Article:

Wang, Bruce and Hughes, Vincent orcid.org/0000-0002-4660-979X (2024) Balancing validity and reliability as a function of sampling variability in forensic voice comparison. Science & justice : journal of the Forensic Science Society. pp. 649-659. ISSN 1355-0306

https://doi.org/10.1016/j.scijus.2024.10.002

Balancing validity and reliability as a function of sampling variability in forensic voice comparison

Abstract

In forensic comparison sciences, experts are required to compare samples of known and unknown origin to evaluate the strength of the evidence assuming they came from the same- and different-sources. The application of **valid** (if the method measures what it is intended to) and **reliable** (if that method produces consistent results) forensic methods is required across many jurisdictions, such as the England & Wales Criminal Practice Directions 19A and UK Crown Prosecution Service and highlighted in the 2009 National Academy of Sciences report and by the President's Council of Advisors on Science and Technology in 2016. The current study uses simulation to examine the effect of number of speakers and sampling variability and on the evaluation of validity and reliability using different generations of automatic speaker recognition (ASR) systems in forensic voice comparison (FVC). The results show that the *state-of-the-art* system had better overall validity compared with less advanced systems. However, better validity does not necessarily lead to high reliability, and very often the opposite is true. Better system validity and higher discriminability have the potential of leading to a higher degree of uncertainty and inconsistency in the output (i.e. poorer reliability). This is particularly the case when dealing with small number of speakers, where the observed data does not adequately support density estimation, resulting in extrapolation, as is commonly expected in FVC casework.

1. Introduction

In forensic comparison sciences, demonstrating the application of valid and reliable methods is required across many jurisdictions, e.g., the USA (Daubert ruling [1993]), the England & Wales Criminal Practice Directions 19A [1] and UK Crown Prosecution Service [2]. The importance of establishing and applying both valid and reliable methods in forensic comparison has also been addressed in the National Academy of Sciences report [3] and President's Council of Advisors on Science and Technology [4]. Validity refers to whether the method does what it is claimed to do (i.e. separate same- and different-source samples), and reliability refers to the consistency of evaluation results if the analyses were repeated by the same (repeatability) or/and different experts and methods (reproducibility). In order to address those concerns, in the field of forensic voice comparison (FVC) [5] as well as some forensic disciplines, such as blood stain [6], bullet and cartridge case comparisons [7], footwear [8] and fingerprint [9], [10], empirical testing has been conducted to assess both the validity and reliability of evaluation methods and results empirically. Studies of this sort use samples where the ground truth of the comparisons is known, but are often 'black-box' in the sense that some steps in the decision-making process are unspecified [9] (i.e., low transparency).

In forensic voice comparison (FVC), the job of the expert is to compare and evaluate recordings, one of an unknown offender, and the other of a known suspect, to assist the trier-of-fact in determining the likelihood that the two speech samples came from the same person or different people. There are broadly two approaches in achieving this: 1) human expert method, which relies on knowledge and experience and involves auditory and/or acoustic analyses; 2) the employment of an automatic speaker recognition (ASR) system [11]. Ultimately, the expert's role is to minimise the probability of a miscarriage of justice, and empirical validation serves as a method to bolster this assurance. Over the past few decades, the likelihood ratio (LR) framework has gained widespread acceptance in the evaluation and reporting of voice evidence, mirroring the approach taken with DNA evidence and reflecting a broader *paradigm shift* in forensic science [12]. DNA typing if often considered the gold standard in forensic evidence evaluation due to its statistical approach computing the probability of matching the offender's sample against the suspect as well as against the relevant population, namely the LR. This has influenced the way voice evidence is assessed. Concurrently, there has been growing debate within the field of FVC about validation. While forensic speech scientists typically provide a verbal LR in FVC, they often do not include numerical estimations of typicality. Nevertheless, there is significant research effort directed towards incorporating numerical LR frameworks into practice [13], [14], [15], [16], aligning with the *paradigm shift*.

While the integration of the LR framework into FVC is a step forward, it does not inherently guarantee high validity and reliability; nor does it, alone, necessarily minimise the probability of a miscarriage of justice. Any approach to FVC involves a series of complex processes and decisions which can, in principle, introduce uncertainty into the pipeline, affecting both the resulting LR in the case and the measurement of system validity. In line with [17], our starting point is that the priority in FVC or indeed any evidence evaluation field, should be to measure and reduce uncertainty, rather than maximising discrimination. Some suggestions, such as proficiency tests and collaborative exercises, have been proposed by the Expert Working Group for Forensic Speech and Audio Analysis [18] of ENFSI (European Network of Forensic Science Institutes) for reliability evaluation and as part of quality control; however, questions about how reliability and uncertainty should be measured and how conclusions should be assessed are not explicitly explained.

