System performance as a function of calibration methods, sample size and sampling variability in likelihood ratio-based forensic voice comparison

Bruce X. Wang and *Vincent Hughes*
{xw961|vincent.hughes}@york.ac.uk
Department of language and linguistic science, University of York

Introduction

In Forensic Voice Comparison (FVC)

VS.

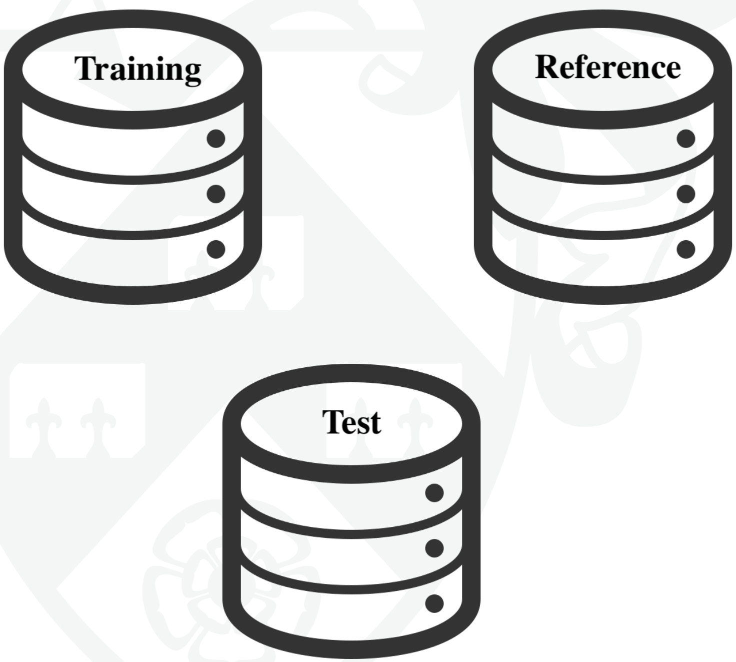
In recent years,

- Likelihood ratio (LR) framework & Growing pressure on experts

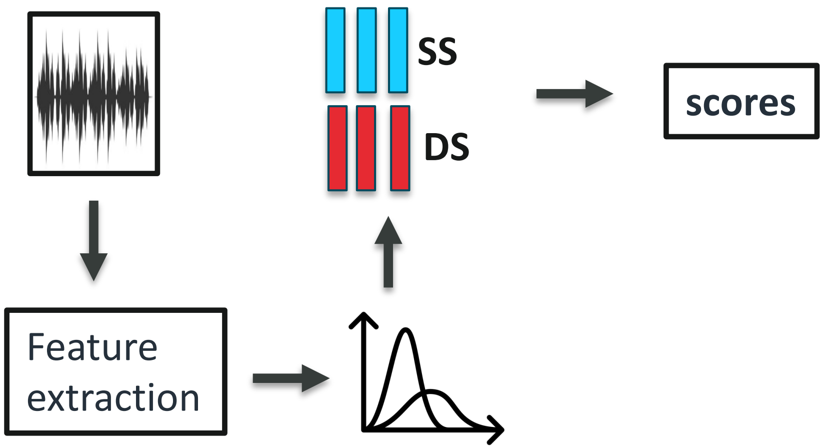
- Established procedures

Introduction

Objectivity *vs.* Subjectivity



Stage 1: *Feature-to-Score*

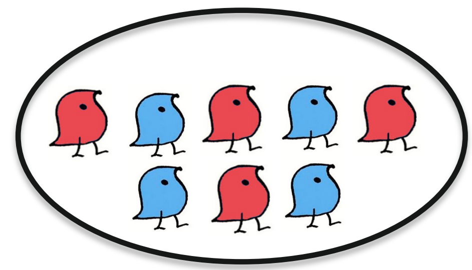


Stage 2: *Score-to-LR*

Introduction

Previous LR-based FVC studies looked in the effect of

- size of training, test and reference data [1,2,3]
- configurations of training, test and reference data [4,5]



Showing that the effect of sampling variability is inevitable regardless of

- the size of training, test and reference data

&

- Configurations of training, test and reference data

Introduction

Therefore, calibration is extremely important for system evaluation and optimisation because one does not want to

