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

Bruce Xiao Wang, Vincent Hughes

Department of Language and Linguistic Science, University of York, UK

{xw961|vincent.hughes}@york.ac.uk

Eligible for the 'Best Student Paper Award': Yes

In likelihood ratio-based (LR) forensic voice comparison (FVC), experts rely on databases to empirically estimate the strength of voice evidence. It is then important to test and empirically validate system performance, because one does not want to obtain extreme LRs that over- or underestimate the strength of evidence, especially when the sample size is small and the density estimation is not well-supported by the data that might lead to extrapolation. Many calibration methods (Brümmer et al., 2007; Brümmer & Swart, 2014; Vergeer et al., 2016; Zadrozny & Elkan, 2002) have been developed to deal with sample size and variability issues. Previous studies (Ali et al., 2015; Morrison & Poh, 2018) have investigated the effectiveness of different calibration methods in dealing with sampling variability and sample size. However, these studies either assumed scores that followed Gaussian distributions with equal variance or used small samples of training data.

The current study simulated skewed scores from an empirical study (Wang et al., 2019) to test the susceptibility of four calibration methods (i.e. logistic regression (Brümmer et al., 2007), regularised logistic (rlogistic) regression (Morrison & Poh, 2018), empirical lower and upper bound (ELUB, Vergeer et al., 2016) and Bayesian model (Brümmer & Swart, 2014)) to score skewness and sample size. Table 1 shows the skewness, kurtosis, mean and standard deviation values derived from scores generated from real data in a FVC study and then used as the basis for simulation. The scores were simulated using skew-t (ST) random sampling function `rst()` (Azzalini, 2020) in R (R core team, 2020). For each set of distribution parameters, the training and test same- (SS) and different-speaker (DS) scores were sampled with increasing sample sizes from 20 to 100 speakers per set, in increments of 10 speakers. The experiments were replicated 100 times for each sample size and calibration method.

Table 1. 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						

In Figure 1, the x-axis indicates the number of training and test speakers used and the y-axis represents the C_{llr}. The different colours represent different levels of skewness. The dashed lines indicate the C_{llr} range and the symbols are the mean C_{llr} across the 100 replications for that sample size. Generally, the C_{llr} range reduces for all four calibration methods when more speakers are used. This means that, predictably, uncertainty reduces as sample size increases. The rlogistic regression and Bayesian model are the least sensitive to skewness and sample size, followed by the logistic regression and ELUB. The C_{llr} range of rlogistic regression and

Bayesian model remain ca. 0.2 or lower across different sample sizes. For logistic regression, the C_{llr} range is more sensitive to score skewness varying between ca. 0.5 and 0.15 across different sample sizes and especially when skewness is higher. The ELUB is the most sensitive to sample size and skewness, i.e. C_{llr} range varies between ca. 0.6 and 0.2 with different sample size for all skewness. For mean C_{llr}, all calibration methods have lower mean C_{llr} when the scores are more skewed. The logistic regression method yields the lowest mean C_{llr} overall, followed by rlogistic regression, Bayesian model and ELUB. Apart from the rlogstic regression, all other three calibration methods have consistent mean C_{llr} given score skewness; however, the mean C_{llr} reduces with larger sample sizes for rlogistic regression, especially when the score skewness is high (i.e. score skewness = -1.4).

The results raise important issues about the trade-off between overall performance and variability, i.e. how much variability is allowed given accuracy (C_{llr} mean) and should we aim for lower mean (higher accuracy) given the system stability (C_{llr} range) varies within certain range? Ultimately, it is our opinion that experts' decisions should be driven by reducing uncertainty, rather than the potential of a very low C_{llr}/absolute system validity.

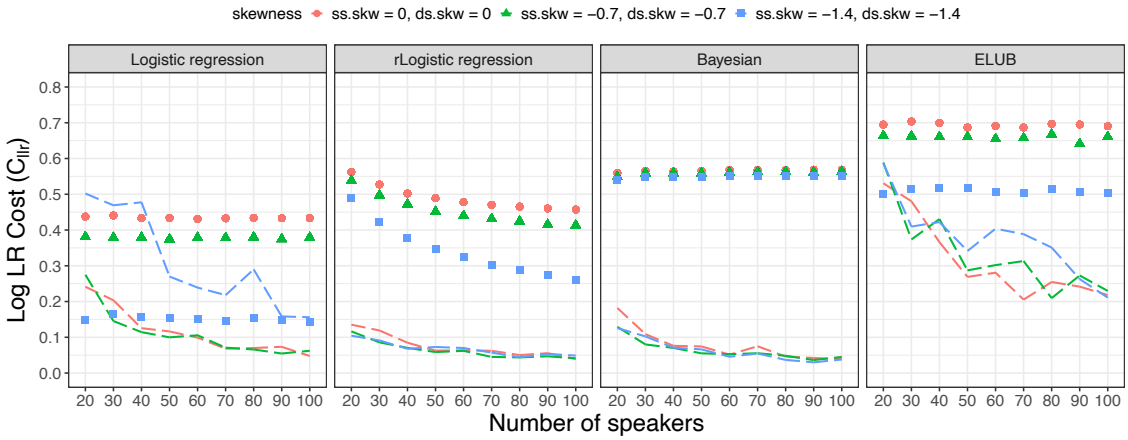

Figure 1 C_{llr} mean and range using different sample sizes and calibration methods.

References

Ali, T., Spreeuwers, L., Veldhuis, R., & Meuwly, D. (2015). Sampling variability in forensic likelihood-ratio computation: A simulation study. *Science & Justice*, *55*(6), 499–508. https://doi.org/10.1016/j.scijus.2015.05.003

Azzalini, A. (2020). *The R package 'sn': The Skew-Normal and Related Distributions such as the Skew-t* (version 1.6-2) [Computer software].

Brümmer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, F., Karafiat, M., van Leeuwen, D. A., Matejka, P., Schwarz, P., & Strasheim, A. (2007). Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 2072–2084. https://doi.org/10.1109/TASL.2007.902870

Brümmer, N., & Swart, A. (2014). Bayesian Calibration for Forensic Evidence Reporting. *Interspeech*, Singapore 388–392.

Morrison, G. S., & Poh, N. (2018). Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors. *Science & Justice*, *58*(3), 200–218. https://doi.org/10.1016/j.scijus.2017.12.005

R, core team. (2020). *RStudio: Integrated Development for R*. RStudio, Inc. http://www.rstudio.com/

Vergeer, P., van Es, A., de Jongh, A., Alberink, I., & Stoel, R. (2016). Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating? *Science & Justice*, *56*(6), 482–491. https://doi.org/10.1016/j.scijus.2016.06.003

Wang, B. X., Hughes, V., & Foulkes, P. (2019). Effect of score sampling on system stability in Likelihood Ratio based forensic voice comparison. *International Congress of Phonetic Sciences*. Melbourne Australia, pp. 3065 – 3069.

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '02*, 694. https://doi.org/10.1145/775047.775151