1.1 Likelihood ratio and validation in FVC

The LR quantifies the strength of evidence under the two mutually-exclusive, competing propositions of the prosecution and defence [19], expressed in its odds form as:

$$LR = \frac{p(E|H_p,I)}{p(E|H_d,I)} \qquad \text{Equation (1)}$$

where $p(E|H_p)$ indicates the probability of observing the difference between the suspect and offender speech samples given the prosecution proposition, i.e., the speech sample comes from the suspect; $p(E|H_d)$ represents the probability of observing the difference between the suspect and offender speech samples given the defence proposition, i.e., the speech sample does not come from the suspect but someone else from the relevant population; I stands for background information about the case. Essentially, the numerator of the LR is an estimation of the similarity between the suspect and offender speech samples, while the denominator is an estimation of their typicality compared to the relevant population.

In order to generate a LR, the expert employs a *system*, defined broadly as the particular course of action that is used to compare the suspect and offender samples and arrive at a conclusion [20], e.g., the data used to represent the relevant population, the linguistic-phonetic variables chosen for analysis, the methods of analysing those variables. Note that there are different methods used within FVC [11], and our focus is on methods that output numerical LRs.

With the development of computational models and automatic speaker recognition (ASR) systems, more and more forensic speech scientists have started using automatic systems [11], [21], [22] for the purpose of FVC casework. There are broadly four stages in forensic ASR system, namely, feature extraction, feature modelling, score generation and LR computation. In stage one, speech features across the entire speech-active portion of a recording are extracted. Typically these features are Mel-frequency cepstral coefficients (MFCCs)[1] or log Mel filterbanks. The extracted features can then be processed in various ways to produce speaker models (e.g., Gaussian Mixture Model-Universal Background Model (GMM-UBM), i-vector, x-vector). Scores are calculated to indicate the similarity (and often typicality) between a pair of recordings. In more modern systems, this is typically done using probabilistic linear discriminant analysis (PLDA) [24] or cosine similarity. In the final stage, scores are converted to interpretable LRs via a process of calibration.

[1] Note that voice activity detection (VAD) in forensic recordings sometimes might not be performed by the system, but segmented by the expert [23].

These ASR systems are evaluated and compared using overall measures of performance on benchmark datasets, which in turn lead to transformations in the algorithms used within systems based on improved validity. In such work, analysts validate their system empirically using data where the ground truth is known, to present the results of validation tests to the end user. Data extracted from pairs of same-speaker (SS) and difference-speaker (DS) recordings taken from the test and training datasets are compared to produce test and training scores which indicate the similarity between the SS and DS samples and assessing typicality with respect to a set of reference data. The training scores are used to train a calibration model (commonly using logistic regression), the coefficients from which are then applied to the test scores to convert them to interpretable LRs. The *system* validity is then typically evaluated using Log LR cost function (C_{llr}) [25], [26]; although accept-reject metrics such as equal error rate (EER) are also commonly used. A C_{llr} between 0 and 1 indicates that the *system* is capturing some useful information, and the closer to 0 the better the system validity is. A C_{llr} of 1 is equivalent to a *system* that consistently produces LRs of 1 irrespective of whether they came from same-speaker (SS) or difference-speaker (DS) comparisons, and a LR of 1 indicates equal supports for prosecution and defence in terms of the strength of evidence. As such, a C_{llr} of higher than 1 indicates that the *system* is not capturing any useful information (and may be affected by miscalibration).

The past two decades have witnessed the development of broadly three generations of ASR systems, namely, GMM-UBM [27], i-vector [28], and DNN-based embedding (e.g. x-vector) [24] [29] systems, each demonstrating improved speaker discrimination performance. The increased use of ASR systems in FVC is likely due to: first, the alignment of ASR systems with the numerical LR framework; second, the operation is less labour-intensive and therefore validation is a more efficient process; third, ASR systems are comparatively more objective than human-centred comparisons; fourth, ASR systems are now demonstrating very good speaker discrimination performance, especially in certain conditions [5]. However, high speaker-discriminatory performance (i.e. high validity) does not necessarily lead to high reliability, and very often it is the opposite [30]. In the context of FVC, [17] used simulated data to demonstrate that as long as the overall performance is considered to be valid, a system which produces more consistent results should be preferred over a system which is less consistent.

