows the rest of the test speakers that are being compared. Each boxplot represents the variation of DS LLRs, e.g., the first boxplot in the top panel indicates the DS LLR ranges of speaker #20 being compared with speaker #8 with different configurations of training and reference speakers across 100 replications. The DS LLRs fluctuate much more markedly than SS LLRs, which is probably caused by the fact that the number of SS LLRs is much smaller than those of DS LLRs. For example, a sample of 30 speakers would only have 30 SS LLRs, but 870 DS LLRs. In terms of contrary-to-fact results, it can be seen that more reliable speakers tend to yield more counterfactual cases. For example, speaker #20 yields the most stable LLRs in DS comparisons across the 100 replications, but this speaker also gives the most contrary-to-fact results when compared with other test speakers. Contrary-to-fact DS LLRs are observed when speaker #20 is being compared with 12 (out of 24) speakers (e.g., #17, #21, #26, #27, #30 etc). Among the 12 speakers, speaker #20 yielded contrary-to-fact results when comparing with three other speakers in all of the 100 replications, i.e., speakers #27, #54, and #56. Clearly, the contrary-to-fact results for speaker #20 would be misleading under a real case scenario. The highest DS LLR goes up to 2.12, which indicates a *moderately strong support* for SS origin (given the speech samples are from different speakers). For consistent-with-fact results in speaker #20, the majority of DS LLRs vary between 0 and -5, which is a difference between *limited* and *very strong support* for DS origin. On the other hand, speaker #51 does not to have contrary-to-fact LLRs when compared with the remaining 24 speakers and the DS LLRs range between ca. -12 and -50 for most of the DS comparisons of speaker #51.

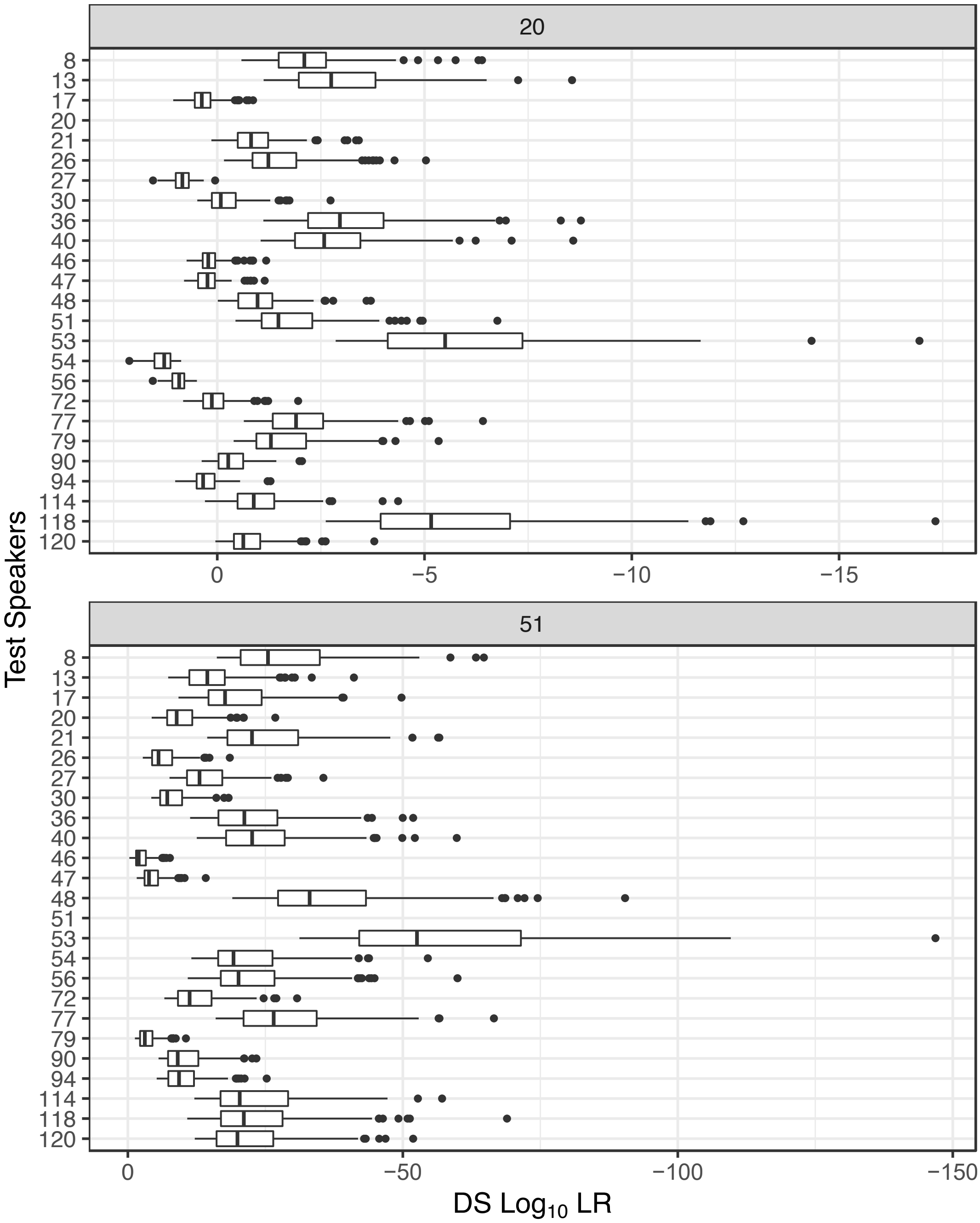


Figure 5.22 DS LLRs of speaker #20 and #51 using different training and reference speakers in the F0123 system.

5.4 Discussion

In this Chapter, 31 systems based on different combinations of input features were evaluated and compared from the perspective of overall performance and individual speakers' behaviour. The performance of all 31 systems suggests that systems with more features in general yield higher validity (i.e., lower mean C_{llr}; Figure 5.4), while some exceptions can be found in the F0123 and F0123DUR systems, i.e., the four-feature system outperforms the five-feature system. Meanwhile, no clear pattern is observed in C_{llr} range (Figure 5.4). Therefore, it is important to acknowledge that adding extra features does not always improve the overall system validity and reliability (and this is especially true when considering individuals).

The overall performance shows a similar pattern to Hughes et al. (2016) in terms of system validity (and of course this is due to the similar data used), where F2 yielded the best discriminatory performance. However, the F2 system also yielded the best system reliability. A comparison between the F0123 and F2 systems shows that the F2 system has a lower C_{llr} range (0.04) than that of the F0123 system (0.15). However, all of the C_{llr} values in the F0123 system across 100 replications are lower than those in the F2 system. Therefore, it is important to understand the trade-off between system validity and reliability, i.e., how much variation are we willing to accept given validity, or the other way round? The trade-off between system validity and reliability should be case-specific, as each case would have more or less different prerequisites, e.g., the amount of speech available in the suspect and offender samples, number of speakers available in training and reference data, and how narrowly or broadly defined is the relevant population. To the best of the author's knowledge, there is no such framework to capture that trade-off in LR-based FVC studies for now. Further, experts have the freedom to make decisions (e.g., which speech feature/background population to be used) and conduct analyses under casework scenarios; therefore, it is important for experts to take researchers' degrees of freedom into consideration as Chapter 5 has shown that different choices of linguistic-phonetic features yield different overall performance and the variability can be substantial in some situations (e.g., system F012 in Figure 5.2).

For individual speakers' behaviour, all speakers yielded consistent-with-fact results in systems with four and five features, while contrary-to-fact results were observed in systems with fewer than four features (Figure 5.6). Furthermore, speakers have different fluctuating patterns in SS and DS comparisons (Figure 5.12). In SS comparisons, speakers are more likely to yield fluctuating performance in systems with more features, indicating that combining multiple features increases within-speaker variation, which makes speakers more sensitive to sampling variability. Moreover, Figure 5.12 shows that speakers have different fluctuating patterns across systems with different numbers of features, e.g., speaker #30 yielded the most fluctuating performance in systems with four features, while speaker #51 yielded the least fluctuating performance. This suggests that the effect of different combinations of features and configurations of training and reference speakers on the reliability of individual test speakers' LLR output is speaker specific.

In DS comparisons, there does not seem to be any general pattern between number of features and individual speaker's reliability (Figure 5.11 bottom panel). However, when compared to systems with the same number of features (Figure 5.12 bottom panel), individual speakers became less fluctuating when four features were used, indicating that different combinations of features and configurations of training and reference speakers had less effect on the reliability of individual test speakers in DS comparisons. A comparison between the most (speaker #51) and least (#20) fluctuating speakers from the F0123 system in the DS comparisons shows that validity and reliability of individual speakers' LLR are likely to be negatively correlated (Figure 5.22).

The overall pattern indicates that both overall performance and individual speakers' behaviour could be affected by specific case conditions, and there is no single rule that applies to all. The patterns observed in the current study are similar to other biometric systems (e.g., fingerprint, iris) for forensic evaluation, i.e., the individual speaker's behaviour is system/case specific (Dunstone & Yager, 2009). It is noteworthy that the corpus used in the current experiments was recorded under well-controlled settings and the speech samples were contemporaneous. Moreover, experiments were conducted within datasets and speaker samples in training and reference sets are likely to be duplicated between each replication. Therefore, the results in this Chapter (and Chapter 4) are necessarily conservative, and more variability in results would be observed if speakers are randomly sampled from other datasets and truly independent samples are used.

5.5 Chapter summary

Chapter 5 explored the effect of sampling variability on overall performance and individual speakers by taking different combinations of linguistic-phonetic features (e.g., only use F1 or use F1 and F2) and different configurations of training and reference speakers into consideration. Followings are the findings of Chapter 5 in bullet points.

- Systems with more features do not necessarily outperform those with fewer features.
- Individual speakers' LRs fluctuate more when more features are included in same-speaker comparisons, but not in different-speaker comparisons.
- Individual speakers are more likely to yield contrary-to-fact results in DS comparisons than SS comparisons.
- The LLR and RMSD are likely to be negatively correlated in DS comparisons regardless of the number of features used, while negative correlation is only observed in the five-feature system for SS comparisons.
- Researchers' degrees of freedom (i.e., different choices of linguistic-phonetic features in the current context) need to be considered as a factor affecting evidence evaluation results.

Chapter 6 Overall performance as a function of score skewness, sample sizes and calibration methods

The previous chapter analysed the validity and reliability of LLR output placing more focus on the *feature-to-score* stage. However, the variability in LLR outputs can be introduced by different sources (e.g., variability in sampling the relevant population; variability due to modelling assumptions; Morrison, 2016) at different stages of LR computation (i.e., *feature-to-score* and *score-to-LR*). This chapter looks into score distributions and overall performance as a function of score skewness, sample size and calibration methods. The following RQs are addressed in the current chapter.

RQs 3.

a. To what extent is overall performance affected by skewed scores?
b. Are certain calibration methods more susceptible to score skewness than others?
c. Would overall performance be improved with larger sample sizes when scores are skewed?

6.1 Introduction

Assuming that systematic variability in LLR output caused by sociolinguistic factors (e.g., regional accent, class, educational background) can be well-controlled at the *feature-to-score* stage, one might still obtain variable and extreme LLR outputs due to random sampling variability and data extrapolation at the tails of score distributions. This is especially true when the sample size is small, and the density estimation is not well-supported by the observed data. Therefore, the choice of calibration method can be extremely important for dealing with the issue of over- or underestimating the strength of evidence (Vergeer et al., 2020).

Wang et al. (2019) investigated the effect of sampling variability on overall performance at the *score-to-LR* stage. They simulated scores under an assumption of normality using the `rnorm()` function in R (R Core Team, 2020). The SS and DS scores were sampled from three sets of normal distributions with different EERs, i.e., where the distance between SS and DS distributions were different. This was designed to mimic variables with different speaker-discriminatory power. 20 SS and 380 DS training and test scores (i.e., 20 training speakers and 20 test speakers) were simulated respectively to produce calibrated LLRs using logistic regression (Brümmer et al., 2007). For each set of distributions, the experiments were replicated 100 times by varying both training and test scores, training scores only, and test scores only. The results show that, under normal distributions with a sample size of 20 speakers, system validity is not necessarily positively correlated with system reliability, e.g., systems having lower mean C_{llr} does not lead to lower C_{llr} range and vice versa; meanwhile, overall performance is less susceptible to sampling variability when only training data is varied compare to when test data is varied. However, Wang et al. (2019) only looked at scores that follow an assumption of normality, while scores are less likely to be normally distributed under real case scenarios due to the reasonable limits of sample size in the real world. A polit study was conducted to investigate the score distributions using linguistic-phonetic features.

6.1.1 Polit study

A pilot study was carried out using six segmental linguistic-phonetic variables (i.e., GOOSE, NORTH, PRICE, TRAP vowels, and two FPs *um* and *uh*) based on 36 SSBE speakers from the DyViS corpus (Nolan et al., 2009). Raw formant data was readily available and retrieved from (Gold & Hughes, 2015). The raw F1, F2 and F3 values were fitted with quadratic curves. Data from each speaker was divided in half to create SS and DS pairs, and the quadratic coefficients were used as the input of MVKD (Aitken & Lucy, 2004) to generate cross-validated scores. Calibration was not conducted given there were only 36 speakers available; meanwhile, the focus here was to investigate score distributions. Figure 6.1 shows the SS (red) and DS (blue) score distributions of the six segmental linguistic-phonetic variables. In general, both SS and DS scores are skewed to some extent, and DS scores are much flatter than SS scores. The SS score distributions seem to be more problematic given that there are fewer data points in SS scores than those in DS scores. As a result, extrapolation effects can be observed at both the tails and centre in SS score distributions, e.g., indicated by dashed circles at the right tail of the PRICE vowel and centre of the TRAP vowel. If scores were modelled following a normality assumption, the theoretical distributions would not be well-supported by the observed data (i.e., histograms in Figure 6.1), which could lead to invalid LR estimates. The question is how much does it matter if scores are not normally distributed? In this chapter, the effect of sampling variability on overall performance is investigated in relation to score skewness; meanwhile, different calibration methods were explored to reduce the level of uncertainty (i.e., reliable overall performance), especially when sample size is small.

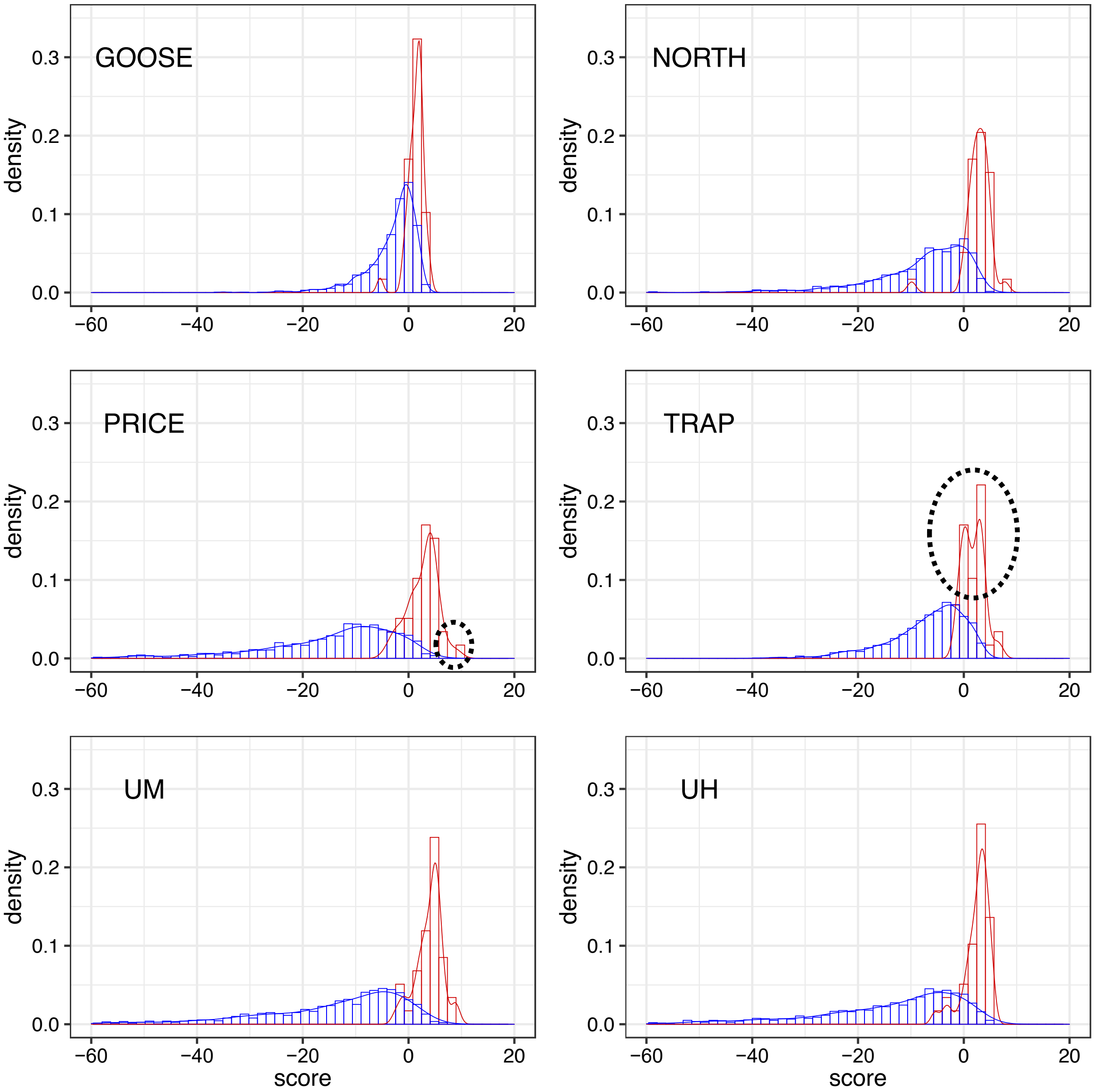


Figure 6.1 Score distributions of six linguistic-phonetic variables (blue curve = DS scores, red curve = SS scores, black dotted circles = examples of extrapolation if normality is assumed).

6.2 Method

The current chapter uses simulated scores to explore overall performance as a function of sample sizes, score distributions and calibration methods. Simulation was carried out based on scores obtained from the FP *um*, previously described in Section 4.4.1, where the test, training and reference speakers were varied across 100 replications. The reason for only using scores from FP *um* is because FP *um* is obtained from a more controlled corpus than Cantonese SFP /a/ and /haijat/. The distributions of SS and DS scores were first investigated, and simulation was carried out using the `skew-t (ST)` function in the `sn()` package (Arellano-Valle & Azzalini, 2013) in R (R Core team, 2020). Four calibration methods were used to test which calibration methods are more or less susceptible to sampling variability. Training and test data were simulated with different numbers of scores to take the effect of sample size into consideration. The following sections firstly give description of the score distributions of the FP *um*, followed by the simulation and calibration procedures. Results and discussion are given in Sections 6.3 and 6.4.

6.2.1 Score distribution

Figure 6.2 (upper panels) shows the distributions of SS and DS log scores of FP *um* by varying test, training and reference speakers. The SS scores are fairly well normally distributed with a cluster between ca. -25 and -50, while the DS scores are more negatively skewed. For the sake of simplicity for the simulation process, a *z-score* of 2 was applied to remove scores that are more than 2 standard deviations from the mean. Figure 6.2 (lower panels) shows the score distributions after the *z-score* was applied. Table 6.1 shows the distribution parameters (i.e., mean, standard deviation, skewness and kurtosis) after *z-score* application. The DS scores have higher skewness and standard deviation values than those of the SS scores, indicating that the DS scores are more skewed and variable than the SS scores. The kurtosis values of both SS and DS scores are slightly higher than 3, suggesting that most of the SS and DS scores are at the centre of the distribution and have a similar tail thickness. The means of the SS and DS scores are 2.6 and -78 respectively.

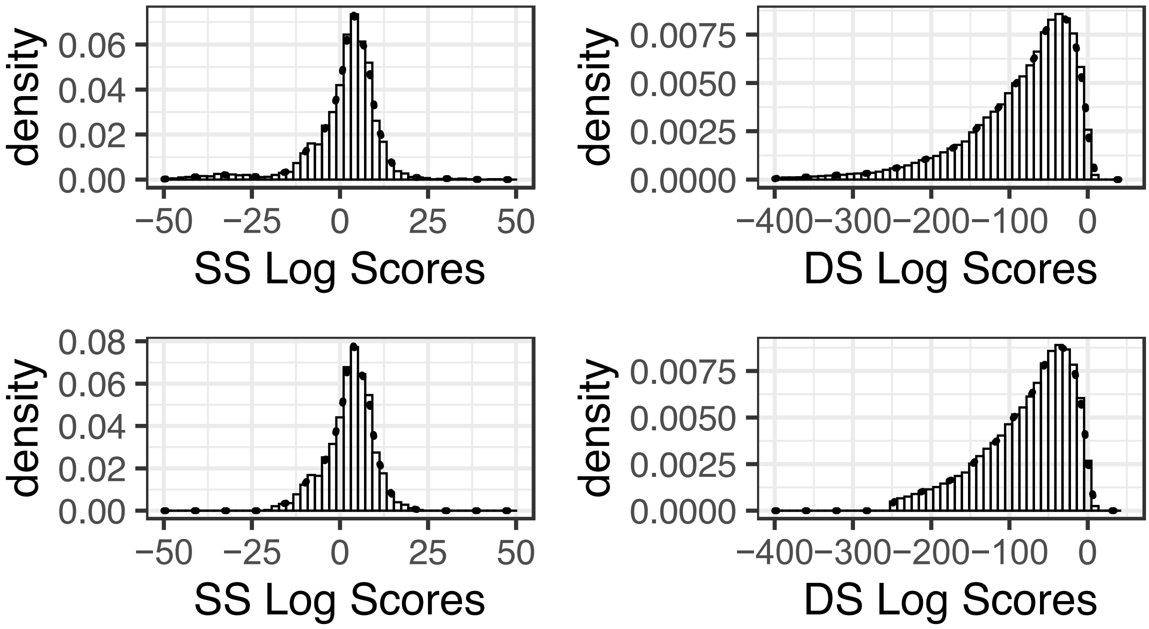


Figure 6.2 Histogram and density distribution of SS and DS Log scores from FP *um* by varying test, training and reference speakers. Upper and lower panels show the distributions of SS and DS scores before and after *z-score* was applied.

***um* log score**	**Mean**	**SD**	**Skewness**	**Kurtosis**
SS	2.6	6.6	-0.5	3.5
DS	-78	56.6	-0.9	3.1

Table 6.1 Distribution parameters of SS/DS log scores from the lower panels in Figure 6.2.

6.2.2 Score simulation and LR computation

The score simulation was carried out using parameters of log scores rather than the raw scores. This was done for two reasons. First, log scores rather than raw scores are normally used for *score-to-LR* computation, and thus it is more sensible to simulate log scores. Second, the raw scores are extremely skewed and less symmetrical. Raw scores only allow for non-negative values; therefore, both SS and DS raw scores are likely to be heavily tailed at the right. Further, the DS raw scores are likely to be stacked close to 0 at x-axis. Simulating the raw scores would further complicate the simulation process and introduce more uncertainty. However, for the sake of exploration, an attempt was also made to simulate raw scores, and the process is demonstrated in Appendix A.

Since the log scores are negatively skewed, it is sensible to simulate scores using skewed distributions. The Skew-t (ST) distribution (Arellano-Valle & Azzalini, 2013) was used for score simulation. The `rst()` function from the R (R Core Team, 2020) package `sn` (Azzalini, 2020) was used. In ST distribution estimation, the mean, standard deviation and kurtosis of SS and DS log scores were adopted from Table 6.1, while the skewness was varied to explore the overall performance as a function of different score skewness. However, the choice and range of skewness relies on the ST distribution model, where four direct parameters (DP) are required, namely ξ (location), ω (scale), α (shape) and ν (thickness at tail). DPs are not as interpretable as 'centred parameters' (CP) and need to be converted from CPs (traditionally known moments such as mean, variance and skewness). Meanwhile, not all the choice of components in CP (i.e., mean, variance and skewness) are admissible to those in DP (i.e., location, scale, shape and thickness at tail), meaning that DP has certain limitations on the possible ranges of the components in CP (see Arellano-Valle & Azzalini, 2013 for detailed discussion). In the current chapter, given the mean, standard deviation and kurtosis (Table 6.1), the maximum absolute value of skewness can be admissible to DP is 1.4. Table 6.2 shows the skewness values used for ST distribution simulation.

ST distribution	**Skewness**		**Kurtosis**		**Mean**		**SD**	
	SS	DS	SS	DS	SS	DS	SS	DS
Set (a)	0	0	3.5	3.1	2.6	-78	6.9	56.6
Set (b)	-0.7	-0.7						
Set (c)	-1.4	-1.4						
Set (d)	0	-0.7						
Set (e)	0	-1.4						
Set (f)	-0.7	0						
Set (g)	-1.4	0						

Table 6.2 Parameter values used for ST distribution simulation. For all the sets, kurtosis, mean and standard deviation were fixed across replications, while only skewness values were varied.

Seven sets of SS and DS skewness values were used for simulation, and only the skewness values were varied for each set. In sets (a), (b) and (c), both SS and DS score skewness were varied, aiming to investigate if overall performance would be degraded when scores were skewed (comparing with normal distributions, i.e., set (a)) in relation to sample size as well as different calibration methods. In sets (d), (e), (f), (g), only the skewness of one of the SS/DS scores was varied, aiming to explore the effect of score distribution mismatch on overall performance. This is because scores are modelled using normal distributions in most of the ASR systems. Figure 6.3 shows the simulated SS (top panel) and DS (bottom panel) log scores (1000 samples per set) by varying the skewness. The dashed vertical lines indicate the mean of the empirical SS and DS log scores, which are 2.6 and -78 respectively.

