Reducing the degree of uncertainty within automatic speaker recognition systems using a Bayesian calibration model

Bruce Xiao Wang, Vincent Hughes
Department of Language and Linguistic Science, University of York
hollamigo@gmail.com/vincent.hughes@york.ac.uk

In data-driven forensic voice comparison (FVC), empirical testing of a system is an essential step to demonstrate validity (i.e. how well the system performs the task) and reliability (i.e. whether the system would yield the same result if the analysis were repeated). The present study focuses on system reliability, aiming to reduce the degree of uncertainty at the *score* space with small sample size and skewed scores (conditions which are typical of the real world). Wang & Hughes (2021) simulated scores to test different calibration methods showing that the Bayesian model (Brümmer & Swart, 2014) outperformed logistic regression in terms of variability in system validity values (i.e. produced less variable results). However, they simulated scores based on a linguistic system using multivariate kernel density (MVKD), which is likely to have worse overall performance than automatic speaker recognition (ASR) systems utilising Mel-frequency cepstral coefficients (MFCCs) or cepstral measures. We simulated scores generated from i-vector and Gaussian Mixture Model – Universal Background Model (GMM-UBM) ASR systems using real speech data to demonstrate the variability in system reliability as a function of score skewness and sample size. Scores were simulated based on parameters of score distribution from Enzinger et al. (2016) and Morrison & Poh (2018).

Scores were simulated using both skewed and non-skewed parameters (i.e., skewness changed to 0) to investigate the effect of score skewness. To account for sample size, training and test speakers were increased from 20 to 100, with 10-speaker increasements. Logistic regression and a Bayesian model were used for calibration and replicated 100 times per sample size. Performance was evaluated using the mean (overall discrimination) and range (overall variability) of the C_{llr}s across the 100 replications.

Figure 1 shows the C_{llr} mean (dots) and range (lines). Using logistic regression, C_{llr} ranges are 1.3 (i-vector) and 0.69 (GMM-UBM) when scores are skewed (panel (a)) and sample size is small (20 speakers), while the C_{llr} ranges are 0.49 (i-vector) and 0.69 (GMM-UBM) when scores follow normal distributions (panel (b)). Score skewness seems to have a less marked effect on system reliability for the GMM-UBM system when sample size is small, principally because GMM-UBM produced less skewed scores. Panels (c) and (d) show that Bayesian calibration improves system reliability considerably when scores are skewed, e.g., the C_{llr} range is ca. 0.3 (Figure 1 (c)) compared with 1.3 (Figure 1 (a)) when 20 speakers are used. For the GMM-UBM system, Bayesian calibration does not seem to improve system reliability as much it does for the i-vector system, and score skewness seems to have less effect on system reliability when 40 or more speakers are used.

The mean C_{llr} stays stable across score skewness and sample size within systems. However, there appears to be a trade-off, such that overall discrimination (i.e. mean C_{llr}) may be slightly poorer where reliability is slightly better. Thus, it is important for experts to consider what the most important metric of system performance to be and what constitutes 'low enough' mean C_{llr} in making decision about which system to use in a forensic case (see Morrison et al., 2021).

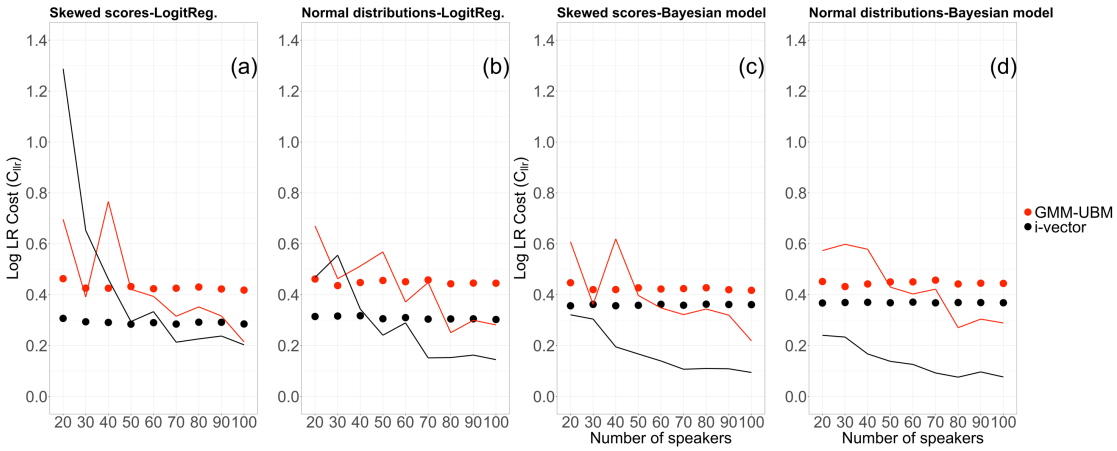

Figure 1. C_{llr} mean and range as a function of score skewness, sample size and calibration methods.

References

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

Enzinger, E., Morrison, G. S., & Ochoa, F. (2016). A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case. *Science & Justice*, *56*(1), 42–57. https://doi.org/10.1016/j.scijus.2015.06.005

Morrison, G., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., Planting, S., Thompson, W. C., van der Vloed, D., J F Ypma, R., & Zhang, C. (2021). Consensus on validation of forensic voice comparison. *Science & Justice*, *61*(3), 229–309. https://doi.org/10.1016/j.scijus.2021.02.002

Morrison, G., & 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

Wang, B. X., & Hughes, V. (2021). System Performance as a Function of Calibration Methods, Sample Size and Sampling Variability in Likelihood Ratio-Based Forensic Voice Comparison. *Interspeech 2021*, 381–385. https://doi.org/10.21437/Interspeech.2021-267