1.2 Understanding measurement and uncertainty

The validation of methods in any type of forensic comparison science can be thought of as a measurement process. The result of the validation is necessarily dependent on elements of that measurement process including, but not limit to, the specifications of the computational equipment to be used, the operations to be performed, the experts who performed the operations and the sequences and conditions in which the operations are executed [31]. In the context of FVC, the measurement process could refer to the systematic procedure used to validate system performance, analyse and compare voice samples to determine the likelihood of same- or different-origin. For instance, computational equipment might involve hardware such as headphones and computers, as well as software such as specialised commercial applications or in-house written scripts. The operations could include recording voice samples, choosing phonetic parameters, extracting relevant acoustic features, and comparing these features using statistical models. The sequences and conditions might involve pre-processing voice samples (e.g., converting between analogue and digital recordings, voice activity detection), accounting for background noises, and considering recording device characteristics.

Any scientific measurement process needs to be repeatable and reproducible to attain a state of **statistical control** [32] before it can "be regarded in any logical sense as measuring anything at all" [30, p.162]. That is to say, the validation results of any forensic comparison need to attain a certain degree of consistency, namely a state of **statistical control**, before it can answer the question of whether it is faithful to what it is intended to measure.

Various factors affect the consistency of scientific measurements, and these factors have been categorised using different terminologies in previous studies, e.g., systematic and random factors [30], tangible and intangible factors [31]. For the sake of simplicity, we will focus on the effects of such factors in terms of *uncertainty*. In defining uncertainty, we follow the points laid out in [31, p.88] that "uncertainty is a broader concept than 'error'" which reflects the "incompleteness of knowledge about how well the test result represents the quantity measured", and "uncertainty can exist even in the absence of error in the sense of 'mistake'". Within exclusively data-driven, quantitative approaches to FVC (i.e. using ASR systems), uncertainty could be introduced at various stages of the measurement process (e.g. selection of training and test data, score and LR computation), affecting the evaluation results and system validation. For example, epistemic uncertainty that is out of experts' control (e.g. sampling variability) [33], [34], [35] and experimental uncertainty that is under experts' control, such as numbers of formants or MFCCs to be used, and numbers of speakers to be sampled into training, test and reference set [36], [37]. Using any automatic systems for FVC, the validation result is not only a reflection of the performance of that specific automatic system, but also the uncertainties introduced at different stages of the operation.

One way of reducing uncertainty in data-driven approaches is to use larger samples, in turn increasing the reliability of any density estimates in probabilistic models. However, the challenge in FVC is that a practitioner may only have access to a limited amount of representative data that accurately reflects real case conditions given the significant challenges around data collection and analysis [38]. Indeed, representative data for a given case may not be available at all. Yet, sufficient representative data that mirrors real case conditions is crucial for the validation of a FVC system and small datasets may misrepresent the potential performance of the system [26]. The primary goal of the current paper is to raise the awareness about the importance of reliability and to balance reliability and discrimination in forensic comparison testing, especially under cases with sparse representative data.

1.3 The current study

The current study aims to address two common factors that limit the consistency of validation results in FVC using ASR system, namely, sample size[2] (i.e., number of speakers) and sampling variability, and underscore the significance of reliability, which is equally crucial if not more so than validity in FVC. The focus on sample size and sampling variability is driven by the real-world paucity of forensically-realistic datasets of recordings, especially those which are representative of the specific conditions of any individual case. Thus, the application of ASR in FVC is likely to always involve relatively small samples [23].

Several studies have investigated the issues of sample size and sampling variability in data-driven FVC, for example, [33] explored the impact of sampling variability on LR computation in relation to different calibration methods, assuming normal and reversed Weibull distributions for scores generated from a i-vector PLDA ASR system [39]. However, this assumption often does not hold in a forensic context, where speech data is rarely normally distributed, especially for comparisons under prosecution proposition, due to the limitations of sample size. [40] examines the effects of sampling variability and sample size on LR outputs, with a specific focus on score skewness and calibration methods within a GMM-UBM ASR system. While this study expanded to involve score skewness, it did not extent to the newer generation of systems. Similarly, [17], [41] investigated various calibration methods to mitigate LR variability in relation to sample size, but this research was confined to the GMM-UBM and i-vector systems. In a recent study, [23] investigated a range of data partitioning techniques with the goal of identifying the most effective technique to reduce the effect of sampling variability on LR outputs, especially in cases where the representative data has a limited number of speakers. Yet, his analyses were limited to the x-vector system, and the choice of x-vector is likely due to the fact that x-vector systems have been shown to have better speaker-discriminatory performance than the GMM-UBM and i-vector systems [5]. However, no studies have collectively compared these three systems in terms of variability in validity and reliability as a function of sampling variability and sample size.

[2] Unless otherwise specified, the term 'sample size' refers to the number of speakers in the training, test, and reference samples.