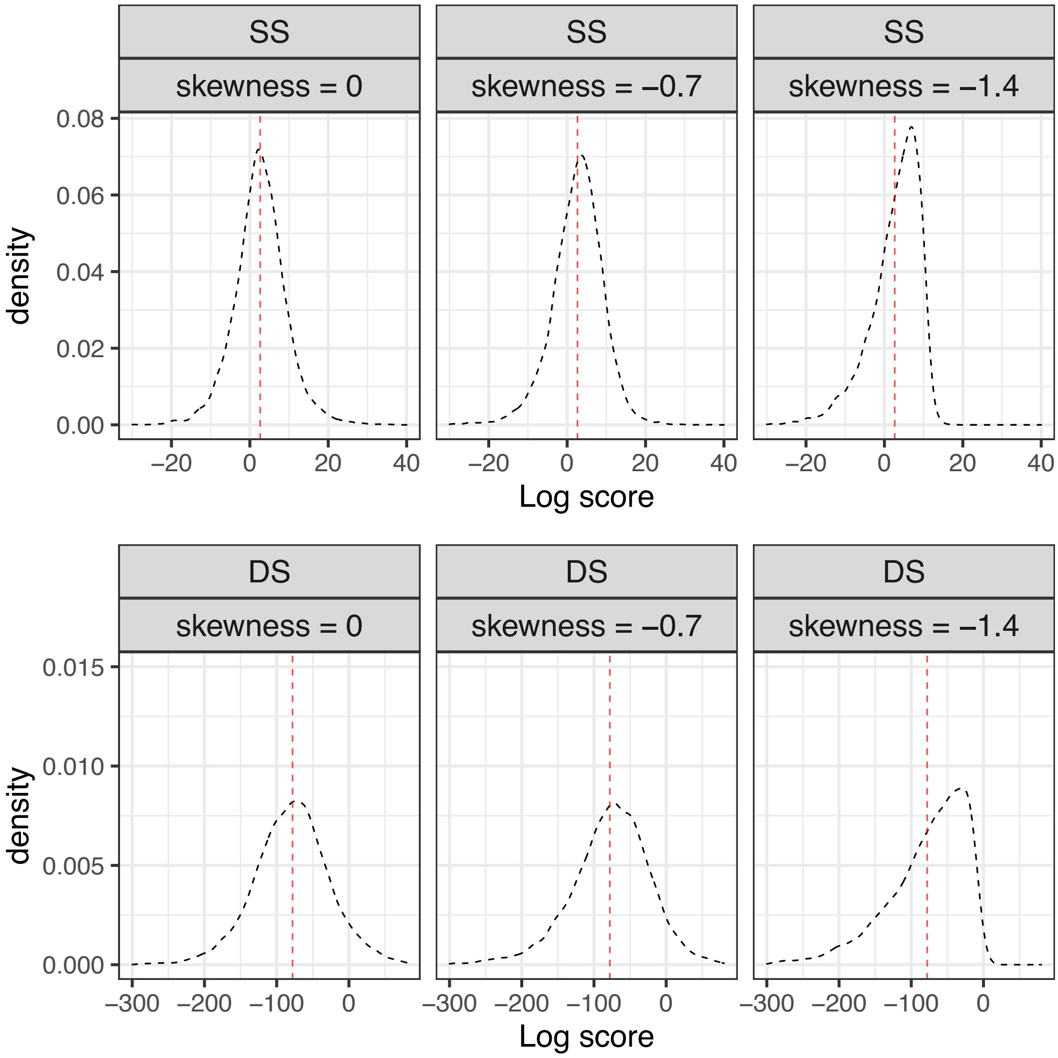


Figure 6.3 Distributions of simulated SS and DS log scores by varying the skewness.

For each set of parameters, the training and test SS and DS scores were sampled with increasing sample sizes, i.e., the number of training and test speakers was increased from 20 to 100 with a 10-speaker increase, i.e., the SS and DS log scores vary from 20 to 100 and 380 to 9900 for training and test data respectively. In order to explore the effect of sampling variability on overall performance when scores have different distributions, the experiment was replicated

100 times within each sample size using independent samples of scores. In this way, the experiments allow us to explore the relationship between sampling variability and sample size as well as which calibration methods are more or less resistant to sample size and sampling variability. A simplified simulation process is shown below in Figure 6.4.

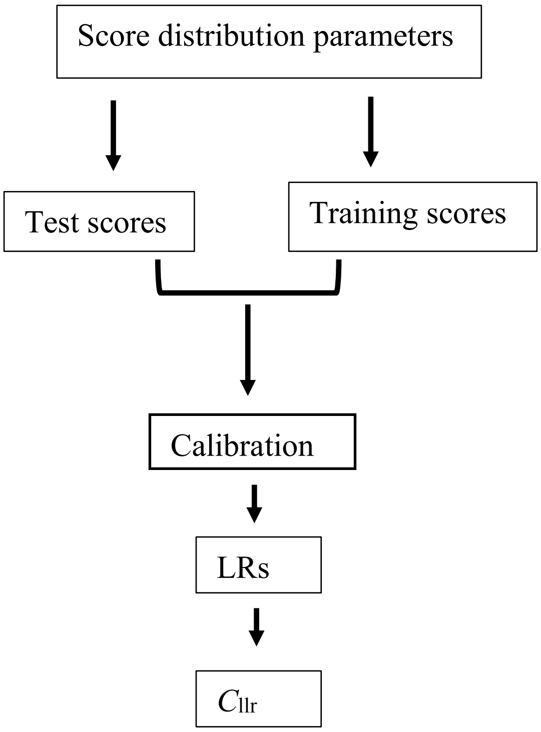


Figure 6.4 Simplified process of *score-to-LR* computation. Test and training scores are sampled respectively for calibration to generate LRs and C_{llr}. The process was replicated 100 times using varying skewness, calibration methods and sample sizes.

The simulated training scores were then used to train calibration models, which were applied to the test data from which system validity was evaluated. Four calibration methods are involved in current chapter, i.e., logistic regression (Brümmer et al., 2007; Morrison, 2013), regularised logistic regression (rLogistic regression) (Morrison & Poh, 2018), empirical lower and upper bound (ELUB, Vergeer et al., 2016) and Bayesian model (Brümmer & Swart, 2014) aiming to explore if certain calibration methods are more or less susceptible to the effect of sampling variability and sample size when the assumption of normality is violated (see Section 3.3.2 for detailed rationale behind these four calibration methods). The overall performance was evaluated using the C_{llr} mean and range (i.e., the difference between the maximum and minimum C_{llr} values across 100 replications).

6.3 Results

6.3.1 Varying the skewness of both SS and DS scores

Figure 6.5 shows the variation in the mean and range of C_{llr}s varying the skewness of both SS and DS scores using different calibration methods. The x-axis indicates the number of speakers used in training and test data respectively and the y-axis represents the C_{llr}. The legends on top indicate the skewness of the score distributions. The dashed lines indicate the C_{llr} range and the circles, triangles and squares are the mean C_{llr}. The general patterns across the four calibration methods show that regardless of the score distribution skewness, the more speakers used, the lower the C_{llr} range is, indicating that including more training and test speakers reduces the sampling variability and improves the system reliability. However, the patterns of the C_{llr} range vary within each calibration method in relation to score skewness and sample size.

Overall, the rLogistic regression yields the most reliable system in terms of C_{llr} range (dashed lines) followed by Bayesian model, logistic regression and ELUB. The C_{llr} range stays stable (ca. 0.1 or lower) regardless of sample size and score skewness for rLogistic regression, and the C_{llr} range for Bayesian model starts to stabilise when the sample size reaches 30 speakers per training and test sets (the C_{llr} range is as low as 0.1). On the other hand, the logistic regression and ELUB calibration methods are more sensitive to score skewness and sample size. For logistic regression, the C_{llr} range varies from ca. 0.25 to ca. 0.05 when sample size increases from 20 to 100 speakers and starts to stabilise when the sample size reaches 40 speakers (per training and test set) when SS/DS score skewness is no higher than -0.7. However, the C_{llr} range shows a much higher fluctuation when the score skewness is -1.4, varying between ca. 0.5 and 0.15 (blue dashed lines) when sample size increases from 20 to 100 speakers. Meanwhile, the ELUB calibration method seems to be the most sensitive to score skewness, sample size and sampling variability. All three dashed lines in ELUB calibration method show marked variability compared with the other three calibration methods.

In terms of system validity, logistic regression consistently yields the lowest mean C_{llr} for each sample size condition given score skewness followed by rLogistic regression, Bayesian model and ELUB. However, logistic regression seems to be more sensitive to score skewness (i.e., *not* sample size). The mean C_{llr} varies from ca. 0.45 when SS/DS are normally distributed, to ca. 0.39 when SS/DS score skewness equals -0.7 and to ca. 0.15 when the SS/DS score skewness equals -1.4. The similar pattern of mean C_{llr} values is also observed in the rLogistic regression and ELUB calibration methods, i.e., the more skewed the scores, the lower the mean C_{llr}. However, the mean C_{llr} stays comparatively consistent (ca. 0.55) using the Bayesian model regardless of the score skewness and sample size. For the four calibration methods, the mean C_{llr} also stays comparatively stable within each score skewness across different sample size; however, one exception is observed using rLogistic regression. The mean C_{llr} reduces with larger sample size and higher skewness for rLogistic regression, varying from ca. 0.5 to ca. 0.25 when SS and DS skewness equals to -1.4 (blue squares).

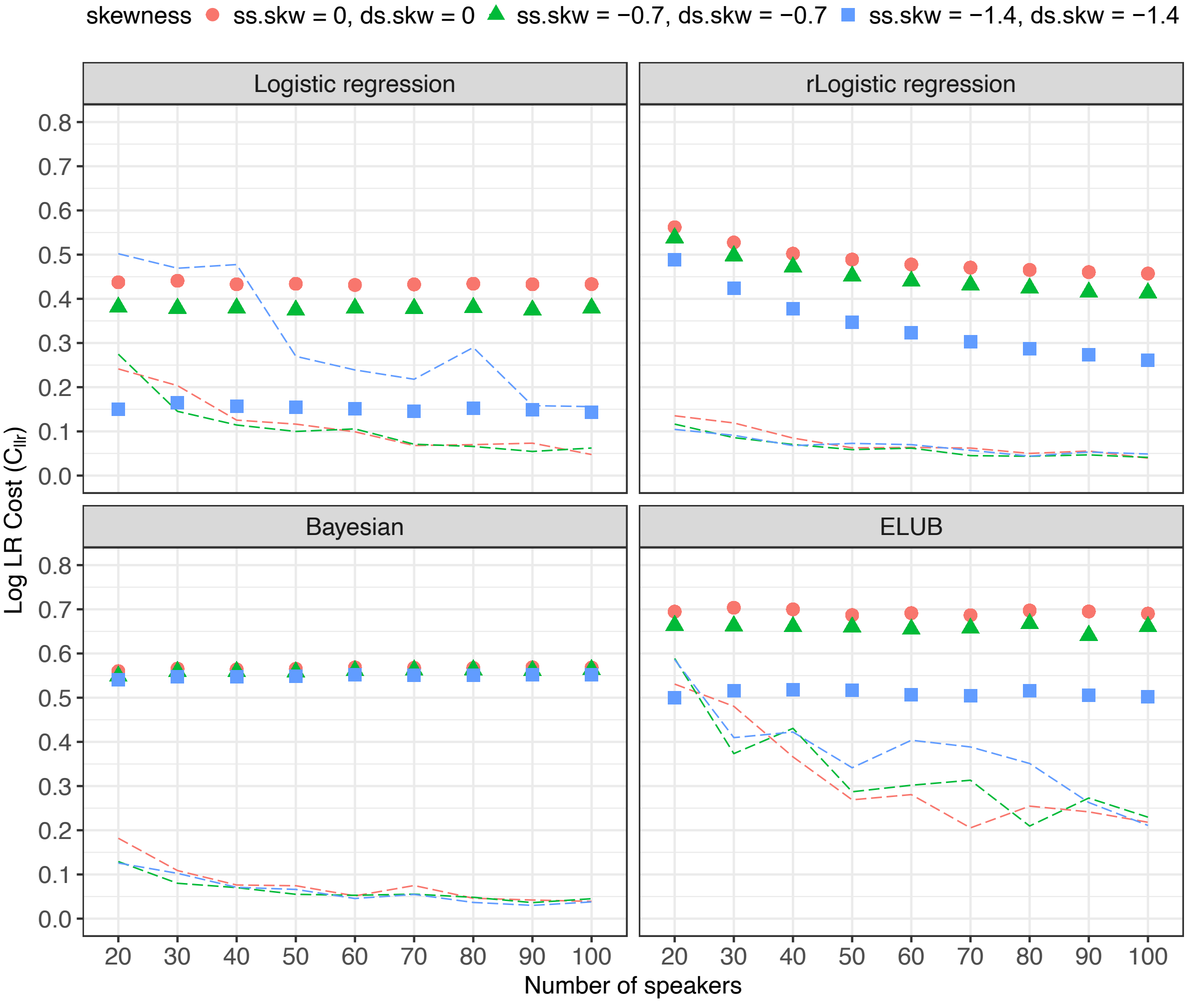


Figure 6.5 Mean (coloured circles, triangles and squares) and range (coloured dashed lines) of C_{llr} values varying the skewness in both SS and DS scores. Each panel shows the performance using different calibration methods across different sample sizes (x-axis). Different colours represent different score skewness values.

6.3.2 Varying only SS score skewness

Figure 6.6 shows the C_{llr} variation when only varying the SS score skewness, while the DS score skewness is kept as 0. In terms of system reliability, rLogistic regression seems to yield the best reliability, followed by the Bayesian model, logistic regression and ELUB. The C_{llr} range starts to stabilise when sample sizes reach 50, 30 and 30 speakers for logistic regression, rLogistic regression and Bayesian model respectively regardless of SS score skewness, and the C_{llr} range is as low as 0.1. On the other hand, overall performance is much more fluctuating using ELUB calibration method. On average, the C_{llr} range varies from ca. 0.55 to ca. 0.25 when sample size increases from 20 to 100 speakers (per training and test sets). In terms of system validity, logistic regression consistently yields the lowest mean C_{llr} (ca. 0.45) across different SS score skewness and sample size, followed by rLogistic regression (ca. 0.5), Bayesian model (ca. 0.55) and ELUB (ca. 0.68). A comparison between Figures 6.5 (i.e., varying both SS and DS score skewness) and 6.6 shows that SS score skewness and sample size do not have much effect on system validity using all four calibration methods as long as DS scores follow the normal distribution. Meanwhile, SS score skewness does not have much effect on system reliability when logistic regression, rLogistic regression and Bayesian model are used for calibration as the dashed lines overlap to a large extent in these three calibration methods. However, the ELUB method is much more susceptible to SS skewness than the other three calibration methods.

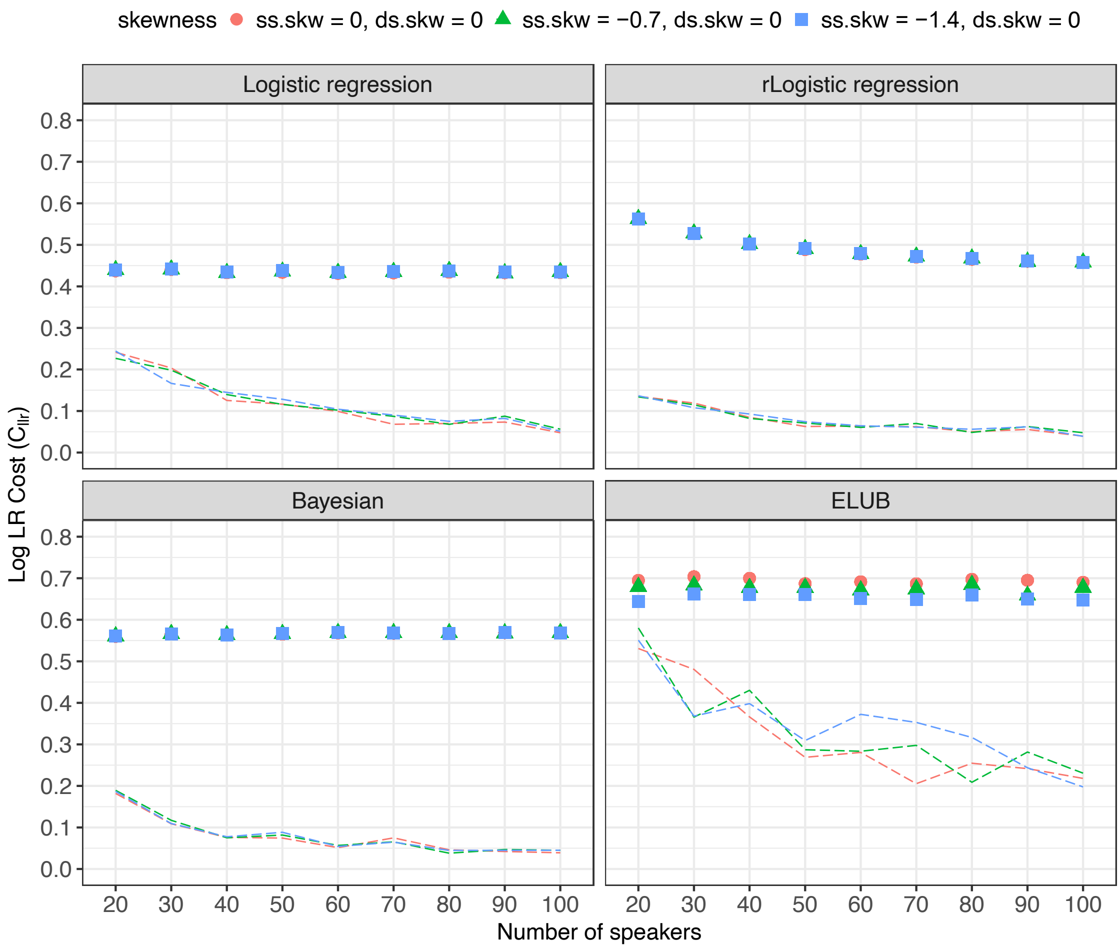


Figure 6.6 Mean (coloured circles, triangles and squares) and range (coloured dashed lines) of *C*llr values only varying the SS score skewness. Each panel shows the performance using different calibration methods across different sample sizes (x-axis). Different colours represent different score skewness values.

6.3.3 Varying only DS skewness

Figure 6.7 shows the *C*llr variation by only varying the DS score skewness, while the SS scores are kept as normal distributions. In terms of system validity, logistic regression consistently yields the lowest mean *C*llr given DS score skewness across different sample sizes, followed by rLogistic regression, Bayesian model and ELUB. Meanwhile, logistic regression is also the most sensitive to DS score distributions, where the mean *C*llr varies between ca. 0.45 and ca. 0.1 when DS score skewness increases from 0 to -1.4. In terms of system reliability, again, logistic regression and ELUB are more sensitive to sample size and DS score skewness than rLogistic regression and Bayesian model, indicated by the distance between dashed lines.

The patterns of C_{llr} mean and range in Figure 6.7 (i.e., only varying DS score skewness) using logistic regression, rLogistic regression and Bayesian model are similar to those in Figure 6.5 (i.e., varying both SS and DS score skewness), reassuring us that the variability observed is mainly caused by skewness in DS scores. However, the C_{llr} range in Figure 6.7 using ELUB shows a different pattern from that in both Figures 6.5 and 6.6, indicating that the variability in C_{llr} range using ELUB is mainly caused by skewness in SS scores.

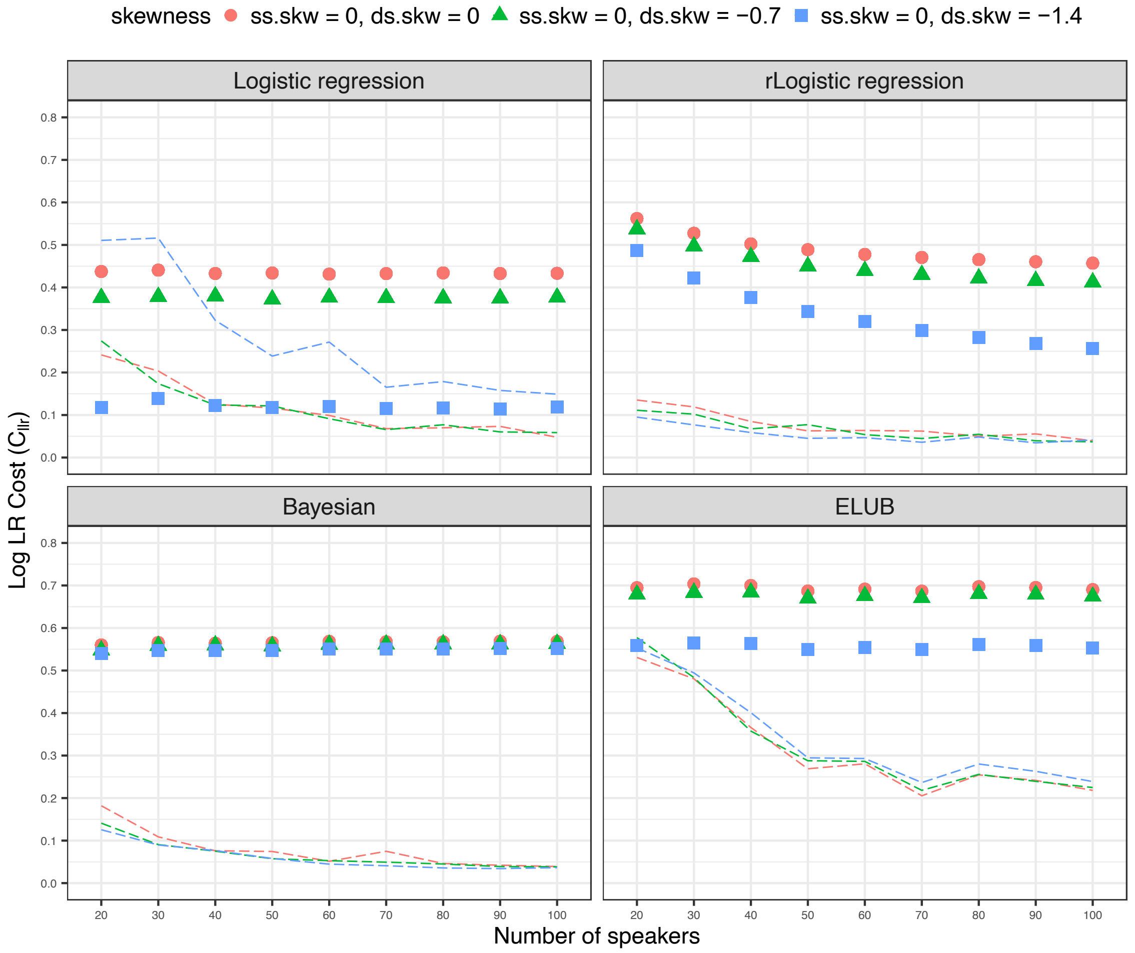


Figure 6.7 Mean (coloured circles, triangles and squares) and range (coloured dashed lines) of C_{llr} values only varying the DS score skewness. Each panel shows the performance using different calibration methods across different sample sizes (x-axis). Different colours represent different score skewness values.

6.4 Discussion

The results of this chapter have shown that overall performance varies to different extents using different calibration methods with different score skewness and different sample sizes. In general, the ELUB calibration method is least preferable as it produces systems that are more sensitive to sampling variability, sample sizes and score skewness than the other three methods. Moreover, Figures 6.5 to 6.7 show that the mean system validity (mean C_{llr}) is mainly affected by skewness in DS scores, while system reliability (C_{llr} range) is mainly affected by skewness in SS scores. Although the C_{llr} range using ELUB reduces when more training and test speakers are included, there remain high levels of C_{llr} variability even with large samples; the C_{llr} range with 100 speakers is equivalent to that produced by rLogistic regression and Bayesian model when using only 20 speakers.

Using 1 as an appropriate threshold for judging C_{llr} (Morrison et al., 2021) (i.e., a good system should yield a C_{llr} as close as to 0, and C_{llr} higher than 1 indicates that the system is not giving any useful information), the wide C_{llr} range using ELUB calibration method suggests that it is the least effective of the four calibration methods. The ELUB calibration method produces C_{llr}s of over 1 across replications, even with 100 speakers in each set. The Bayesian model is less affected by sampling variability, sample sizes and score skewness and should generally be preferred, especially when sample size is small. The disadvantage, however, is that priors (see Section 3.3) need to be pre-specified when using the Bayesian model. The priors within the Bayesian model need to be specified based on the mean and variance of the training data, which could be different from case to case in the real world. Comparatively, the system reliability is less affected by sampling variability and sample size using rLogistic regression; however, the system mean validity (mean C_{llr}) is affected by sample size and score skewness, especially when DS scores are skewed. Higher DS score skewness and larger sample sizes lead to lower mean C_{llr} using rLogistic regression. Similar to the Bayesian model, a κ^{ψ} value (see Section 3.3) needs to be specified using rLogistic regression, and different κ^{ψ} values of the rLogistic regression method need to be specified depending on the purpose of calibration, i.e., lower κ^{ψ} values deal with complete separation issues and higher κ^{ψ} values deals with extreme LR output issues (Morrison & Poh, 2018). For using logistic regression, system validity and reliability are more affected by score skewness, especially when DS scores are skewed, and should not be preferred when high skewness value is observed in DS score distributions.

In real world FVC, we might be dealing with small sample sizes – especially when using linguistic features, given the significant challenges around data collection and analysis (Gold & Hughes, 2014). The results of the current chapter show that logistic regression consistently yielded lower C_{llr} mean but higher C_{llr} range than rLogistic regression and the Bayesian model when using smaller numbers of training and test speakers. The C_{llr} mean becomes even lower and C_{llr} range becomes higher when score skewness is higher. It is therefore extremely important to understand the trade-off between C_{llr} mean and C_{llr} range, i.e., how much variability is allowed given validity (C_{llr}) and should we aim for lower C_{llr} mean (higher validity) as long as the system reliability (C_{llr} range) varies within a certain range? Ultimately, it is the author's opinion that experts' decisions should be driven by reducing uncertainty, rather than the absolute validity (i.e., the potential of a very low C_{llr}). Although it is difficult to set a generalised trade-off framework for all cases in the real world given the complexity and uniqueness of each individual case, results from the current thesis suggests that systems need to be tested multiple times with different sets or configurations of training and test data before applying it in real cases.

