

The effect of speaker randomisation in Likelihood-ratio based forensic voice comparison

Bruce Xiao Wang*, Vincent Hughes & Paul Foulkes
{xw961|vincent.hughes|paul.foulkes}@york.ac.uk

1. Introduction

Likelihood-ratios have been employed in FVC to test:

- speaker-specificity of individual vowels and phonetic sequences ([1],[2]).
- precision and accuracy of FVC system ([3]).
- effects of sample size on FVC systems ([4],[5]).

However, few studies have considered the variability in system performance as a function of the make-up (rather than size) of the training, test, and reference data used, i.e:

- **Who do we select?**
- **Which data group (test/training/reference) do we assign them to?**
- **Does it matter?**

2. Research Questions

1. What happens if speakers are arranged differently in the test, training & reference data?
2. Can we use one LR-based FVC system for multiple cases?
3. What is the sensitivity to individual speakers' Log10 LR if different training and reference speakers are used?

3. Method

3.1 Corpus and tokens

- telephone recordings from central Guangdong Cantonese speakers
- sentence final particle /a/ 'ah' (155 speakers)
- disyllabic word /hea/ 'yes/yeah' (64 speakers)
- mean of 14 tokens per speaker per variable

3.2 Raw data elicitation

- target variables manually segmented in Praat ([6])
- 9 measurements from F1 and F2 via Praat script

3.3 Data processing

- quadratic & cubic polynomials fitted to /a/ and /hea/
- polynomial coefficients used as the input of MVKD for cross-validated LR computation (scores)
- score calibration: logistic regression

4. Experiments

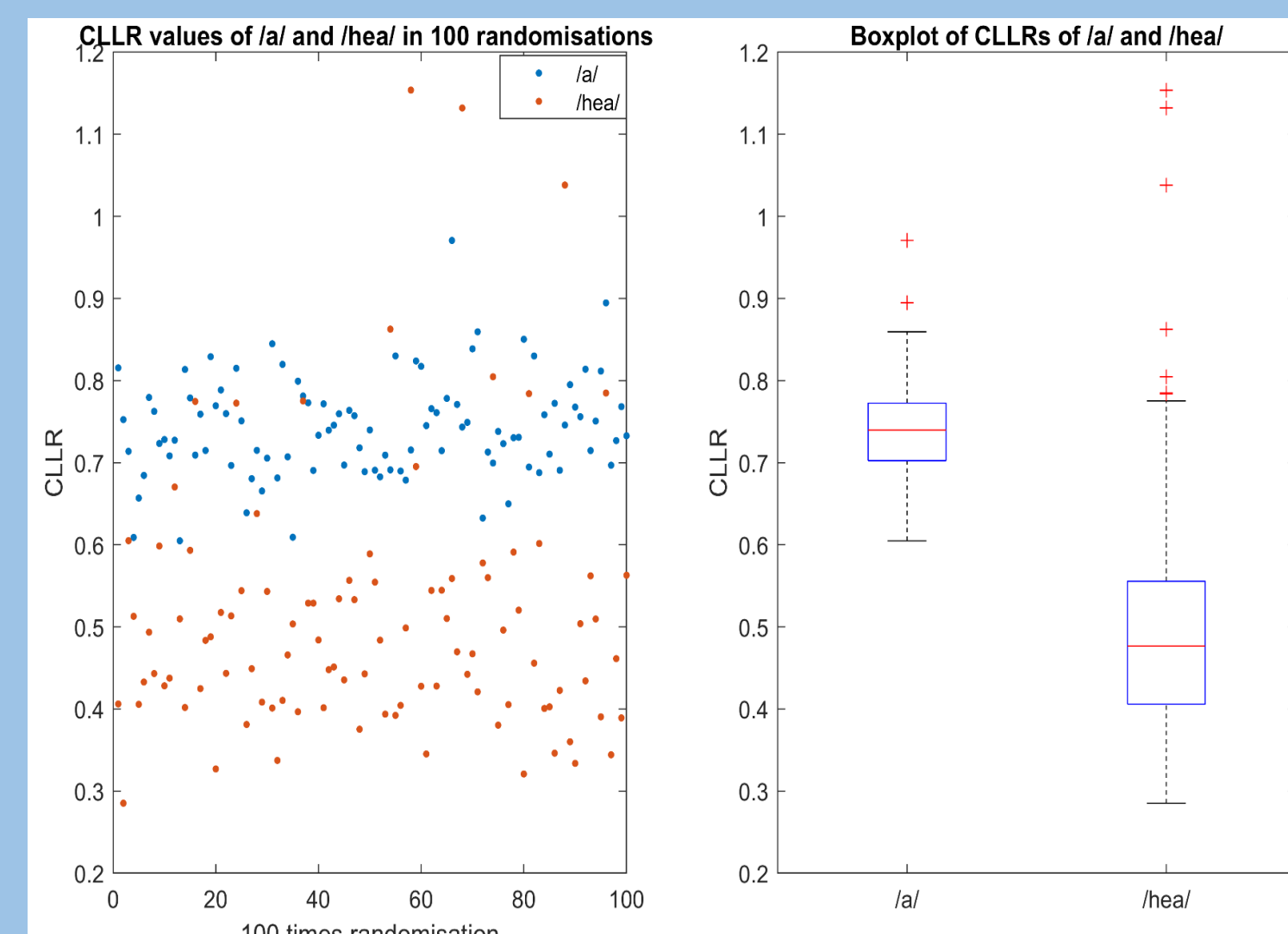
- 30 (/a/) and 15 (/hea/) speakers used in each data group (test, training and reference)
- 3 experiments were replicated 100 times with pre-defined speaker randomisation rules in R ([7]).

