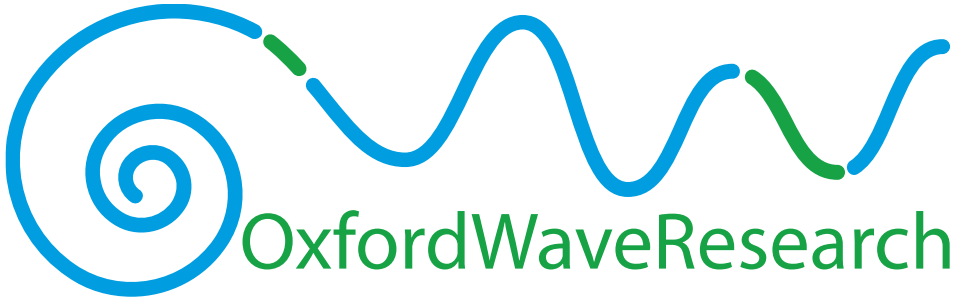
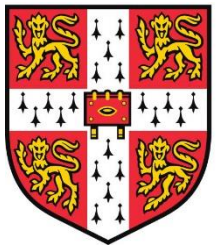




UNIVERSITY OF CAMBRIDGE

How does the perceptual similarity of the relevant population to a questioned speaker affect the likelihood ratio?

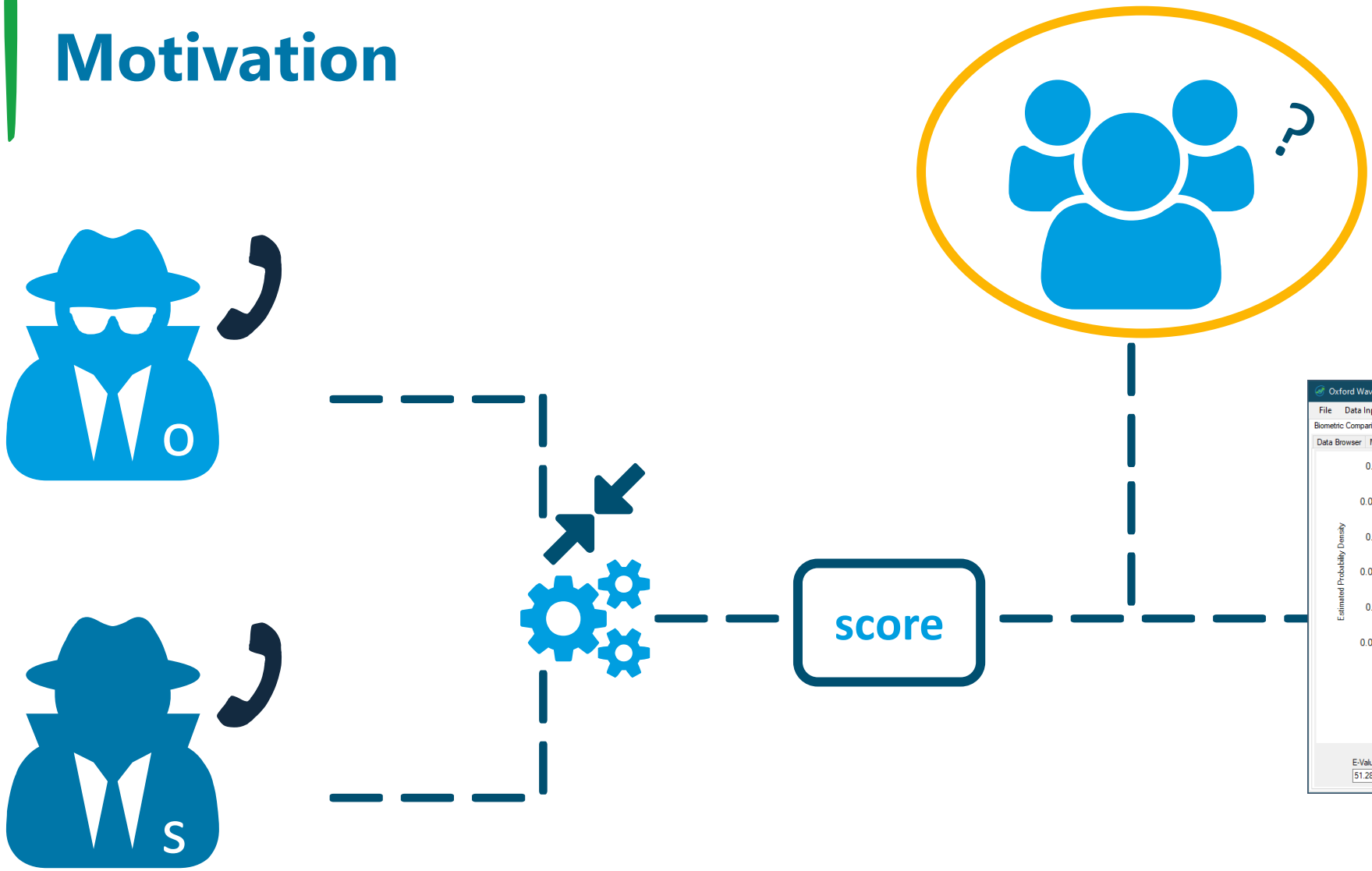
Linda Gerlach[1, 2], Tom Coy [2], Finnian Kelly[2], Kirsty McDougall[1], and Anil Alexander[2]

