Reducing uncertainty at the score-to-LR stage in likelihood ratio-based forensic voice comparison using automatic speaker recognition systems

Bruce Xiao Wang, Vincent Hughes

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

bruce.wang@alumni.york.ac.uk,vincent.hughes@york.ac.uk

Abstract

In data-driven forensic voice comparison (FVC), empirical testing of a system is an essential step to demonstrate validity and reliability. Numerous studies have focused on improving system validity, while studies of reliability are comparatively limited. In the present study, simulated scores were generated from i-vector and GMM-UBM automatic speaker recognition systems using real speech data to demonstrate the variability in system reliability as a function of score skewness, sample size, and calibration methods (logistic regression or a Bayesian model). Using logistic regression with small samples of skewed scores, C_{llr} range is 1.3 for the i-vector system and 0.69 for the GMM-UBM system. When scores follow a normal distribution, C_{llr} ranges reduce to 0.49 (i-vector) and 0.69 (GMM-UBM). Using the Bayesian model, the C_{llr} ranges are 0.31 and 0.60 for i-vector and GMM-UBM systems respectively when scores are skewed, and the C_{llr} range remains stable when scores follow a normal distribution irrespective of sample size. The results suggests that score skewness has a substantial effect on system reliability. With this in mind, in FVC it may be preferable to use an older generation of system which produces less variable results, but slightly weaker discrimination, especially when sample size is small.

Index Terms: forensic voice comparison, likelihood-ratio, logistic regression, Bayesian model, uncertainty

1. Introduction

1.1. Likelihood ratios and validation

Forensic voice comparison (FVC) typically involves the analysis of two speech samples; one from a known suspect and the other from an unknown offender. There is now an overwhelming consensus that experts should express their conclusions in FVC cases in the form of a likelihood ratio (LR), which is a measure of the strength of the voice evidence in light of the competing propositions of the prosecution and defence. In FVC, the expert needs to employ a *system*, defined broadly as the particular courses of action that are used to compare the suspect and offender samples and arrive at a conclusion [1]. For an end user (e.g. jury and/or the court) to be able to interpret that conclusion provided by the expert appropriately, it is essential to understand the validity (i.e., how well the system performs the task that it is designed to do) and reliability (i.e., whether the system would yield the same result if the analysis were repeated) of the system used. There is now considerable regulatory pressure on experts to validate systems under conditions reflective of forensic casework. Within exclusively data-driven, quantitative approaches to FVC (e.g. using automatic speaker recognition (ASR) systems), there are established procedures to empirically validating methods using the LR-based framework. This normally involves three sets of data (i.e., training, test and reference) and two stages (i.e., *feature-to-score* and *score-to-LR*) [1].

1.2. Understanding uncertainty

Any approach to FVC involves a series of processes and decisions which can, in principle, introduce uncertainty into the system, affecting both the resulting LR in the case and the measure of system validity. The priorities for forensics differ from those of other applications of speaker recognition. Within the field of speaker recognition, systems are evaluated and compared using overall measures of performance on benchmark datasets, which drive paradigm shifts in the algorithms used (e.g. from GMM-UBM to i-vector to x-vector). Conversely, in our opinion, measuring and ultimately attempting to minimise uncertainty (at every level within a system) should be the principle focus of forensic experts, as this directly impacts of the probability of miscarriages of justice. Thus, a system which produces more consistent results (i.e. one that has less uncertainty in its output) should be preferred over a system which is less consistent, even if, on average, the more consistent system produces poorer performance – it is the variability that matters, not the mean. The choice of system itself may depend on the specific voices under analysis in a case.

Numerous FVC studies have explored factors which introduce uncertainty into the process and affect system validity and reliability, e.g., sample size [2], channel mismatch [3], sampling variability [4]–[6], accent-mismatch [7]. Various techniques have been proposed to reduce the degree of uncertainty at both the *feature-to-score* and *score-to-LR* stages to improve system validity and reliability, e.g., using cepstral-mean subtraction [8] for channel mismatch compensation, obtaining reference data that is better matched for accent [9], and using recordings with conditions reflecting those of a real FVC case [10]. Sample size is a common factor that affects system validity and reliability especially when estimating within speaker variability, i.e. the distribution of within speaker variability is likely to be heavy-tailed (i.e., skewed) and less likely to follow normal distribution when the sample size is small. [11] proposed a method/framework for incorporating uncertainty into LR computation in the *feature* space using glass fragment data. They used a heavy-tailed Student's t distribution to model within-source variability showing that their proposed models outperformed MVKD [12] model which uses Gaussian distributions to model within-source variability. However, the challenges of reducing uncertainty in the *feature* space are that the feature space is often complex and highly multidimensional, and thus requires models with much high number of parameters. In comparison,