6.5 Chapter summary

The current chapter investigated the overall performance as a function of score skewness, sample sizes and calibration methods. The results suggest that score skewness has a marked effect on system performance (especially when sample size is small) and using regularised logistic regression and Bayesian model can shrink the LR output and reduce the degree of uncertainty when sample size is small. The key results are summarised using bullet point below,

- Overall performance is least affected by score skewness and sample sizes using the Bayesian model.
- Using rLogistic regression, the mean system validity is more affected by skewness in DS scores, while the mean system reliability is generally not affected.
- Using logistic regression, the mean system validity and reliability are affected by skewness in DS scores, especially when the DS score skewness is high (i.e., -1.4 in current chapter).
- Using ELUB, the mean system validity is more affected by skewness in DS scores, while the mean system reliability is more affected by skewness in SS scores.

Chapter 7 Discussion and Conclusion

The results presented in Chapters 4 to 6 have shown that sampling variability can be introduced at both *feature-to-score* and *score-to-LR* stages and that both can affect overall system performance and LRs for individual comparisons substantially. Moreover, empirical results from this thesis have important implications for issues of uncertainty, decision-making throughout the evaluation process, and subjectivity and objectivity in data-driven LR-based FVC. Specifically, they provide new insights on how best to go about testing and validating data-driven LR-based FVC systems in casework in order for the results to be useful to both the expert, and more importantly, to the trier-of-fact. This chapter first discusses sampling variability introduced at both stages in relation to findings in Chapters 4 to 6. The implications for research, casework and future work are then discussed in Sections 7.3 and 7.4.

7.1 Sampling variability at *feature-to-score* stage

The results of Chapter 4 show that that considerable variation can be observed in overall performance if different configurations of training, test and reference speakers are used (RQ 1a) and the overall performance is primarily caused by different configurations of test speakers. Chapter 4 shows that C_{llr} values vary from 0.6 to 0.97, 0.29 to 1.15 and 0.13 to 1.22 for /a/, /haijat/ and *um* respectively across 100 replications when varying the speakers in all three sets. The variability in overall performance is therefore as wide as the variability one would expect to see between variables and between populations; the difference between what would be considered very good performance versus very bad performance. Therefore, these results show that caution should be exercised when judging the speaker discriminatory power of a variable based on a single configuration of speakers in the training, test and reference sets. This is especially true since many data-driven LR-based studies use the same number of (or fewer) speakers than were used in Chapter 4. Using different, but still representative, speakers could affect overall performance substantially, depending on who exactly those speakers are. However, further examination of the results from the final three experiments in Chapter 4 show that variability in overall performance is almost exclusively due to the effects of varying the test speakers. It is noted that *um* produced a higher C_{llr} range in Experiment 4 (varying training speakers) than in Experiment 2 (varying test speakers) due to extreme outliers. By comparison, the variability in C_{llr} as a function of the configurations of the reference and training sets was fairly small. Thus, as long as the sets are of a sufficient size (in these experiments, over 25 speakers), the specific speakers used for the training and reference data have little effect on the overall performance of the system.

Chapter 5 explored how overall performance (RQ 2a) and individual speakers' LR (RQ 2b) are affected by different combinations of linguistic-phonetic features as well as how individual speakers are affected by different configurations of training and reference speakers (RQ 2c). The results show that given a sufficient sample size and the training and reference speakers are sampled from the relevant population, sampling variability affects overall performance to different extents using different combinations of features. Using FP *um,* systems which include F2 in general outperform systems without F2. Moreover, combining more features does not necessarily improve the overall performance. For individual speakers' behaviour, sampling variability has a limited effect on individual test speakers in SS comparisons (within one LLR magnitude in terms of strength of evidence, indicated by RMSD values; Figure 5.11), while individual test speakers are more sensitive to sampling variability in DS comparisons. However, most of the individual test speakers tend to be less affected by sampling variability in DS comparisons when four or more features are used (Figure 5.12).

The results of Chapters 4 and 5 suggest that sampling variability at the *feature-to-score* stage is multi-dimensional. First, overall performance depends on exactly *which* speaker is used in *which* set (especially for test speakers). Second, the effect of sampling variability on overall performance varies depending on which features or combinations of features are being used. For FP *um*, F2 and systems with F2 involved outperform those without. Third, sampling variability affects individual test speakers depending on the number of features being used, e.g., Chapter 5 shows that the LR outputs for individual speakers are more stable when four or more features are used.

7.2 Sampling variability at *score-to-LR* stage

Assuming that sampling variability can be controlled or predicted at the *feature-to-score* stage, Chapter 6 aims to investigate if overall performance is affected by score skewness (RQ 3a) and if the degree of uncertainty can be reduced using certain calibration methods when sample size is small (RQs 3b and 3c). The results show that sampling variability affects overall performance to different extents in relation to different sample sizes and when score distributions violate normality assumption. Section 6.2.3 reveals that the mean system validity (*C*llr mean) is more affected by DS score skewness, while the system reliability (*C*llr range) is more affected by SS score skewness. However, using different calibration methods can potentially reduce the effect of sampling variability to different extents by shrinking the LRs, e.g., the Bayesian model calibration method is the most robust to sampling variability, as both *C*llr mean and range stay stable across different sample size and score skewness. Meanwhile, the *C*llr mean is more susceptible to sampling variability using rlogistic regression (while the *C*llr range is not much affected) across different sample sizes, especially when DS scores are skewed. Using logistic regression for calibration, the *C*llr mean is more affected by sampling variability across different score skewness (especially when DS scores are skewed), while the *C*llr range is more sensitive to sampling variability across both score skewness and sample sizes. Similarly, the *C*llr mean is more affected by sampling variability when DS scores are skewed (rather than SS score skewness or sample sizes) using ELUB, while the *C*llr range is more affected by sampling variability across both different samples sizes and when SS scores are skewed.

The patterns of *C*llr mean and range across different sample sizes and score skewness suggest that it is important to acknowledge the trade-off between system validity and reliability. Moreover, score distribution skewness needs to be taken into consideration in system evaluation, as sampling variability does not have a fixed effect on the overall performance when scores are not normally distributed. The logistic regression yields the lowest and stable mean *C*llr across different sample sizes given skewness; however, the *C*llr mean and range varies to a large extent when DS scores are skewed. On the other hand, the Bayesian model yields higher *C*llr mean across samples sizes and score skewness; nevertheless, the *C*llr mean and range stay stable. Under a real case scenario, the mean *C*llr obtained using logistic regression is likely to be misleading in system evaluation due to the variability in LR outputs when DS scores are skewed to different extents. Ultimately, the goal in system evaluation (for FVC and other forensic evidence) is to reduce the uncertainty under different case conditions, rather than aiming for the lowest *C*llr.

7.3 Direct Implications

This section discusses the direct implications of the current study for data-driven LR-based FVC. The general finding (of Chapter 4) relating to C_{llr} variability when using different sets of training and reference speakers seems to be positive for casework. The training and reference sets are the core elements of any system; the test data are intended to be 'unseen' representative speakers to assess how well the system performs, and are not part of the system *per se*. The fact that overall performance tested in the thesis are so insensitive to different speaker configurations in the training and reference sets means that the expert can be relatively confident about the transferability. However, in the real world we would expect more variability due to low-quality and non-contemporaneous forensic data, channel mismatch, different stylistic and situational contexts etc. Moreover, the data used in the current thesis was not fully independent across replications, and whether these results would extend to more real-world conditions is an empirical question.

However, the key source of uncertainty in overall performance derives from the configurations of the test set. Therefore, it is essential that the expert carefully considers the speakers that are used for testing, since this can have a substantial effect not only on the system that an expert decides to use in a case, but it may also lead to over- or under-estimation of the true validity to the court, potentially leading to incorrect decisions being made by the trier-of-fact. One potential way to deal with this issue is to use data that are more representative of the voices in the case. The issue here is not one of finding the best configuration of test speakers to produce the optimal performance, but rather to find a set of test speakers that produces a validity measure that is representative for the case. It has been argued that systems should be evaluated using recordings that reflect the conditions of the case at trial (Enzinger et al., 2016; Enzinger & Morrison, 2017). In terms of speaker characteristics, this is taken to mean that the speakers used are representative of the relevant population, often defined broadly by sex and language (Rose, 2004). Clearly, based on the variability reported in Chapters 4 and 5, this alone is insufficient, especially where the variable(s) can potentially provide good speaker discrimination and/or the number of speakers is small. Chapter 5 suggests that there is no single rule that applies to all, and there are always exceptions in individual speakers no matter what generalisations are drawn. Not only for speech evidence, but also for other biometric systems, individual behaviour is system specific. Using linguistic features, the variation in individual test speakers' LLR outputs observed in Chapter 5 is mainly due to sampling variability (or random effects) in the training and reference speakers from the relevant population. However, the role of random effects in individual speakers' behaviour remains unknown. Identifying a subset of a database of speakers who are in some way 'more similar' to the offender (akin to the suggestion in Morrison et al., 2012, and procedures in some automatic systems) is likely to produce more representative results. However, this involves making pragmatic but subjective decisions: either based on narrower demographic properties of the offender (although, of course, we cannot know these properties for certain, since the identity of the offender is the very question at stake) or using some measure to define speaker similarity. This is not necessarily problematic. As highlighted by the court in R v T [2010] "the probability that is quoted (by the expert as a conclusion in a case) … will inevitably be a personal probability and the extent to which the data influence that probability will depend on expert judgement" (at para 80). Having tailored test data is, in the author's view, the preferred approach to reducing the uncertainty in the performance of a system and LR outputs of individual test speakers. This allows the expert to have a better sense of how the system will perform in the specific case. However, as highlighted in Hughes and Foulkes (2015) and discussed in Section 1.4, there will always be some mismatch between the data used for building and evaluating a system and the case data. It is likely, therefore, to be fruitful to examine ways to further reduce uncertainty by incorporating it into the LR computation itself (e.g., Brümmer, 2013); examples based on sample size, score distribution and calibration methods are demonstrated in Chapter 6 showing that different calibration methods need to be considered when scores are skewed and sample sizes are small. Alternatively, a minimal requirement might be that both researchers and experts undertake speaker sampling of the kind described in this study (i.e., run replications) in order to understand the potential range within which a system performs. This may not provide case-specific information, but will provide insights into how certain we can be about the performance of a system in general. For instance, the range of values produced for *um* in this study means that we would need to be extremely cautious about making generalisations about speaker discriminatory power or the usefulness of such a system in casework. Ultimately, researchers should aim to answer the question of the reliability of overall performance in the real world instead of the reliability of overall performance based on data samples at hand (i.e., training, test and reference data).

7.4 Wider Implications

In all forms of FVC studies, experts have the degrees of freedom to make decisions in order to conduct an analysis, e.g., data-driven/non-data-driven LR-based or auditory acoustic based. Such decisions can affect the overall result in the same way as the data-driven methods used in this thesis. It is, therefore, crucial for experts to recognise and acknowledge where there is subjectivity and uncertainty within the process. On the basis of the variability in the current thesis, it is the author's opinion that experts' decisions should be driven by reducing uncertainty in evidence evaluation rather than trying to maximise discrimination or the potential of producing a high validity (i.e., a very low C_{llr} in current context; e.g., Rose, 2013a). Using data-driven approaches allow us to explicitly measure/describe and possibly deal with uncertainty. For example, Chapter 4 demonstrated the variability in system output caused by different configurations of training, test and reference speakers, Chapter 5 showed how systems and individual speakers were affected using different linguistic-phonetic features, and Chapter 6 explored possible calibration methods to reduce the degree of variability in system output by incorporating uncertainty into LR computation. However, the challenges in the data-driven approach lie in the fact that implementation requires complex mathematical models and data. Meanwhile, the data-driven approach also involves the issue of explaining and interpreting systems, a general challenge for machine learning (Molnar, 2019), and results to an end-user (e.g., the court).

Uncertainty does not go away just because a data-driven approach isn't employed. In auditory acoustic analysis, experts can assess the typicality based on one's expertise or relevant literature of the regional accent involved. This would result in different degrees of uncertainty in typicality assessment depending on the experts' knowledge/experience about that particular regional accent or how relevant the literature is (e.g., year of publication, journals or conference proceedings, whether peer-reviewed or not). When it comes to acoustic analysis of the similarity between suspect and offender samples, different choices of linguistic-phonetic features and methods for measurement would lead to different evaluation results. This has been shown by Roettger (2019) in quantitative-based phonetic studies and in the current study as well. It is then important to acknowledge uncertainty and subjectivity, and whether compensation is made for that in some way, e.g., the analyst attaches less weight to a variable when the number of tokens is small. Future work is needed to deal with the issue of uncertainty more systematically.

7.5 Future studies

The current study has explored the effect of sampling variability in data-driven LR-based FVC at different stages, namely *feature-to-score* and *score-to-LR*. At the *feature-to-score* stage, it has shown the importance of acknowledging and reporting uncertainty, specifically in the test set, which can be established either by using a more tailored subset of test speakers or, minimally, reporting the range of values produced through speaker sampling of the sort described in Chapters 4 and 5. In this way, the analyst can provide an estimate of the range of validity values the system can produce, and thus provide a means to record the uncertainty in the LR calculations. Providing such information is critical for the trier-of-fact to evaluate the evidence provided by the expert. Chapter 5 also reveals that sampling variability affects both system and individual speakers to different extents using different combinations of linguistic-phonetic features. Therefore, caution should be exercised when giving generalisations of overall performance and individual speakers' behaviour based on a single set of features. More importantly, the experiment procedure in Chapter 5 can be used as a starting point for the analysis of individual speakers' behaviour. Taking automatic speaker recognition systems for example, same set of speakers can be firstly tested under different system settings (e.g., using different features, different statistical models). Then, individual speakers can be grouped (using zoo plot or any other kind) based on LRs which give insight into which speaker is more or less sensitive to different settings of the system. Specific speakers can then be located for further analysis/investigation, e.g., why does some speakers have stable behaviour under different system settings and others do not?

Chapter 6 investigated the overall performance focusing on the *score-to-LR* stage, it has been shown that the overall performance is affected by sampling variability to different extents in relation to different score skewness and sample sizes. Using the Bayesian model and rlogistic regression can potentially reduce the variation in LR output; however, priors and κ^{ψ} values need to be pre-specified arbitrarily using these two models. One solution for that is to empirically test a range of priors and κ^{ψ} values that reflects different case conditions. Considerably more work is required reduce the degree of uncertainty as well as to incorporate uncertainty into LR computation, and this could be implemented at both *feature-to-score* and *score-to-LR* stages. As demonstrated in the current thesis, different calibration methods can be used to reduce the degree of uncertainty at *score-to-LR* stage, and Bayesian model and regularised logistic regression can effectively shrink LRs when sample size is small. However, future studies should consider the underlying rational of using different priors for calculating LRs (note that one shall not confuse the prior used for calculating the LRs with the *prior odds* for calculating the posterior odds in the Bayes' theorem.) Taking Bayesian model for example, only the Jeffrey prior was used in the current thesis; however, there are also the Haldane prior and Laplace prior and Brümmer (2011) has shown that using Haldane prior is appropriate for DNA evidence. While this is outside the scope of the current study, future studies should investigate the mechanism behind different priors and the suitability of using different priors for speech evidence. To incorporate uncertainty into LR computation at the *feature-to-score* stage, Ramos et al (2021) proposed two models, i.e., heavy-tailed with warped Gaussian mixtures model and heavy-tailed with variational autoencoder model, which incorporate uncertainty into LR computation at the feature space for glass comparisons. They have demonstrated that their proposed models outperform Aitken & Lucy's (2004) MVKD model which does not incorporate uncertainty into LR computation. Future studies can potentially investigate the feasibility of applying models proposed by Ramos et al. for speech data.

7.6 Conclusion

The aim of this study is to investigate the effect of sampling variability on overall performance and individual speakers' behaviour in data-driven LR-based FVC. Numerous studies (e.g., Gwo & Wei, 2016; Morrison, 2016; Morrison & Enzinger, 2016) have discussed the validity and reliability issues in forensic evidence evaluation; however, the variability in LR output observed in the current study has not previously been addressed in the field of FVC. Specifically, the performance of overall system is markedly affected by sampling variability as well as the use of different input features. The results have shown that LR output is dependent on how speakers are arranged in the training, test and reference samples and which linguistic-phonetic features are used. Individual speakers' behaviour also fluctuates depending on the specific linguistic-phonetic features and training and reference speakers used. However, this is not to suggest that similar patterns would be expected using other systems. It is possible that more stable overall performance and individual speakers' behaviour could be observed if high dimensionality features are used (e.g., MFCC). Given that all training, test and reference speakers are sampled from the relevant population following certain database selection guidelines (e.g., Morrison et al., 2012), the overall performance and individual speakers' behaviour could still vary to different extents. On one hand, the variability observed in overall performance and individual speakers' behaviour is partially due to statistical issues (e.g., sample size, data extrapolation). However, a bigger question for forensic phoneticians is whether there are any systematic linguistic patterns that can be observed to predict or reduce the variability in LR output, e.g., whether a more tailored subset of the relevant population can be selected based on systematic linguistic patterns that would reduce the variability in LR output. It is hoped that this thesis can provide some practical guide for FVC casework and help to improve the validation procedure, i.e., the overall performance should be tested multiple times under different conditions (e.g., using different sets/configurations of training and reference data) and the individual speakers' behaviour should be investigated and reported as part of the system testing.

Appendix

Appendix A - Simulating raw scores

Attempts were made to simulate raw scores and this section demonstrates the simulation process. Table A.1 shows the mean, standard deviation, skewness and kurtosis of SS and DS raw scores. Probable theoretical distribution candidates were fitted to raw scores using Maximum likelihood estimation (MLE), where the estimated distribution parameters were evaluated using *goodness-of-fit* statistics and criteria.

um	Mean	SD	Skewness	Kurtosis
SS	5.63e+05	3.09e+06	7.81	59.51
DS	0.77	29.28	59.52	4121.94

Table A.1 Summary statistics of SS and DS raw scores.

MLE uses a likelihood function to estimate the distribution parameters given data, and MLE is when the likelihood function finds the parameter values that maximise the likelihood of obtaining the observed data. A generalised likelihood function is expressed in Equation (A1);

$$\Lambda(\theta) = \prod_{i=1}^{N} f(y_i; \theta) \qquad \text{(Equation. A1)}$$

(Eliason, 1993)

where y_i is the observed data and θ stands for distribution parameters. It is worth noting that MLE aims to estimate the parameters of the 'true' distribution given data, rather than the distribution of the observed data. Before the distribution parameters can be estimated, possible theoretical distribution candidates need to be selected. Table A.1 shows that raw scores are both right skewed. Therefore, right skewed distributions seem to be good candidates for parameter estimation. However, gamma, Weibull, lognormal and exponential are all theoretically right-skewed distributions. In order to eliminate less probable distributions that the empirical data might come from, the Cullen and Frey (CF) graph is used here. The CF graph uses kurtosis and the square of skewness to describe the distributions from empirical data among a set of theoretical distributions. It is worth noting that the CF graph is used for indicative purposes only, not the selection of the most probable distribution. Figure A.1 shows the CF graph of SS and DS raw scores. The blue dots in the graph show the location of the empirical data, and the yellow circles are bootstrapped values of kurtosis and square of skewness. Bootstrap sampling is applied to take uncertainty into consideration in the estimation of kurtosis and skewness from empirical data (Efron & Tibshirani, 1994), because kurtosis and skewness are not robust moments. The bootstrap sampling process is carried out by random sampling from empirical data with replacement (See Delignette-Muller and Dutang, 2015, p. 5 for details).

For both raw SS and DS scores, the normal distribution is at the top left corner, which is far away from the empirical data (raw SS and DS scores). This confirms that the normal distribution is one of the least probable theoretical distributions that the empirical data might come from. Similarly, the exponential distribution does not seem to be a close candidate for the empirical data either. On the basis of GF, gamma and Weibull distributions are the two most probable theoretical distributions of the empirical data that might be observed. Therefore, gamma and Weibull are selected for parameter estimation using MLE. It is noted that the empirical data does not lie exactly on any of the theoretical distributions. However, the aim here is to eliminate less probable theoretical distributions that the empirical data might come from, rather than selecting the 'right' theoretical distribution for the empirical data.

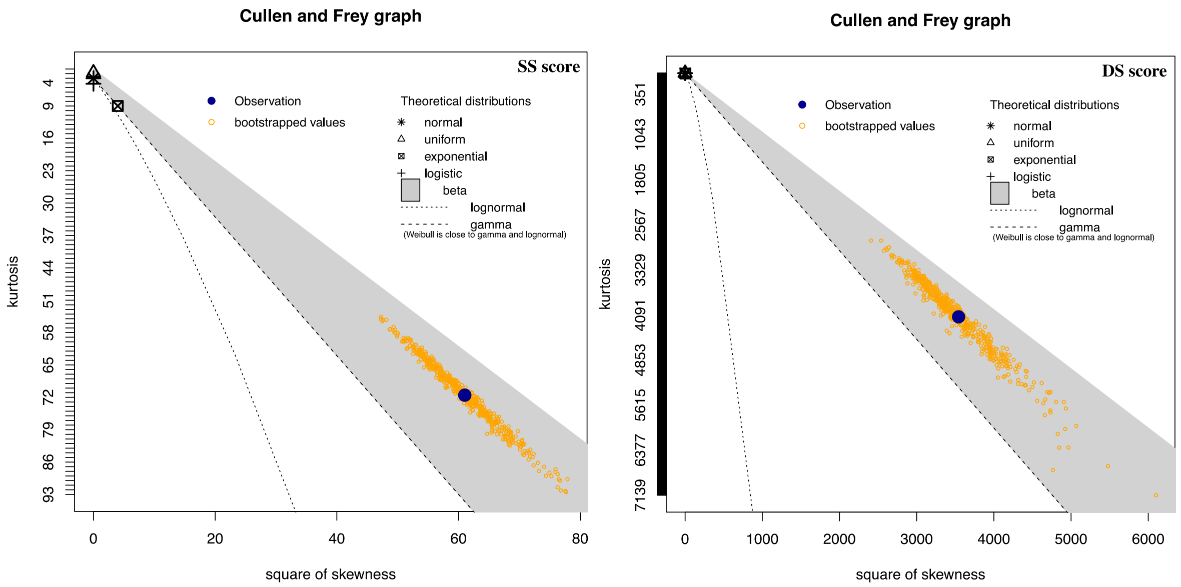


Figure A.1 CF graph of SS and DS empirical scores (plot produced using *fitdistrplus* package (Delignette-Muller and Dutang, 2015) in R (RStudio Team 2020)).

Following the CF graph, the parameters of Weibull and gamma distributions are estimated using MLE given the empirical scores, and Table A.2 shows the parameters of fitted Weibull and gamma distributions.

um	Weibull		gamma	
	scale	shape	rate	shape
SS	0.02	0.18	0.016	0.09
DS	0.002	0.16	0.009	0.074

Table A.2 Parameter values of fitted Weibull and gamma distributions from MLE.

The goodness-of-fit statistics and criteria are used for the evaluation of score fitting under two competing theoretical distributions, i.e., Weibull and gamma. The goodness-of-fit statistics measures the distance between the fitted theoretical distributions and the observed data (Delignette-Muller & Dutang, 2015, p.9). The lower the value, the closer the theoretical distributions fit the observed data (Vose, 2008, p. 284). For continuous distributions, three goodness-of-fit statistics are often considered, Kolmogorov-Smirnov (KS), Cramer-von Mises (CvM) and Anderson-Darling (AD) (Delignette-Muller and Dutang, 2015, p. 10). The formulas of these three goodness-of-fit measures are shown below:

KS	$\max[\lvert F_n(x) - F(x)\rvert]$
CvM	$\int_{-\infty}^{\infty} \lvert F_n(x) - F(x)\rvert^2 \psi(x) f(x) dx;\ \psi(x) = 1$
AD	$\int_{-\infty}^{\infty} \lvert F_n(x) - F(x)\rvert^2 \psi(x) f(x) dx; \psi(x) = \frac{n}{F(x)\{1-F(x)\}}$

Where,

n is the number of observed data points,

$F(x)$ is the distribution function of the fitted theoretical distribution,

$f(x)$ is the density function of the fitted theoretical distribution,

$F_n(x)$ = *i/n* and *i* is the cumulative rank in observed data points.

(ibid)

Among the three goodness-of-fit statistics, AD is used to select the best fitted theoretical distribution. There are several justifications for the decision behind it. First, the value of KS relies on the largest vertical difference between the cumulative distribution function (CDF) of the fitted theoretical distribution and the observed data (Vose, 2008, p. 291), meaning that it does not consider the lack of fit across the rest of the distribution. Second, CvM seems to be a better option than KS, and both CvM and AD take the lack of fit across the rest of the distribution into consideration, because $f(x)$ "weights the observed difference by the probability that a value will be generated at that x value" (Vose, 2008, p. 293). However, the difference in the suitable function $\psi(x)$ in CvM and AD means that when different theoretical distributions are fitted to observed data, AD gives compensations to different variances of the vertical distances between different fitted theoretical distributions (ibid), while CvM treats different fitted theoretical distributions equally (because the suitable function $\psi(x) = 1$ in CvM). Moreover, it is claimed (Delignette-Muller and Dutang, 2015, p. 10; Vose, 2008) that AD places equal weight on fitting the central and tail of a distribution.

As for the goodness-of-fit criteria, the Akaike's Information Criterion (AIC; (Parzen et al., 1998) and Bayesian Information Criterion (BIC; (Schwarz, 1978) are widely used for the evaluation of goodness-of-fit for statistical models. These two information criteria can be expressed using a unified log-likelihood function with different penalties attached (Dziak et al., 2012). The formula of information criterion can be expressed as:

$$-2\ell \ + A_n p \qquad \text{(Equation A2)}$$

(Atkinson, 1980)

Where ℓ is the log-likelihood,
A_n is either a constant or function of the sample size n,
p is the number of parameters in the model.