Speaker randomised	Expt. 1	Expt. 2	Expt.3
Test speakers	✓	✓	
Training speakers	✓		✓
Reference speakers	✓		✓

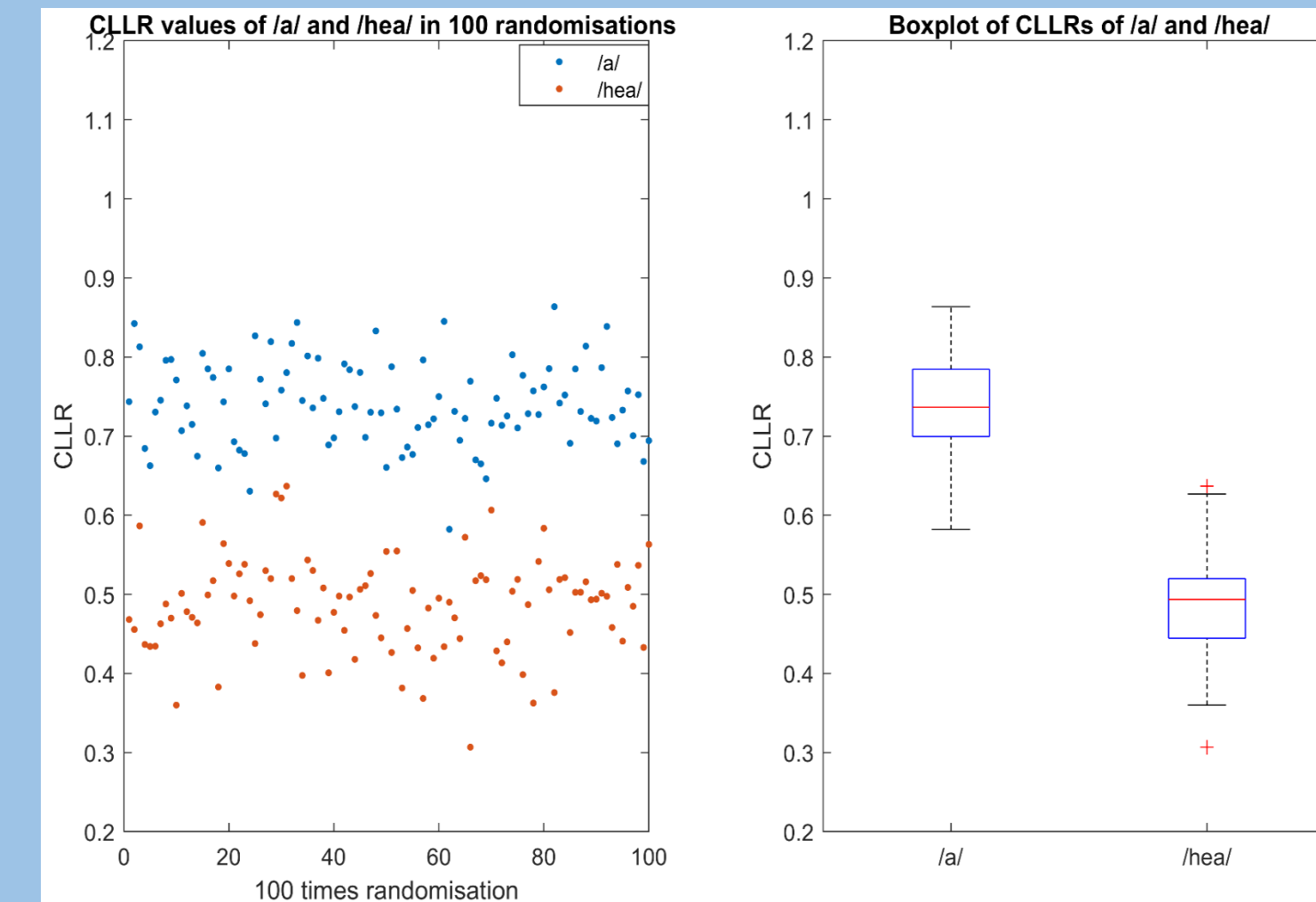
Table 1. speakers randomised in each experiment.