reducing uncertainty in the *score* space is less complex in terms of data dimensionality (using univariate score data) and the number of parameters required for the model is typically relatively low. A range of calibration methods have been proposed to incorporate uncertainty into LR computation at the *score-to-LR*, e.g., Bayesian model [13], empirical lower and upper bound [14], regularised logistic regression [14]. [16] used simulated scores to test the effectiveness of these calibration methods showing that Bayesian model [13] outperformed logistic regression in terms of system reliability when sample size is small and scores are skewed. However, they only compared the performance of different calibration methods using skewed scores and did not examine the extent uncertainty is affected when comparing skewed scores with normally distributed scores. Moreover, they simulated scores in [16] were based on a linguistic-phonetic system using MVKD, which is likely to have poorer overall performance than ASR systems using MFCCs.

1.3. The current study

The current study investigates the feasibility of using the Bayesian calibration model [13] within the *score* space as a means of reducing uncertainty introduced by sample size and score skewness. This is because the Bayesian model has previously yielded promising performance when sample size is small [16]. Moreover, it has been shown that scores are less likely to follow normal distribution when sample size is limited [4, 15, 16]. We also test different ASR systems (GMM-UBM, i-vector-PLDA) as different systems are likely to produce scores which are distributed in different ways and thus are likely to have different levels of uncertainty within them. Training and test scores used in the current study were simulated using score distribution parameters obtained from [10, 15] where the original scores were generated using MFCC and deltas with i-vector PLDA and GMM-UBM systems respectively. In our study, Logistic regression [17] is used to serve as a baseline model as it is one of the most widely used calibration methods for ASR systems. The baseline, logistic-regression calibration results were then compared with results using Bayesian calibration and evaluated in terms of both discrimination (mean C_{llr}) and uncertainty (C_{llr} range).

2. Method

2.1. Score simulation

Scores were simulated based on score distribution parameters obtained from [10, 15]. Each score indicates the similarity and typicality between SS and DS comparisons, and equivalent scores were simulated from both an i-vector system and GMM-UBM system. We recognise that neither of these systems are currently state-of-the-art; however, the point here is to compare relative rather than absolute patterns, focusing principally on uncertainty rather than discrimination as well as the interaction between the two. Given the priorities in forensics, in some circumstances, it may be preferable to use older system that produces less variability (i.e. less uncertainty) rather than the state-of-the-art.

Table 1 shows the score distribution parameters, i.e., skewness, kurtosis, mean and standard deviation. For both GMM-UBM and i-vector approaches, SS and DS scores are skewed, with SS scores having higher skewness than DS scores. Since both SS and DS scores are skewed to some extent, skew-t (ST) [18] distributions were chosen for simulation using the `rst()` function from the R (R core team, 2020) package `sn` [19]. In order to investigate if overall performance would be affected when scores were skewed (comparing with normal distributions), the skewness for both SS and DS scores were also changed to 0 (i.e., normal distribution) while the kurtosis, mean and standard deviation were kept fixed. Figure 1 shows examples of simulated i-vector (top panels) and GMM-UBM (bottom panels) scores by varying score skewness.

Table 1: *Score distribution parameters from an i-vector PLDA system using real speech data from 111 male Australian speakers.*

	i-vector		GMM-UBM	
	SS	DS	SS	DS
Skewness	-1.36	-0.69	0.56	-0.31
Kurtosis	8.47	3.64	4.06	3.99
Mean	-56.78	-223.23	0.04	-0.04
Standard deviation	34.79	83.50	0.04	0.04

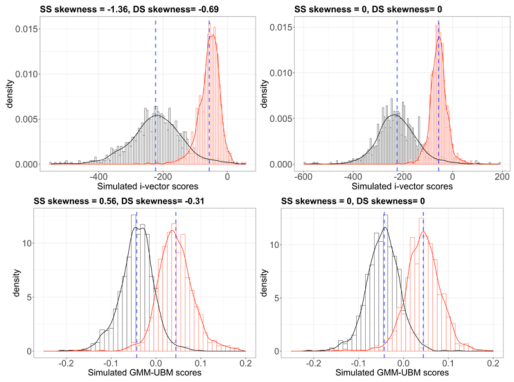

Figure 1: *Examples of simulated i-vector (top panel) and GMM-UBM (bottom panels) scores using parameters from Table 1, sample size = 1000 in each of the SS and DS scores. Blue dashed lines indicate the mean.*