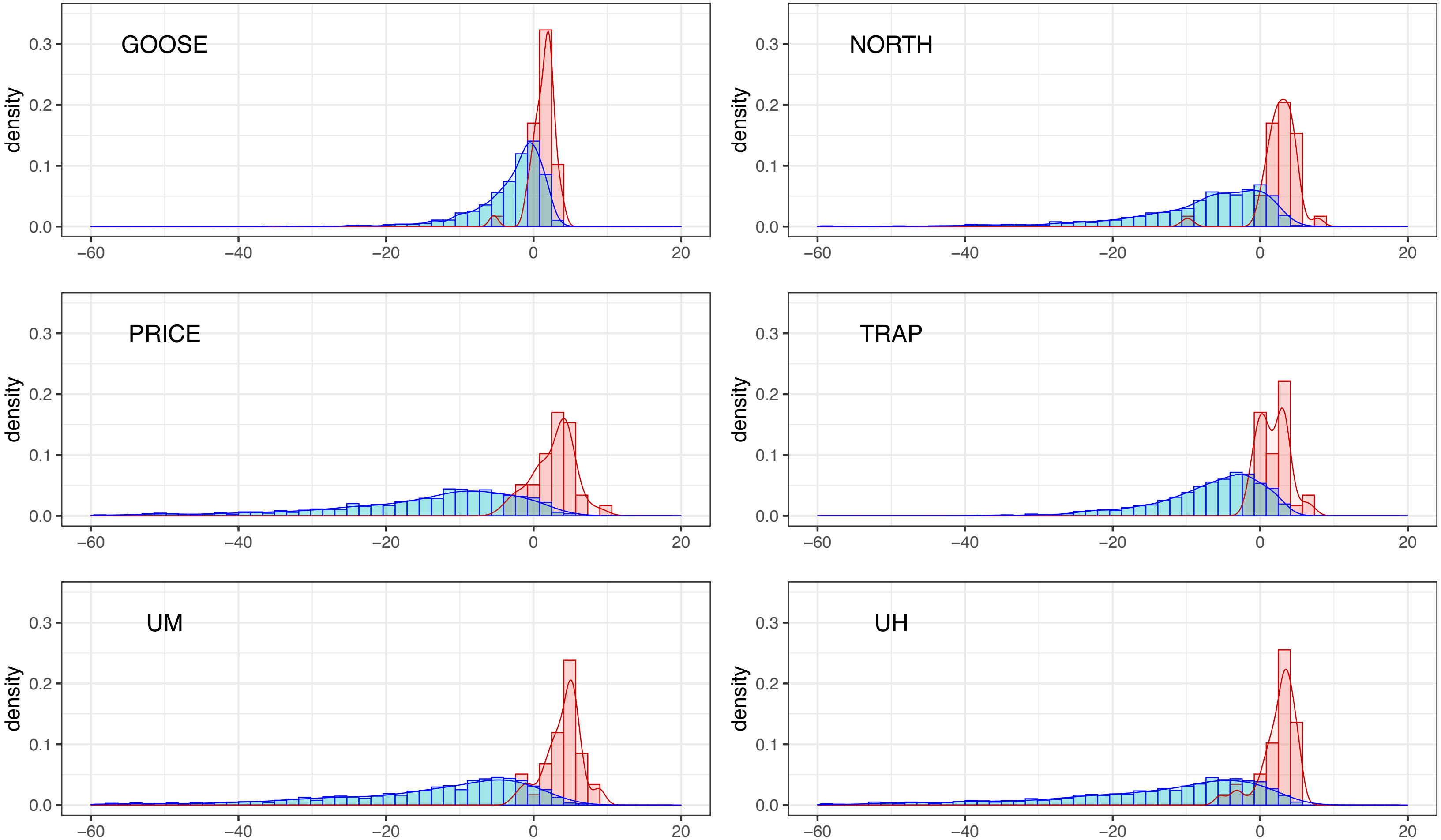
- give extreme LRs that over- or underestimate the strength of voice evidence
- give false information to the court that leads to miscarriage of justice

Previous studies [6, 7] have tested the effectiveness of different calibration methods. However,

- limited sets of data
- scores → Gaussian distribution

Questions

Pilot: segmental features based on 36 SSBE speakers (vowel formant data from [15])



Questions

Given the limit of sample size in the real world, how does system perform when

- scores are skewed

&

- sample size is limited ?

Can we incorporate uncertainty into the LR itself in LR computation?

Can certain calibration methods reduce the level of uncertainty when the sample size is small?