In the present study, we simulated scores, derived from real speech data, from three generations of ASR systems (i.e. GMM-UBM, i-vector, x-vector), demonstrating the effect of sampling variability in relation to validity and reliability, replicating real-world casework conditions (i.e. small sample size). Logistic regression was first used to calibrate the simulated scores and convert them into LRs, as is the typical procedure in current automatic FVC systems [42], [43], [44]. Further, Bayesian model was used for calibration as previous studies [17], [41] have demonstrated the efficacy of Bayesian model in reducing variability in LRs, particularly in situations with limited sample sizes. However, it is worth noting that the motivation of the current study is not to compare different calibration methods nor to suggest data partitioning methods described in [23], but to highlight the importance of reliability, focusing on measuring uncertainty, and a conceptual shift in balancing reliability and validity in FVC systems that output numerical LRs. System validity and reliability were assessed using the C_{llr} and 95% credible intervals (CI) of the LRs. The simulation used in the current study allows us to focus on the sample size and sampling variability factors, rather than the reliability of data extraction or/and data used to train background models (e.g., DNNs).

2. Methods

For this study, we used scores, generated by GMM-UBM, i-vector and x-vector systems, from previous studies [41], [45] [3]. The same subset of a speech corpus containing male Australian speakers was used to generate scores using the three automatic systems. For each speaker, the corpus contains a landline telephone call with background office noise and a pseudo police interview with background ventilation system noise. There are multiple recordings for each speaker, resulting in 111 SS scores and 9720 DS scores. For the GMM-UBM and i-vector scores, MFCCs and deltas were used as the input speech features, whereas log Mel filterbanks were used for the x-vector system [46]. Three systems had different scoring methods. In the GMM-UBM system, the scores are the likelihood of the measurement vector of the offender given the suspect model and UBM, where the UBM was trained with 512 Gaussian components using speech data exclude the suspect and offender data, and the suspect model was trained using MAP adaptation to the UBM model [47]. For both i-vector and x-vector systems, the scores were computed from the corresponding vectors (i.e., i- or x-vector) first using linear discriminant analysis (LDA) for domain mismatch compensation (e.g., channel mismatch), followed by probabilistic linear discriminate analysis (PLDA) calculating the likelihood of the two vectors assuming they came from the same speaker or different speakers. The differences between the i-vector and x-vector lie in their respective extraction methods and underlying models. The i-vector is a low-dimensional representation derived from a high-dimensional GMM supervector, which is trained based on a UBM using mean-only Maximum a Posteriori (MAP) adaptation [48]. In contrast, the extraction of an x-vector relies on a pre-trained DNN, capturing more complex and non-linear patterns in the speech data [49].

The parameters of extracted score distributions are given in Table 1, and Figure 1 shows simulated SS and DS score distributions with a sample of 100 data point in each distribution. The blue dashed lines indicate the mean.

[3] GMM-UBM, i-vector and x-vector scores are available at https://geoff-morrison.net/