To account for sample size, training and test scores were simulated with randomly selected sets of speakers increasing from 20 to 100 in 10-speaker increments, i.e., the SS and DS scores vary from 20 to 100 and 380 to 9900 for training and test data respectively. The simulated scores were calibrated using logistic regression and a Bayesian model respectively. Training and test scores were also simulated and calibrated 100 times per sample size to take the effect of sampling variability into consideration. The overall performance was evaluated using the mean (overall discrimination) and range (overall variability) of the C_{llr}s across the 100 replications. A C_{llr} lower than 1 indicates that the system is capturing useful information, and systems with better overall performance should yield both lower C_{llr} mean and range.

2.2. Calibration

Logistic regression is one of the most widely used calibration methods in FVC and has been employed widely in previous studies [3, 20–22], and it has been suggested that logistic regression is more robust to violations of the assumption of normality [1]. For these reasons, it is used as a baseline condition in our current experiments. Logistic regression involves fitting a sigmoidal curve is fitted to SS and DS training scores using maximum likelihood function in probability space[1], [23]. The sigmoidal curve is then

transformed from the probability space to the log-odds space to generate the linear relationship between SS and DS scores. Once the linear relationship is obtained, the shift and scale values (i.e., regressions coefficients) can then be added and multiplied to the test scores respectively to generate the calibrated LRs.

The Bayesian model involves the use of priors (i.e. hyperparameters) to reduce the magnitude of the LRs when uncertainty is high [13, 15]. Note that the priors (i.e., hyperparameters) here are those used for LR computation, not the *prior probability* in Bayes' theorem as applied to the case itself. Those hyperparameters, i.e., the prior belief and the strength of the belief for the mean and variance of the training scores, need to be specified. However, to our knowledge, no study has investigated the use of different hyperparameters in FVC and more importantly the rationale behind the usage of different hyperparameters. Therefore, as in [15], Jeffreys reference uninformative priors were used. Moreover, it has been shown that uninformative priors yield more constrained Bayes factors (BF, the Bayesian counterpart of the frequentist LR) than informative priors [24]. Once the hyperparameters are specified, training scores are used to train the Bayesian model, and the likelihood of the Bayesian model is evaluated using the test scores [13] using Equation 1.

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

Where t is a t distribution, n is the sample size, x is the test score, $\hat{\mu}$ and $\hat{\sigma}2$ are the sample mean and variance of the training score. The calculation of BF is then the ratio between the likelihood of the Bayesian models evaluated using test scores (Equation 2).

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

[15] pointed out that monotonicity is not guaranteed if the t distribution is used for both numerator and denominator. Therefore, certain constraints need to be imposed to reduce the extent of non-monotonicity. We then follow [15] and use pooled sample variance ($\hat{\sigma}^2$), rather than the variance of only training scores. The degrees of freedom ($n_{ss}+n_{ds}-2$) are therefore adjusted to take the pooled variance calculation into consideration and the $\bar{n}$ is the sum of SS and DS samples divided by 2. The implementation of calibration methods was conducted using a Matlab script [15].

[15] pointed out that monotonicity is not guaranteed if the t distribution is used for both numerator and denominator. Therefore, certain constraints need to be imposed to reduce the extent of non-monotonicity. We then follow [15] and use pooled sample variance ($\hat{\sigma}^2$), rather than the variance of only training scores. The degrees of freedom ($n_{ss}+n_{ds}-2$) are therefore adjusted to take the pooled variance calculation into consideration and the $\bar{n}$ is the sum of SS and DS samples divided by 2. The implementation of calibration methods was conducted using a Matlab script [15].

3. Results

Figure 2 shows the C_{llr} mean (dots; discrimination) and range (lines; uncertainty) across the 100 replications for different sample sizes, score skewness and calibration methods respectively. The x-axis represents the number of training and test speakers and the y-axis shows C_{llr}. The GMM-UBM results are in red and the i-vector results are in black. Predictably, average C_{llr} values are lower across all sample size conditions for the i-vector system, compared with the GMM-UBM system. However, there are a series of interesting patterns related to uncertainty (C_{llr} range) meaning that the i-vector system is not necessarily the optimal choice for FVC.