The formula A2 can be interpreted as finding the lowest value of -2ℓ plus a penalty, where A_n and p together serve as the penalty. Therefore, BIC and AIC differ in the selection of A_n, where the A_n of BIC is equal to the natural log of sample size n (*Ln(n)*), and the A_n of AIC is a constant number 2. For both BIC and AIC, the lower the value, the better the statistical model fits to empirical data, but it is worth noting that BIC and AIC values are not interpretable when given alone. In the current simulation experiments, only BIC is considered. This is because BIC gives similar weight to both Type I (false positive) and Type II (false negative) errors (Schwarz, 1978), while AIC puts more weight on Type II errors than Type I (Dziak et al., 2012). In score simulation, both Type I and Type II errors are undesirable. High Type I error rate would lead to overfitting, which could possibly overestimate the system variability, i.e., the actual overall performance should be less variable; while high Type II error rate would lead to underfitting, which leads to the underestimation of system variability, i.e., the actual overall performance should be more variable.

Apart from AD and BIC values, Figures A.2 and A.3 give visualisation of goodness-of-fit over the centre and tail of the empirical scores. Figure A.2 shows the probability density function (PDF) and cumulative distribution function (CDF) plots of Weibull and gamma (coloured dotted lines) fitted to empirical scores (histograms). The PDF and CDF plots give the overall goodness-of-fit to the empirical scores, and more overlap between empirical and estimated data indicates a better goodness-of-fit. The histograms show that most of the raw SS scores are stacked between 0 and ca. 25 with a few spreads out from 25 onwards, and the majority of the raw DS scores are packed between 0 and ca. 200 with a few distributed between 200 and 1000 and even fewer spread out from 1000 onwards. The PDF and CDF plots show that gamma and Weibull perform equally well in terms of overall goodness-of-fit.

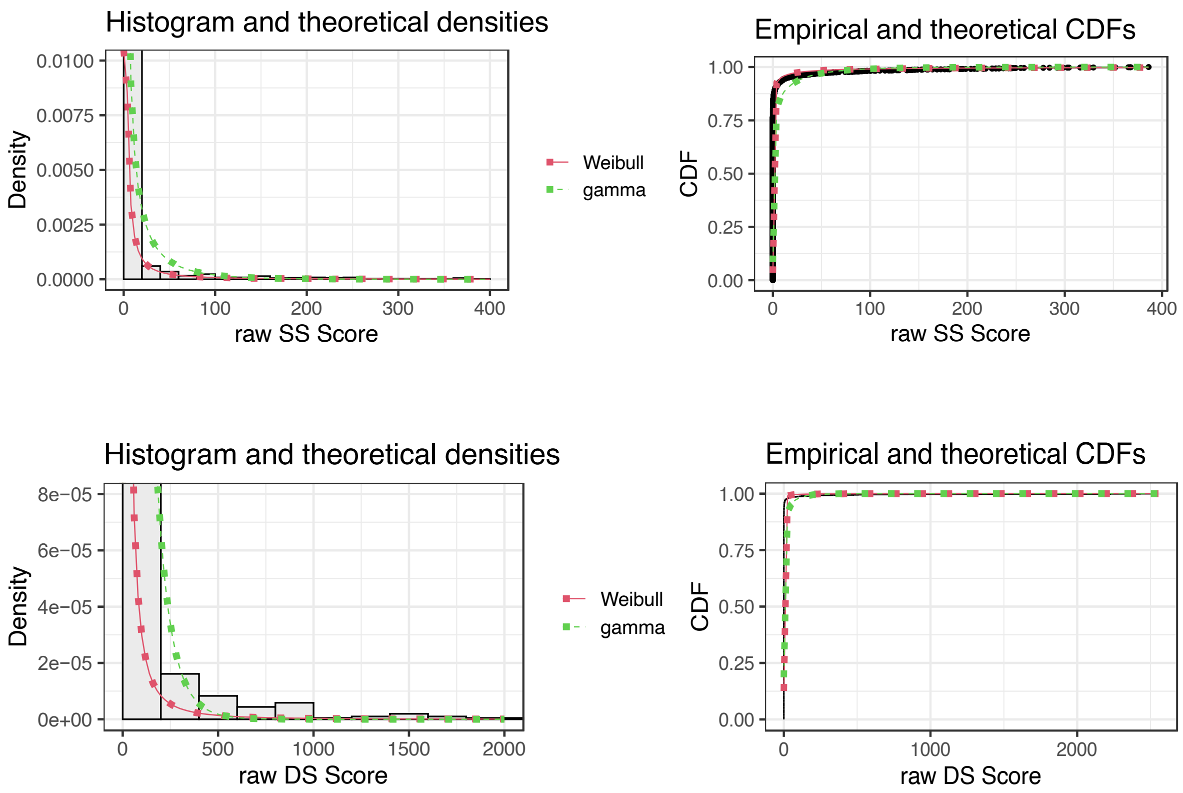


Figure A.2 PDFs (left panels) and CDFs (right panels) showing the goodness-of-fit of gamma and Weibull distributions fitted to FP *um* empirical raw SS and DS scores. Coloured dotted lines are theoretical distributions.

Figure A.3 shows the P-P and Q-Q plots of theoretical gamma and Weibull distributions fitted to SS (upper panel) and DS (lower panel) empirical scores. In the P-P plot, each data point from the empirical distribution is plotted against the theoretical distribution function, which gives more emphasis to the lack-of-fit at the distribution centre. The Q-Q plot is a representation of the empirical quantiles plotted against the theoretical quantiles, which gives more emphasis to the lack-of-fit at the tails of the distributions. For SS scores, the P-P plot shows that Weibull is better for the centre of the empirical data, while the Q-Q plot shows that none of the theoretical distributions can fit the tail of the SS scores accurately, even gamma gives a slightly better description at the right end. For DS scores, Weibull also seems to have a better goodness-of-fit at the distribution centre. However, the tail of DS scores seems to be much more variable than SS scores, and none of the theoretical distributions seem to be close to the distribution tails of DS scores.

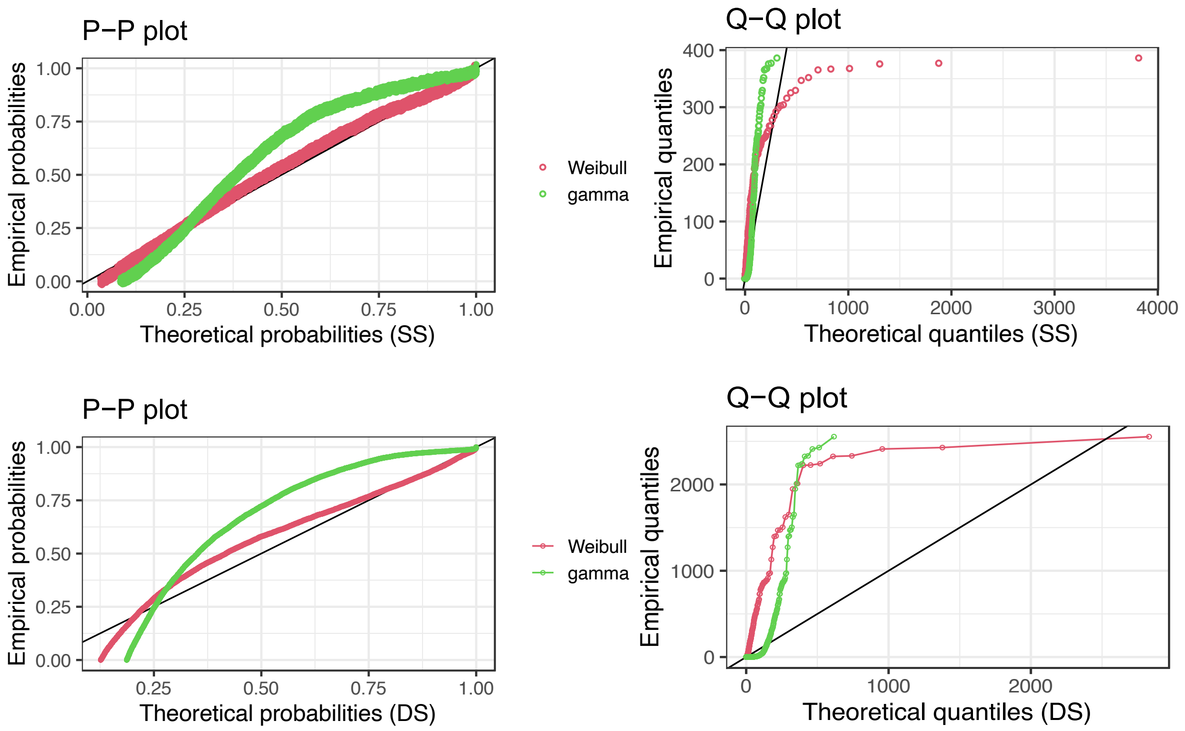


Figure A.3 P-P and Q-Q plots of gamma and Weibull distributions fitted to FP *um* empirical raw SS and DS scores. P-P plots indicate the lack-of-fit at the distribution centre and Q-Q plots show the lack-of-fit at the distribution tails.

um	SS		DS	
	gamma	Weibull	gamma	Weibull
AD	315.91	21.46	1281.09	215.65
BIC	-31790.51	-33615.10	-113740.1	-120906.3

Table A.3 AD and BIC values of gamma and Weibull fitted to empirical SS and DS scores.

Table A.3 shows the AD and BIC values for the gamma and Weibull distributions fitted to the SS and DS empirical scores. The Weibull distribution function gives lower AD and BIC values for both SS and DS empirical scores, which is consistent with Figures A.2 and A.3, suggesting that the Weibull distribution function is a more probable candidate for the empirical scores. Since the empirical scores at the tails are extremely variable and cannot be fully fitted using the Weibull distribution or any of the theoretical distributions, the estimated Weibull parameters can be used as a reference for raw score simulation. Similar to the ST distribution simulation, different sets of Weibull parameters can be used to test the effects of sampling variability on overall performance when Weibull distribution has different degree of variabilities at the tails. However, only the scale parameter would need to be varied. This is because the tail in the Weibull distribution is mostly affected by the scale parameter. If the scale parameter is increased, the Weibull distribution would be stretched to the right leading to a heavier tail, and if the scale parameter is decreased, the Weibull distribution would be pushed to the left close to 0 resulting in a lighter tail. Figure A.4 gives an example of using different scale parameters (the shape parameter is fixed), showing that larger scale parameter would stretch the Weibull distribution to the right and result in more variability at the tails (scale parameter increased from 0.02 to 2).

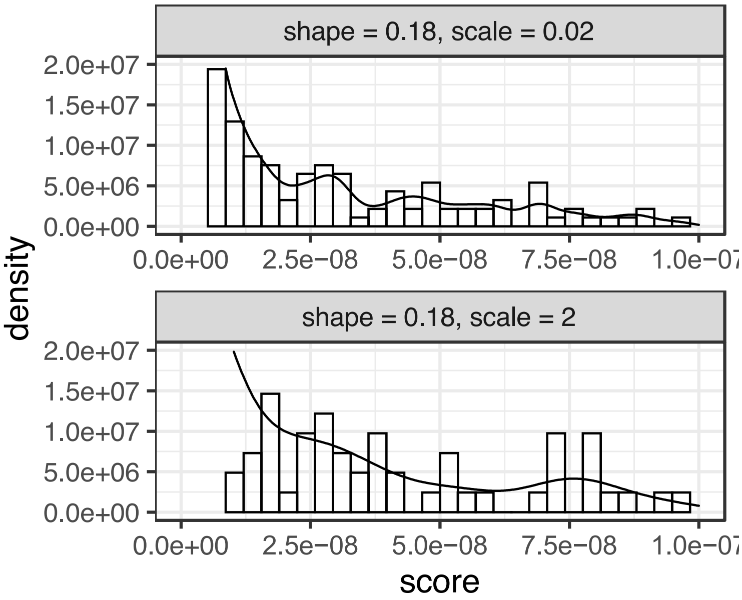


Figure A.4 Simulated score distributions with different scale parameters.

Table A.4 shows the three sets of Weibull scale and shape parameters that could be used for SS and DS raw score simulation. The scale and shape parameters in set (a) are the estimated ones using MLE from empirical scores. In set (b) and set (c), the scale parameters are increased to 0.2 and 2 and 0.02 and 0.2 for SS and DS scores respectively, while the shape parameters are fixed.

	λ (scale)	*k* (shape)		λ (scale)	*k* (shape)
SS set (a)	0.02	0.18	DS set (a)	0.002	0.16
SS set (b)	0.2	0.18	DS set (b)	0.02	0.16
SS set (c)	2	0.18	DS set (c)	0.2	0.16

Table A.4 Three sets of scale and shape parameters for SS and DS score simulation.

Once the scale and shape values are specified, raw scores can be sampled from Weibull distributions and used for LR computation. However, score simulation in Chapter 6 was conducted using the log scores, rather than the raw scores demonstrated here. As stated in section 6.2, the process of simulating the raw scores demonstrated here (i.e., Appendix A) is for the sake of exploration.

Appendix B - Zoo plots of SS and DS LLRs of 25 test speakers in 31 systems, and strip texts on top of the plots indicate specific system.

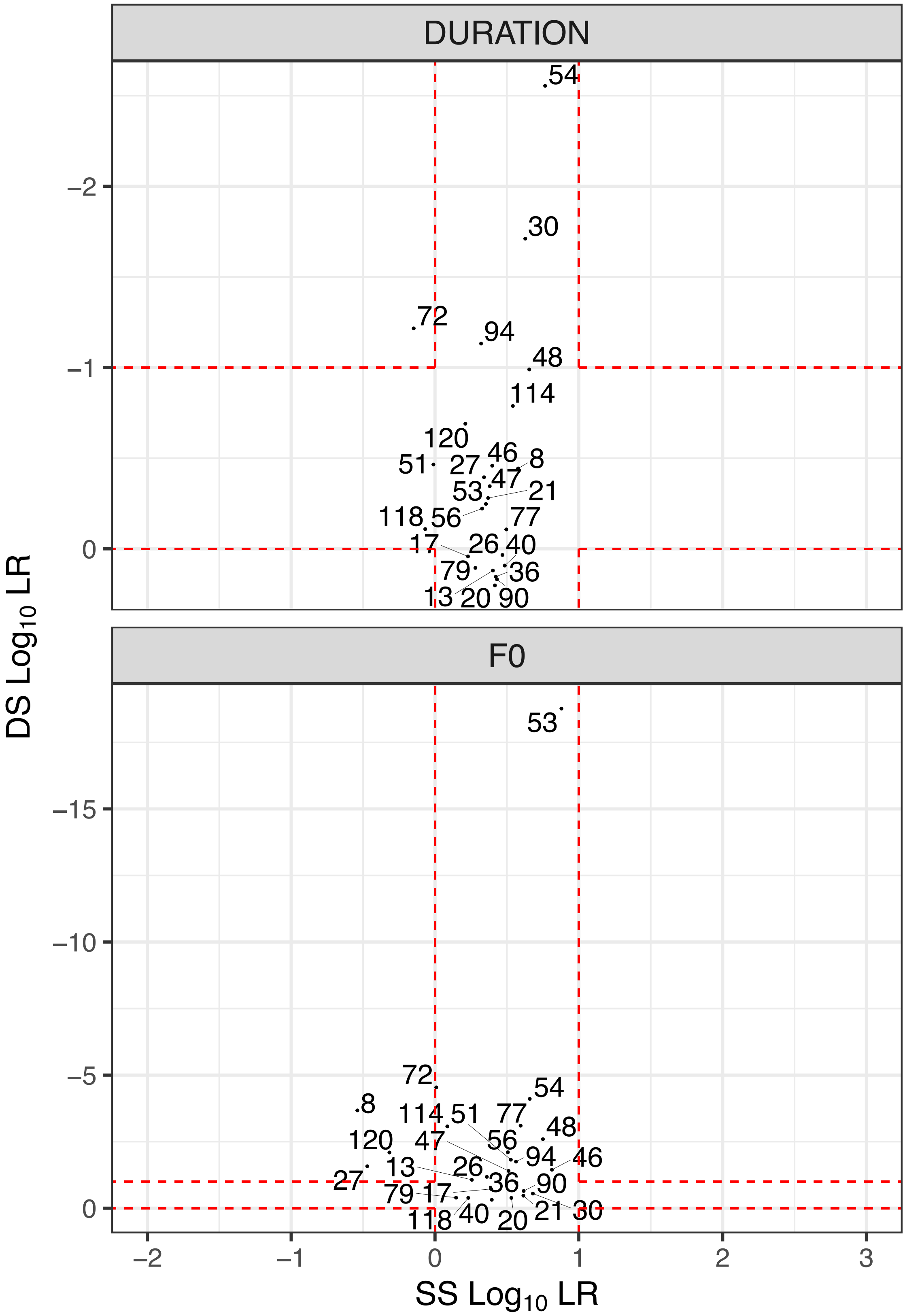


Figure B.1 Zoo plots of SS and DS LLRs of 25 test speakers in DURATION and F0 systems.

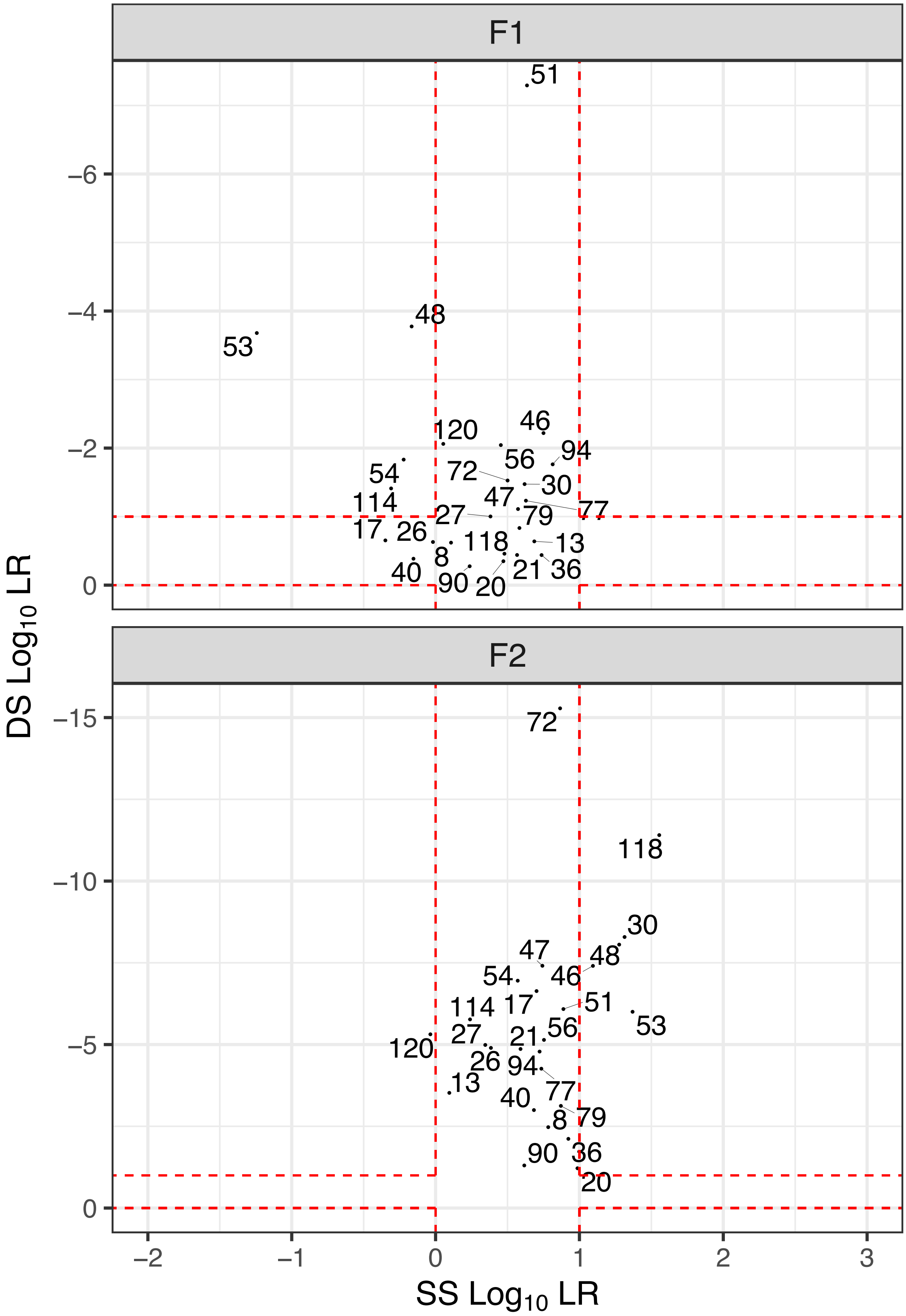


Figure B.2 Zoo plots of SS and DS LLRs of 25 test speakers in F1 and F2 systems.

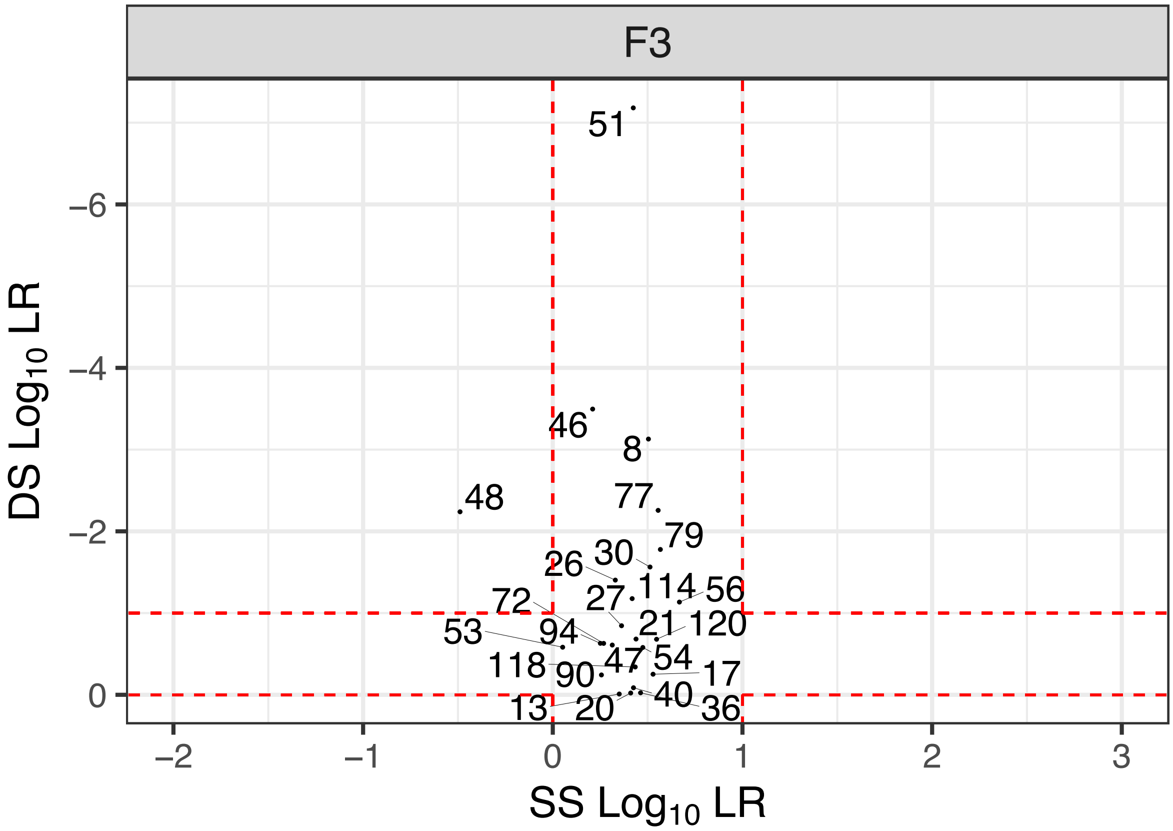


Figure B.3 Zoo plot of SS and DS LLRs of 25 test speakers in F3 system.

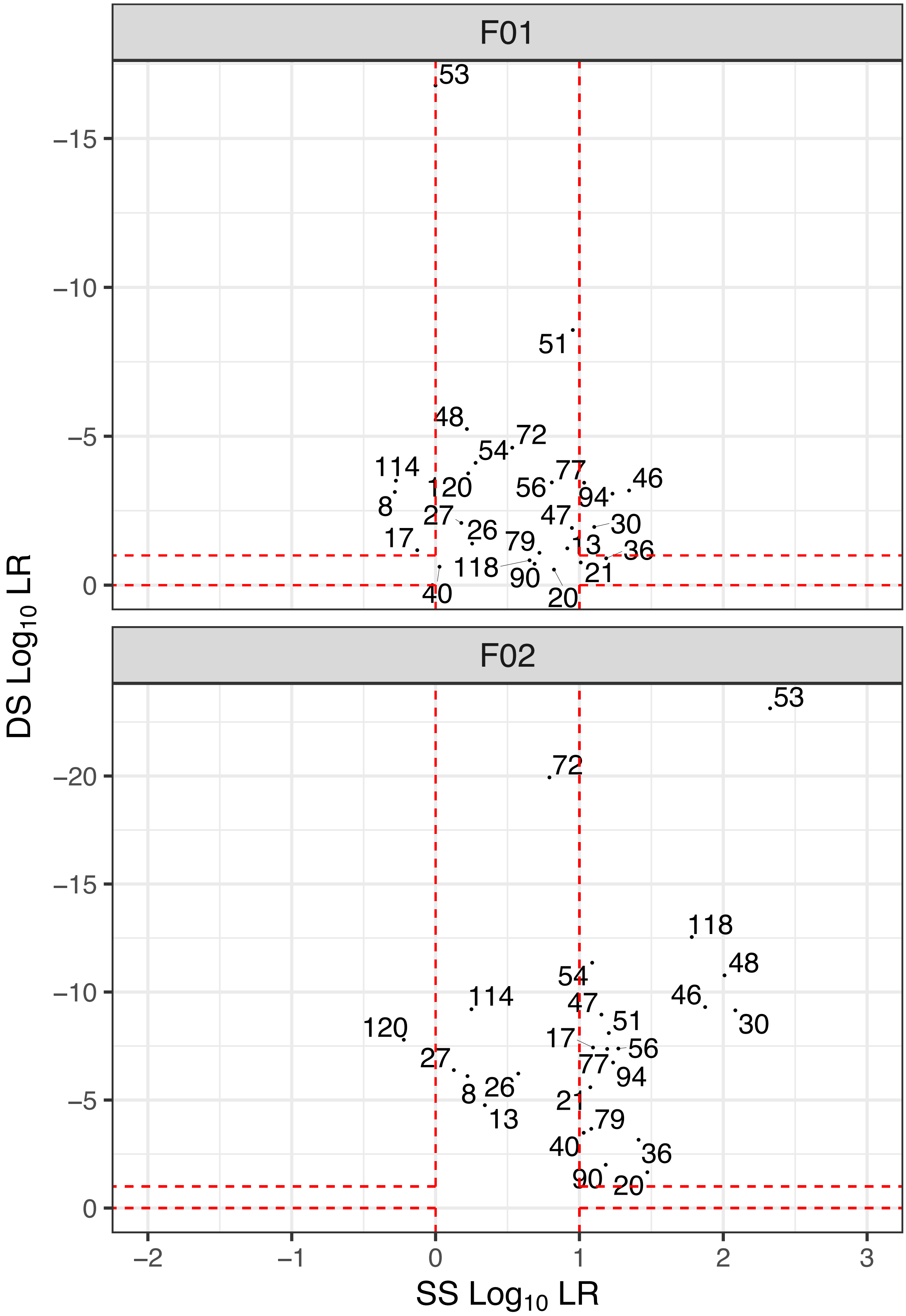


Figure B.4 Zoo plots of SS and DS LLRs of 25 test speakers in F01 and F02 systems.

