Effect of Score Sampling on System Stability in Likelihood Ratio based Forensic Voice Comparison

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

1. Introduction

Forensic Voice Comparison involves

- The analysis of questioned speech and known speech [6 7]

The LR framework concerns

- Similarity: how similar the questioned and know speech are [6 7]
- Typicality: how typical in a wider population [6 7]

LR framework involves two stages

- Feature to score (formants, F0, MFCC, long term F0, voice quality) [10]
- Score to LR (aka. Calibration [2] [10])

Previous studies show that system performance varies due to:

- Number of speakers used [5 8]
- Demographically matched or mismatched speakers [4]
- Speaker configurations in training, test and reference data [13]
- Different speech features used [12 14]

2. Aim

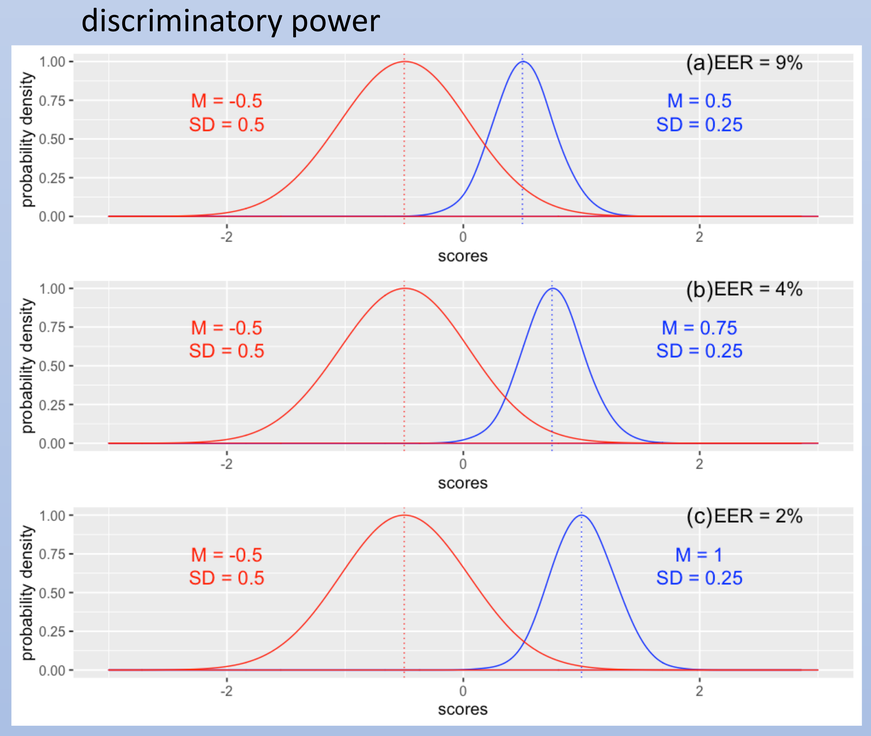
Use simulated data (highly controlled) to test the effect of speaker sampling on system performance (i.e. to avoid potential variations from sociolinguistic factors, channel mismatch, and recording devices etc.)

- How robust is a system's performance to changing which speakers are used in the training and test sets?
- Do some phonetic variables provide more or less stable LR output according to the specific sample of speakers used?

3. Method & Experiments

Data simulation

- Scores for 1000 speakers were computed from three sets of normal distributions in R [1 11]. These are referred to as set (a), set (b) and set (c)
- Different means and standard deviations used for same-/different-speaker comparisons in each set
- Different equal error rates (EER) to mimic variables with different speaker-discriminatory power



In each 100 times sampling

Sampling	Training speakers	Test speakers
Expt. 1	✓	✓
Expt. 2	✓	
Expt. 3		✓

20 test speakers

20 training speakers

Calibration[2 10]

Test scores

Training scores

LRs

C_{llr} [3 9]

4. Results

Experiment 1. Sampling training and test scores

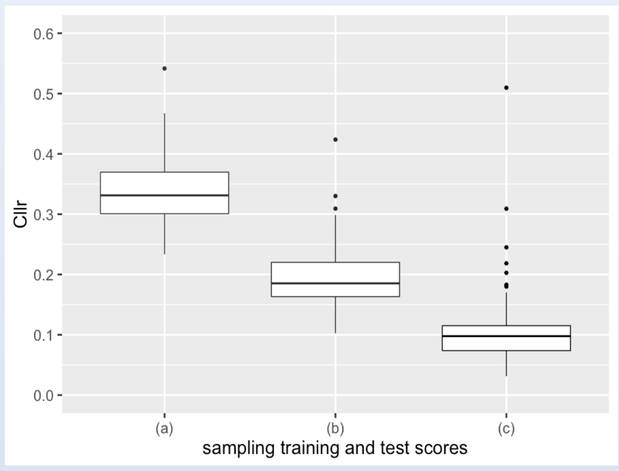


Figure 1. Variation of C_{llr}s by sampling training and test scores from pseudo-datasets (a), (b) and (c).

C_{llr}	(a)	(b)	(c)
Min.	0.23	0.10	0.03
1st Qu.	0.30	0.16	0.07
Median	0.33	0.19	0.10
3rd Qu.	0.37	0.22	0.12
Max.	0.54	0.42	0.51
IQR	0.07	0.06	0.05
OR	0.31	0.32	0.48

Table 1. C_{llr} minimum, 1st quartile, median, 3rd quartile, maximum, IQR and OR of sets (a), (b) and (c) in experiment 1.