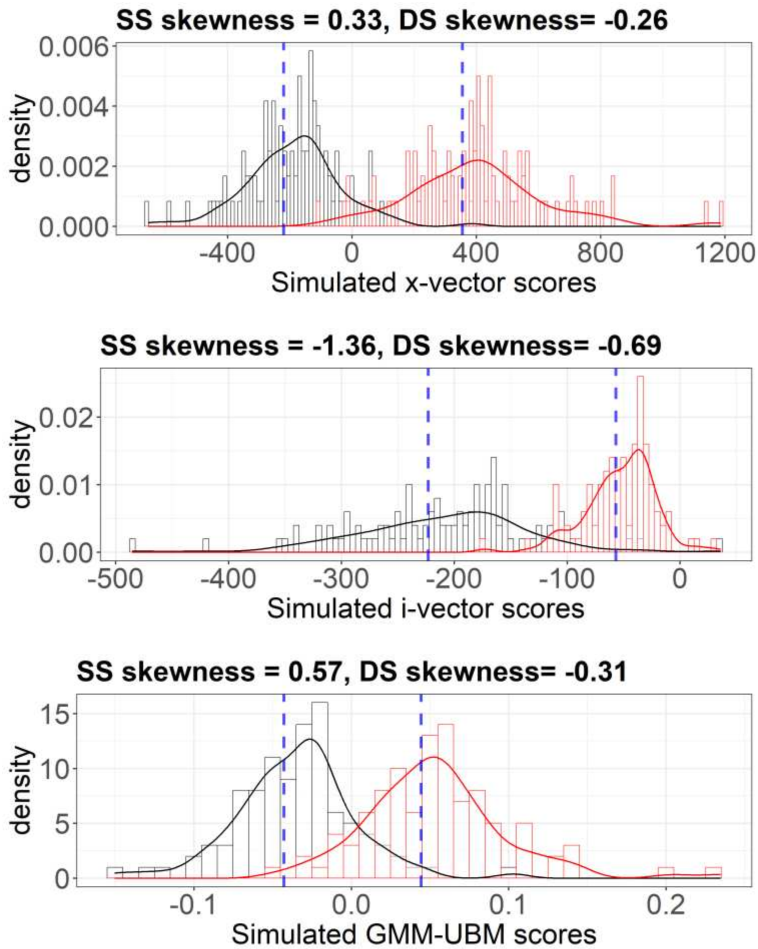


Figure 1. Examples of simulated x-vector, i-vector and GMM-UBM scores using parameters from Table 1, sample size = 100 in each of the SS (red) and DS (black) score distributions.

For all three systems, both scores are skewed to some extent, with SS scores having higher skewness than DS scores. As such we simulate scores from the skew-t (ST) [50] distributions using the `rst()` function from the R [51] package `sn` [52]. To account for uncertainty introduced by sample size, the training and test scores were sampled with increasing numbers of speakers from 20 to 100 with a 10-speaker increase, namely, the SS and DS scores vary from 20 to 100 and 380 to 9900 for training and test data respectively. A brand-new dataset was created for each sample size each time. The simulated training scores were then used to train calibration models, which were applied to the test data from which system validity was evaluated. For calibration models training, we employed two methods. First, logistic regression [53] which is widely employed in previous studies and in commercial ASR systems [43], [54], [55], [56]. Second, the Bayesian model [57] (see Appendix for a brief explanation for Bayesian model). The choice of the logistic regression is due to its wide acceptance and popularity [20], while the Bayesian model was chosen for its ability to reduce uncertainty and variability in LRs [41], particularly in situations with limited sample sizes [17]. To account for uncertainty introduced by sampling variability, the experiment was conducted 100 times within each sample size using independent samples of scores, a typical bootstrapping method used for system evaluation in other biometric recognition systems [58]. This allows us to explore the relationship between sampling variability and sample size in relation to different generations of ASR systems.

	x-vector		i-vector		GMM-UBM	
	SS	DS	SS	DS	SS	DS
Skewness	0.33	-0.26	-1.36	-0.69	0.56	-0.31
Kurtosis	2.57	3.58	8.47	3.64	4.06	3.99
Mean	335.37	-219.57	-56.78	-223.23	0.04	-0.04
Standard deviation	211.09	157.18	34.79	83.50	0.04	0.04

Table 1. Score distribution parameters extracted from x-vector, i-vector and GMM-UBM systems using real speech data from previous studies [41], [45].

Following the consensus on the validity evaluation in FVC [26], C_{llr} was used as the main metric to access system validity (see Appendix for C_{llrMin} and C_{llrCal}). We used the mean C_{llr} across 100 replications for validity evaluation and overall range (OR; i.e., the difference between the maximum and minimum C_{llr} values across 100 replications) to access the reliability. In addition, we calculated 95% credible intervals (CI) of the LRs, often used in FVC [59], to measure the reliability of LR output. Unlike confidence intervals, credible intervals treat the boundaries (two intervals) as fixed variables while the estimated LR is treated as a random variable [61, 62]. The CI is the region of a posterior distribution within which one can be reasonably certain that the true LR value lies. Note that under the LR framework, any LR obtained is an estimate of the true unknown LR, and there is ongoing debate within the field about how best to measure reliability [5]. For both C_{llr} OR and 95% CI of the LR, the wider the values the less reliable the LR estimate. In a forensic context, it is crucial to consider the entire range of outputs a system can produce, as any extreme output has the potential to substantially impact the probability of a miscarriage of justice. This means that general performance indicators, such as mean LRs, are less informative for forensic purposes, where the focus is on understanding the potential for errors across all possible outcomes.

3. Results

Figure 2 shows the C_{llr} mean (validity; top panel) and range (reliability; bottom panel) across 100 replications as a function of sample size and sampling variability using scores from three generations of ASR systems, calibrated using logistic regression and the Bayesian model respectively. The x-axis indicates the number of speakers in training and test data respectively and y-axis gives the C_{llr}. Figure 3 shows the C_{llr} distribution across different systems, with a sample size of 20 training and test speakers, using logistic regression and Bayesian model for calibration respectively.