Current study

We simulated scores from skewed distributions to test four calibration methods:

- Logistic regression [8]
- Empirical lower and upper bound [9]
- Regularised logistic regression [7]
- Bayesian model [10]

Claimed incorporate uncertainty into the LR itself, such that LRs will be closer to 1 when uncertainty is high (i.e. when sample size is small).

Aiming to investigate:

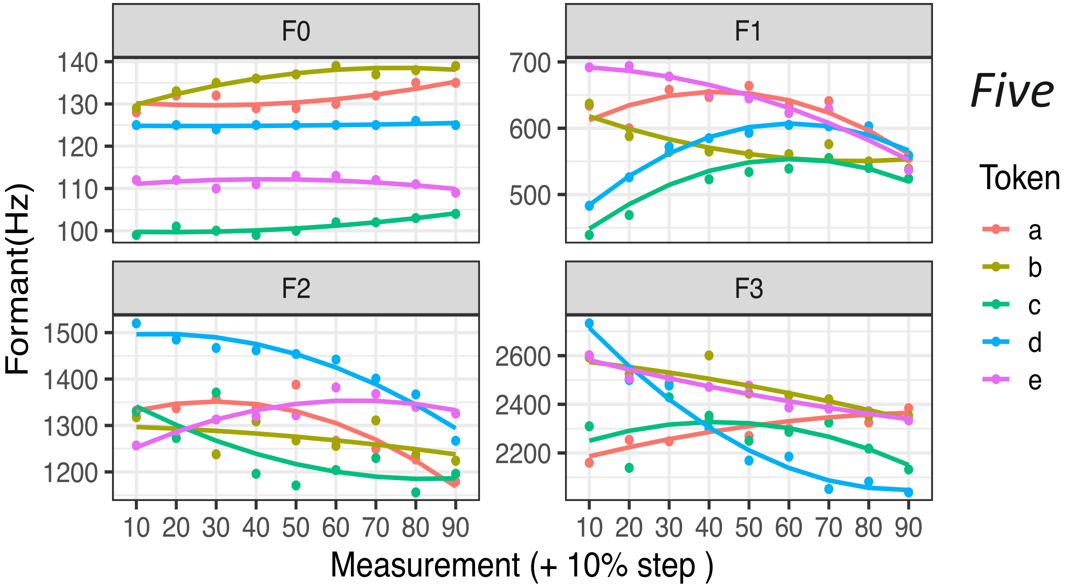
a. overall system validity
and
b. the reliability of the system validity

Methods

Data

Parameters for score distribution simulations were derived from the acoustics of filled pause ***um***,

- 90 SSBE speakers from DyViS [11]
- Quadratic curves → F1, F2, F3 and f0
- Multivariate kernel density (MVKD) [12]



Five tokens, speaker 114 DyViS [11].