5. Results

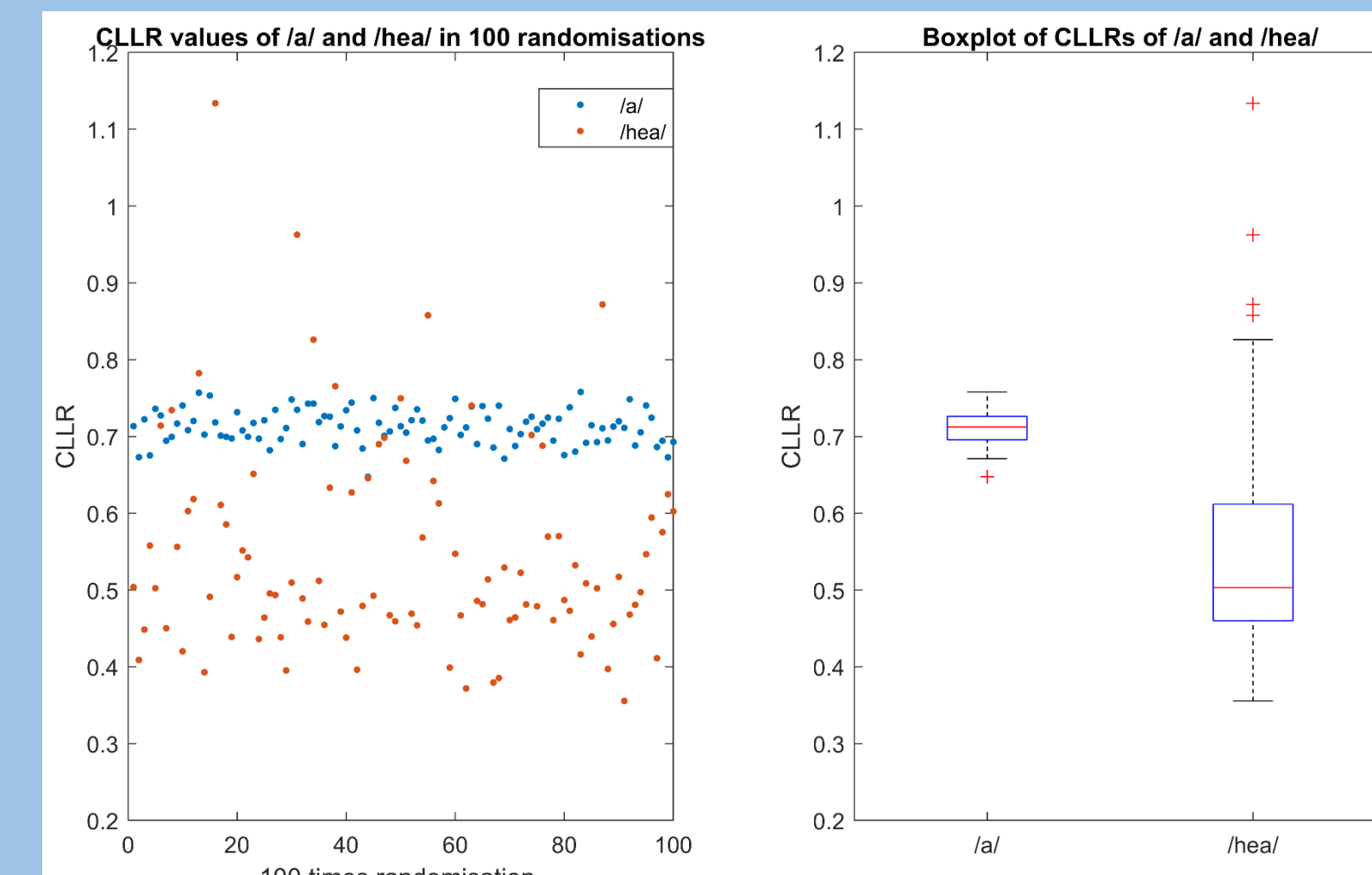
5.1 overall performance (Cllr)



Expt. 1. Varying test, training & reference speakers.