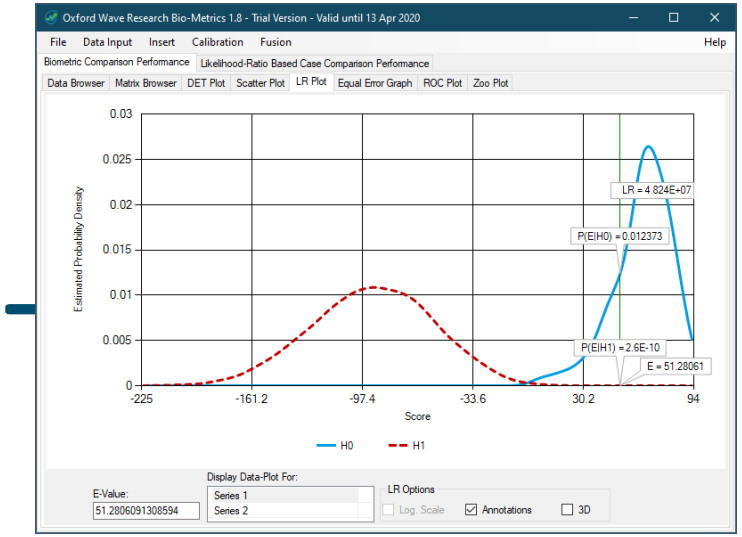
[1]University of Cambridge, [2]Oxford Wave Research

{lg589|kem37@cam.ac.uk, linda|tom.coy|finnian|anil@oxfordwaveresearch.com}

IAFPA conference Marburg (online), 22.-25.08.2021

Motivation





$$LR = \frac{p(E|H_s)}{p(E|H_d)}$$

UNIVERSITY OF CAMBRIDGE

Background – Selecting a relevant population

- To assess the degree of typicality, a **relevant population of a certain size** must be selected by the practitioner.
- **Logical relevance of the relevant population**

The relevant population should be representative of the offender sample in terms of:

- Speaker sex
- Language
- Accent
- Socio-demographic factors
- Channel conditions
- Recording device
- …

(see e.g. Hughes 2014, van der Vloed et al. 2020)

- **Perceived voice similarity of the relevant population**

It may improve system performance to use a relevant population that is perceptually similar to the offender sample.

(see Morrison et al. 2012)

Background – Using perceived voice similarity?

- Morrison et al. (2012):
  - Lay listener judgements of the overall similarity of the offender sample to candidates for a relevant population
  - Proof-of-concept: automatic speaker recognition (ASR) system as a substitute (GMM-UBM, MFCC)
  - Improved system performance compared to randomly selected relevant populations

- Criticism (Gold & Hughes 2014):
  - Influence of listeners' linguistic background on judgements of perceived voice similarity
  - Individual differences in the perception of voice similarity
  - Replicability of voice similarity ratings not certain
  - Assessment of perceived similarity is an expert matter

Background – Using perceived voice similarity?

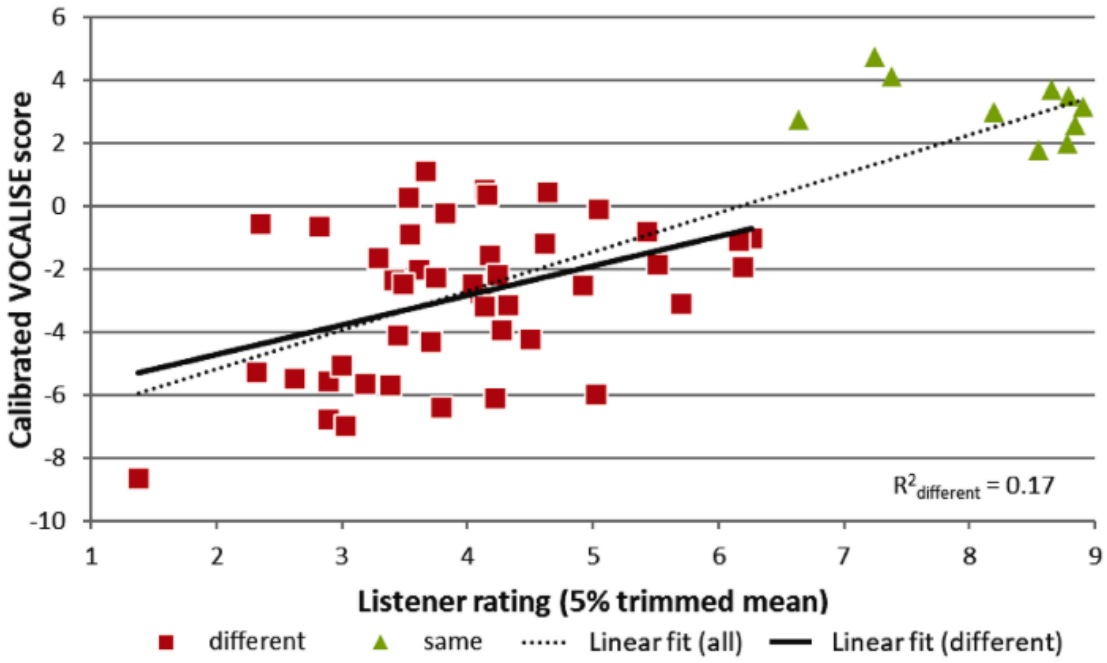
- Morrison et al. (2012):
  - …or a
  - …M,
- Cr
  - Assessment of perceived similarity is an expert matter