Methods

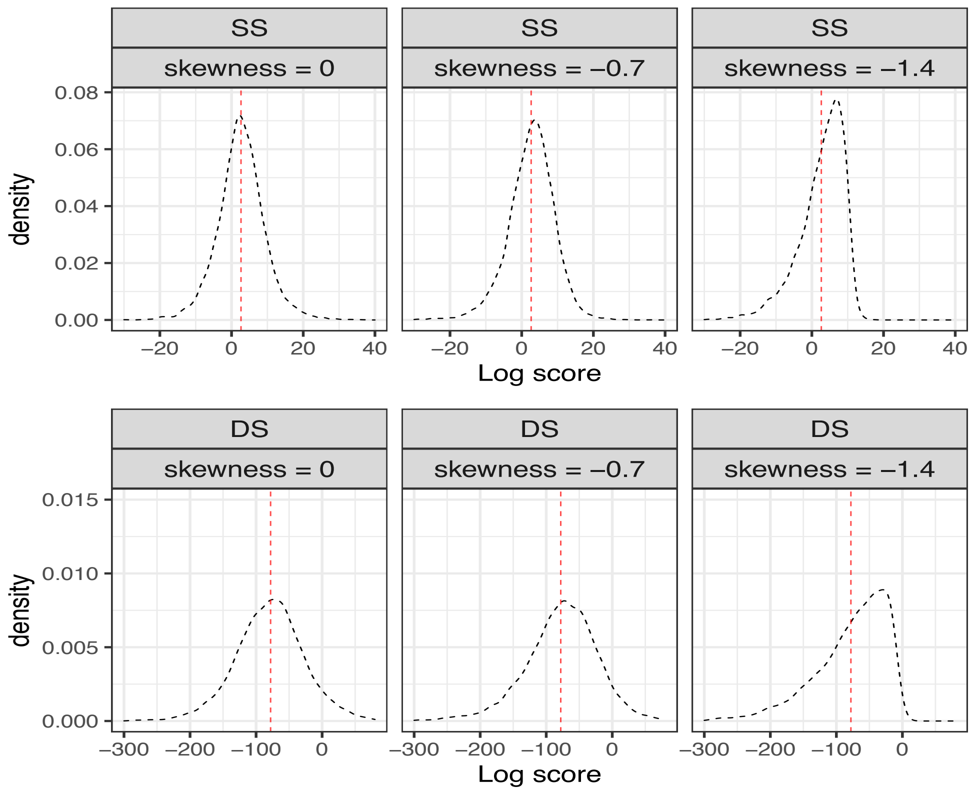
Data

Based on the parameters from scores generated from real data,

Score distribution parameters used for simulation.

Distribution parameters	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	6.6
Set (b)	-0.7	-0.7						
Set (c)	-1.4	-1.4						

Scores were simulated using `sn`[13] function in R[14].



Methods

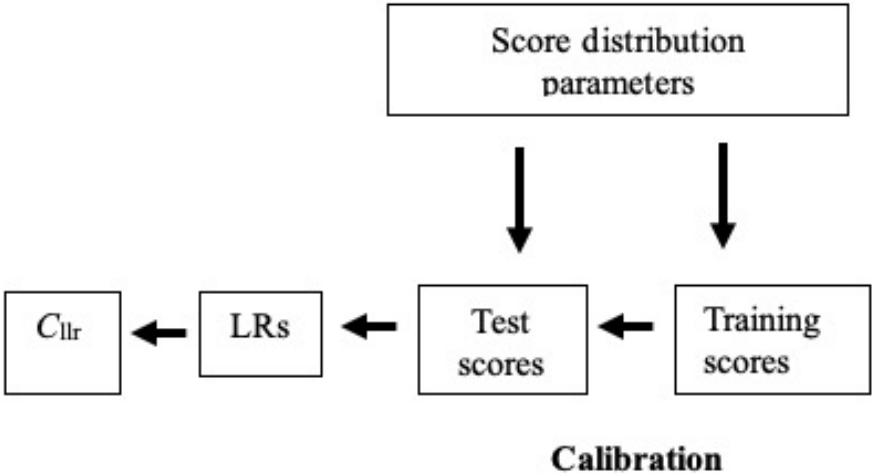
Sample size

Training and test scores were sampled with increasing sample sizes:

- 20 to 100 speakers
- 10-speaker increasements
- SS scores 20 ~ 100
- DS scores 380 ~9900

Experiments were replicated 100 times for each sample size and calibration method.

Schematic of the simulation process using score distribution parameters, replicated 100 times for each sample size.