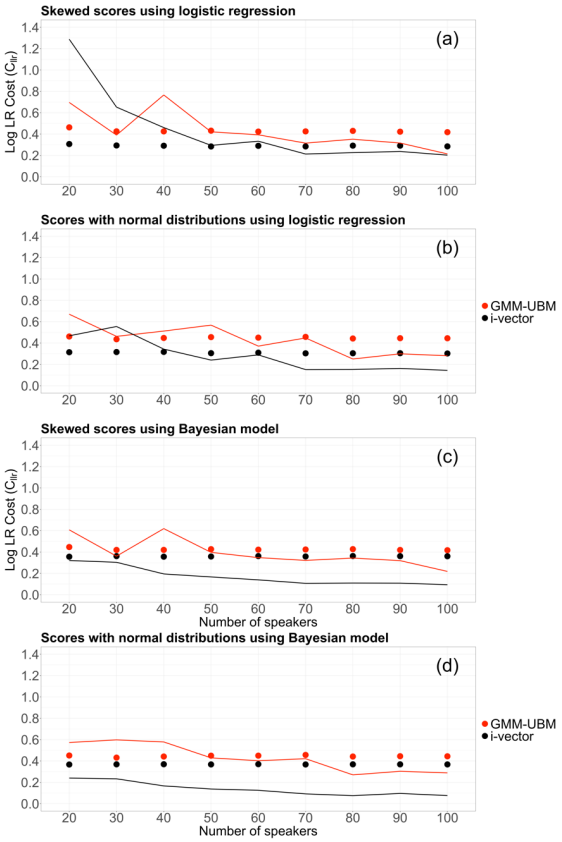

Figure 2: *C_{llr} mean and range as a function of score skewness, sample size and calibration methods.*

3.1 Baseline systems with logistic regression

When scores are sampled from skewed distributions (i.e. using parameters from score distributions actually output by these ASR systems) and logistic regression is used for calibration (Figure 2. (a)), as is the typical situation in current FVC systems, the C_{llr} range fluctuates substantially for both GMM-UBM and i-vector systems, especially when sample size is small. The i-vector system produces a very high C_{llr} range with only 20 speakers, despite producing lower average C_{llr}. This is driven by a small number of replications that produce very high C_{llr} when sample size is low. This is problematic for FVC as it means there is much greater uncertainty overall for the i-vector system despite better levels of average discrimination, compared with the GMM-UBM system. For both i-vector and GMM-UBM systems, C_{llr} range decreases as sample size increases. After the inclusion of 40 speakers, the C_{llr} range for the i-vector system is lower than for the GMM-UBM system. The mean C_{llr} also reduces from 0.48 to 0.40 between 20 and 100 speakers for GMM-UBM while it remains at around 0.3 for i-vector.

The C_{llr} ranges for both i-vector and GMM-UBM systems are less fluctuating as a function of sample size when scores are sampled from normal distributions (Figure 2. (b)), rather than the skewed distributions (Figure 2. (a)). The C_{llr} range varies from 0.49 to 0.18 (i-vector) and from 0.69 to 0.29 (GMM-UBM) when sample sizes increase from 20 to 100 speakers. This indicates that score skewness has a marked

effect on system reliability, such that the greater the skew in the underlying score distributions, the greater the uncertainty, when using logistic regression (especially when sample size is small). This would also explain the pattern in Figure 2 (a), since the GMM-UBM system produces inherently less skewed scores than the i-vector system.

