System performance as a function of calibration methods, sample size and sampling variability 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

Abstract

In data-driven forensic voice comparison, sample size is an issue which can have substantial effects on system output. Numerous calibration methods have been developed and some have been proposed as solutions to sample size issues. In this paper, we test four calibration methods (i.e. logistic regression, regularised logistic regression, Bayesian model, ELUB) under different conditions of sampling variability and sample size. Training and test scores were simulated from skewed distributions derived from real experiments, increasing sample sizes from 20 to 100 speakers for both the training and test sets. For each sample size, the experiments were replicated 100 times to test the susceptibility of different calibration methods to sampling variability. The C_{llr} mean and range across replications were used for evaluation. The Bayesian model and regularized logistic regression produced the most stable C_{llr} values when the sample size is small (i.e. 20 speakers), although mean C_{llr} is consistently lowest using logistic regression. The ELUB calibration method generally is the least preferred as it is the most sensitive to sample size and sampling variability (mean = 0.66, range = 0.21-0.59).

Index Terms: likelihood ratio, forensic voice comparison, calibration, sample size, sampling variability

1. Introduction

1.1. Developing and testing LR-based systems

In likelihood ratio-based (LR) forensic voice comparison (FVC), as well as automatic and semi-automatic speaker recognition more generally, analysts rely on databases of speakers to estimate empirically the strength of the voice evidence. It is then essential to test and empirically validate system performance, i.e. how good or bad the system is at separating same- (SS) and different-speaker (DS) pairs. Normally, this involves two stages (i.e. *feature-to-score* and *score-to-LR*) [1] requiring three datasets (i.e. training, test and reference). At the *feature-to-score* stage, acoustic data are extracted from recordings and used to generate speaker models. Pairs of speakers models (both SS and DS) are compared to generate LR-like scores which capture the similarity and typicality [2] between samples. The scores from the training data are then used at the *score-to-LR* stage (i.e. calibration) to convert the test scores to interpretable, calibrated LRs. This involves generating a calibration model from the training scores and then applying that model to the scores from the test data. System validity is evaluated on the basis of the calibrated test scores. Many methods for assessing validity are available, with the log LR cost function (C_{llr}) [3] generally considered the most appropriate, especially in the forensic context. Forensic experts and laboratories are now under increasing national and international regulatory pressure to demonstrate system validity to the trier of fact.

1.2. Sample size and sample variability

Variability in the LRs computed by a system can be introduced by different sources (e.g. variability in the relevant population, choice of features, statistical modelling) [4] at both stages. This can, of course, be problematic in terms of the precision of the specific LR in a given FVC case [4] but also in terms of misrepresenting the overall performance of the system (both in a validation exercise and in research). Sample size can be a substantial source of uncertainty in LR computation. This is especially true of FVC based on linguistic input features, where samples are generally small; many FVC studies use just 20 speakers in each data set. Numerous studies have investigated system validity as a function of sample size and sampling variability. For example [4,5] explored the effect of the number of reference speakers on system performance, showing that the C_{llr} varies between ca. 0.4 and 0.7 when the number of reference speakers varies between 10 and 120. Similarly, [7] explored system performance as a function of the number of training, test and reference speakers, suggesting that system stability can be achieved using minimally 30 training and reference speakers and minimally 15 reference speakers. Moreover, [8] showed that different configurations of training, test and reference speakers also have an effect on overall system performance. They replicated the experiments 100 times by randomly sampling 25 speakers (from a relevant population) into the training, test and reference sets respectively. Results show that the C_{llr} varies between 0.32 and 1.33 across 100 replications. Taken together, previous studies show that no matter how many or which speakers are used, variability in system performance is still inevitable. However, this is especially true when the sample size is small, and the density estimation is not well-supported by the data leading to extrapolation at the tails of the distributions.

1.3. Calibration methods

Many calibration methods are available and some may provide a potential solution to issues of sample size. The choice of calibration method is extremely important for system evaluation and optimisation because one does not want to obtain extreme LRs that over- or underestimate the strength of evidence [9]. Different calibration methods (e.g. logistic regression [10], pool adjacent violators [11], Bayesian model [12], scoring method [13]–[15]) have been developed and the performance has been compared. For example, [16] explored the effectiveness of three calibration methods (i.e. kernel density estimation, logistic regression, pool adjacent violators) in dealing with sampling variability with three sizes of the training scores. Results show that logistic regression is the least sensitive to sampling variability when the sample size of the training data is large. However, the size of test data was not taken into consideration and only three sets of sample size of the training data were considered. Similarly, [17] used simulated scores to explore the effectiveness of different calibration methods in shrinking LR output and tested the generalizability using data from real cases. However, this work only compared the effectiveness of different calibration methods using scores that follow Gaussian distributions with equal variance and did not take skewness into consideration. This is, however, important as the score distributions in real cases are less likely to follow Gaussian distributions with equal variance due to the reasonable limits of sample size in the real world.