Methods

Evaluation

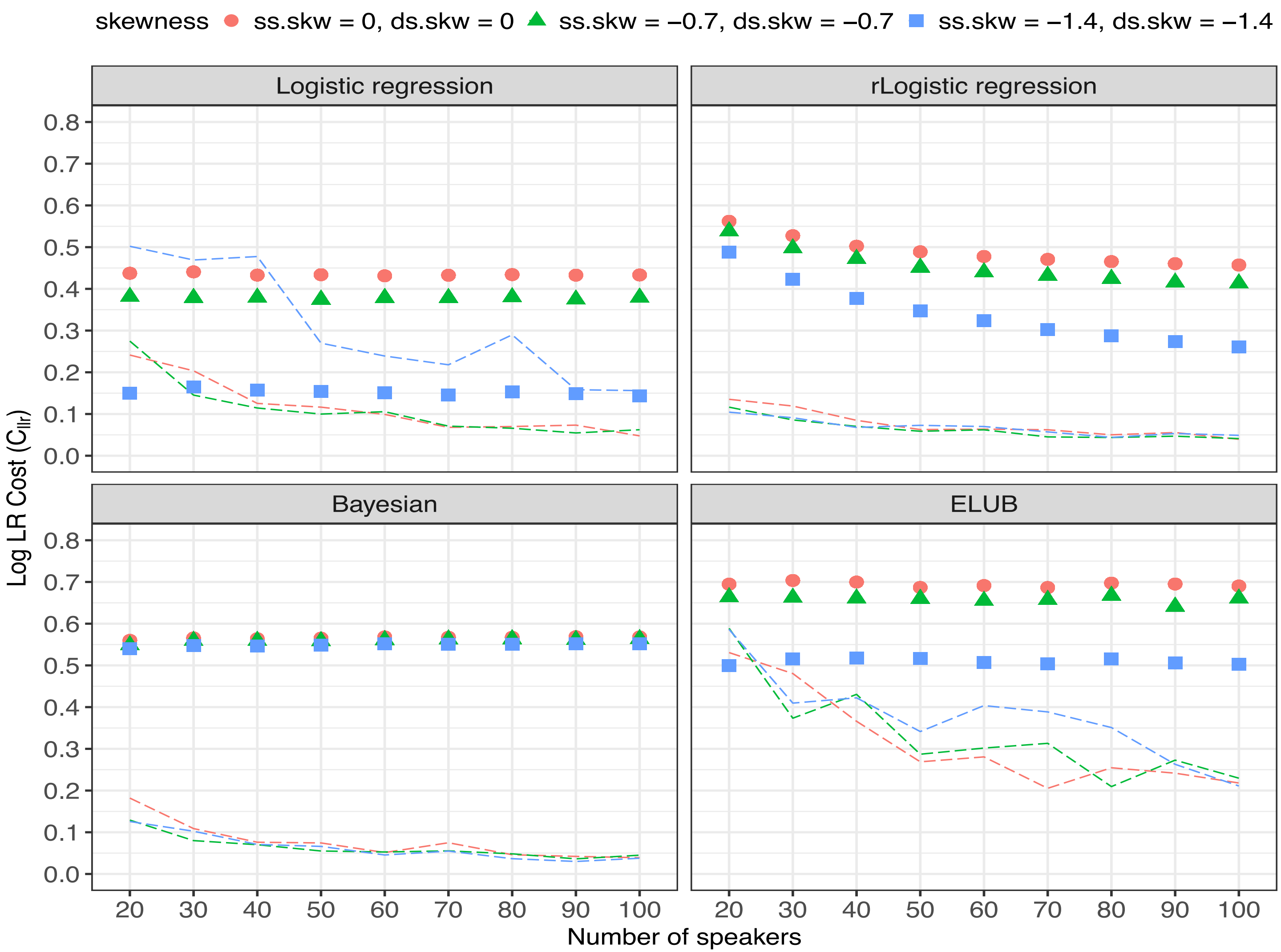
System validity: C_{llr} mean of 100 replications.

System reliability: C_{llr} range, Max. C_{llr} – Min. C_{llr} in 100 replications.

A C_{llr} of less than 1 indicates that the system is capturing useful information.

Systems with better performance should yield both lower C_{llr} mean and range.

Results

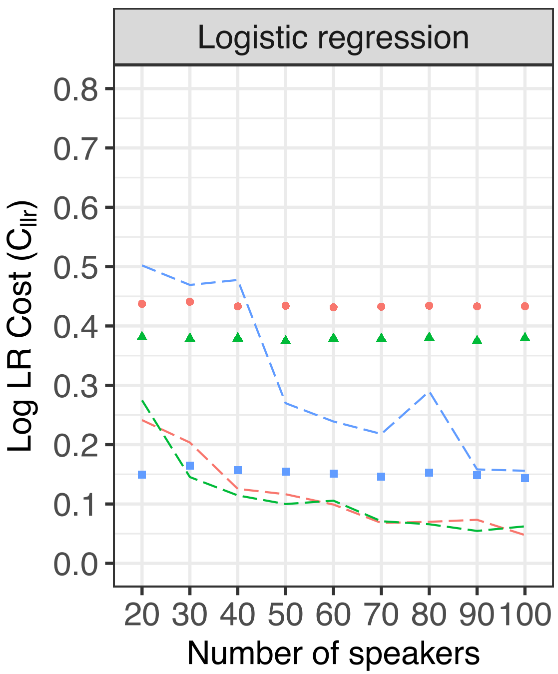


skewness ● ss.skw = 0, ds.skw = 0 ▲ ss.skw = −0.7, ds.skw = −0.7 ■ ss.skw = −1.4, ds.skw = −1.4
Logistic regression
rLogistic regression
Bayesian
ELUB
Log LR Cost (C_{llr})
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
20 30 40 50 60 70 80 90 100
Number of speakers