Predictably, regardless of calibration method, the mean C_{llr} values are the lowest using scores simulated from the x-vector (c. 0.21 – 0.25) system across all sample size conditions, followed by the i-vector (c. 0.29 – 0.36) and GMM-UBM systems (c. 0.42 – 0.46). This pattern indicates that the x-vector system yields the best overall validity compared to the i-vector and GMM-UBM systems across all sample size conditions and calibration methods. Nevertheless, a range of intriguing trends associated with reliability (C_{llr} OR and CI) emerge when comparing different systems that utilises different sample sizes.

For reliability evaluation, all three systems yield wide C_{llr} ORs across the 100 replications, especially when the sample size is small, as is the typical situation in real FVC cases. When using 20 to 30 training and test speakers and logistic regression for calibration, the more advanced i-vector (C_{llr} OR = 1.29) and x-vector (C_{llr} OR = 1.01) systems had larger C_{llr} OR than the less advanced GMM-UBM (C_{llr} OR = 0.70) system (Figure 2 bottom panel). Figure 3 shows that C_{llr} values from the i-vector and x-vector systems are higher than 1 for some replications when using 20 speakers, while the GMM-UBM system has C_{llr}s lower than 1 in all replications (see Appendix for C_{llr} distribution using different number of speakers). Similar patterns can be observed using 95% CI of LRs. Specifically, the 95% CI of LRs is 2.31 using logistic regression with 20 speakers in the GMM-UBM system (Table 2). This value increases to 3.88 and 2.7 in the i-vector and x-vector systems respectively. Both the OR and CI values indicate that there is much greater uncertainty in the validation output of the i-vector and x-vector systems despite better levels of average discrimination, compared with the GMM-UBM system. For all three systems, the C_{llr} OR decreases as sample size increases. After the inclusion of 40 speakers, the C_{llr} OR of the i-vector and x-vector systems is lower than that of the GMM-UBM system. Using 40 or 60 speakers, the i-vector system had a higher C_{llr} OR than the x-vector system; however, as the number of speakers increased to 70 and beyond, the i-vector system had a more stable C_{llr} OR than the x-vector system (Figure 2, Table 2). Similarly, 95% CI values in general decrease as sample size increases across all three systems.

A similar pattern is observed when employing small sample sizes and Bayesian model calibration, namely, the more advanced x-vector system in general exhibits a lower mean C_{llr} but a higher C_{llr} OR and 95% CIs compared to the less advanced GMM-UBM and i-vector systems, particularly with 20 and 30 training and test speakers. The x-vector system only yields a lower C_{llr} OR than that of the GMM-UBM system when 40 training and test speakers are used, but higher C_{llr} OR when 50 or more speakers are used. Additionally, when comparing the x-vector to the i-vector systems, the x-vector consistently yields higher C_{llr} OR values across various sample sizes. For the 95% CI of LRs, the x-vector system consistently yielded higher CI values than those of GMM-UBM and i-vector systems across all sample sizes.