1.4. The current study

The current study uses simulated scores from skewed distributions, derived from real data, to investigate the effectiveness of four calibration methods (i.e. logistic regression [10], empirical lower and upper bound (ELUB) [18], Bayesian model [12] and regularised logistic regression [17]) at dealing with issues relating to sample size and sampling variability. With the exception of logistic regression, the calibration methods tested all incorporate uncertainty into the LR itself, such that LRs will be closer to 1 when uncertainty is high (i.e. when sample size is low). Simulation was carried out based on distribution parameters of log scores obtained from an empirical study [8] where acoustic-phonetic features extracted from the English filled pause (FP) *um* were used as input. The focus of the current study is to investigate overall system validity and the stability of the system validity.

2. Method

2.1. Acoustic features derived from filled pauses

The simulated score distributions were derived from testing conducted with acoustic features extracted from filled pauses (FP) *um.* The original data consisted of 90 Standard Southern British English (SSBE) speakers and two samples per speaker from the DyViS corpus [19]. The first three formants and F0 of the vocalic portion of *um* were exacted and fitted with quadratic polynomial curves. The polynomials coefficients were then used for LR computation. 30 speakers were randomly sampled into training, test and reference sets, and the experiment was replicated 100 times showing that the C_{llr} varies between 0.13 and 1.22 across 100 replications.

2.2. Simulation

Table 1 shows the distribution parameters of log scores used for simulation. The mean, standard deviation, skewness and kurtosis values were derived from the original scores (somewhat surprisingly, both SS and DS scores were negatively skewed, although generally the extent of the skew was limited). It is acknowledged that the skewness could vary from case to case, but only one set of skewness is used here for the purpose of current study (examining the effect of different levels of skew is planned for future study).

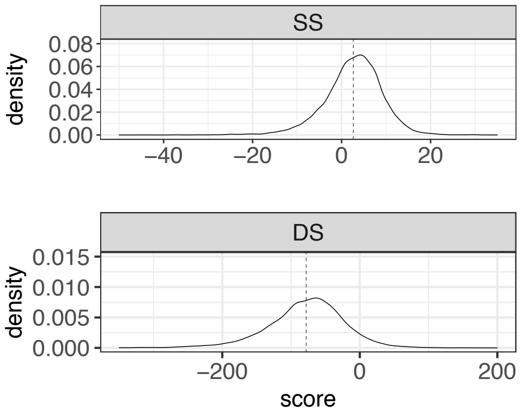
Table 1: *Distribution parameters for score simulation.*

***um* score**	**Mean**	**SD**	**Skewness**	**Kurtosis**
Log SS	2.6	6.6	-0.7	3.5
Log DS	-78	56.6	-0.7	3.1

Score simulation was carried out based on log scores because firstly log scores are normally used for *score-to-LR* computation. Secondly, the raw scores only allow for non-negative values where the SS and DS score distributions are extremely skewed and less symmetric. Both SS and DS raw scores are likely to be heavily tailed with the raw DS scores stacked between 0 and 1. Therefore, simulating the raw scores would further complicate the simulation process and introduce more uncertainty.

Since the SS and DS log scores derived from the real data are negatively skewed, the skew-t (ST) distribution [20] was considered appropriate for score simulation. The `rst()` function from the R [21] package *`sn`* [22] was used. Figure 1 shows the simulated SS (top panel) and DS (bottom panel) score distributions using parameters from Table 1. The dotted lines indicate the mean of the simulated SS and DS log scores (i.e. 2.6 and -78).