Expt. 2. Varying test speakers



Expt. 3. Varying training & reference speakers

5.2 Log10 LR of individual speakers in experiment 3.

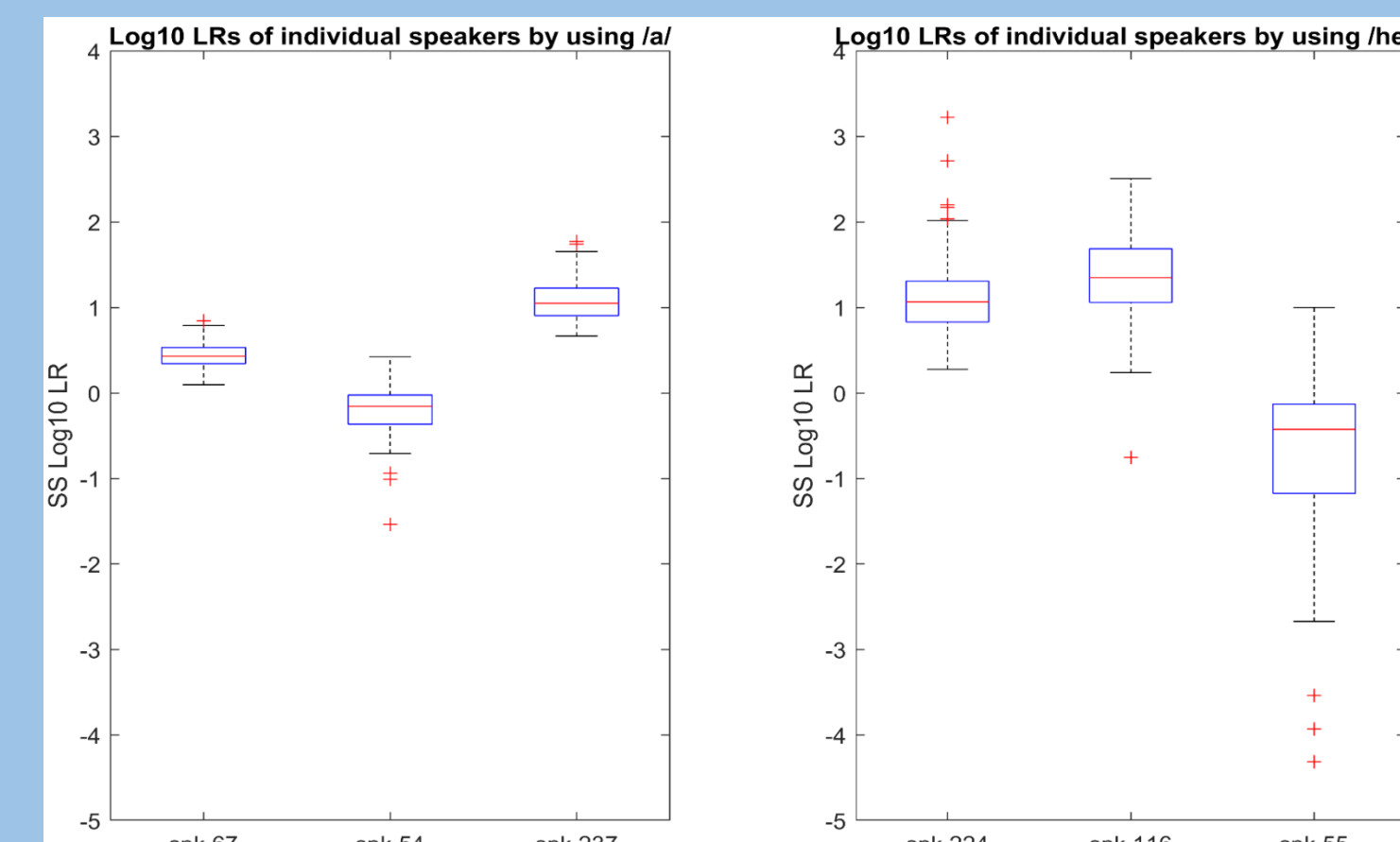


Figure 4. Log10 LR of test speakers with highest, mean, and lowest mean formant values in expt. 3.