Experiment 2. Only sampling training scores

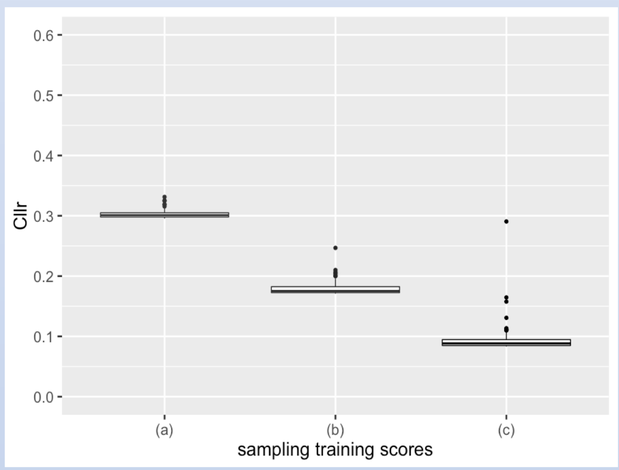


Figure 2. Variation of C_{llr}s by sampling training scores from pseudo-datasets (a), (b) and (c).

C_{llr}	(a)	(b)	(c)
Min.	0.30	0.17	0.08
1st Qu.	0.30	0.17	0.08
Median	0.3	0.17	0.09
3rd Qu.	0.31	0.18	0.09
Max.	0.33	0.25	0.29
IQR	0.01	0.01	0.01
OR	0.03	0.08	0.21

Table 2. C_{llr} minimum, 1st quartile, median, 3rd quartile, maximum, IQR and OR of sets (a), (b) and (c) in experiment 2.

Experiment 3. Only sampling test scores

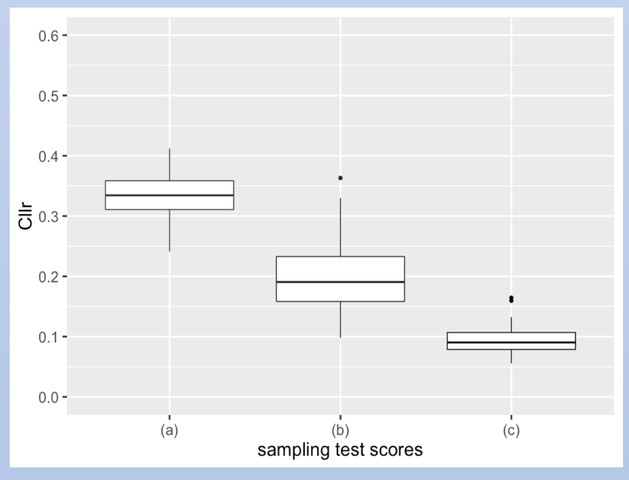


Figure 3. Variation of C_{llr}s by sampling test scores from pseudo-datasets (a), (b) and (c).

C_{llr}	(a)	(b)	(c)
Min.	0.24	0.10	0.06
1st Qu.	0.31	0.15	0.08
Median	0.33	0.19	0.09
3rd Qu.	0.36	0.23	0.11
Max.	0.41	0.36	0.16
IQR	0.05	0.08	0.03
OR	0.17	0.26	0.10

Table 3. C_{llr} minimum, 1st quartile, median, 3rd quartile, maximum, IQR and OR of sets (a), (b) and (c) in experiment 3.

5. Discussion

Experiment 1 vs Experiment 2 vs Experiment 3

- System stability varies considerably if different sets of SS and DS scores are used in each replication (Figure 1)
- Varying training scores has a limited effect on system stability (Figure 2)
- Sampling test scores has more effect on the system stability than sampling training scores (Figure 2 and Figure 3)

Set (a) vs Set (b) vs Set (c)

- Mean C_{llr} decreases as EER goes down in experiments 1, 2, and 3.
- Set (c) yielded more outliers than sets (a) and (b) in experiments 1, 2 and 3 regardless of speaker-discriminatory power (lower EER)
- The lower the EER, the lower the Cllr IQR, but not OR [Set (c)]

6. Conclusion

- Score-sampling has a marked effect on system stability regardless of speaker-discriminatory power of the feature being used
- Variables with higher speaker-discriminatory power do not necessarily yield higher system stability

8. Reference

[1] Ahrens, J. H., Dieter, U. (1973). Extensions of Forsythe's method for random sampling from the normal distribution. *Mathematics of Computation,* 27, 927-937. [2] Brümmer, N. et al. (2007) Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST SRE 2006. *IEEE Transactions on Audio Speech and Language Processing, 15*, pp. 2072-2084. [3] Brümmer, N., Du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2-3), 230-275. [4] 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. [5] Hughes, V. (2017). Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough?. *Speech Communication*, *94*, 15-29. [6] Hughes, V., & Wormald, J. H. (2019). Sharing innovative methods, data and knowledge across sociophonetics and forensic speech science. *Linguistics Vanguard*. [7] Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass*, 2(4), 671-711. [8] Kinoshita, Y., & Ishihara, S. (2014). Background population: how does it affect LR-based forensic voice comparison?. *International Journal of Speech, Language & the Law*, *21*(2). [9] Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, *51*(3), 91-98. [10] 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. [11] R Core Team (2018). R: A language and environment for statistical computing. R Foundatin for statistical Computing, Vienna, Austria. http://www.R-project.org/ [12] 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. [13] Wang, B. X., Hughes, V., Foulkes, P. (2018) A preliminary investigation of the effect of speaker randomisation in likelihood-ratio based forensic voice comparison. *IAFPA 2018*. University of Huddersfield, 125-125. [14] Zhang, C., Morrison, G. S., & Thiruvaran, 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.