Problem:

- Varying views on how to select a relevant population with insufficient insight into the impact on the strength of evidence
- Collection of lay listener ratings for selecting a relevant population is time-consuming and costly
- Features other than MFCCs that have a closer relationship with perceived voice similarity should be investigated

ASR and perceived voice similarity

- Gerlach et al. (2020, 2021):
  - Exploration of the relationship between lay listener judgements of perceived voice similarity and automatically obtained comparison scores using an automatic speaker recognition system
  - Promising results using automatically extracted phonetic (auto-phonetic) features to approximate perceived voice similarity
  - Auto-phonetic features include features that have been found to be perceptually relevant (F1 to F4; F1 to F4 and F0)

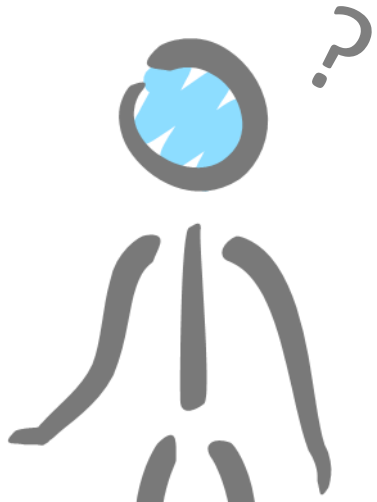


(Gerlach et al. 2020, p. 91)

UNIVERSITY OF CAMBRIDGE

Research questions

- What impact does the size of the relevant population have on the strength of evidence?
- What impact does the perceived voice similarity (assessed by an ASR system) of the relevant population have on the strength of evidence?
- How do the size of the relevant population and degree of perceived voice similarity of the relevant population to the questioned speaker interact?



GBR-ENG Database (2019)

- 6000 telephone recordings (landline or mobile)
- 600 male and female speakers
- Recordings of 3-6 minutes in duration
- Spontaneous speech in English
- Recordings made across three regions of England:
  - North, Midlands, South
  - Denoting location of the speaker at the time of the call (not their dialect)
- Metadata includes:
  - Speaker identity
  - Gender (biological sex)
  - Age
  - Region of upbringing, with a bias towards the South of England

Methodology - Dataset Preparation

- Male speakers
- Landline condition
- Requirements:
  - Minimum net speech post-VAD: 20s
  - 2 files per speaker
- Mock test set of 15 speakers:
  - 5 x North; 5 x Midlands; 5 x South
- Remaining 167 speakers formed relevant population superset

Methodology - Control Experiment

- Control experiment to assess the stability of log likelihood ratio cost (Cllr)
  - Cllr: metric of magnitude of system errors
  - Randomly-sampled relevant populations
  - Size increments of 5; from 15 speakers to 100 speakers
  - 10 random samples per increment; 200 total comparisons
- VOCALISE: x-vector, MFCC, VAD on
  - 15 mock offender vs 15 mock suspect files
  - Relevant population speakers used as calibration set
  - 15 same-speaker (SS) scores and 210 different-speaker (DS) scores per comparison
- Van Leeuwen & Brümmer (2013): calibrated scores from ASR system can be interpreted as LRs
  - Results loaded into Bio-Metrics performance metrics software to obtain Cllrs