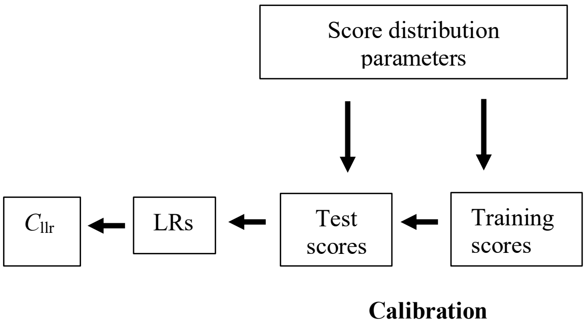
Figure 1: *Distributions of simulated SS and DS log scores (1000 samples per set).*



Sets of training and test scores were simulated with increasing sample sizes from 20 to 100 speakers in 10-speaker increasements. This means that the number of SS and DS log scores varies from 20 to 100 and 380 to 9900 respectively for both the training and test data. The experiment was also replicated 100 times within each sample size (i.e. number of speakers) using independent samples of scores. In this way, the experiments allow us to explore the relationship between sampling variability and sample size as well as which calibration methods is more or less resistant to sample size and sampling variability. Figure 2 shows a schematic representation of the simulation process.

The simulated training scores were then used to train calibration models which were applied to the test data, from which system validity was evaluated. Overall system performance was evaluated using the C_{llr} mean and range (i.e. the difference between the maximum and minimum C_{llr} values across 100 replications). Systems with better performance should yield lower C_{llr} mean and range. A C_{llr} of 1 is equivalent to a system that consistently produces LRs of 1 irrespective of whether they came from SS or DS comparisons. As such, a C_{llr} of less than 1 indicates that the system is capturing useful information.

Figure 2: *Schematic of the simulation process using score distribution parameters, replicated 100 times for each sample size.*



2.3. Calibration

Four calibration methods were tested, implemented using a set of Matlab functions from [17].

2.3.1 Logistic regression and regularised logistic regression

The rationale behind logistic regression [10] and regularised logistic regression (rlogistic regression) [17] are similar, i.e. the training scores are used to train a logistic regression model and the coefficients are then applied to the test scores to generate calibrated LRs (equation 1):

$$Calibrated\,LR = \alpha + \beta s \tag{1}$$

where α and β are the shift and scale values that are added and multiplied to the test scores (s) to generate calibrated LRs. Both logistic regression and rlogistic regression are claimed to be robust to violations of assumptions of normality and equal variance [23], i.e. both logistic regression and rlogistic regression are robust when scores are skewed. The difference between logistic regression and rlogistic regression is that the weight (κ^{ψ}) of an uninformative distribution needs to be applied in rlogistic regression. Depending on the value of κ^{ψ}, a smaller κ^{ψ} (≤ 0.1) value deals with the issues of complete separation and a larger κ^{ψ} (≥ 1) value deals with extreme LR outputs [17]. In the current study, we follow [17] and use a κ^{ψ} value of 5 for the purpose of reducing the variability of LR outputs.

2.3.2 Bayesian model

The fully Bayesian approach involves the use of priors (i.e. hyperparameters) to reduce the magnitude of the LRs when uncertainty is high [12,17]. The fully Bayesian calibration models need to be estimated using SS and DS training scores respectively. The likelihood of the Bayesian models are then evaluated using test scores [12]. Meanwhile, the prior belief and the strength of the belief for the mean and variance of the training scores need to be specified. However, due to the nature of FVC, the ground truth is impossible to know and it has been shown that uninformative priors yield more constrained Bayes factors than informative priors [24]. We therefore follow [17] using the Jeffreys reference priors. The formula for Bayesian model estimation can then be simplified to:

$$\lambda^{B} = t_{n-1}(x|_{\hat{\mu}}, \frac{n+1}{n-1}\hat{\sigma}^{2}) \tag{2}$$

where t is a t distribution, n is the sample size, x is the test score, μ and σ are the sample mean and pooled variance of the training score [17, p.203]. The calculation of Bayes factors is then the ratio between the likelihood of the Bayesian models evaluated using test scores,

$$\log(BF) = \log\left(\frac{t_{n_{ss}+n_{ds}-2}\left((x \mid \widehat{\mu_{ss}}, \frac{\bar{n}+1}{\bar{n}-1}\hat{\sigma}^{2})\right)}{t_{n_{ss}+n_{ds}-2}\left((x \mid \widehat{\mu_{ds}}, \frac{\bar{n}+1}{\bar{n}-1}\hat{\sigma}^{2})\right)}\right) \tag{3}$$

where the subscripts *ss* and *ds* indicate that the data come from the SS and DS training scores respectively and $\bar{n}$ is the sum of SS and DS samples divided by 2 [17, p.204].