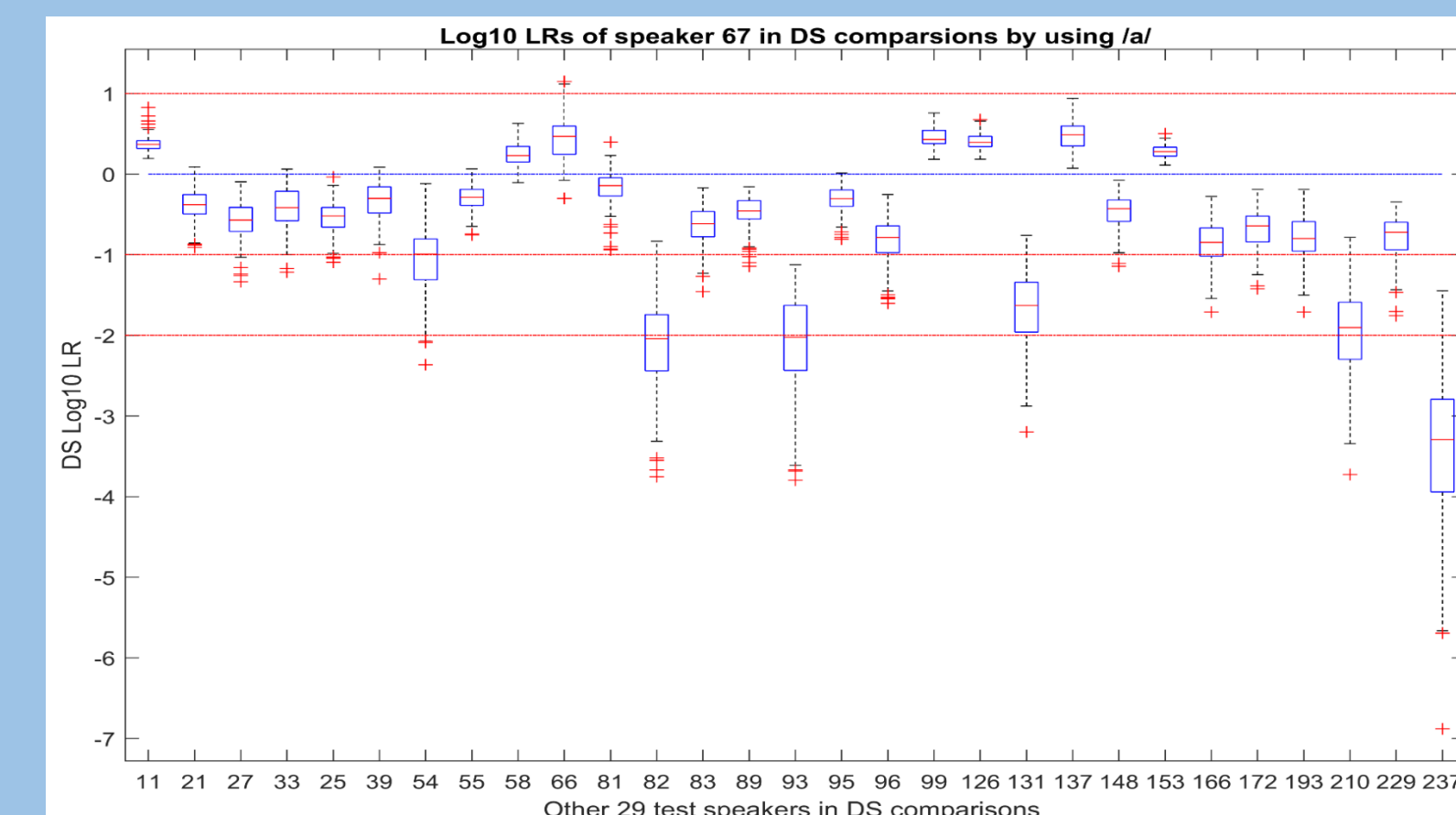


Figure 5. Log10 LR of speaker 67 in DS sessions in expt. 3.

6. Conclusion

- Cllr varies from 0.6 to 0.97 for /a/ and 0.29 to 1.15 for /hea/ when using different sets of speakers from a matched dialectic group (**expt.1**)
- One system could potentially be used for multiple FVC cases as long as the system is well-calibrated (**expt.2**). Cllr varies from 0.58 to 0.86 for /a/ and 0.31 to 0.64 for /hea/.
- Using different sets of training and reference speakers causes variation in individual speakers' Log10 LR which might give misleading information to the trier of fact (**expt.3**)
- Forensic practitioners need to run the system multiple times before implementing for real cases.

Reference

- [1] Zhang, C., Morrison, G. S., & Thiruvanan, T. (2011, August). Forensic voice comparison using Chinese/iau. In *Proceedings of the 17th International Congress of Phonetic Sciences* (Vol. 17, p. 21). City University of Hong Kong: Organizers of ICPhS XVII at the Department of Chinese, Translation and Linguistics.
- [2] 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. *Odyssey 2016*, 326-333.
- [3] Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, 51(3), 91-98.
- [4] Hughes, V. (2014). *The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison* (Doctoral dissertation).
- [5] Hughes, V., & Foulkes, P. (2015). The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age. *Speech Communication*, 66, 218-230.
- [6] Boersma, P. and Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.36.
- [7] Lo, J. (2018). fvlrr: Likelihood Ratio Calculation and Testing in Forensic Voice Comparison [unpublished R package] version 0.1.0