Methodology - Similarity Experiment

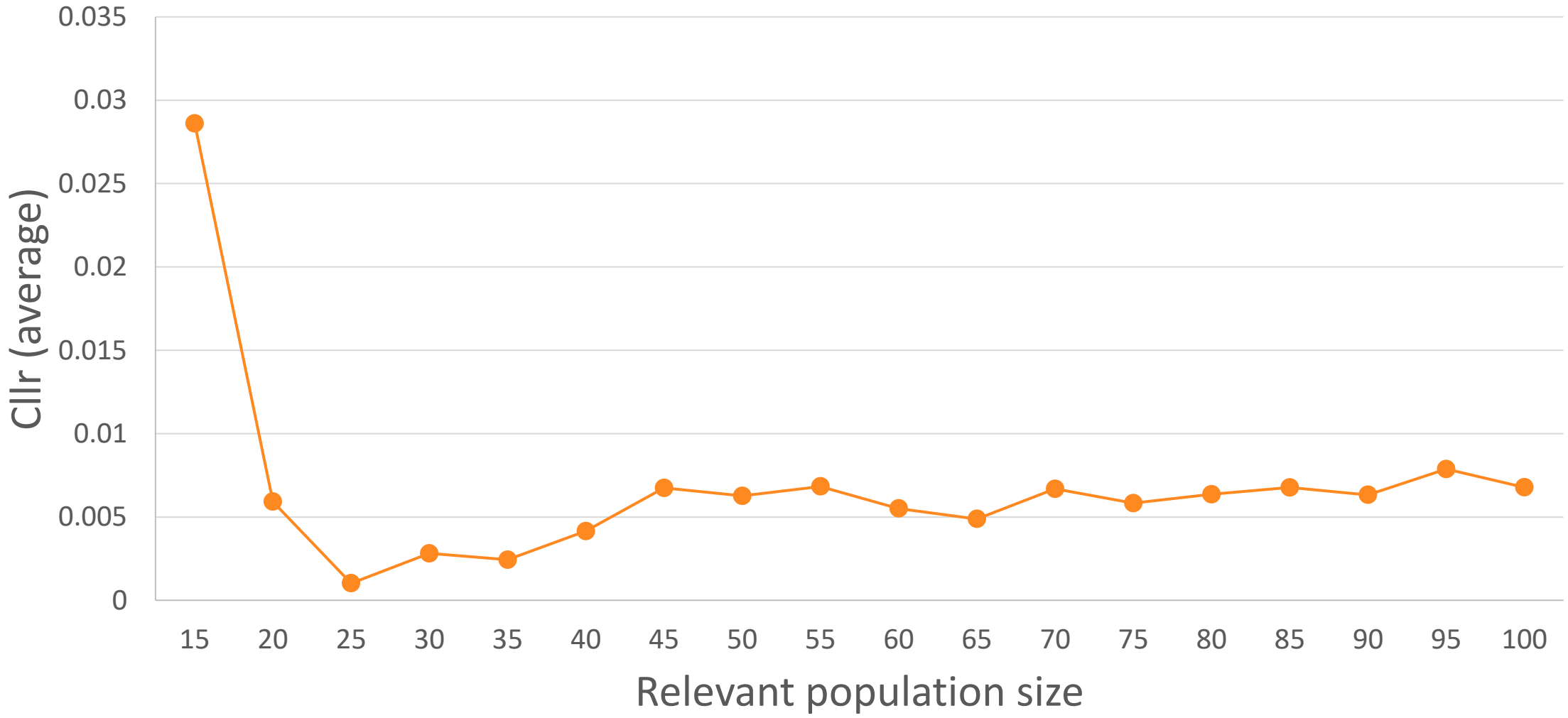
Speaker 852 

Example “top” 

Example “bottom” 

- VOCALISE auto-phonetic (F1 to F4), x-vector, VAD on
  - 15 mock offenders vs 167 population superset
  - Scores ranked for each mock offender
- Similar (“top”), dissimilar (“bottom”) and random (“random”) relevant populations
  - Population sizes: 30, 40, 50, 60, 70, 80, 90, 100
  - 21 comparisons per mock offender – (top/bottom/random x 7 population sizes)
  - 315 total comparisons
- VOCALISE MFCC, x-vector, VAD on
  - 15 mock offender vs 15 mock suspect files
  - 1 SS score (mock offender in questions), 14 DS scores per comparison
  - 15 SS scores, 210 DS scores per condition (e.g. 40 top or 60 random)