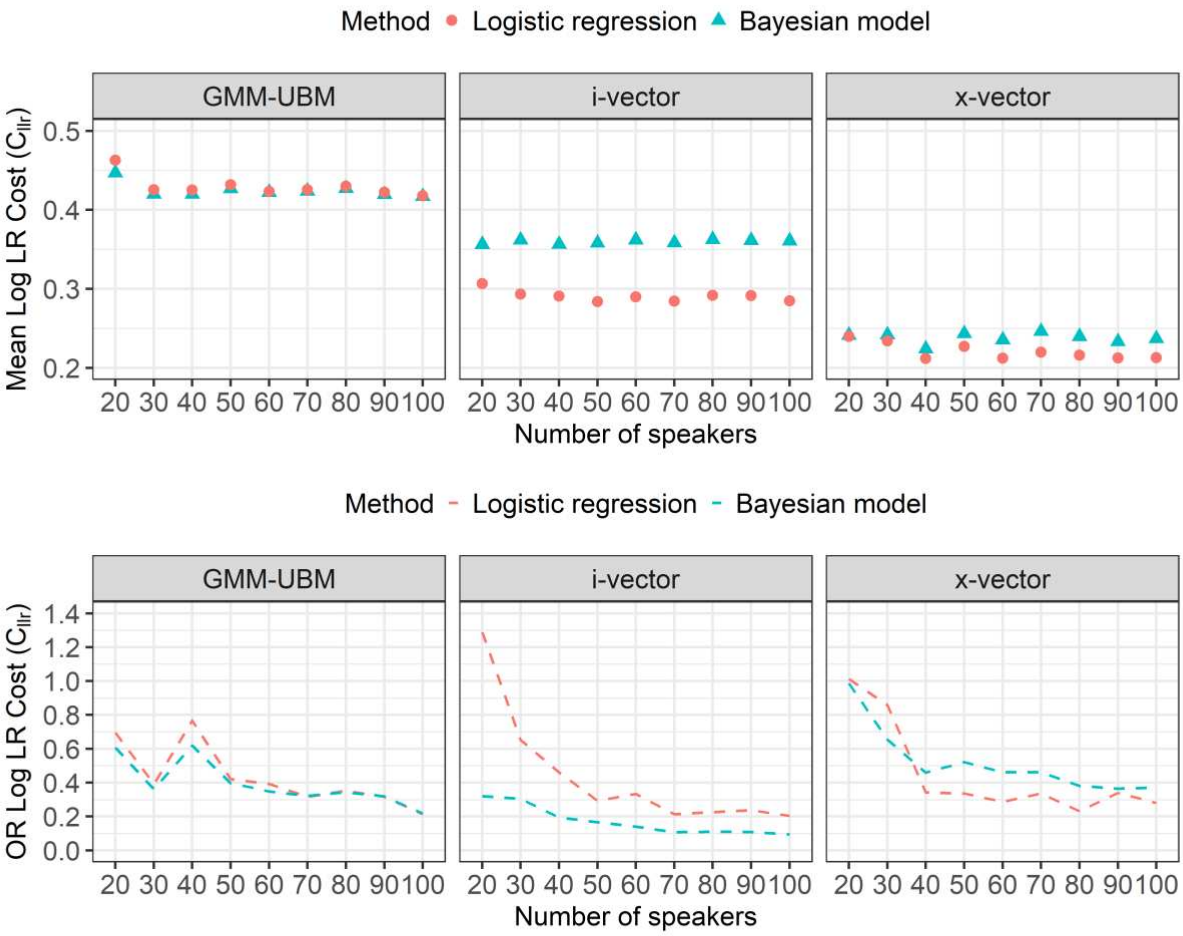


Figure 2. C_{llr} mean (top panel) and overall range (bottom panel) of three ASR systems using validation as a function of sample size and sampling variability across three generations of ASR systems.

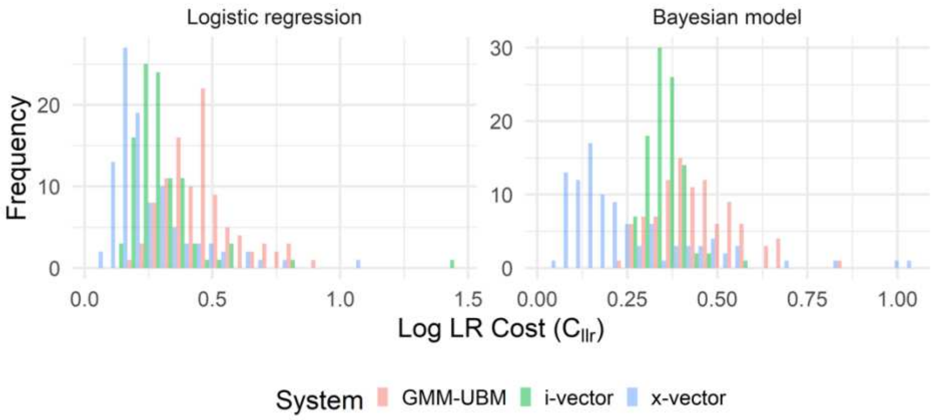


Figure 3. C_{llr} distribution across 100 replications using scores from three ASR systems and 20 training and test speakers.

Logistic regression									
	GMM - UBM			i-vector			x-vector		
Sample size	C_{llr} Mean	C_{llr} OR	± 95% CI	C_{llr} Mean	C_{llr} OR	± 95% CI	C_{llr} Mean	C_{llr} OR	± 95% CI
20	0.46	0.7	2.31	0.31	1.29	3.88	0.24	1.01	2.70
30	0.43	0.39	2.20	0.29	0.65	3.46	0.23	0.86	2.65
40	0.43	0.77	2.08	0.29	0.46	3.60	0.21	0.34	2.45
50	0.43	0.42	2.17	0.28	0.29	3.53	0.23	0.34	2.43
60	0.42	0.39	2.04	0.29	0.33	3.45	0.21	0.29	2.43
70	0.43	0.32	2.09	0.28	0.21	3.52	0.22	0.34	2.30
80	0.43	0.35	2.05	0.29	0.23	3.49	0.22	0.23	2.25
90	0.42	0.32	2.04	0.29	0.24	3.52	0.21	0.34	2.32
100	0.42	0.21	2.07	0.28	0.2	3.43	0.21	0.28	2.25