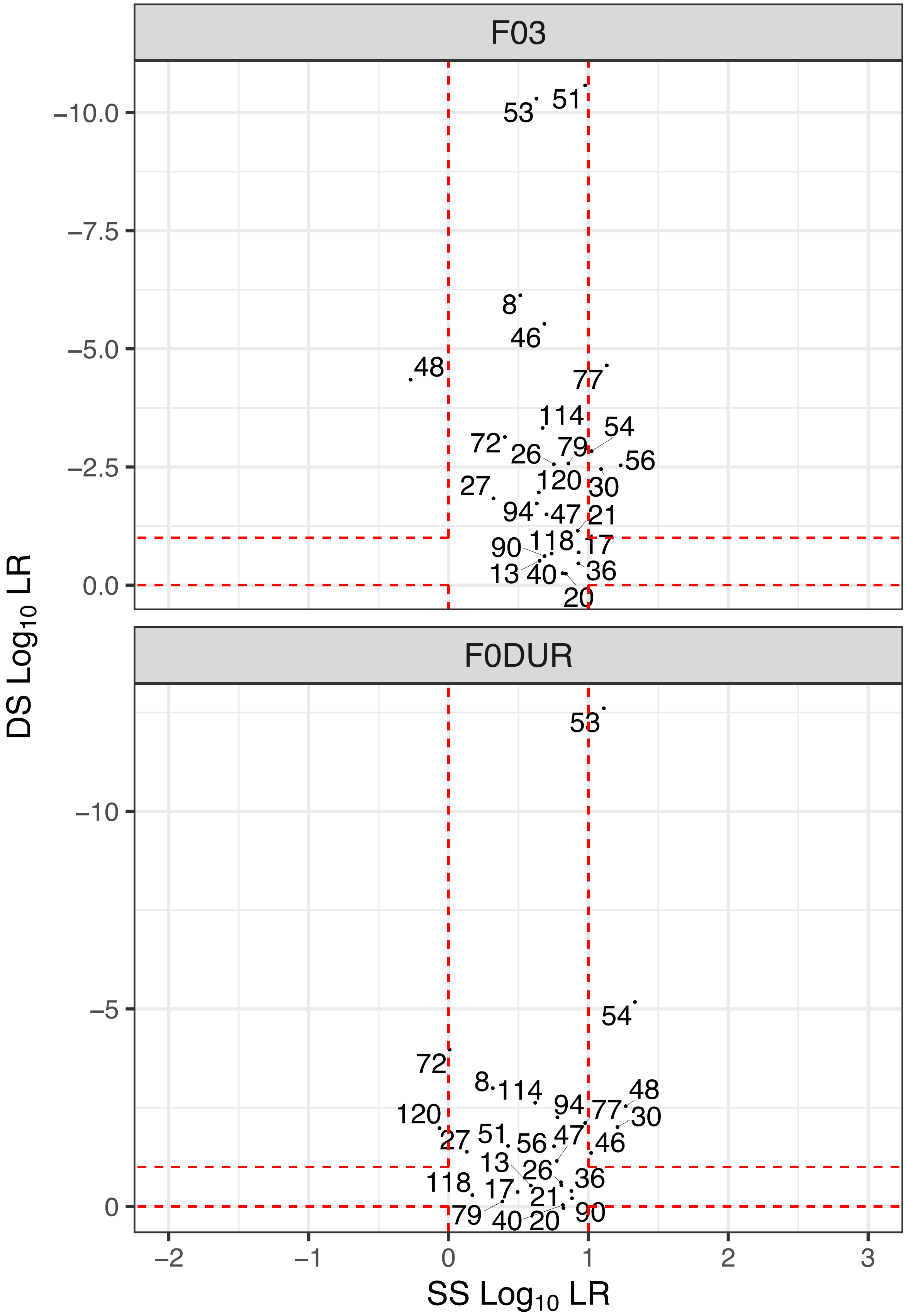


Figure B.5 Zoo plots of SS and DS LLRs of 25 test speakers in F03 and F0DUR systems.

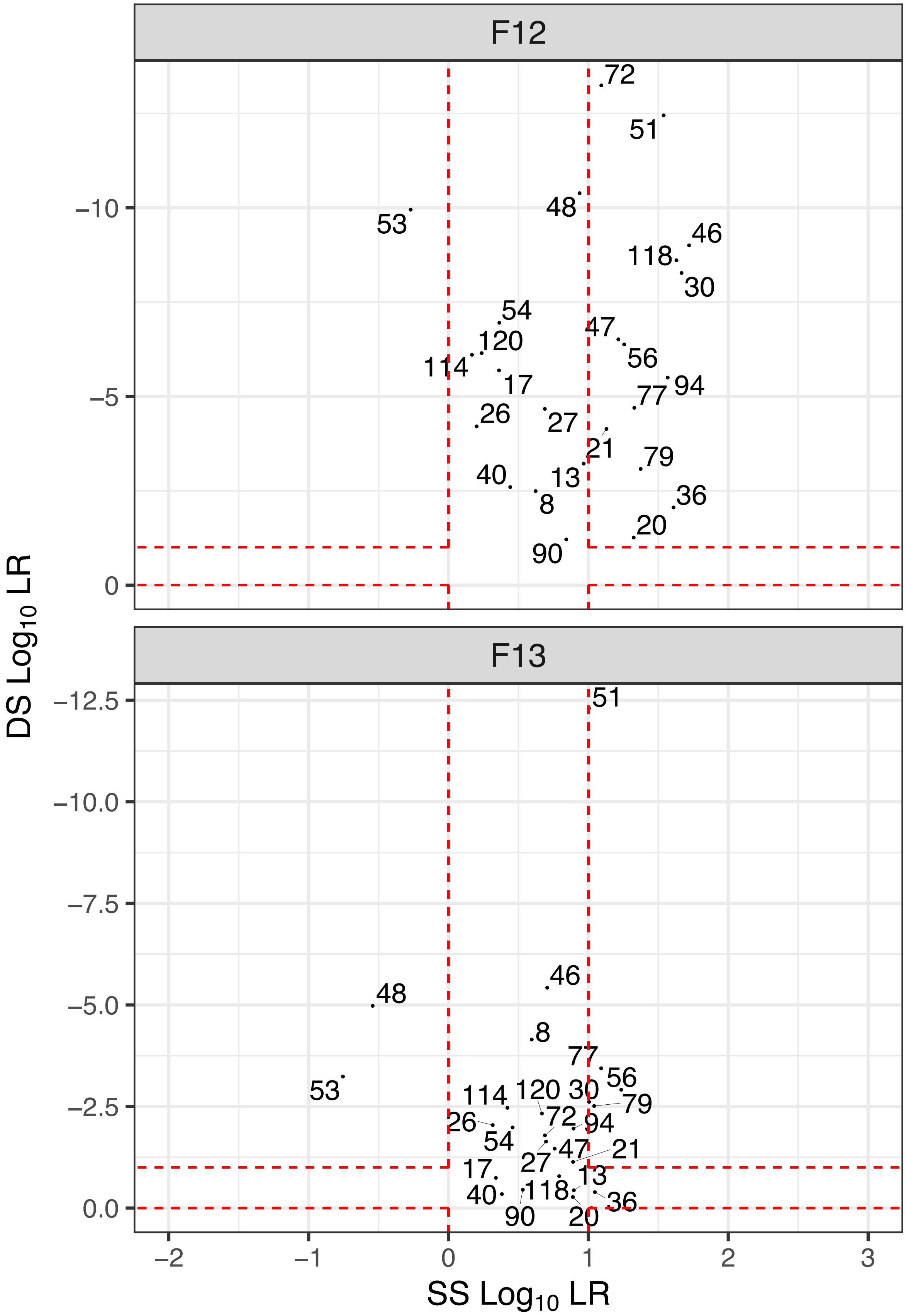


Figure B.6 Zoo plots of SS and DS LLRs of 25 test speakers in F12 and F13 systems.

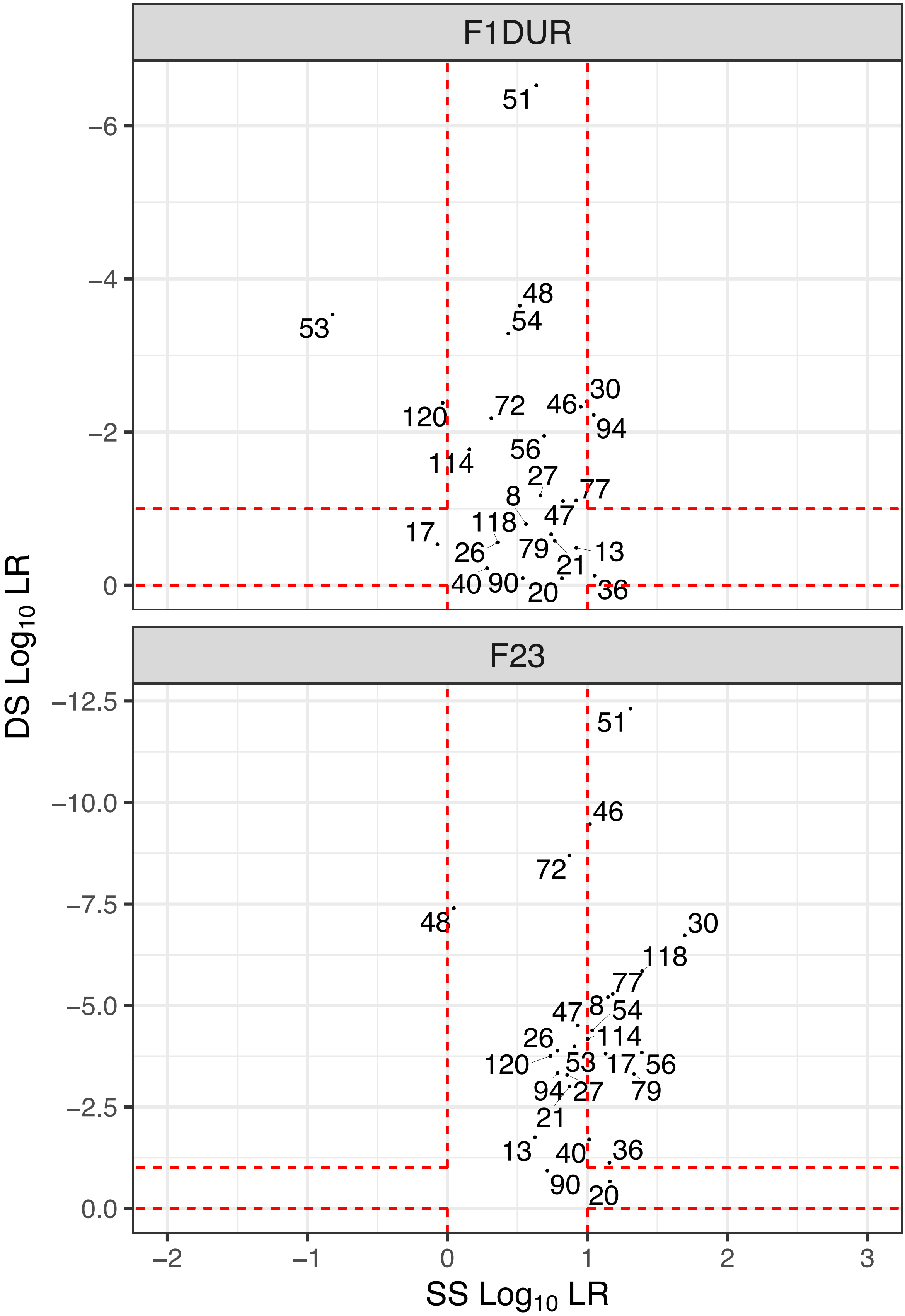


Figure B.7 Zoo plots of SS and DS LLRs of 25 test speakers in F1DUR and F23 systems.

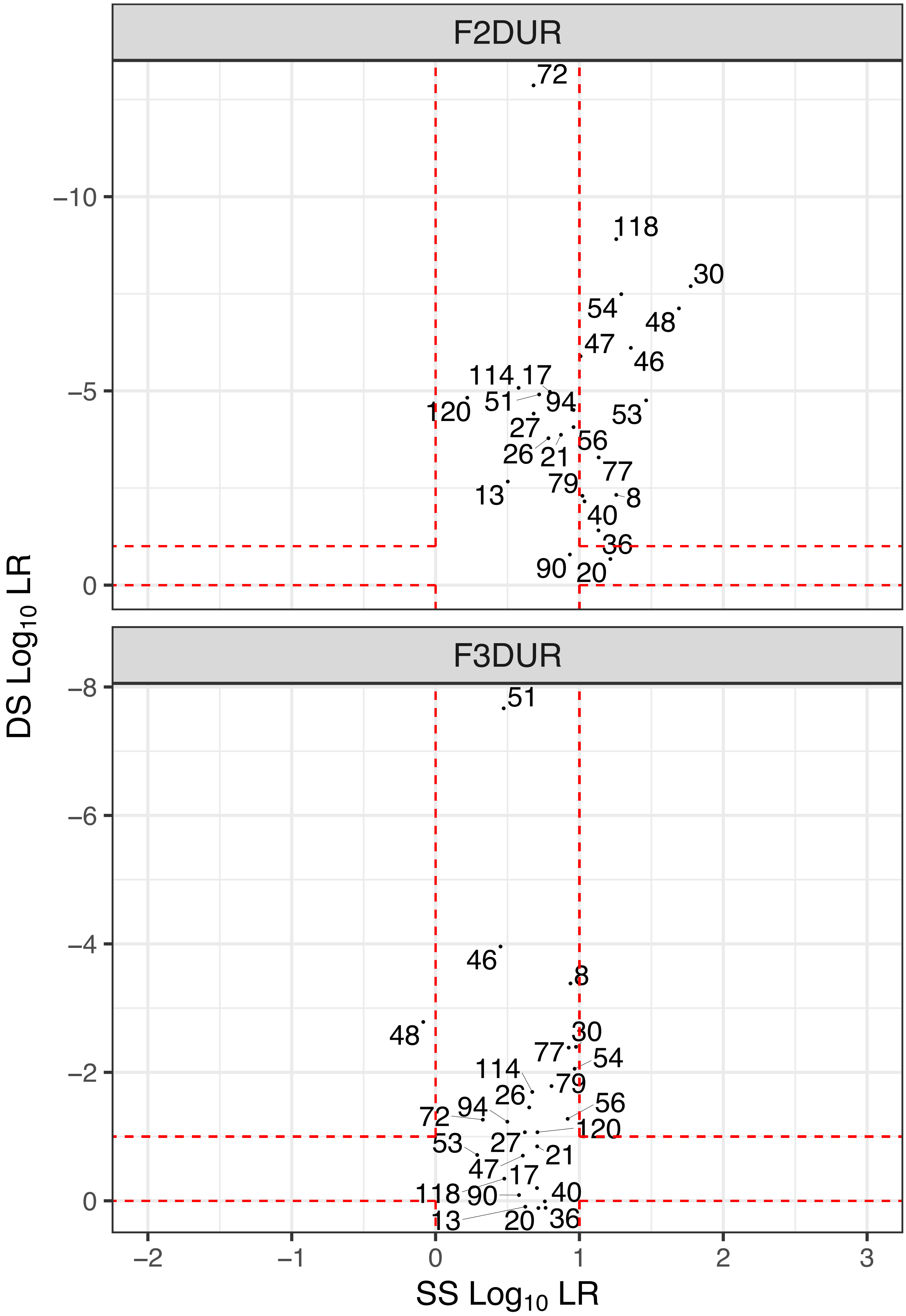


Figure B.8 Zoo plots of SS and DS LLRs of 25 test speakers in F2DURand F3DUR systems.

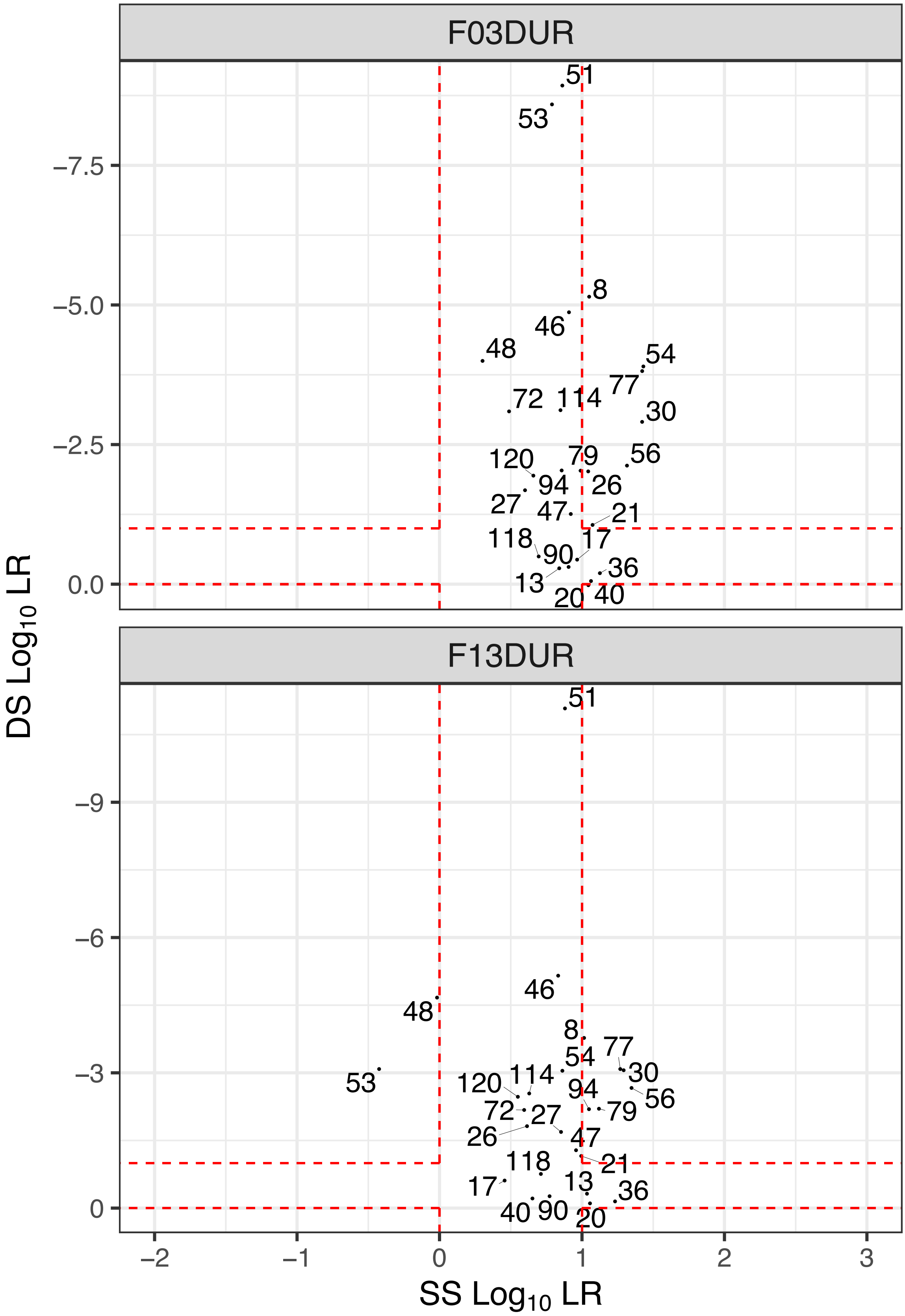


Figure B.9 Zoo plots of SS and DS LLRs of 25 test speakers in F03DUR and F13DUR systems.

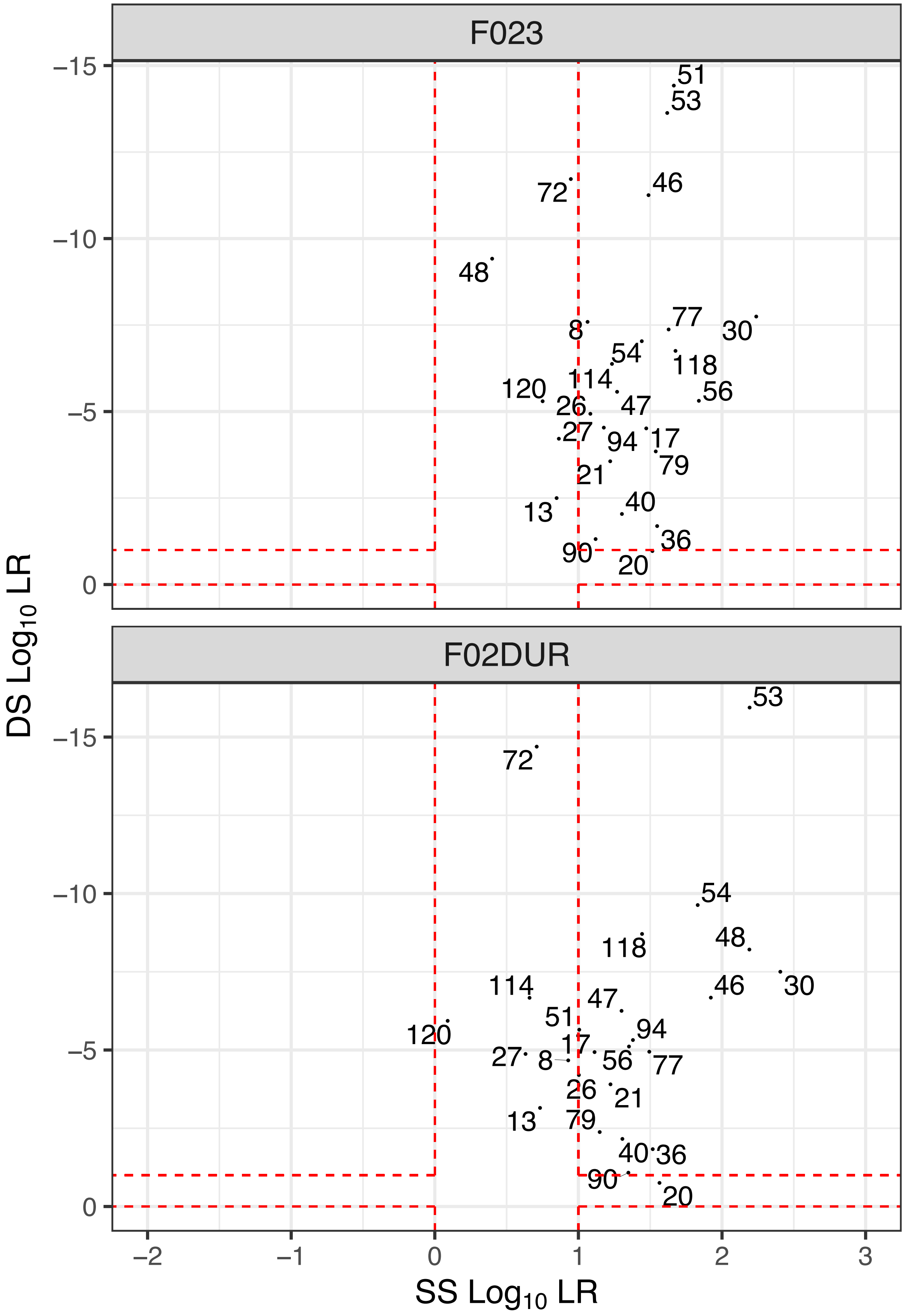


Figure B.10 Zoo plots of SS and DS LLRs of 25 test speakers in F023 and F02DUR systems.