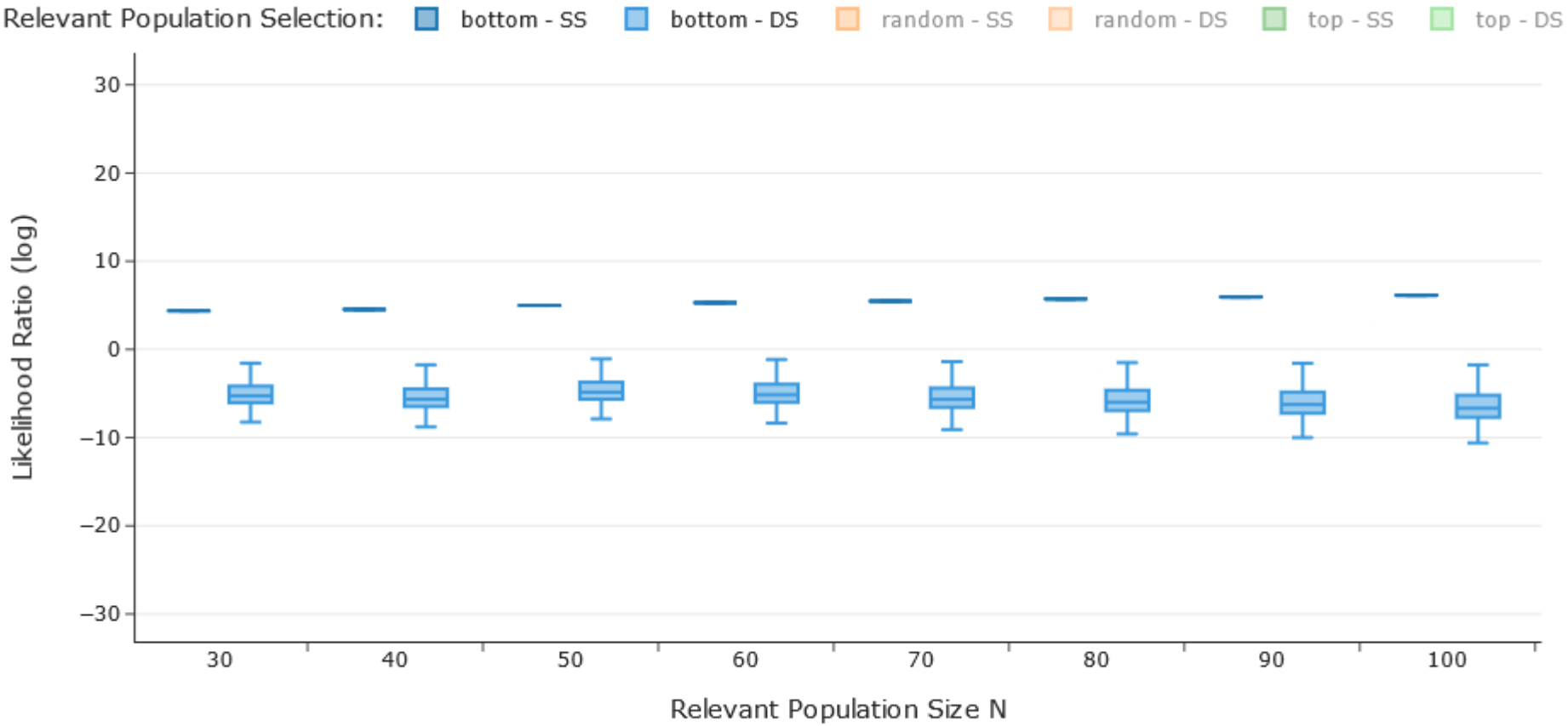
Results – Control Experiment



UNIVERSITY OF CAMBRIDGE

Results – Similarity Experiment: Individual LRs

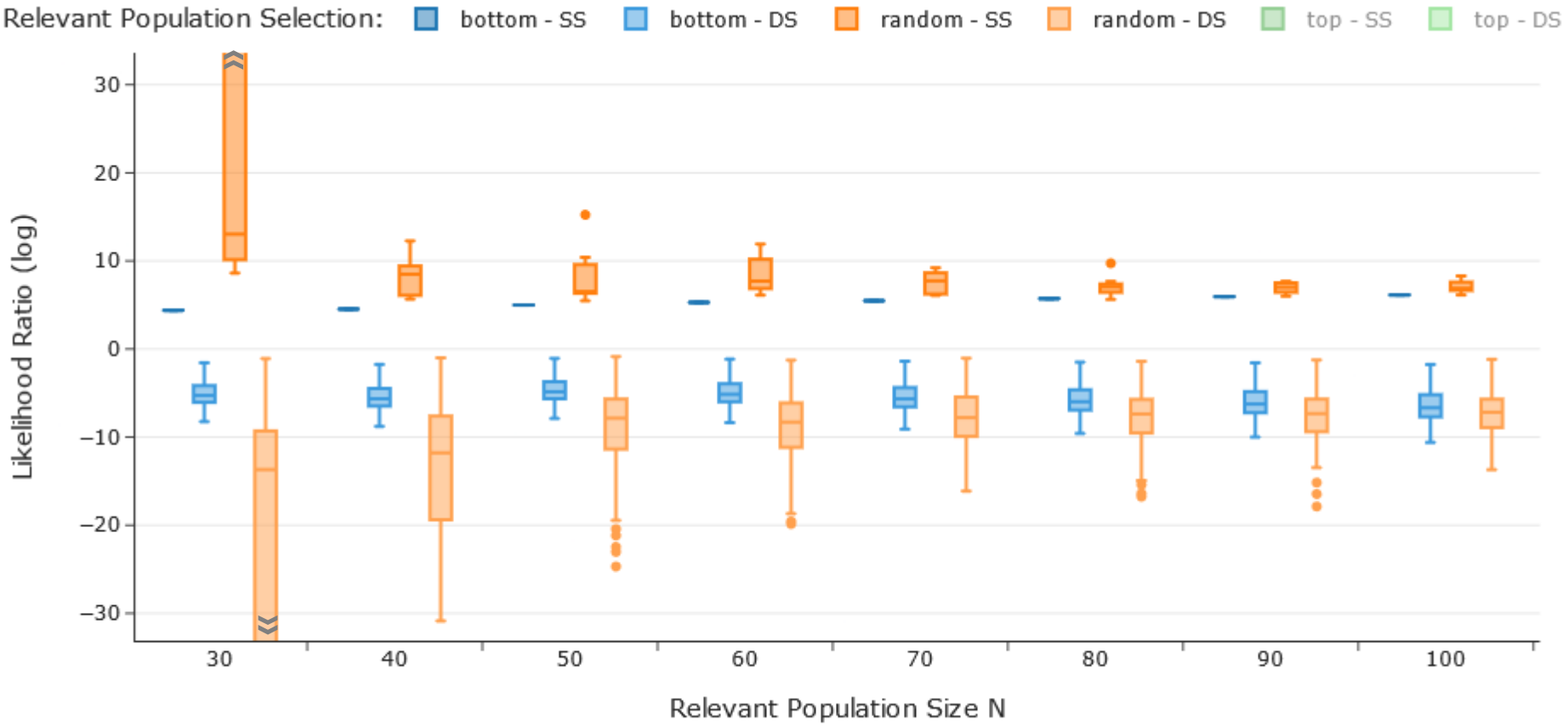
GB-SOU, speaker 852



UNIVERSITY OF CAMBRIDGE

Results – Similarity Experiment: Individual LRs

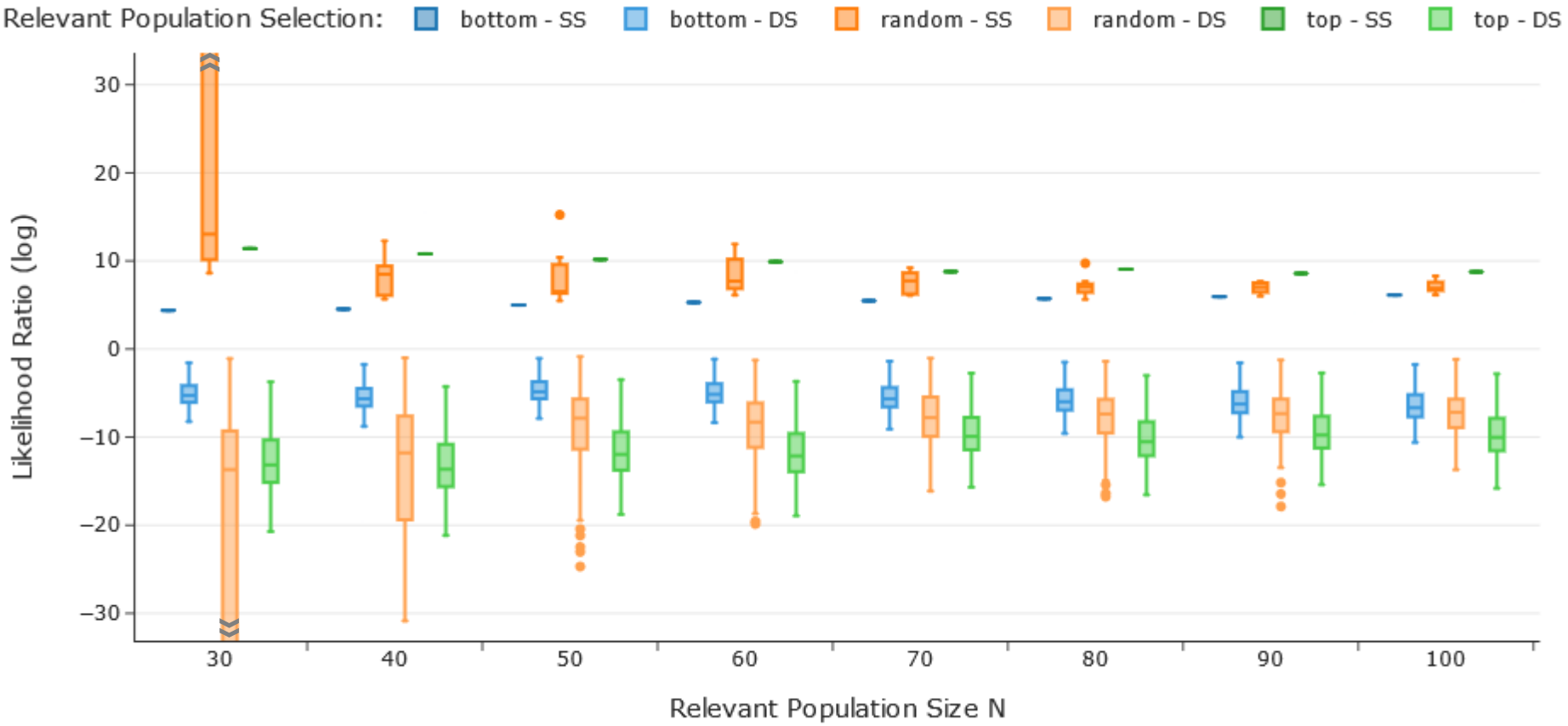
GB-SOU, speaker 852



UNIVERSITY OF CAMBRIDGE

Results – Similarity Experiment: Individual LRs

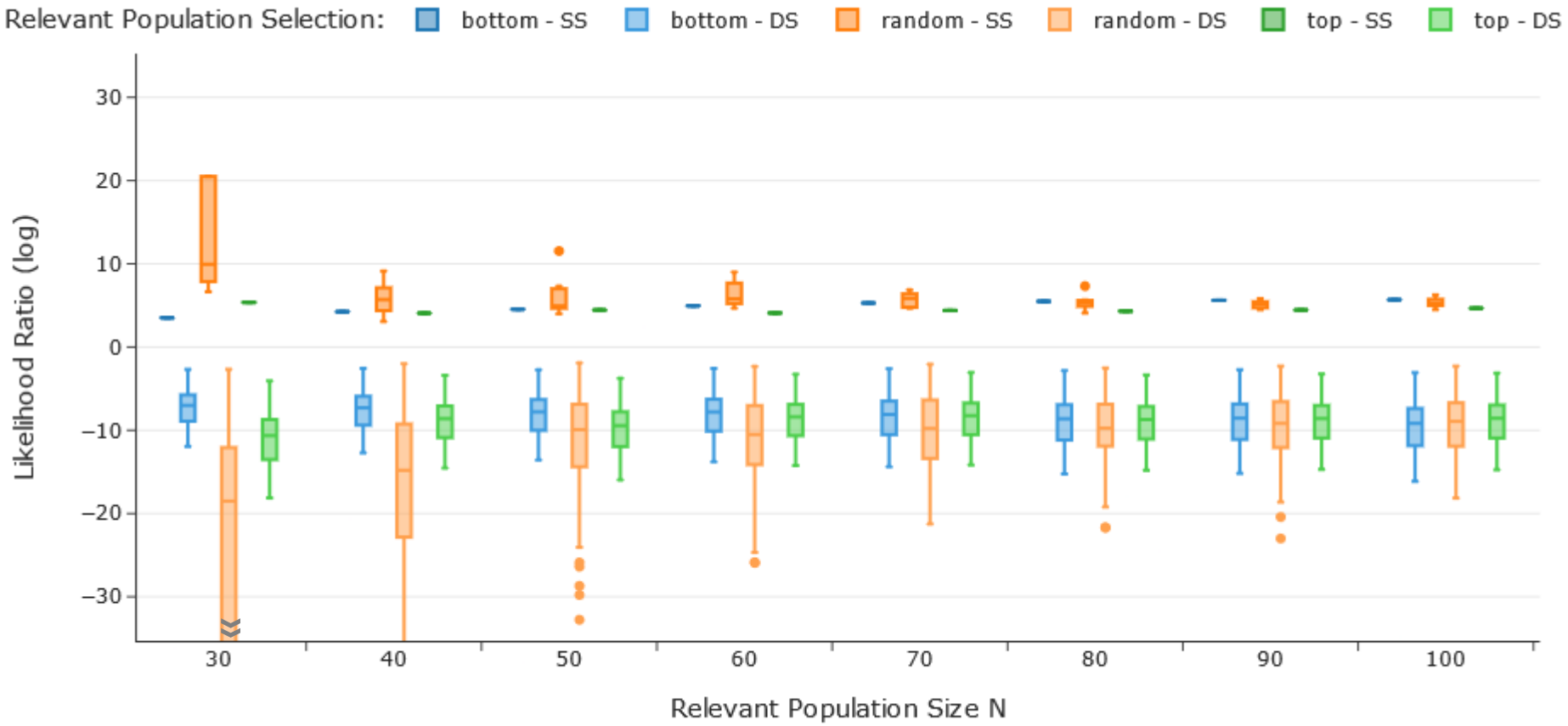
GB-SOU, speaker 852



UNIVERSITY OF CAMBRIDGE

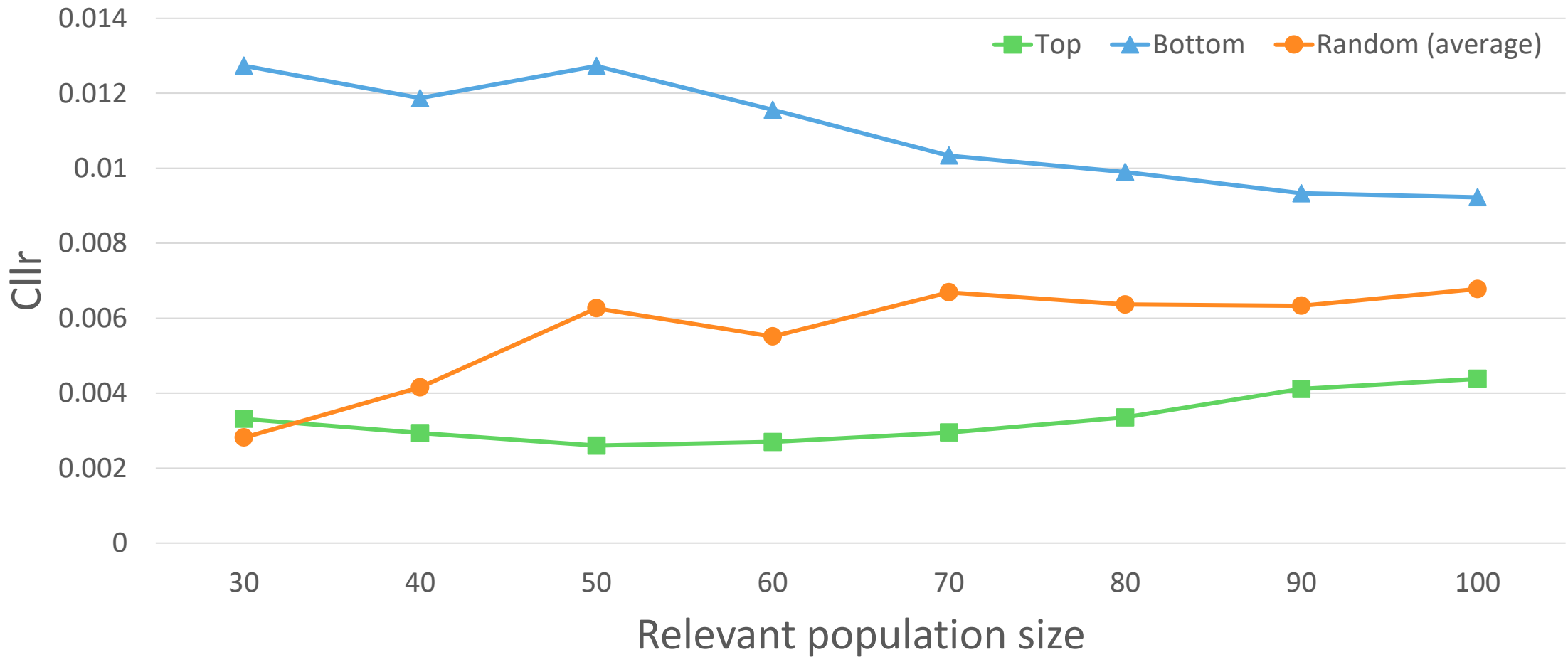
Results – Similarity Experiment: Individual LRs

GB-SOU, speaker 669



UNIVERSITY OF CAMBRIDGE

Results – Similarity Experiments: Broader Trends



UNIVERSITY OF CAMBRIDGE

Findings

- Impact of the size of the relevant population:
  - Overall very low Cllrs
  - Use of small relevant populations leads to unreliable Cllrs
  - Average Cllrs seem to level off from a relevant population size of 45

- Impact of perceived voice similarity on strength of evidence:
  - Better distinction between same- and different-speaker comparisons with “top” relevant population for majority of individual mock cases (9 out of 15)
  - Consideration of other factors playing a bigger role than perceived voice similarity necessary

- Interaction between perceived voice similarity and relevant population size:
  - Lowest Cllrs almost entirely during use of “top” relevant population
  - Decreasing impact of the perceived similarity on Cllrs with increasing relevant population size

Conclusion and future work

- Should the practitioner rely only on perceived voice similarity (as judged by lay listeners) for the selection of the relevant population?
  - No, for various reasons (see Gold & Hughes 2014).

  - **BUT:** Using the most similar speakers from a preselected set of representative speakers may increase the strength of evidence.
  - **AND:** An automatic speaker recognition approach relying on phonetic features may reduce the time needed for the similarity assessment, while possibly providing greater reliability.

- Future work:
  - Increase sample size
  - Use lower quality audio
  - Explore the impact of different ‘degrees’ of similarity
  - Explore results for female speakers

References

- Bio-Metrics 1.8 performance metrics software, Oxford Wave Research Ltd., https://www.oxfordwaveresearch.com/products/bio-metrics, last accessed 10.06.2021.
- GBR-ENG database (2019). A telephonic speech database collected for the UK Government for evaluating speech technologies. Further details on application.
- Gerlach, L., McDougall, K., Kelly, F., Alexander, A., & Nolan, F. (2020). Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features. *Speech Communication* 124: 85-95.
- Gerlach, L., McDougall, K., Kelly, F., & Alexander, A. (2021, this conference). How do automatic speaker recognition systems 'perceive' voice similarity?. In *Proc. 29th Annual Conference of the International Association for Forensic Phonetics and Acoustics*.
- 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.
- Hughes, V. (2014). The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison (Doctoral dissertation). University of York.
- Kelly, F., Forth, O., Kent, S., Gerlach, L., & Alexander, A. (2019). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. In *Audio Engineering Society (AES) Forensics Conference 2019, Porto, Portugal.*
- Van Leeuwen, D. A., & Brümmer, N. (2013). The distribution of calibrated likelihood-ratios in speaker recognition*. arXiv preprint arXiv: 1304.1199.*
- Morrison, G. S., Ochoa, F., & Thiruvaran, T. (2012). Database selection for forensic voice comparison. In *Proc. Odyssey Speaker and Language Recognition Workshop.* 62-77.
- van der Vloed, D., Kelly, F., & Alexander, A. (2020). Exploring the effects of device variability on forensic speaker comparison using VOCALISE and NFI-FRIDA: A forensically realistic database. In *Proc. Odyssey Speaker and Language Recognition Workshop*. 402-407.

Questions?