2.3.3 Empirical lower and upper bounds (ELUB)

The ELUB [18] method uses empirical data to set maximum and minimum values for the LRs that a given system can output based on the training set. Then all other LRs produced by the test data are limited to within that range. ELUB is claimed to be robust to extreme LRs caused by data extrapolation where the estimated distribution density is not well supported by the observed data at the distribution tails [18]. The rationale behind the ELUB calibration method lies in a rule of the thumb that the LRs should be smaller than the sample size of the training data for the defence hypothesis and no larger than 1 divided by the size of training data for prosecution hypothesis [18]. The implementation of ELUB is carried out using the expected utility ratio (EU ratio; [13, pp. 204]):

$$EUratio = \frac{\begin{cases} 1 & if\ LR_{th}>1 \\ LR_{th} & if\ LR_{th} \leq 1 \end{cases}}{\frac{nLR_{s} \leq LR_{th} + 1}{nLR_{s}+1} + LR_{th} \times \frac{nLR_{d} > LR_{th}+1}{nLR_{d}+1}} \tag{4}$$

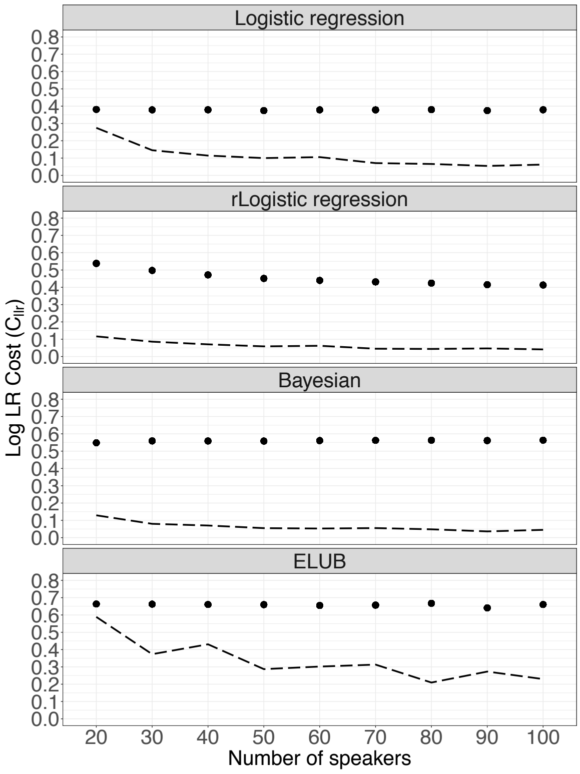
where the numerator is the neutral system LR that is no larger than 1 and the exact value depends on LR_{th} (a threshold LR) [18]. For the denominator, nLR_{s} and nLR_{d} are the number of the SS and DS LRs in the training data respectively. The $nLR_{s} \leq LR_{th}$ represents the number of SS LRs that is no larger than LR_{th} and $nLR_{d} > LR_{th}$ represents the number of DS LRs that is higher than LR_{th}. The upper and lower boundaries obtained from EUratio are then applied to test data to shrink the LR output [17].

3. Results

Figure 3 shows the mean (black dots) and range (dashed lines) of C_{llr} values as a function of sample size, across the 100 replications for each of the calibration methods. The x-axis indicates the number of speakers used in training and test data respectively and the y-axis represents the C_{llr} values.

The C_{llr} mean and range using different calibration methods show different patterns across different sample sizes. Logistic regression consistently yields the lowest mean C_{llr} across different sample sizes (varying between 0.37 and 0.38) followed by rlogistic regression (0.41 to 0.54), Bayesian (0.54 to 0.56) and ELUB (0.65 to 0.66). With the exception of rlogistic regression, the mean C_{llr} values of the other three calibration methods are fairly consistent (black dots are fairly flat) within each calibration method across different sample sizes showing that the mean C_{llr} does not become lower with larger sample sizes. For rlogistic regression, the mean C_{llr} decreases from 0.54 using 20 speakers to 0.41 using 100 speakers.

Figure 3: C_{llr} *mean and range using different sample size and calibration methods.*



In terms of C_{llr} ranges, the rlogistic regression and Bayesian methods appear to be less affected by sample size and sampling variability, i.e. the C_{llr} range remains low (ca. 0.1 or lower) and stable regardless of the number of speakers used. Contrastively, logistic regression seems to be more sensitive to sampling variability when the number of training and test speakers is less than 30. The C_{llr} range using logistic regression reduced from 0.27 to 0.06 when the number of training and test speakers increased from 20 to 100. Among the four calibration methods, EULB is most sensitive to sample size and sampling variability, and the C_{llr} range fluctuates considerably. For example, the C_{llr} range is 0.59 using 20 training and test speakers and lowered to 0.37 using 30; however, it increases to 0.43 when 40 training and test speakers are used.