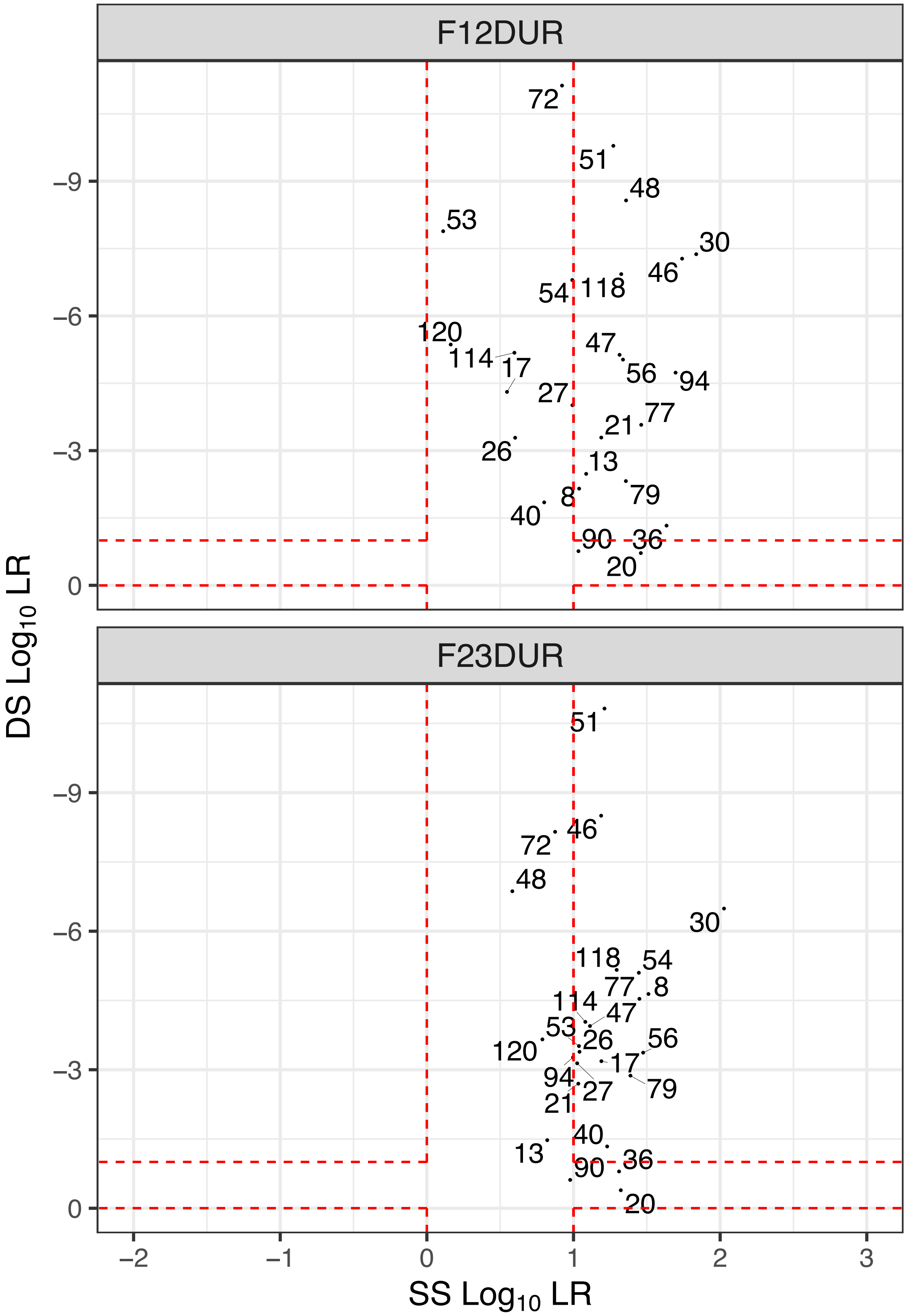


Figure B.11 Zoo plots of SS and DS LLRs of 25 test speakers in F12DUR and F23DUR systems.

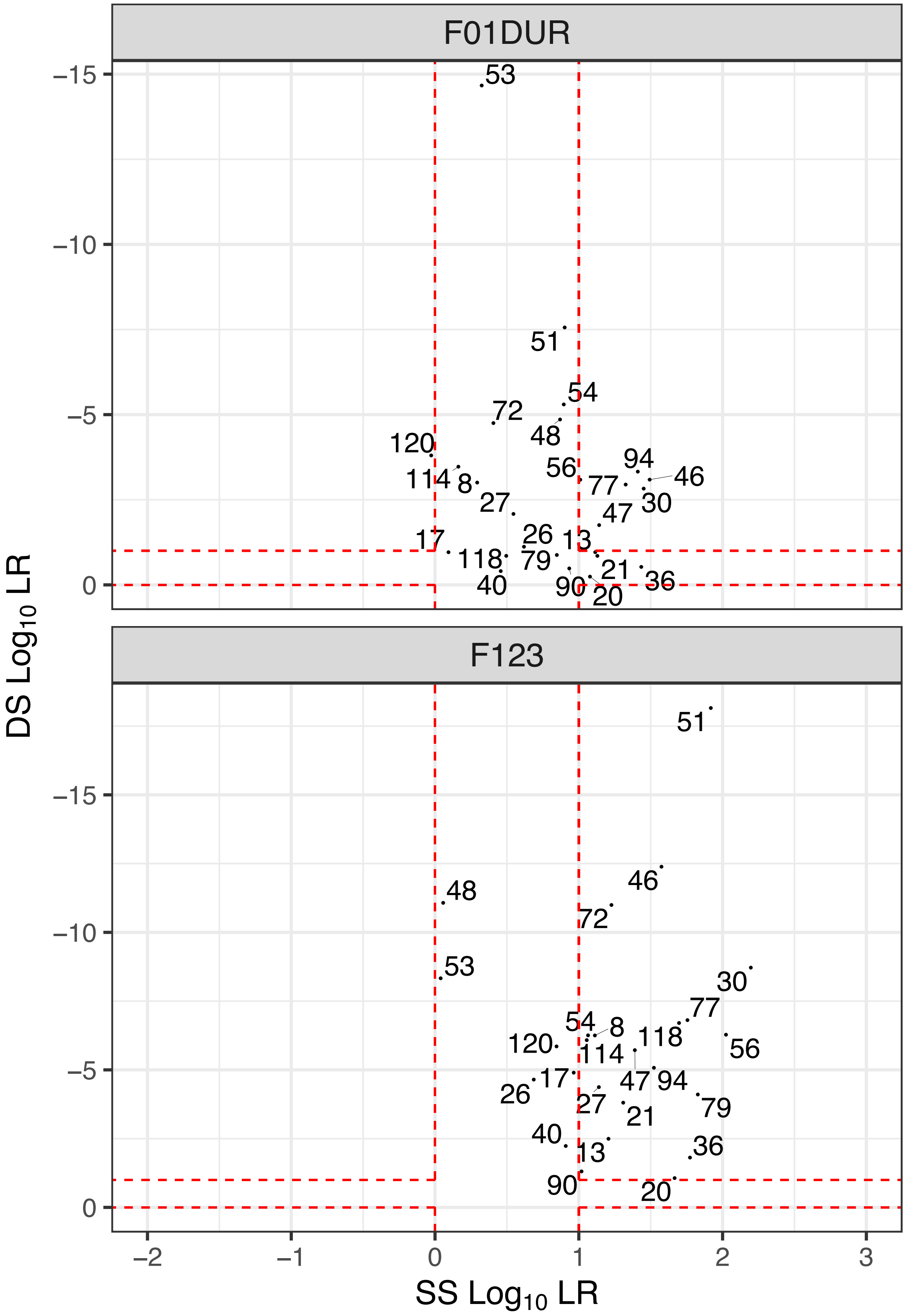


Figure B.12 Zoo plots of SS and DS LLRs of 25 test speakers in F01DUR and F123 systems.

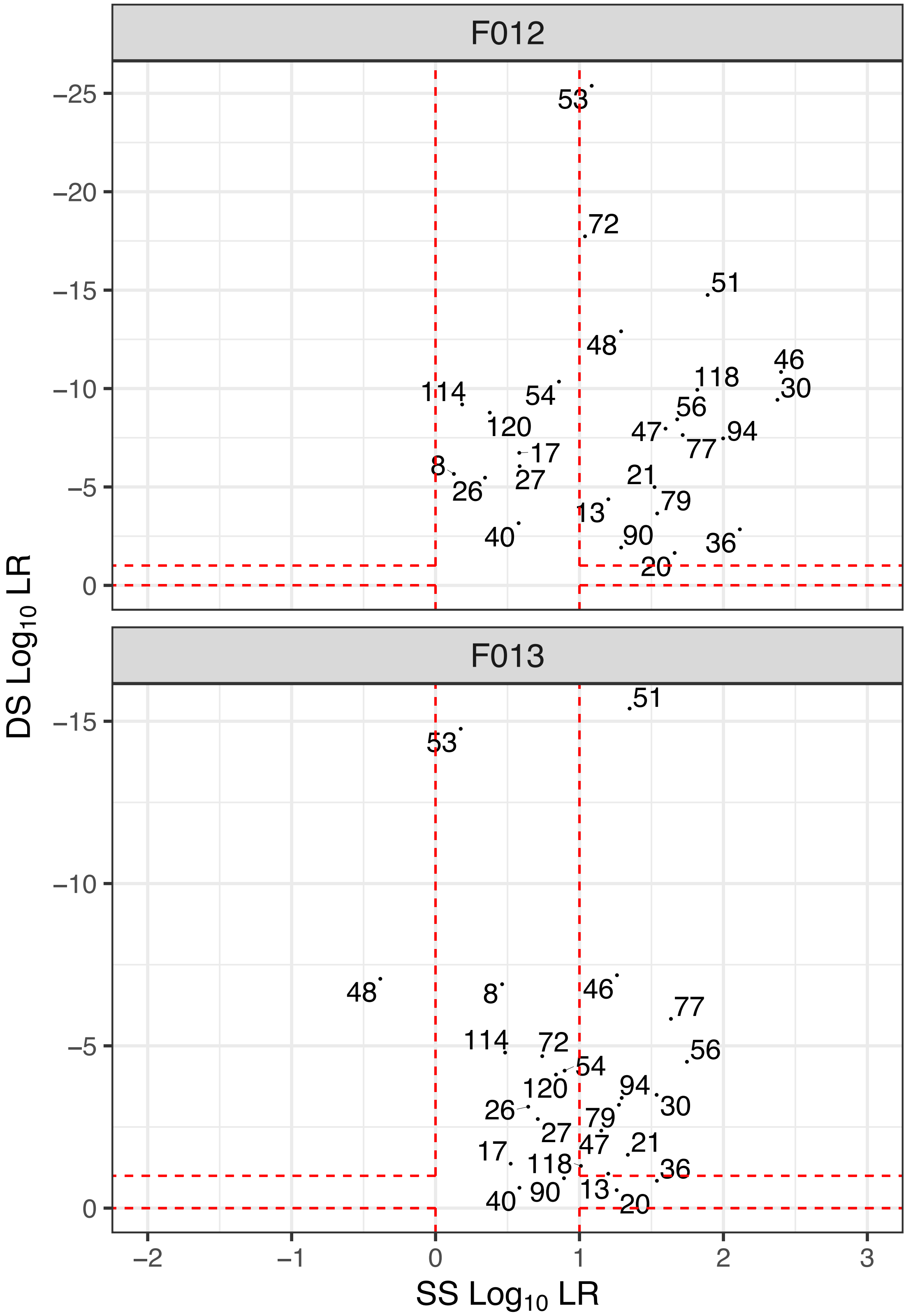


Figure B.13 Zoo plots of SS and DS LLRs of 25 test speakers in F012 and F013 systems.

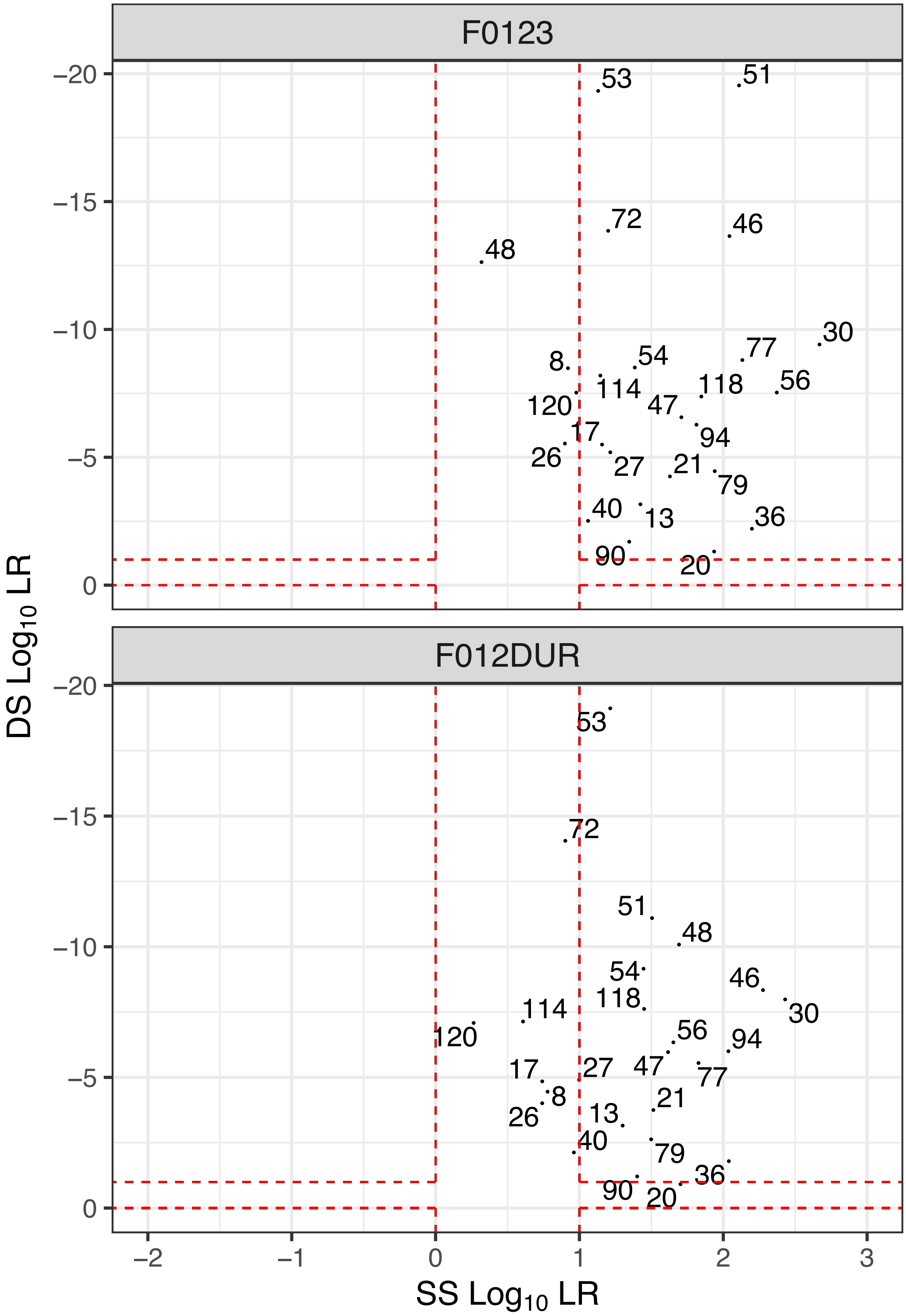


Figure B.14 Zoo plots of SS and DS LLRs of 25 test speakers in F0123 and F012DUR systems.

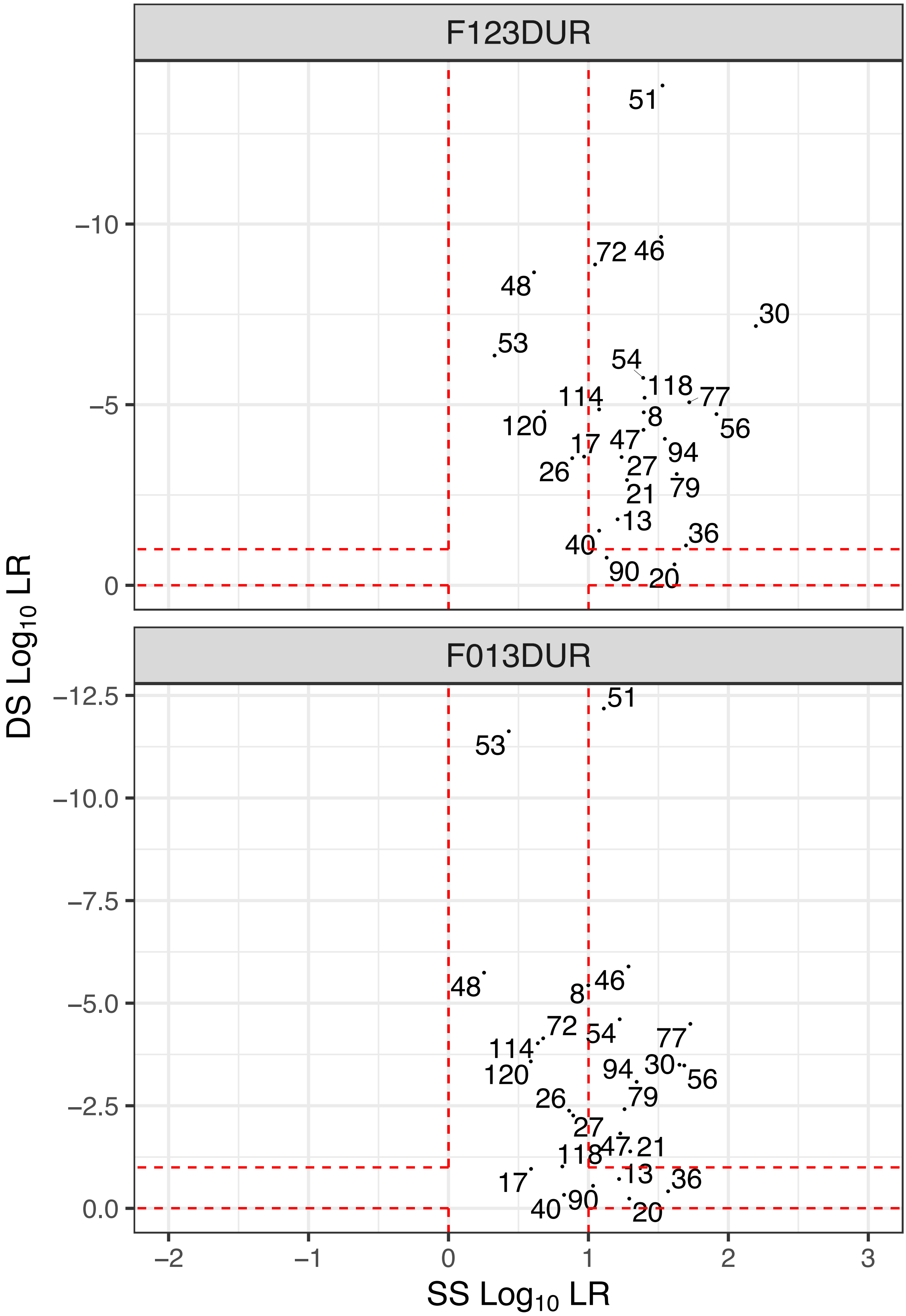


Figure B.15 Zoo plots of SS and DS LLRs of 25 test speakers in F123DUR and F013DUR systems.

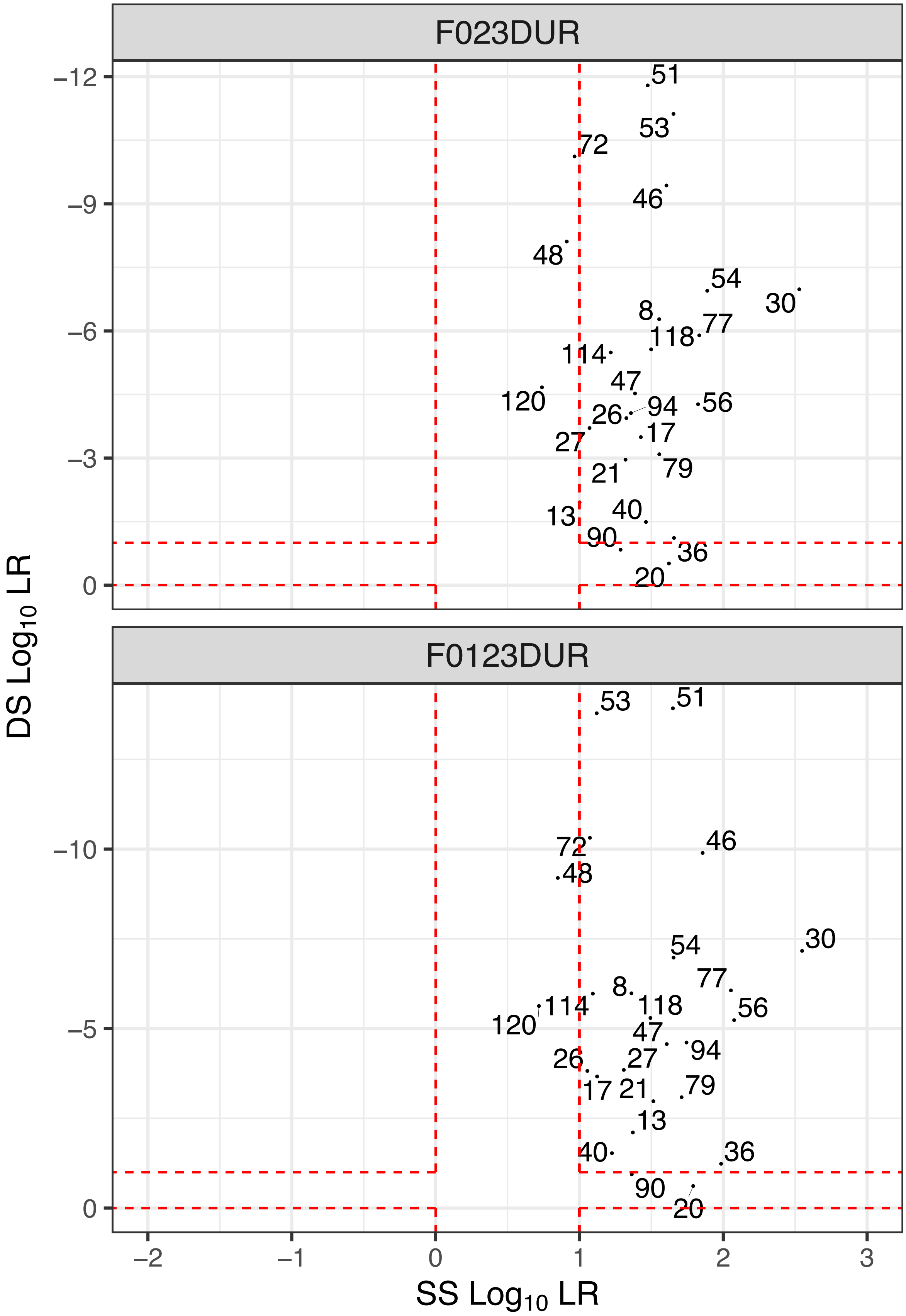


Figure B.16 Zoo plots of SS and DS LLRs of 25 test speakers in F023DUR and F0123DUR systems.

Appendix C - Plots of mean SS and DS LLR of individual speakers across 31 systems. Strip texts on top indicate speaker ID.

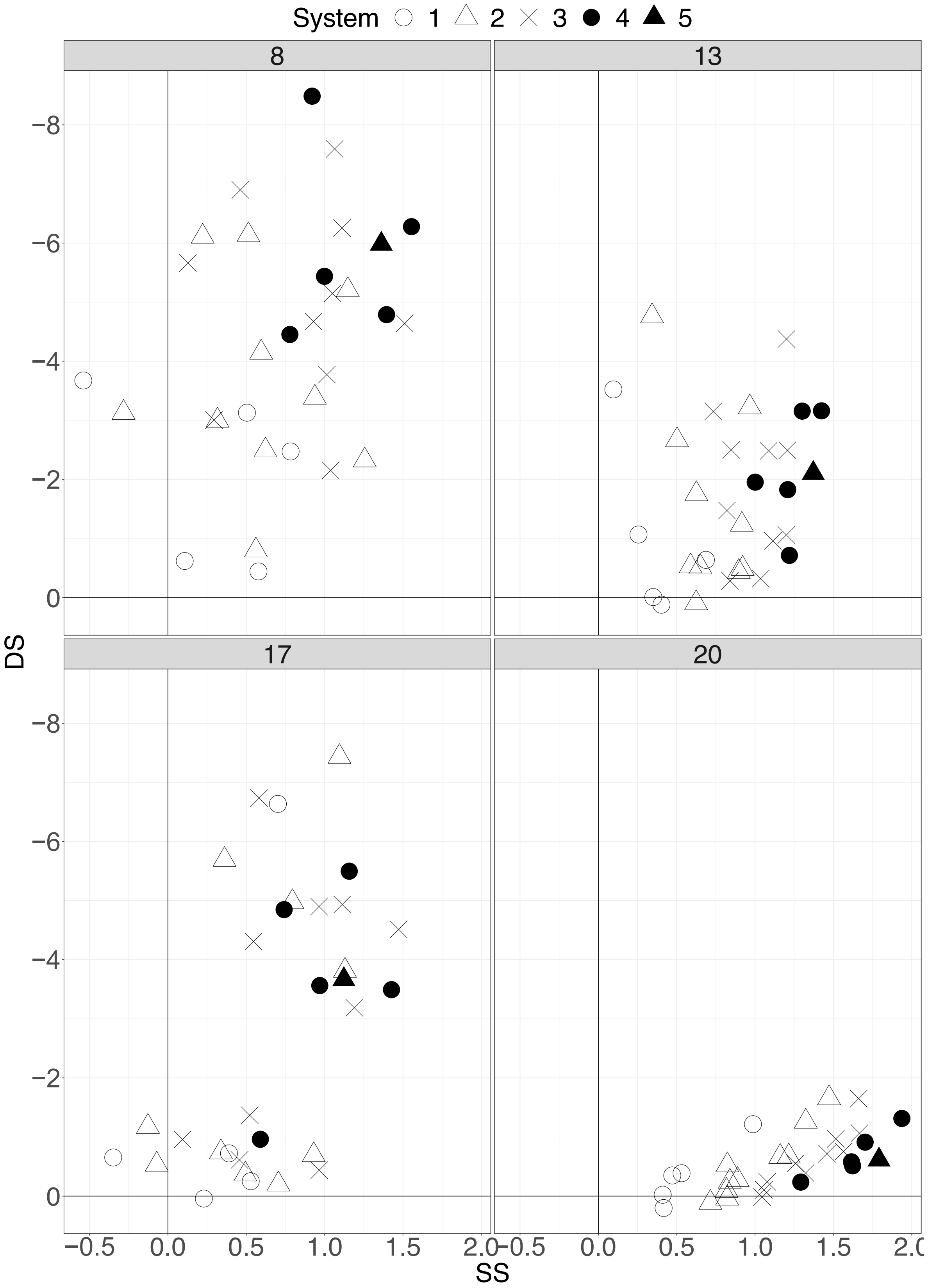


Figure C.1 Mean SS and DS LLR of speakers 8, 13, 17 and 20 across 31 systems.