3.2. Systems with Bayesian model

Using the Bayesian calibration model on the skewed data (Figure 2. (c)), less fluctuation and lower absolute values for C_{llr} range are observed across different sample sizes compared with using logistic regression (Figure 2. (a)). Again, C_{llr} range decreases as sample size increases. C_{llr} range varies from 0.60 to 0.31 and from 0.31 to 0.15 for GMM-UBM and i-vector systems respectively when sample size increases from 20 to 100 speakers. Overall, the i-vector system consistently produced the lowest C_{llr} mean and range, compared with GMM-UBM. The trade-off in terms of improved and less variable C_{llr} range is a slight increase in C_{llr} mean (i.e. a loss in discrimination) compared with logistic regression. When scores are sampled from normal distributions (Figure 2. (d)), the C_{llr} range (0.22 to 0.19) for the i-vector system shows less fluctuation compared with the skewed scores. The C_{llr} range for the i-vector system remains lower than 0.20 when the sample size is over 40 speakers; however, the difference between the C_{llr} fluctuations in the Figure 2. (c) and Figure 2. (d) is very small. Similarly, there is little difference between the C_{llr} range from the GMM-UBM system using normally distributed scores (C_{llr} range = ca. 0.59 to 0.30; Figure 2. (d)) and those using skewed scores (C_{llr} range = ca. 0.60 to 0.31; Figure 2. (c)). Irrespective of the score skewness, the mean C_{llr} values remain stable across different sample sizes for both GMM-UBM (mean C_{llr} = ca. 0.48) and i-vector mean (C_{llr} = ca. 0.38) systems.

4. Discussion

4.1 Baseline system

The results show that score skewness has a more marked effect on system reliability than on system validity. When sample size is small and logistic regression is used, the C_{llr} variability is high for both systems (Figure 2. (a) and (b)), but especially for the i-vector system. For example, the C_{llr} range in the i-vector system is almost 1.3 when scores are skewed (Figure 2. (a)) and 20 speakers are used. However, C_{llr} range is around 0.5 with the same sample size (i.e., 20 speakers) when scores follow a normal distribution (Figure 2. (b)). Meanwhile, score skewness seems to have a less marked effect on system reliability for the GMM-UBM system when sample size is small. This is principally because GMM-UBM produced less skewed scores (Table 1).

Within calibration methods, mean system validity (mean C_{llr}) stays stable regardless of score skewness and sample size. This is unsurprising because system validity is mostly dependent on the amount of overlap between SS and DS score distributions which is depend on the distance between the mean of SS and DS score distributions as well as the variance. Since skewness was the only parameter adjusted in the current study, the mean system validity should not be substantially affected across normally distributed and skewed scores.

4.2 Bayesian calibration as a solution

The results in Figure 2 (c) and (d) show that Bayesian calibration improves system reliability considerably when scores are skewed (as is commonly found in the real world). Using scores for 20 speakers simulated from the i-vector system and the Bayesian calibration model, the C_{llr} range is just above 0.3 (Figure 2 (c)) compared with 1.3 (Figure 2 (a)) for logistic regression. C_{llr} range further decreases when more speakers are used. For scores simulated from the GMM-UBM system, the Bayesian calibration does not seem to improve system reliability as much it does in the i-vector system. Figure 2 (c) and (d) shows that score skewness has much less effect on system reliability using Bayesian model compared with the use of logistic regression for the i-vector system, i.e., the differences between the C_{llr} range (black lines in Figure 2 (c) and (d)) are less than 0.1 across different sample sizes. For the GMM-UBM system, score skewness seems to have less effect on system stability when 40 or more speakers are used, i.e., the difference between the C_{llr} range values in Figure 2 (c) and (d) (red lines) become smaller when 40 or more speakers are used.

The mean C_{llr}s obtained using Bayesian calibration do not fluctuate considerably when scores are sampled from skewed and normal distributions respectively (Figure 2. (c) and (d)). This is similar to the pattern observed in mean C_{llr}s using logistic regression (Figure 2 (a) and (b)). However, for the i-vector system, the mean C_{llr} is slightly higher using Bayesian calibration (mean C_{llr} = 0.38) than using logistic regression (mean C_{llr} = 0.30), while the mean C_{llr} remains stable irrespective of calibration methods for GMM-UBM. It is therefore important for experts to consider what they consider a 'low enough' mean C_{llr} to be in making decision about which system to use in a forensic case and the potential trade-offs between discriminability and uncertainty (see [25]).

5. Conclusion

The current study investigated the effect of score skewness and sample size on overall performance of different types of ASR systems and the feasibility of reducing uncertainty at the *score* space via Bayesian calibration. On the surface, it seems that using Bayesian model can effectively reduce the level of uncertainty and fluctuation in overall performance caused by score skewness. The trade-off is a slight increase in mean C_{llr}. Taken together, our results indicate that, in the forensic context, it is not the case that the modern 'state-of-art' system which is capable of the best validity is necessarily the optimal choice. Rather, this choice is dependent on sample size, score skewness, and the choice of calibration method (likely amongst of considerations).