4. Discussion

The results of this study have shown that overall system performance varies to different extents using different calibration methods with different sample sizes. In general, the ELUB calibration method is less preferable as it produces systems that are more sensitive to sampling variability and sample sizes than the other three. Although the C_{llr} range using ELUB reduces when more training and test speakers are included, there remains high levels of C_{llr} variability even with large samples; the C_{llr} range with 100 speakers is equivalent to that produced by logistic regression when using only 20 speakers (Figure 3, bottom panel). Using 1 as an appropriate threshold for judging C_{llr} [25] (i.e. a good system should yield a C_{llr} as close as to 0, and C_{llr} higher than 1 indicates that the system is not giving any useful information), the wide C_{llr} range using ELUB calibration suggests that it is the least effective of the four calibration methods as it produces C_{llr}s of over 1 across replications, even with 100 speakers in each set. The Bayesian model and rlogistic regression are less affected by sampling variability and should generally be preferred, especially when sample size is small. The disadvantage, however, is that priors or a κ^{ψ} value need to be pre-selected when using these two models. The priors within the Bayesian model need to be specified based on the mean and variance of the training data, which could be different from case to case in the real world. Similarly, different κ^{ψ} values of the rlogistic regression method need to be specified depending on the purpose of calibration, i.e. lower κ^{ψ} values deal with complete separation issues and higher κ^{ψ} values deals with extreme LR output issues [17]. Logistic regression should be preferred when the number of training and test speakers reaches more than 40 as the C_{llr} range tends to be lower when sample size is larger, and the C_{llr} mean is consistently lower than other three calibration methods.

In real world FVC, we are often dealing with small sample sizes – especially when using linguistic features, given the not insignificant challenges around data collection and analysis [26]. In our testing, logistic regression yielded lower C_{llr} mean but higher C_{llr} range than rlogistic regression or the Bayesian model when using smaller numbers of training and test speakers were used. It is therefore extremely important to understand the trade-off between C_{llr} mean and C_{llr} range, i.e. how much variability is allowed given accuracy (C_{llr}) and should we aim for lower C_{llr} mean (higher accuracy) as long as 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 absolute validity (i.e. the potential of a very low C_{llr}). Although it is difficult to set a generalised trade-off framework for all cases in the real world given the complexity and uniqueness of each individual case, we suggest that systems need to be tested multiple times with different sets or configurations of training and test data before applying it in real cases.

5. Conclusion

In the current study, four calibration methods have been compared in relation to sampling variability and sample size. Scores were simulated based on skewed distributions which provides some novelty in the testing of calibration method in LR-bared FVC. Although the results show that the C_{llr} range is relatively low using logistic regression, rlogistic regression and Bayesian model when score distributions are skewed, the current study only simulated scores using one set of skewness. It is likely that the score distributions are much more variable in real cases, especially when the sample size is small. In future work, we will test the robustness of calibration methods using scores derived different distributions.

6. References

[1] G. S. Morrison, 'Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio', *Australian Journal of Forensic Sciences*, vol. 45, no. 2, pp. 173–197, Jun. 2013, doi: 10.1080/00450618.2012.733025.

[2] G. S. Morrison and E. Enzinger, 'Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality', *Science & Justice*, vol. 58, no. 1, pp. 47–58, Jan. 2018, doi: 10.1016/j.scijus.2017.06.005.

[3] N. Brümmer and J. du Preez, 'Application-independent evaluation of speaker detection', *Computer Speech & Language*, vol. 20, no. 2–3, pp. 230–275, Apr. 2006, doi: 10.1016/j.csl.2005.08.001.

[4] G. S. Morrison, 'Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate', *Science & Justice*, vol. 56, no. 5, pp. 371–373, Sep. 2016, doi: 10.1016/j.scijus.2016.05.002.

[5] S. Ishihara and Y. Kinoshita, 'How Many Do We Need? Exploration of the Population Size Effect on the Performance of Forensic Speaker Classification'. In *Interspeech,* Brisbane Australia, 2008, p.1941 - 1944.