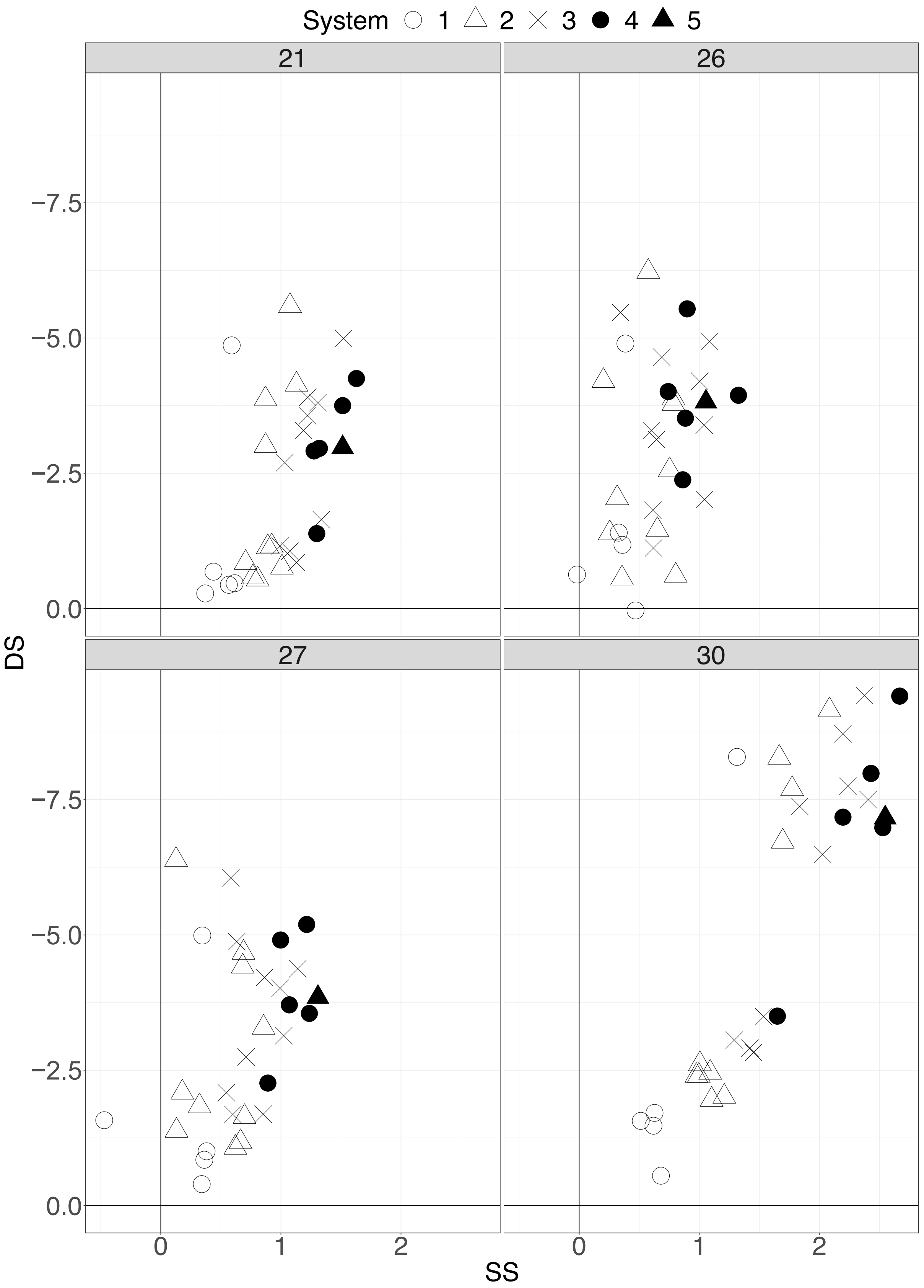


Figure C.2 Mean SS and DS LLR of speakers 21, 26, 27 and 30 across 31 systems.

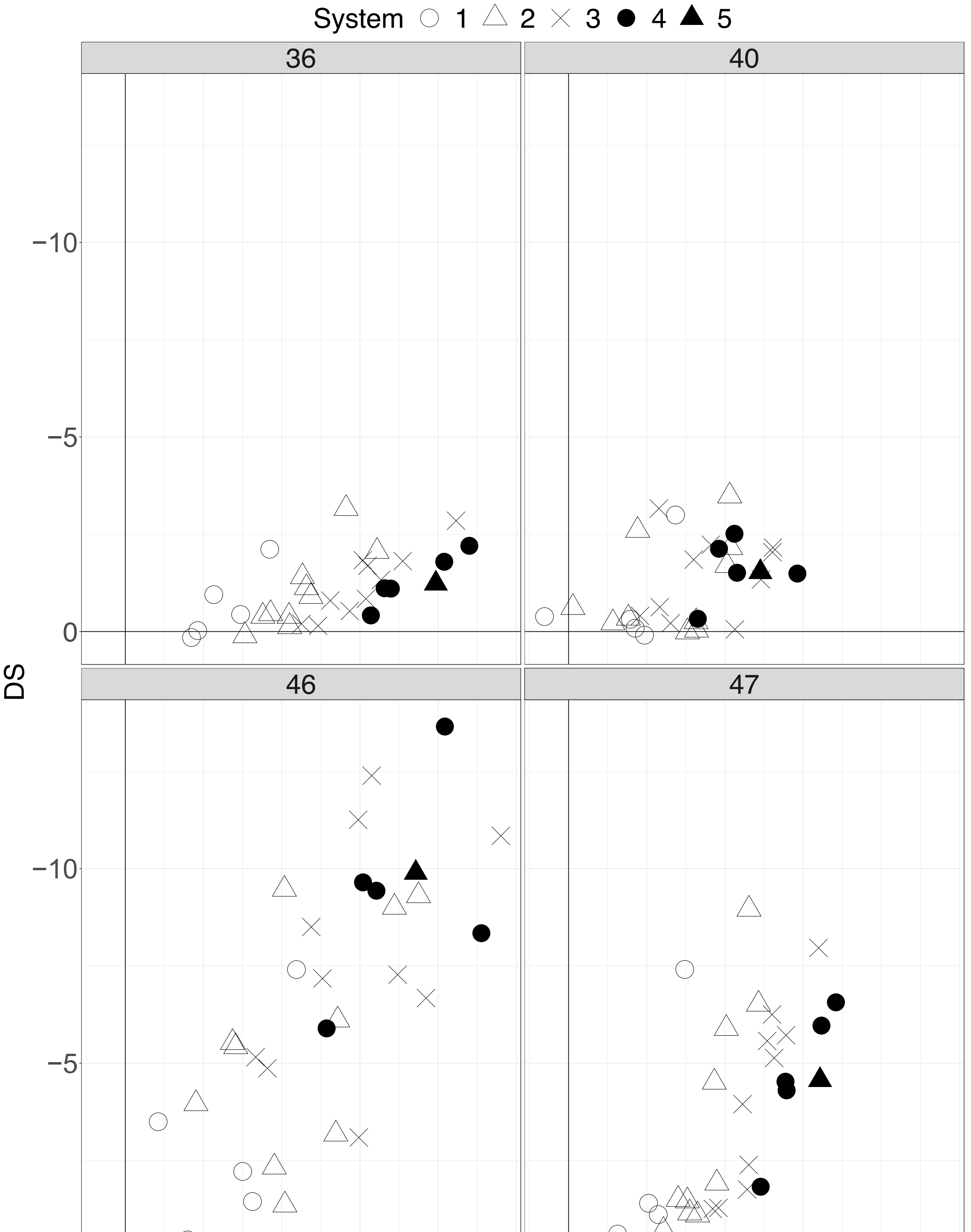


Figure C.3 Mean SS and DS LLR of speakers 36, 40, 46 and 47 across 31 systems.

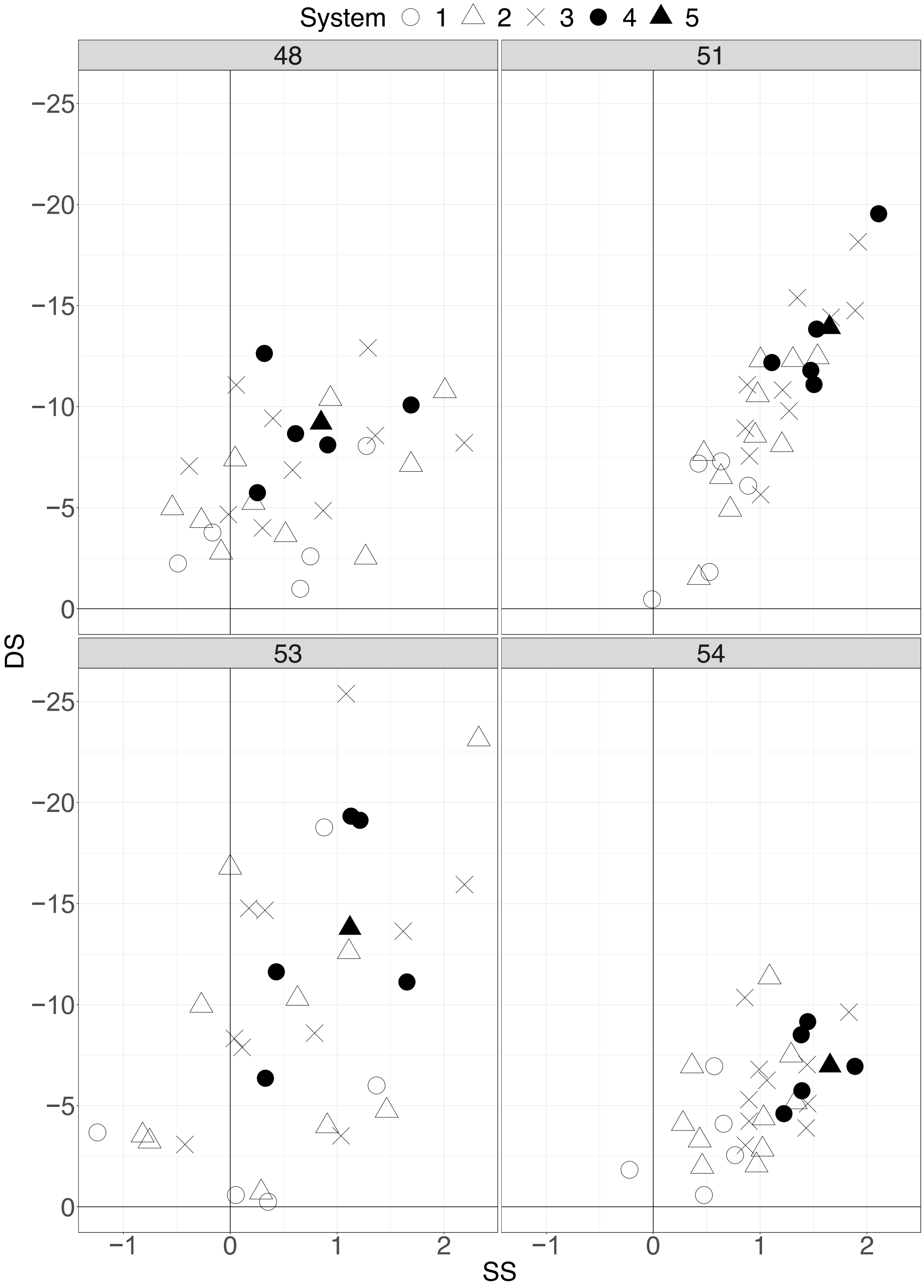


Figure C.4 Mean SS and DS LLR of speakers 48, 51, 53 and 54 systems.

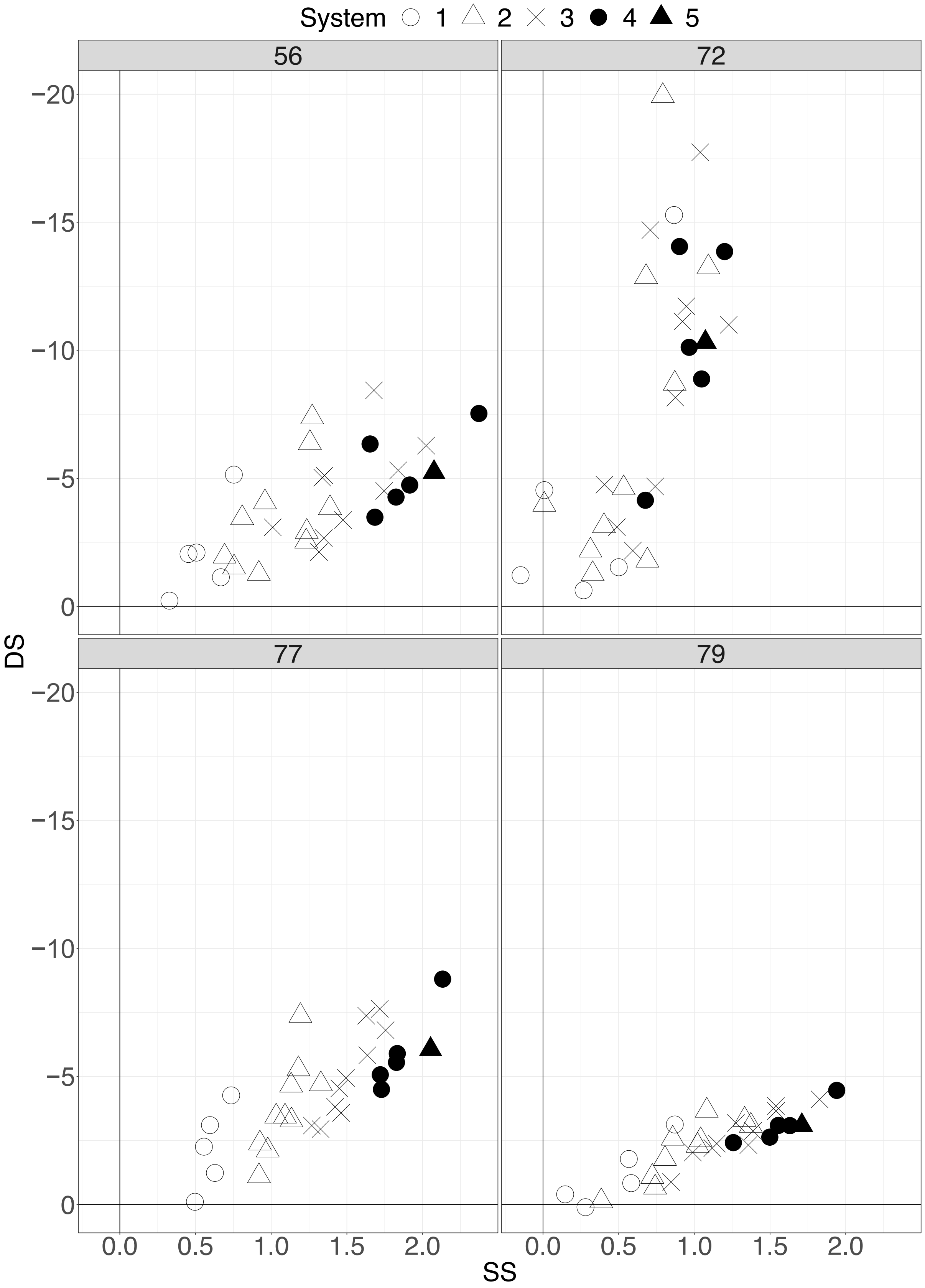


Figure C.5 Mean SS and DS LLR of speakers 56, 72, 77 and 79 across 31 systems.

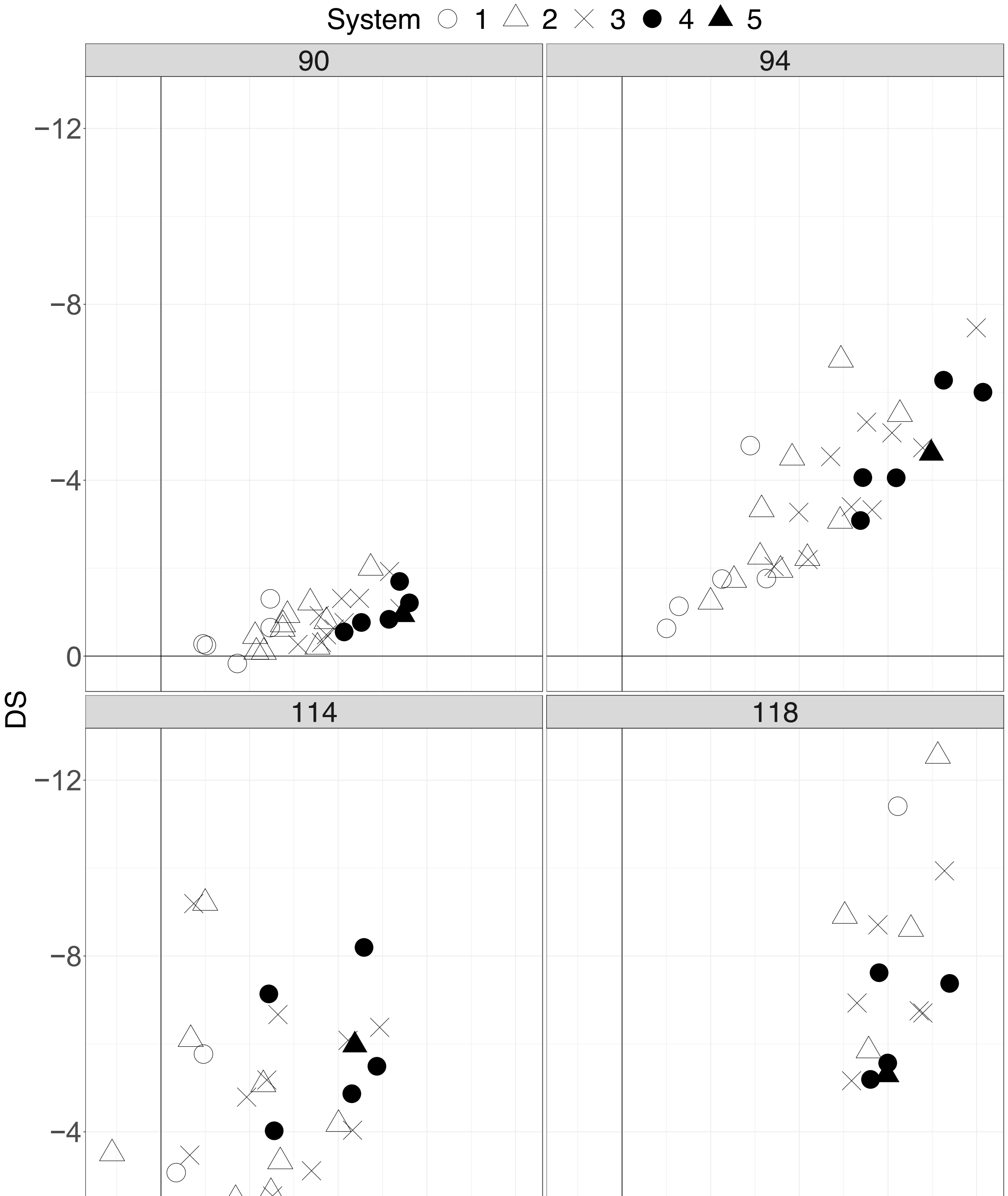


Figure C.6 Mean SS and DS LLR of speakers 90, 94, 114 and 118 across 31 systems.

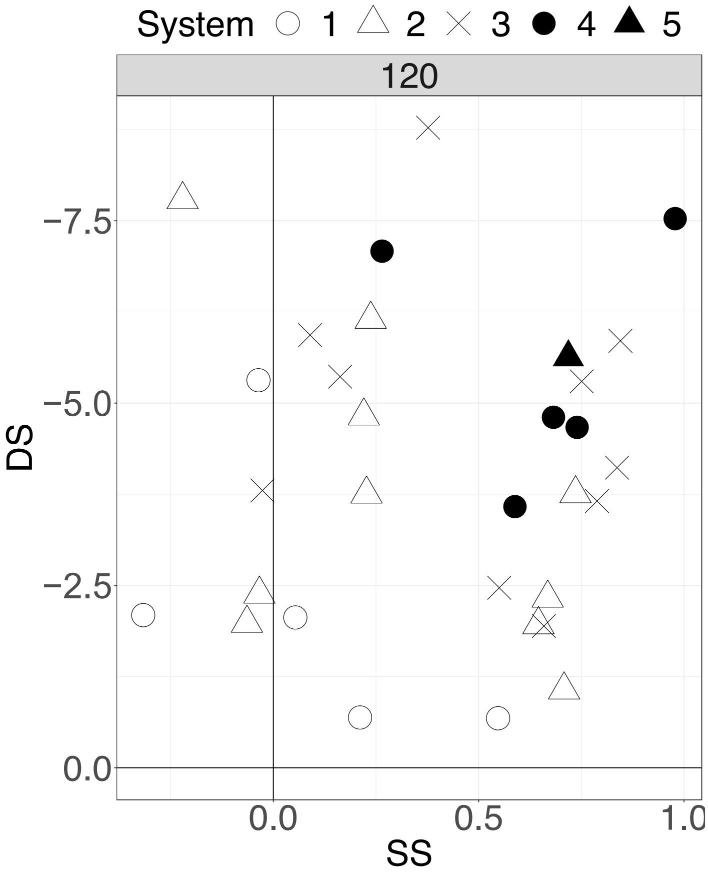


Figure C.7 Mean SS and DS LLR of speaker 120 across 31 systems.

List of abbreviations

|	Given
AcPA	Acoustic phonetic analysis
ASR	Automatic speaker recognition
AuPA	Auditory phonetic analysis
BF	Bayes factor
CI	Credible interval
Cllr	Log likelihood ratio cost
CP	Centred parameter
CPD	Criminal Practice Directions
CPS	Crown Prosecution Service
DCT	Discrete cosine transform
DP	Direct parameter
DS	Different speaker
DyViS	Dynamic Variability in Speech
E	Evidence
EER	Equal error rate
ELUB	Empirical lower and upper bound
ENFSI	European Network of Forensic Science Institutions
EU	Expected utility
F(1-5)	Formant (1st to 5th)
F0	Fundamental frequency
FP	Filled pause
FSAAWG	Forensic Speech and Audio Analysis Working Group
FSS	Forensic speech science
FVC	Forensic voice comparison
GMM	Gaussian mixture model
GSM	Global system for mobile communications
H	Hypothesis
HASR	Human-assisted automatic speaker recognition
Hd	Defence hypothesis
Hp	Prosecution hypothesis
I	Background information

IAFPA	International Association for Forensic Phonetics and Acoustics
IARPA	Intelligence Advanced Research Projects Activity
IPA	International Phonetic Alphabet
IQR	Interquartile range
KDE	Kernel density estimation
KS	Known sample
LLR	Log$_{10}$ LRs
LR	Likelihood ratio
LTF0	Long-term fundamental frequency
MAP	Maximum a posteriori
MFCC	Mel-frequency cepstral coefficients
MVKD	Multivariate kernel density
p	Probability
PAV	Pool adjacent violator
PDF	Probability density function
QS	Questioned sample
RMSD	Root-mean-square-deviation
RQ	Research question
SD	Standard deviation
SFP	Sentence final particle
SS	Same speaker
SSBE	Standard Southern British English
TUULS	The Use and Utility of Localised Speech Forms
UBM	Universal background model
UKPS	UK Position Statement

Legal cases

Daubert v Merrell Dow Pharmaceuticals [1993] 509 US 579

People v Collions, 68 Cal. 2d 319 (Cal. 1968)

R v T [2010] EWCA Crim 2439.

Bibliography

Alexander, A., & Drygajlo, A. (2004). Scoring and Direct Methods for the Interpretation of Evidence in Forensic Speaker Recognition. *Proceedings of Interspeech.* Jeju Island, South Korea. pp. 2397-2400, doi: 10.21437/Interspeech.2004-540

Aitken, C. G. G. (1995). *Statistics and the Evaluation of Evidence for Frensic Scientists*. John Wiley & Sons.

Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*(1), 109–122. https://doi.org/10.1046/j.0035-9254.2003.05271.x

Aitken, C. G. G., & Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists* (Second Edition). John Wiley & Sons.

Ali, T., Spreeuwers, L., Veldhuis, R., & Meuwly, D. (2015). Sampling variability in forensic likelihood-ratio computation: A simulation study. *Science & Justice*, *55*(6), 499–508. https://doi.org/10.1016/j.scijus.2015.05.003

Andrus, T., Dubinski, E., Fiscus, J., Gillies, B., Harper, M., T, H., Hefright, B., Jarrett, A., Lin, W., Ray, J., Rytting, A., Shen, W., Tzoukermann, E., & Wong, J. (2016). *IARPA Babel Cantonese Language Pack IARPA-babel101b-v0.4c LDC2016S02.* LDC. https://catalog.ldc.upenn.edu/LDC2016S02

Arellano-Valle, R. B., & Azzalini, A. (2013). The centred parameterization and related quantities of the skew-t distribution. *Journal of Multivariate Analysis*, *113*, 73–90. https://doi.org/10.1016/j.jmva.2011.05.016

Azzalini, A. (2020). *The R package 'sn': The Skew-Normal and Related Distributions such as the Skew-t* (version 1.6-2) [Computer software].

