Effects of variability on the outcome of numerical likelihood ratios for forensic voice comparison

Vincent Hughes
Department of Language and Linguistic Science, University of York
vh503@york.ac.uk

THE UNIVERSITY of York

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

1. Background:

- Forensic voice comparison (FVC) = comparison of speech patterns in recording of the suspect and a recording of the criminal
- Applying the likelihood ratio (LR) framework to FVC requires pragmatic decisions relating to sources of linguistic variability when compiling a database (non-case specific) and/or defining the _relevant population_ (case-specific) in order to assess typicality
- However, generally the only factors controlled are **speaker sex**, **broad regional dialect** (often defined by language) and **non-contemporaneity** [6,7,8]
- But why are these factors favoured over others? The magnitude of their effect on the LR isn't well understood
- No agreement over practical factors such as how many speakers or tokens per speaker are required for the reference sample

2. Research Question:

To what extent are numerical LRs affected by (i) the number of reference speakers, (ii) the number of tokens per reference speaker and (iii) dialect mismatch between test and reference data?

3. Data/ Method:

- /uː/ (e.g. goose, _boot_) vowels manually segmented and dynamic time-normalised measurements of F1 and F2 extracted [4]
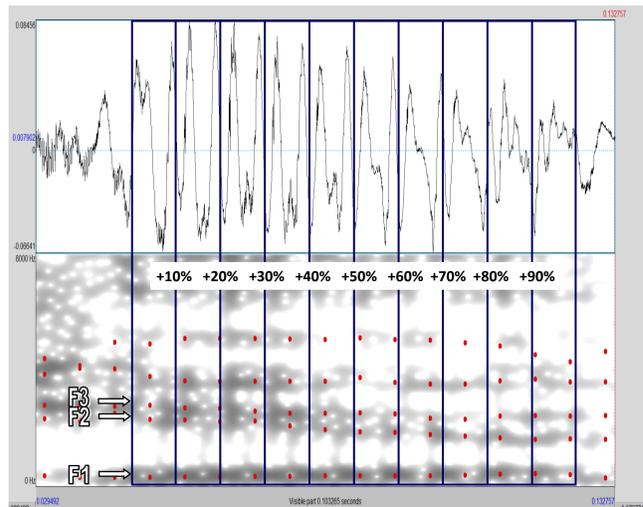
- Reference data: **New Zealand English** (NZE)
  - 102 male speakers/ 13 tokens per speaker
- Test data: **ONZE (NZE)**/ **Manchester (MAN)**/ **Newcastle (NCL)**/ **York (YRK)**
  - 8 male speakers per set/ 16 tokens per speaker
- /uː/ = **not a regional stereotype**
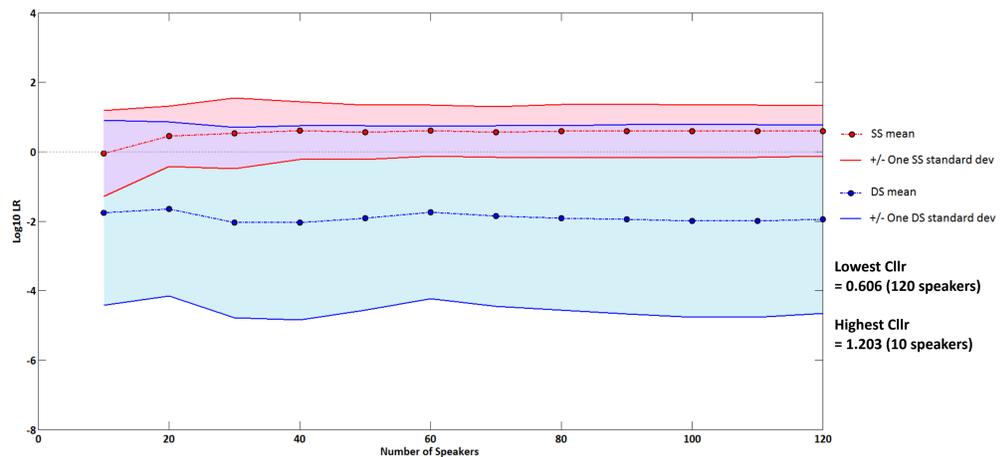 - Linguistically the 4 sets are relatively homogeneous:
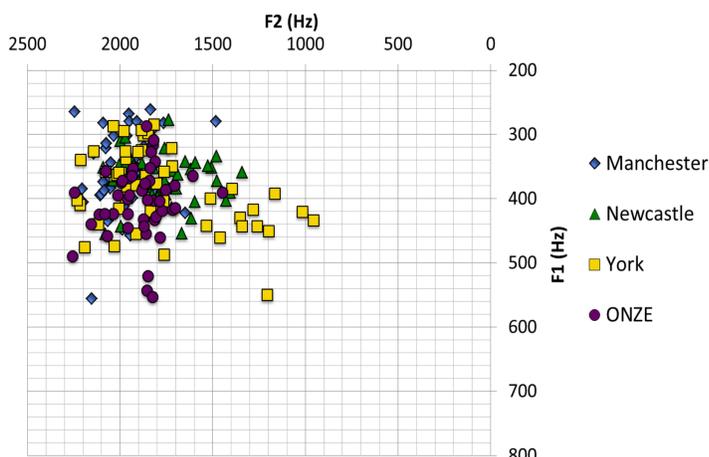
- Raw formant trajectories fitted with quadratic polynomials
- LRs calculated using MVKD formula [1,5]
- Raw LRs converted to log10/ natural log LRs
- Error quantified using **Log-LR Cost Function** (Cllr) [2]
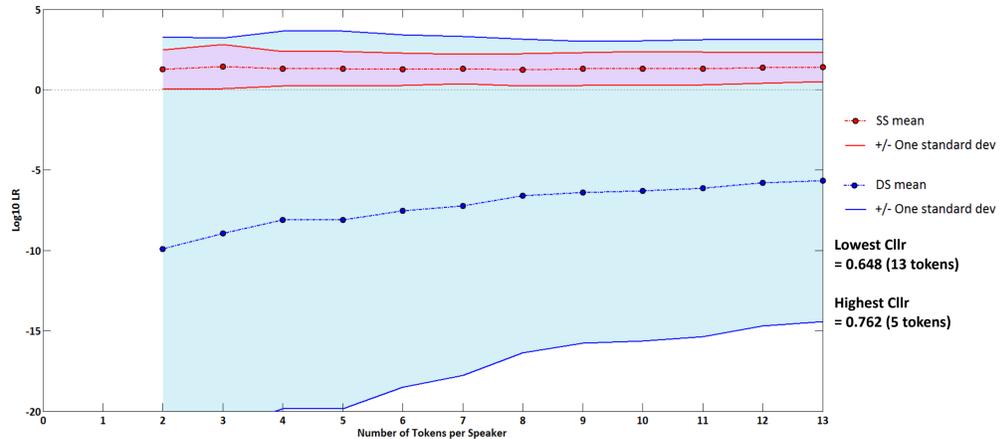- Results analysed with Champod and Evett's verbal scale [3]:

Log10 LR	Verbal Expression
+/- 5	Very Strong Evidence
+/- 4	Strong Evidence
+/- 3	Moderately Strong Evidence
+/- 2	Moderate Evidence
+/- 1	Limited Evidence

4. Results

(i) Number of reference speakers:

Lowest Cllr = 0.606 (120 speakers)

Highest Cllr = 1.203 (10 speakers)

(ii) Number of tokens per reference speaker:

Lowest Cllr = 0.648 (13 tokens)

Highest Cllr = 0.762 (5 tokens)

(iii) Dialect mismatch:

	Cllr:
MAN	1.151
NCL	1.203
YRK	0.470
NZE	0.586
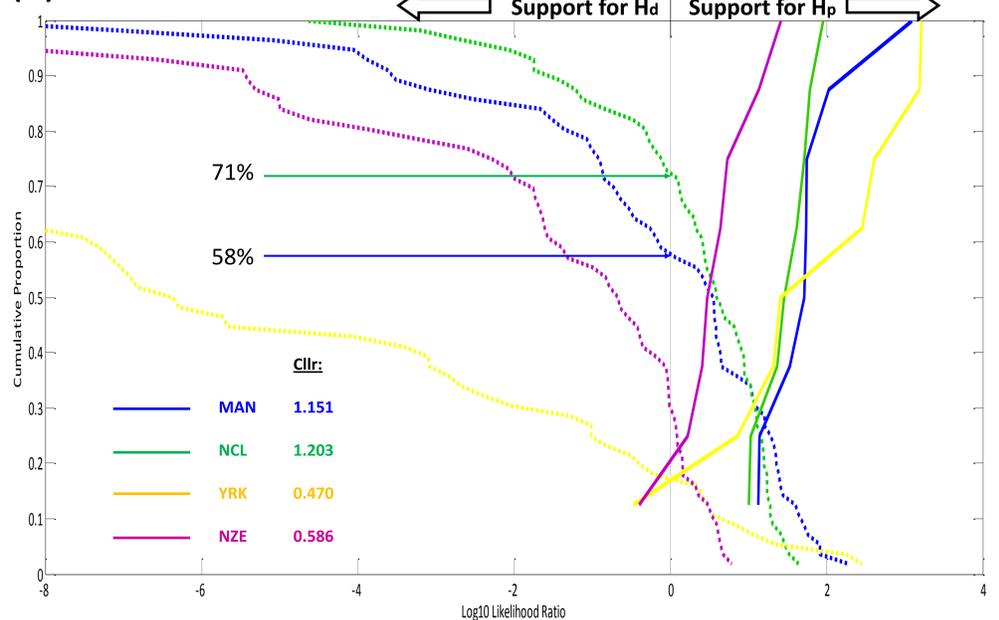
- Same-speaker (SS) mismatch evidence overestimated by an average of **one verbal category** (limited to moderate support for H$_p$)
- High levels of contrary to fact support for H$_p$ for different-speaker (DS) pairs from MAN and NCL
- Poor Cllr performance for MAN and NCL
- York = high levels of within- and between-speaker variation

5. Conclusion:

- Yes **dialect matters**: even for segmental features which aren't expected to vary considerably according to region
- BUT more significantly the extent to which regional dialect is an issue **depends on the feature** under investigation
- We also need to consider other sources of social and stylistic variation, such as **class**, **ethnicity**, **age**, **interlocutor**, **topic** etc. for some features at least

References:
[1] Aitken, C. G. G. and Lucy, D. (2004) Evaluation of trace evidence in the form of multivariate data. _Applied Statistics_ 54: 109-122.
[2] Brümmer, N. and du Preez, J. (2006) Application of independent evaluation of speaker detection. _Computer Speech and Language_ 20: 230-275.
[3] Champod, C. and Evett, I. W. (2000) Commentary on A.P.A. Broeders (1999). _Forensic Linguistics_ 7(2): 238-243.
[4] McDougall, K. (2004) Speaker-specific formant dynamics: an experiment on Australian English /aɪ/. _IJSLL_ 13(1): 89-126.
[5] Morrison, G. S. (2007) MatLab implementation of Aitken and Lucy's (2004) likelihood ratio software using multivariate kernel-density estimation.
[6] Morrison, G. S., Rose, P. and Zhang, C. (2012, to appear) Protocol for the collection of databases of recordings for forensic-voice-comparison research and practise. _Australian Journal of Forensic Sciences._
[7] Rose, P. (2004) Technical forensic speaker identification from a Bayesian linguist's perspective. _Odyssey._ 31st May—3rd June 2004. 3-10.
[8] Rose, P., Kinoshita, Y. and Alderman, T. (2006) Realistic extrinsic forensic speaker discrimination with the diphthong /aɪ/. _Proceedings of the 11th Australasian International Conference on Speech Science and Technology._ 6-8 December 2006. University of Auckland, NZ. 329-334.
[9] Rose, P. and Morrison, G. S. (2009) A response to the UK position statement on forensic speaker comparison. _IJSLL_ 16(1): 139-163.