[6] Y. Kinoshita and S. Ishihara, 'Background population: how does it affect LR based forensic voice comparison?', *International Journal of Speech, Language and the Law*, vol. 21, no. 2, pp. 191–224, 2015, doi: 10.1558/ijsll.v21i2.191.

[7] V. Hughes, 'Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough?', *Speech Communication*, vol. 94, pp. 15–29, 2017, doi: 10.1016/j.specom.2017.08.005.

[8] B. X. Wang, V. Hughes, and P. Foulkes, 'The effect of speaker sampling in likelihood ratio based forensic voice comparison', *International Journal of Speech, Language and the Law*, vol. 26, no. 1, pp. 97–120, Aug. 2019, doi: 10.1558/ijsll.38046.

[9] P. Vergeer, Y. van Schaik, and M. Sjerps, 'Measuring calibration of likelihood-ratio systems_ a comparison of four metrics, including a new metric devPAV', *Forensic Science International*, p. 35, 2020, doi: https://doi.org/10.1016/j.forsciint.2021.110722.

[10] N. Brümmer *et al.*, 'Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006', *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 7, pp. 2072–2084, Sep. 2007, doi: 10.1109/TASL.2007.902870.

[11] B. Zadrozny and C. Elkan, 'Transforming classifier scores into accurate multiclass probability estimates', in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, Edmonton, Alberta, Canada, 2002, p. 694, doi: 10.1145/775047.775151.

[12] N. Brümmer and A. Swart, 'Bayesian Calibration for Forensic Evidence Reporting', in *Interspeech*, Singapore, 2014, pp. 388–392.

[13] D. Meuwly and A. Drygajlo, 'Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modelling (GMM)', in *Proceedings of Odyssey*, Crete, Creece, Jun. 2001, p. 6.

[14] A. Drygajlo, D. Meuwly, and A. Alexander, 'Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition', in *Proceedings of Eurospeech*, Geneva, Switzerland, 2003, pp. 689–692.

[15] A. Alexander and A. Drygajlo, 'Scoring and Direct Methods for the Interpretation of Evidence in Forensic Speaker Recognition', in *Interspeech*, Jeju, Korea, Oct. 2004, p. 4.

[16] T. Ali, L. Spreeuwers, R. Veldhuis, and D. Meuwly, 'Sampling variability in forensic likelihood-ratio computation: A simulation study', *Science & Justice*, vol. 55, no. 6, pp. 499–508, Dec. 2015, doi: 10.1016/j.scijus.2015.05.003.

[17] G. S. Morrison and N. Poh, 'Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors', *Science & Justice*, vol. 58, no. 3, pp. 200–218, May 2018, doi: 10.1016/j.scijus.2017.12.005.

[18] P. Vergeer, A. van Es, A. de Jongh, I. Alberink, and R. Stoel, 'Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating?', *Science & Justice*, vol. 56, no. 6, pp. 482–491, Dec. 2016, doi: 10.1016/j.scijus.2016.06.003.

[19] F. Nolan, K. McDougall, G. De Jong, and T. Hudson, 'The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research', *International Journal of Speech, Language and the Law*, vol. 16, no. 1, pp. 31–57, Sep. 2009, doi: 10.1558/ijsll.v16i1.31.

[20] R. B. Arellano-Valle and A. Azzalini, 'The centred parameterization and related quantities of the skew-t distribution', *Journal of Multivariate Analysis*, vol. 113, pp. 73–90, Jan. 2013, doi: 10.1016/j.jmva.2011.05.016.

[21] Core team R, *RStudio: Integrated Development for R*. RStudio, Inc., 2020.

[22] A. Azzalini, *The R package 'sn': The Skew-Normal and Related Distributions such as the Skew-t*. 2020.

[23] H. Tibshirani, T. R, and F. J, *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[24] G. S. Morrison, J. Lindh, and J. M. Curran, 'Likelihood ratio calculation for a disputed-utterance analysis with limited available data', *Speech Communication*, vol. 58, pp. 81–90, Mar. 2014, doi: 10.1016/j.specom.2013.11.004.

[25] G. Morrison *et al.*, 'Consensus on validation of forensic voice comparison', *Science & Justice*, Mar. 2021, doi: 10.1016/j.scijus.2021.02.002.

[26] E. Gold and V. Hughes, 'Issues and opportunities: The application of the numerical likelihood ratio framework to forensic speaker comparison', *Science & Justice*, vol. 54, no. 4, pp. 292–299, Jul. 2014, doi: 10.1016/j.scijus.2014.04.003.