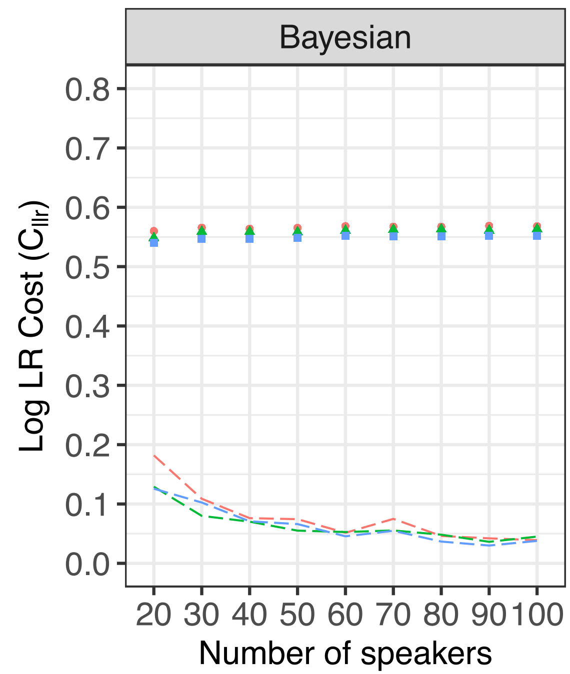
Take-home message

Direct implications:

- Score skewness ***vs.*** calibration methods ***vs.*** sample size
  - e.g., logit reg.



- Validity (C_{llr} mean) vs. reliability (C_{llr} range)
  - e.g., logit reg. ***vs.*** Bayesian model



Take-home message

Wider implications:

- Experts’ decisions
    - Degree of freedom
    - lower uncertainty > higher validity
    - all forms of FVC casework

Corresponding author:

Bruce Wang

Email: xw961@york.ac.uk
Twitter: https://twitter.com/P0rfav0r

References

[1] G. S. Morrison, 'Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate', *Science & Justice*, vol. 56, no. 5, pp. 371–373, Sep. 2016, doi: 10.1016/j.scijus.2016.05.002.

[2] S. Ishihara and Y. Kinoshita, 'How Many Do We Need? Exploration of the Population Size Effect on the Performance of Forensic Speaker Classification'. In *Interspeech,* Brisbane Australia, 2008, p.1941 - 1944.

[3] V. Hughes, 'Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough?', *Speech Communication*, vol. 94, pp. 15–29, 2017, doi: 10.1016/j.specom.2017.08.005.

[4] B. X. Wang, V. Hughes, and P. Foulkes, 'The effect of speaker sampling in likelihood ratio based forensic voice comparison', *International Journal of Speech, Language and the Law*, vol. 26, no. 1, pp. 97–120, Aug. 2019, doi: 10.1558/ijsll.38046.

[5] 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*, *0*(0). https://doi.org/10.1558/ijsll.41466

[6] T. Ali, L. Spreeuwers, R. Veldhuis, and D. Meuwly, 'Sampling variability in forensic likelihood-ratio computation: A simulation study', *Science & Justice*, vol. 55, no. 6, pp. 499–508, Dec. 2015, doi: 10.1016/j.scijus.2015.05.003.

[7]G. S. Morrison and N. Poh, 'Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors', *Science & Justice*, vol. 58, no. 3, pp. 200–218, May 2018, doi: 10.1016/j.scijus.2017.12.005.

[8] N. Brümmer *et al.*, 'Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006', *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 7, pp. 2072–2084, Sep. 2007, doi: 10.1109/TASL.2007.902870.

[9] P. Vergeer, A. van Es, A. de Jongh, I. Alberink, and R. Stoel, 'Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating?', *Science & Justice*, vol. 56, no. 6, pp. 482–491, Dec. 2016, doi: 10.1016/j.scijus.2016.06.003.

[10]N. Brümmer and A. Swart, 'Bayesian Calibration for Forensic Evidence Reporting', in *Interspeech*, Singapore, 2014, pp. 388–392.

[11] F. Nolan, K. McDougall, G. De Jong, and T. Hudson, 'The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research', *International Journal of Speech, Language and the Law*, vol. 16, no. 1, pp. 31–57, Sep. 2009, doi: 10.1558/ijsll.v16i1.31.

[12] 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

[13] A. Azzalini, *The R package 'sn': The Skew-Normal and Related Distributions such as the Skew-t*. 2020.

[14] Core team R, *RStudio: Integrated Development for R*. RStudio, Inc., 2020.

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