Future work should consider the effects of different priors on Bayesian calibration. The Jefferys prior used in the current study is a member of the *beta distribution* (i.e., continuous distributions with intervals between 0 and 1). Similarly, the Haldane (i.e., $\alpha = 0$) and Laplace (i.e., $\alpha = 1$) priors are also members of the *beta distributions*, and [13] pointed out that Haldane prior is more appropriate for evaluating DNA evidence. A further empirical question then would be *which* prior gives the best overall performance; however, a scientific question would be *why* certain prior yields the best overall performance and the rationale behind the choice of different priors for LR computation, which needs to be explored more in future studies.

6. References

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

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

[3] V. Hughes, P. Harrison, P. Foulkes, P. French, and A. J. Gully, 'Effects of formant analysis settings and channel mismatch on semiautomatic forensic voice comparison', in *International Congress of Phonetic Sciences*, Melbourne, Australia, Aug. 2019, pp. 3080–3084.

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

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

[6] B. Wang, V. Hughes, and P. Foulkes, 'Effect of score sampling on system stability in Likelihood Ratio based forensic voice comparison', in *International Congress of Phonetic Sciences*, Melbourne, Australia, Aug. 2019, p. 5.

[7] D. Watt *et al.*, 'Assessing the effects of accent-mismatched reference population databases on the performance of an automatic speaker recognition system', *Int. J. Speech Lang. Law*, vol. 27, no. 1, Jul. 2020, doi: 10.1558/ijsll.41466.

[8] S. Furui, 'Cepstral analysis technique for automatic speaker verification', *IEEE Trans. Acoust. Speech Signal Process.*, vol. 29, no. 2, pp. 254–272, Apr. 1981, doi: 10.1109/TASSP.1981.1163530.

[9] V. Hughes and P. Foulkes, 'The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age', *Speech Commun.*, vol. 66, pp. 218–230, a 2015, doi: 10.1016/j.specom.2014.10.006.

[10] E. Enzinger, G. S. Morrison, and F. Ochoa, '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', *Sci. Justice*, vol. 56, no. 1, pp. 42–57, Jan. 2016, doi: 10.1016/j.scijus.2015.06.005.

[11] D. Ramos, J. Maroñas, and J. Almirall, 'Improving calibration of forensic glass comparisons by considering uncertainty in feature-based elemental data', *Chemom. Intell. Lab. Syst.*, vol. 217, p. 104399, Oct. 2021, doi: 10.1016/j.chemolab.2021.104399.

[12] C. G. G. Aitken and D. Lucy, 'Evaluation of trace evidence in the form of multivariate data', *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 53, no. 1, pp. 109–122, Jan. 2004, doi: 10.1046/j.0035-9254.2003.05271.x.

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

[14] 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?', *Sci. Justice*, vol. 56, no. 6, pp. 482–491, Dec. 2016, doi: 10.1016/j.scijus.2016.06.003.

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

[16] B. X. Wang and V. Hughes, 'System Performance as a Function of Calibration Methods, Sample Size and Sampling Variability in Likelihood Ratio-Based Forensic Voice Comparison', in *Interspeech 2021*, Aug. 2021, pp. 381–385. doi: 10.21437/Interspeech.2021-267.

[17] 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.

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

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

[20] P. Rose and Z. Cuiling, 'Conversational Style Mismatch: its Effect on the Evidential Strength of Long- term F0 in Forensic Voice Comparison', in *Proc. 17th Australasian Int'l conf. on Speech Science and Technology*, Sydney, 2018, pp. 157–160.

[21] F. Kelly, A. Fröhlich, V. Dellwo, O. Forth, S. Kent, and A. Alexander, 'Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)', *Speech Commun.*, vol. 112, pp. 30–36, Sep. 2019, doi: 10.1016/j.specom.2019.06.005.

[22] L. Gerlach, K. McDougall, F. Kelly, A. Alexander, and F. Nolan, 'Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features', *Speech Commun.*, vol. 124, pp. 85–95, Nov. 2020, doi: 10.1016/j.specom.2020.08.003.

[23] D. Ramos, 'Forensic evaluation of the evidence using automatic speaker recognition systems', *Univ. Autónoma Madr.*, 2007.

[24] G. S. Morrison, J. Lindh, and J. M. Curran, 'Likelihood ratio calculation for a disputed-utterance analysis with limited available data', *Speech Commun.*, 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', *Sci. Justice*, vol. 61, no. 3, pp. 229–309, Mar. 2021, doi: 10.1016/j.scijus.2021.02.002.