Blatchford, H., & Foulkes, P. (2006). Idenfication of voices in shouting. *International Journal of Speech Language and the Law*, *13*(2), 962. https://doi.org/10.1558/ijsll.2006.13.2.241

Broeders, A. (1999). *Some observations on the use of probability scales in forensic identification*. https://doi.org/10.1558/SLL.1999.6.2.228

Brümmer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, F., Karafiat, M., van Leeuwen, D. A., Matejka, P., Schwarz, P., & Strasheim, A. (2007). Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker

Recognition Evaluation 2006. *Proceedings of IEEE Transactions on Audio, Speech, and Language*, *15*(7), 2072–2084. https://doi.org/10.1109/TASL.2007.902870

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, *20*(2–3), 230–275. https://doi.org/10.1016/j.csl.2005.08.001

Brümmer, N. (2010). *Measuring, Refining and Calibrating Speaker and Language Information Extracted from Speech.* Unpublished Doctoral dissertation. University of Stellenbosch.

Brümmer, N., & Swart, A. (2014). Bayesian Calibration for Forensic Evidence Reporting. *Proceedings of Interspeech*, Singapore, pp.388–392.

Brümmer, N. (2013). Tutorial for Bayesian forensic likelihood ratio. *arXiv preprint arXiv:1304.3589*.

Byrne, C., & Foulkes, P. (2004). The ‘Mobile Phone Effect’ on vowel formants. *International Journal of Speech Language and the Law*, *11*(1), 83–102.

Champod, C., & Evett, I. W. (2000). Commentary on A.P.A. Breoders (1999) ‘Some observations on the use of probability in forensic identification’. *Forensic Linguistics*, *7*(2), pp.238–243.

Champod, C., & Meuwly, D. (2000). The inference of identity in forensic speaker recognition. *Speech Communication*, *31*(2–3), 193–203. https://doi.org/10.1016/S0167-6393(99)00078-3

Chen, A., & Rose, P. (2012). Likelihood Ratio-based Forensic Voice Comparison with the Cantonese Triphthong /iau/. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, Macquarie, Australia, pp.197–200.

Clark, H. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Coleman, R. F., & Walls, H. J. (1974). The evaluation of scientific evidence. *Criminal Law Review*, pp. 276–287.

CPD. (2015). *England & Wales Criminal Practice Directions*. https://www.justice.gov.uk/courts/procedure-rules/criminal/docs/2015/crim-practice-directions-V-evidence-2015.pdf

CPS. (2019). *UK Crown Prosecution Service*. https://www.cps.gov.uk/legal-guidance/expert-evidence

Curran, J. M. (2005). An introduction to Bayesian credible intervals for sampling error in DNA profiles. *Law, Probability and Risk*, *4*(1–2), pp. 115–126. https://doi.org/10.1093/lpr/mgi009

Curran, J. M., Buckleton, J. S., Triggs, C. M., & Weir, B. S. (2002). Assessing uncertainty in DNA evidence caused by sampling effects. *Science & Justice*, *42*(1), 29–37. https://doi.org/10.1016/S1355-0306(02)71794-2

Curran, J. M. (2016). Admitting to uncertainty in the LR. *Science & Justice, 56*, 380-382. http://dx.doi.org/10.1016/j.scijus.2016.05.005

Delignette-Muller, M. L., & Dutang, C. (2015). An R Package for Fitting Distributions. *Journal of Statistical Software*, *64*(4). https://doi.org/10.18637/jss.v064.i04

Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D. (1998). SHEEP, GOATS, LAMBS and WOLVES: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation. *Proceedings of International Conference of Spoken Language*, Sydney, Australia, paper 0608.

Dunstone, T., & Yager, N. (2009). *Biometric system and data analysis: Design, evaluation, and data mining*. Springer.

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). Sensitivity and Specificity of Information Criteria. *The Pennsylvania State University Technical Report Series 12* (119).

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap.* Chapman and Hall/CRC Press.

Eliason, S. R. (1993). *Maximum Likelihood Estimation: Logic and Practice*. SAGE.

ENFSI. (2015). *Guideline for evaluative reporting in forensic science*. https://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf

Enzinger, E., & Morrison, G. (2017). Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case. *Forensic Science International*, *227*, pp. 30–40.

Enzinger, E., & Morrison, G. S. (2012). The importance of using between- session test data in evaluating the performance of forensic-voice- comparison systems. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*. Macquarie, Australia. pp. 137–140.

Enzinger, E., Morrison, G. S., & Ochoa, F. (2016). A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case. *Science & Justice*, *56*(1), 42–57. https://doi.org/10.1016/j.scijus.2015.06.005

Enzinger, E., Zhang, C., & Morrison, G. S. (2012). Voice source features for forensic voice comparison – an evaluation of the GLOTTEX software package. *Proceedings of Odyssey: The Speaker and Language Recognition Workshop.* pp.78–85.

Fecher, N., & Watt, D. (2013). Effects of Forensically-Realistic Facial Concealment on Auditory-Visual Consonant Recognition in Quiet and Noise Conditions. *Proceedings of Auditory-Visual Speech*, pp.81–86.

Finkelstein, M. O., & Fairley, W. B. (1970). A Bayesian Approach to Identification Evidence. *Harvard Law Review*, *83*(3), 489–517. https://doi.org/10.2307/1339656

Foulkes, P., Carrol, G., & Hughes, S. (2004, July). Sociolinguistic and acoustic variability in filled pauses. *Presentation at International Association for Forensic Phonetics and Acoustics*, Helsinki, Finland.

Foulkes, P., & French, P. (2001). Forensic phonetics and sociolinguistics. In *Concise Encyclopedia of Sociolinguistics* (pp. 329–332). Elsevier.

Foulkes, P., & French, P. (2012). Forensic Speaker Comparison: A Linguistic–Acoustic Perspective. In *The Oxford Handbook of Language and Law* (pp. 557–572). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199572120.013.0041

French, P., & Harrison, P. (2006). Investigative and evidential applications of forensic speech science. In *Witness Testimony: Psychological, Investigative and Evidential Perspectives* (pp. 247–262). Oxford University Press.

French, P., & Harrison, P. (2007). Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases, with a foreword by Peter French and Philip Harrison. *International Journal of Speech Language and the Law*, *14*(1), pp. 39-59. https://doi.org/10.1558/ijsll.v14i1.137

Gold, E. (2009). *The effects of video and voice recorders in cellular phones on vowel formants and fundamental frequency*. Unpublished MSc dissertation. University of York.

Gold, E., & French, P. (2011). International Practices in Forensic Speaker Comparison. *International Journal of Speech Language and the Law*, *18*(2), 293–307. https://doi.org/10.1558/ijsll.v18i2.293

Gold, E., & French, P. (2019). International practices in forensic speaker comparisons: Second survey. *International Journal of Speech Language and the Law*, *26*(1), 1–20. https://doi.org/10.1558/ijsll.38028

Gold, E., & Hughes, V. (2014). Issues and opportunities: The application of the numerical likelihood ratio framework to forensic speaker comparison. *Science & Justice*, *54*(4), 292–299. https://doi.org/10.1016/j.scijus.2014.04.003

Gold, E., & Hughes, V. (2015). Front-end approaches to the issue of correlations in forensic speaker comparison. *Proceedings of 18th International Congress of Phonetic Sciences*. Glasgow, UK.

Greisbach, R., Esser, O., & Weinstock, C. (1995). Speaker identification by formant contours. In A. Braun, & O. Köster (eds). *Studies in forensic phonetics* (pp. 49-55). Trier: Wissenschatlicher Verlag.

Grigoras, C., Smith, J., Morrison, G., & Enzinger, E. (2013). Forensic audio analysis – Review: 2010-2013. *Proceedings of the 17th International Science Managers' Symposium*, pp.621–637.

Guillemin, B. J., & Watson, C. (2009). Impact of the GSM Mobile Phone Network on the Speech Signal – Some Preliminary Findings. *International Journal of Speech Language and the Law*, *15*(2), 193–218. https://doi.org/10.1558/ijsll.v15i2.193

Gwo, C.-Y., & Wei, C.-H. (2016). Shoeprint retrieval: Core point alignment for pattern comparison. *Science & Justice*, *56*(5), 341–350. https://doi.org/10.1016/j.scijus.2016.06.004

Home Office. (2003). *Criminal Justice Act (Chapter 44)*. Her Majesty's Stationery Office.

Huang, P. P., & Kok, G. P. (1999). *Speak Cantonese, Book One*. Yale University Press.

Hughes, V. (2014). *The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison*. Unpublished Doctoral dissertation. University of York.

Hughes, V. (2017). Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough? *Speech Communication*, *94*, 15–29. https://doi.org/10.1016/j.specom.2017.08.005

Hughes, V., Cardoso, A., Foulkes, P., French, P., Gully, A., & Harrison, P. (2019). Forensic voice comparison using long-term acoustic measures of laryngeal voice quality. *Proceedings of 19th International Congress of Phonetic Sciences*. Melbourne, Australia.

Hughes, V., & Foulkes, P. (2015a). The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age. *Speech Communication*, *66*, 218–230. https://doi.org/10.1016/j.specom.2014.10.006

Hughes, V., & Foulkes, P. (2015b). Variability in analyst decisions during the computation of numerical likelihood ratios. *International Journal of Speech Language and the Law*, *21*(2), 279–315. https://doi.org/10.1558/ijsll.v21i2.279

Hughes, V., Harrison, P., Foulkes, P., French, P., & Gully, A. J. (2019). Effects of formant analysis settings and channel mismatch on semiautomatic forensic voice comparison. *Proceedings of International Congress of Phonetic Sciences*. Melbourne, Australia. pp. 3080–3084.

Hugues, V. and Kinoshita, Y. (under revision) Likelihood ratio analysis, reference population, and the presentation of conclusions. To appear in Nolan, F., Hudson, T. and McDougall, K. (eds.) *Oxford handbook of Forensic Phonetics.* Oxford: OUP.

Hughes, V., Wood, S., & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech Language and the Law*, *23*(1), 99–132. https://doi.org/10.1558/ijsll.v23i1.29874

Ishihara, S., & Kinoshita, Y. (2008). How Many Do We Need? Exploration of the Population Size Effect on the Performance of Forensic Speaker Classification. *Proceedings of 9th Annual Conference of the International Speech Communication Association*, pp.1941–1944.

Jessen, M. (2008). Forensic Phonetics. *Language and Linguistics Compass*, *2*(4), 671–711. https://doi.org/10.1111/j.1749-818X.2008.00066.x

King, J. (2012). *Assessing the discriminatory power of vocalic hesitation markers for forensic speaker comparison casework*. Unpublished MSc Dissertation. University of York.

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

Kinoshita, Y. (2005). Does Lindley's LR estimation formula work for speech data? Investigation using long-term f0. *International Journal of Speech, Language and the Law*, *12*(2), 235–254. https://doi.org/10.1558/sll.2005.12.2.235

Kinoshita, Y., & Ishihara, S. (2014). Background population: How does it affect LR based forensic voice comparison? *International Journal of Speech Language and the Law*, *21*(2), 191–224. https://doi.org/10.1558/ijsll.v21i2.191

Kinoshita, Y., & Ishihara, S. (2010). F0 Can Tell Us More: Speaker Verification Using the Long Term Distribution. *Proceedings of the 13th Australasian International Conference on Speech Science and Technology.*, pp.50–53.

Kinoshita, Y., Ishihara, S., & Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *The International Journal of Speech, Language and the Law*, *16*(1), 21. https://doi.org/doi : 10.1558/ijsll.v16i1.91

Kinoshita, Y., Osanai, T., & Clermont, F. (2018). Forensic voice comparison using sub-band cepstral distances as features: A first attempt with vowels from 306 Japanese speakers under channel mismatch conditions. *Proceedings of 17th Australasian International Conference on Speech Science and Technology*, pp.45–48.

Kinoshita, Y., & Wagner, M. (2014). LR-based forensic voice comparison under severe test-data scarcity. *Proceedings of Annual Conference of the International Speech Communication Association.* pp. 16–19.

Kirchhübel, C. & Brown, G (2021). Competency Testing: Opportunities and Challenges. *Presentation at International Association for Forensic Phonetics and Acoustics*. Marburg, Germany.

Künzel, H. J. (2001). Beware of the 'telephone effect': The influence of telephone transmission on the measurement of formant frequencies. *The International Journal of Speech, Language and the Law*, *8*(1), 80–99. https://doi.org/10.1558/ijsll.v8i1.80

Kwok, H. (1984). *Sentence particles in Cantonese*. Centre of Asian Studies, University of Hong Kong.

Law, A. (2002). Cantonese sentence-final particles and the CP domain*. *UCL Working Papers in Linguistics*, *14*, 375–397.

Lennon, R., Plug, L., & Gold, E. (2019). A comparison of multiple speech tempo measures: Inter-correlations and discriminating power. *Proceedings of the 19th International Congress of Phonetic Sciences*. Melbourne, Australia. pp.785–789.

Leung, C. (1992). *香港粤语语助词的研究 [A Study of the Utterance Particles in Cantonese as Spoken in Hong Kong].* Unpublished M. Phil Dissertation. Hong Kong Polytechnic University.

Leung, W.-M. (2009). A Study of the Cantonese Hearsay Particle wo from a Tonal Perspective. *International Journal of Linguistics*, *1*(1), 1. https://doi.org/10.5296/ijl.v1i1.204

Li, J., & Rose, P. (2012). Likelihood Ratio-based Forensic Voice Comparison with F-pattern and Tonal F0 from the Cantonese /oy/ Diphthong. *Proceedings of the 14th*

Australasian International Conference on Speech Science and Technology. pp.201–204.

Lindblom, B. (1963). Spectrographic Study of Vowel Reduction. *The Journal of the Acoustical Society of America*, *35*(11), 1773–1781. https://doi.org/10.1121/1.1918816

Liu, X. M. (2006). *刑事侦查程序理论与改革研究 [Criminal investigation theory and reform].* China Legal Publishing House.

Lo, J. (2018). *fvclrr: Likelihood Ratio Calculation and Testing in Forensic Voice Comparison* (2.0.1) [Computer software]. https://github.com/justinjhlo/fvclrr

Lo, J. (2021). Seeing the trees in the forest: Diagnosing individual performance in likelihood ratio based forensic voice comparison. *Presentation at XVII Associazione Italiana Scienza Della Voce Annual Conference*.

Marquis, R., Biedermann, A., Cadola, L., Champod, C., Gueissaz, L., Massonnet, G., Mazzella, W. D., Taroni, F., & Hicks, T. (2016). Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings. *Science & Justice*, *56*(5), 364–370. https://doi.org/10.1016/j.scijus.2016.05.009

McDougall, K. (2004). Speaker-specific formant dynamics: An experiment on Australian English /ai/. *International Journal of Speech Language and the Law*, *11*(1), 103–130.

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law*, *13*(1), 89–126. https://doi.org/10.1558/sll.2006.13.1.89

McDougall, K., & Nolan, F. (2007). Discrimination of speakers using the formant dynamics of /u:/ in British English. *Proceedings of 16th International Congress of Phonetic Science*, 1825–1828.

McGehee, F. (1937). The Reliability of the Identification of the Human Voice. *The Journal of General Psychology*, *17*(2), 249–271.

Menard, S. (2010). *Logistic Regression: From Introductory to Advanced Concepts and Applications—SAGE Research Methods*. SAGE. http://methods.sagepub.com/book/logistic-regression-from-introductory-to-advanced-concepts-and-applications

Molnar, C. (2019). *Interpretable machine learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/

Morrison, G., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., Planting, S., Thompson, W. C., van der Vloed, D., J F Ypma, R., & Zhang, C. (2021). Consensus on validation of forensic voice comparison. *Science & Justice*, *61*(3), 229–309. https://doi.org/10.1016/j.scijus.2021.02.002

Morrison, G., & Poh, N. (2018). Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors. *Science & Justice*, *58*(3), 200–218. https://doi.org/10.1016/j.scijus.2017.12.005

Morrison, G. S. (2009a). Forensic speaker recognition using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aI/. *International Journal of Speech Language and the Law*, *15*(2), 249–266. https://doi.org/10.1558/ijsll.v15i2.249

Morrison, G. S. (2009b). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *The Journal of the Acoustical Society of America*, *125*(4), 2387–2397. https://doi.org/10.1121/1.3081384

Morrison, G. S. (2011a). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM). *Speech Communication*, *53*(2), 242–256. https://doi.org/10.1016/j.specom.2010.09.005

Morrison, G. S. (2011b). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, *51*(3), 91–98. https://doi.org/10.1016/j.scijus.2011.03.002

Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, *45*(2), 173–197. https://doi.org/10.1080/00450618.2012.733025

Morrison, G. S. (2016). Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate. *Science & Justice*, *56*(5), 371–373. https://doi.org/10.1016/j.scijus.2016.05.002

Morrison, G. S., & Enzinger, E. (2016). What should a forensic practitioner's likelihood ratio be? *Science & Justice*, *56*(5), 374–379. https://doi.org/10.1016/j.scijus.2016.05.007

Morrison, G. S., Lindh, J., & Curran, J. M. (2014). Likelihood ratio calculation for a disputed-utterance analysis with limited available data. *Speech Communication*, *58*, 81–90. https://doi.org/10.1016/j.specom.2013.11.004

Morrison, G. S., Ochoa, F., & Thiruvaran, T. (2012). Database selection for forensic voice comparison. *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 62–77.

Morrison, G. S., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., & Goemans Dorny, C. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International*, *263*, 92–100. https://doi.org/10.1016/j.forsciint.2016.03.044

Morrison, G. S., Thiruvaran, T., & Epps, J. (2010). Estimating the Precision of the Likelihood-Ratio Output of a Forensic-Voice-Comparison System. *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 63–70.

Morrison, G. S., Zhang, C., & Rose, P. (2011). An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic Science International*, *208*(1–3), 59–65. https://doi.org/10.1016/j.forsciint.2010.11.001

Nolan, F. (2001). Speaker identification evidence: Its forms, limitations, and roles. *In Proceedings of the Conference Law and Language: Prospect and Retrospect*, 1–19. http://www.ling.cam.ac.uk/francis/LawLang.doc

Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech Language and the Law*, *16*(1), 31–57. https://doi.org/10.1558/ijsll.v16i1.31

Ommen, D. M., Saunders, C. P. & Neumann, C. (2016). An argument against presenting interval quantifications as a surrogate for the value of evidence. *Science & Justice, 56,* 383-387. http://dx.doi.org/10.1016/j.scijus.2016.07.001

Pang, J., & Rose, P. (2012). Likelihood Ratio-Based Forensic Voice Comparison with the Cantonese Diphthong /ei/ F-Pattern. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, 205–208.

Parzen, E. (1962). *On estimation of a Probability Density Function and Mode*. *33*(3), 1065–1076.

Parzen, E., Tanabe, K., & Kitagawa, G. (Eds.). (1998). *Selected Papers of Hirotugu Akaike*. Springer New York. https://doi.org/10.1007/978-1-4612-1694-0

R, core team. (2020). *RStudio: Integrated Development for R*. RStudio, Inc. http://www.rstudio.com/

Ramos-Castro, D. (2007). *Forensic evaluation of the evidence using automatic speaker recognition systems.* Unpublished doctoral dissertation, Universidad Autónoma de Madrid.

Ramos, D., Maroñas, J., & Almirall, J. (2021). Improving calibration of forensic glass comparisons by considering uncertainty in feature-based elemental data. *Chemometrics and Intelligent Laboratory Systems*, *217*, 104399.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, *10*(1–3), 19–41. https://doi.org/10.1006/dspr.1999.0361

Robertson, B., Vignaux, G. A., & Berger, C. E. H. (2016). *Interpreting evidence: Evaluating forensic science in the courtroom* (Second edition). John Wiley and Sons, Inc.

Rodman, R., McAllister, D., Bitzer, D., Cepeda, L., & Abbitt, P. (2002). Forensic speaker identification based on spectral moments. *Forensic Linguistics, 9*(1), 22-42. DOI: 10.1558/sll.2002.9.1.22

Roettger, T. B. (2019). Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *10*(1), 1. https://doi.org/10.5334/labphon.147

Rose, P. (2002). *Forensic Speaker Identification*. Taylor & Francis.

Rose, P. (2013a). More is better: Likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends. *International Journal of Speech Language and the Law*, *20*(1), 77–116. https://doi.org/10.1558/ijsll.v20i1.77

Rose, P. (2013b). Where the science ends and the law begins: Likelihood ratio-based forensic voice comparison in a $150 million telephone fraud. *International Journal of Speech Language and the Law*, *20*(2), 277–324. https://doi.org/10.1558/ijsll.v20i2.277

Rose, P. (2004). Technical Forensic Speaker Identification from a Bayesian Linguist's Perspective. *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*.

Rose, P., & Cuiling, Z. (2018). Conversational Style Mismatch: Its Effect on the Evidential Strength of Long- term F0 in Forensic Voice Comparison. *Proceedings of 17th Australasian International Conference on Speech Science and Technology*, pp.157–160.

Rose, P., Lucy, D., & Osanai, T. (2004). Linguistic-acoustic forensic speaker identification with likelihood ratios from a multivariate hierarchical random effects models: A 'non-idiot's Bayes' approach. *Proceedings of the 10th Australian International Conference on Speech Science and Technology*, 492–497.

Rose, P., & 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. https://doi.org/10.1558/ijsll.v16i1.139

Rose, P., & Wang, X. (2016). Cantonese forensic voice comparison with higher-level features: Likelihood ratio-based validation using F-pattern and tonal F0 trajectories over a disyllabic hexaphone. *Proceedings of Odyssey: The Speaker and Language Recognition Workshop,* 326–333. https://doi.org/10.21437/Odyssey.2016-47

Saks, M. J., & Koehler, J. J. (2005). The Coming Paradigm Shift in Forensic Identification Science. *Science*, *309*, 892–895. https://doi.org/10.1126/science.1111565

San Segundo, E., & Yang, J. (2019). Formant dynamics of Spanish vocalic sequences in related speakers: A forensic-voice-comparison investigation. *Journal of Phonetics*, *75*, 1–26. https://doi.org/10.1016/j.wocn.2019.04.001

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375

Sybesma, R., & Li, B. (2007). The dissection and structural mapping of Cantonese sentence final particles. *Lingua*, *117*(10), 1739–1783. https://doi.org/10.1016/j.lingua.2006.10.003

Taylor, D., Hicks, T., & Champod, C. (2016). Using sensitivity analyses in Bayesian Networks to highlight the impact of data paucity and direct future analyses: A contribution to the debate on measuring and reporting the precision of likelihood ratios. *Science & Justice*, *56*(5), 402–410. https://doi.org/10.1016/j.scijus.2016.06.010

Thompson, W. C., & 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. https://doi.org/10.1007/BF01044641

Tschäpe, N., Trouvain, J., Bauer, D., & Jessen, M. (2005, August). Idiosyncratic patterns of filled pauses. In *14th Annual Conference of the International Association for Forensic Phonetics and Acoustics*. IAFPA, Marrakesh, Morocco.

Vergeer, P., van Es, A., de Jongh, A., Alberink, I., & Stoel, R. (2016). Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating? *Science & Justice*, *56*(6), 482–491. https://doi.org/10.1016/j.scijus.2016.06.003

Vergeer, P., van Schaik, Y., & Sjerps, M. (2020). Measuring calibration of likelihood-ratio systems_ a comparison of four metrics, including a new metric devPAV. *Forensic Science International*, 35. https://doi.org/10.1016/j.forsciint.2021.110722

Vose, D. (2008). *Risk analysis: A quantitative guide* (3rd ed). Wiley.

Wakefield, J. C. (2011). *The English equivalents of Cantonese sentence-final particles: A contrastive analysis* [Hong Kong Polytechnic University]. https://theses.lib.polyu.edu.hk/handle/200/6084

Wang, B., Hughes, V., & Foulkes, P. (2019). Effect of score sampling on system stability in Likelihood Ratio based forensic voice comparison. *Proceedings of International Congress of Phonetic Sciences*. Melbourne, Australia.

Wang, C. Y., & Rose, P. (2012). Likelihood Ratio-based Forensic Voice Comparison with Cantonese /i/ F-pattern and Tonal F0. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, 209–212.

Watt, D., Harrison, P., Hughes, V., French, P., Llamas, C., Braun, A., & Robertson, D. (2020). Assessing the effects of accent-mismatched reference population databases on the performance of an automatic speaker recognition system. *International Journal of Speech Language and the Law*, *27*(1). https://doi.org/10.1558/ijsll.41466

Watt, D., Llamas, C., French, P., Braun, A., & Robertson, D. (2018). A new corpus of Northern Englishes: Building the TUULS database for sociolinguistic and forensic research on phonetic variation in the Northeast of England. *The 8th Northern Englishes Workshop*, 21. https://blogs.ncl.ac.uk/northernenglishes8/files/2018/03/NEW_8_Book_of_Abstracts_final.pdf

Wood, S., Hughes, V., & Foulkes, P. (2014, September 31). Filled pauses as variables in speaker comparison: Dynamic formant analysis and duration measurements improve performance. *International Association for Forensic Phonetics and Acoustics Annual Conference*. Zürich, Switzerland.

Yager, N., & Dunstone, T. (2010). The Biometric Menagerie. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(2), 220–230. https://doi.org/10.1109/TPAMI.2008.291

Zadrozny, B., & Elkan, C. (2002a). *Transforming Classifier Scores into Accurate Multiclass Probability Estimates*. 6.

Zadrozny, B., & Elkan, C. (2002b). Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the Eighth ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining - KDD '02, 694. https://doi.org/10.1145/775047.775151

Zhang, C., Morrison, G. S., Enzinger, E., & Ochoa, F. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – Female voices. *Speech Communication*, *55*(6), 796–813. https://doi.org/10.1016/j.specom.2013.01.011

Zhang, C., Morrison, G. S., & Rose, P. (2008). Forensic Speaker Recognition in Chinese: A Multivariate Likelihood Ratio Discrimination on /i/ and /y/. *Proceedings of Interspeech*, 1937–1940. doi: 10.21437/Interspeech.2008-512

Zhang, C., Morrison, G. S., & Thiruvaran, T. (2011). Forensic voice comparison using Chinese /iau/. *Proceedings of International Congress of Phonetic Sciences, Hong Kong*, 2280–2283.