Bayesian model									
	GMM - UBM			i-vector			x-vector		
Sample size	C_{llr} Mean	C_{llr} OR	± 95% CI	C_{llr} Mean	C_{llr} OR	± 95% CI	C_{llr} Mean	C_{llr} OR	± 95% CI
20	0.45	0.61	1.91	0.36	0.32	1.70	0.24	0.99	2.90
30	0.42	0.36	1.93	0.36	0.3	1.70	0.24	0.65	3.08
40	0.42	0.62	1.95	0.36	0.19	1.73	0.22	0.46	3.03
50	0.43	0.4	1.97	0.36	0.17	1.71	0.24	0.52	3.06
60	0.42	0.35	1.94	0.36	0.14	1.70	0.24	0.46	3.11
70	0.42	0.32	1.97	0.36	0.11	1.71	0.25	0.46	3.11
80	0.43	0.34	1.96	0.36	0.11	1.71	0.24	0.38	3.07
90	0.42	0.32	1.96	0.36	0.11	1.72	0.23	0.36	3.11
100	0.42	0.22	1.97	0.36	0.09	1.71	0.24	0.37	3.11

Table 2. C_{llr} mean and overall range of three ASR systems as a function of sample size and sampling variability using Bayesian model for calibration.

4. Discussion

The debate about whether and how the reliability of LRs should be measured and reported to the courts is ongoing and controversial. As discussed in [60], some believe that the concept of measuring the reliability of LRs is not appropriate, while the others believe it is essential. The main contention lies in whether the reliability of LRs should be incorporated into the LR itself or reported separately using extra metrics (e.g., OR and CI used in the current paper). Nevertheless, it is crucial to acknowledge sources of potential variability [34], [36], [47], [59], [61], [62] and to establish methods to measure the variability, whether by incorporating it into LR itself or reporting it via supplementary metrics. The current study investigates sampling variability in the known source in relation to sample size and different generations of ASR systems.

The results show that system validity and reliability of three generations of ASR systems varies to different extents due to sampling variability and different sample size. As expected, the state-of-the-art, x-vector system yielded the best overall validity (i.e., lowest mean C_{llr}) compared to the i-vector and GMM-UBM systems. This is likely because the x-vector system captures more speaker specific information in the representation or getting additional discrimination benefit through PLDA over the GMM-UBM system or pre-trained DNNs over the i-vector system. Further, mean C_{llr} for each ASR system remains relatively stable across different sample sizes. This is likely because validity is dependent on the distance between the mean of SS and DS score distributions, which is consistent in the current study because scores were sampled from the same underlying distributions. However, our results show that better system validity and higher discriminability have the potential of leading to higher degree of uncertainty and inconsistency. Using 1 as an appropriate threshold for judging C_{llr} [63], the C_{llr} OR from three systems indicates that less advanced systems are preferable when the sample size is small. For the GMM-UBM and i-vector systems, the results show that Bayesian model should be the preferred calibration method as it reduces uncertainty and yields lower C_{llr} OR than that of the logistic regression across different number of training and test speakers. Surprisingly, the x-vector system produced a lower C_{llr} OR when calibrated with a Bayesian model using 20 to 30 training and test speakers, in contrast to using logistic regression for calibration. However, this pattern reversed with the inclusion of a larger pool of training and test speakers, where the Bayesian model calibration resulted in a higher C_{llr} OR compared to logistic regression calibration.

In real world FVC, we are often dealing with small sample sizes [38]. Based on the variability in C_{llr} values reported in the current study, it is suggested that researchers' aim in system validation should not be driven by obtaining better validity and higher discrimination, but better reliability or reducing the uncertainty, namely, a system producing reliable and consistent performance under various conditions (e.g., different number or/and configurations of speakers) should be preferred, other than a system (e.g., the state-of-the-art system) that has the potential of obtaining a very good validity (i.e., low C_{llr}) under one condition but not under other conditions.

The results also show that it is difficult to predict the direction of the trend in terms of the consistency of evaluation results. There is a general trend for reduced variability with larger samples, but of course we still see some random fluctuations, e.g., the OR for the GMM-UBM system is lower with 30 speakers than with 40 speakers, the OR for the x-vector system is lower with 60 speakers than with 70 speakers. However, the very question is "what is the *tolerable variation*?". This question cannot be addressed by one expert employing one comparison system (e.g., the state-of-the-art system) under one condition (e.g., one sample size) in one laboratory, but requires the cooperation and communication between different laboratories (nationally or/and internationally) and legal parties (e.g., police, jury, judge). It is essential to establish a *measurement process* and the *statistical control* [32] (i.e., to attain a certain degree of consistency) for reliability evaluation in the field of FVC. However, the challenges to reliability evaluation do not just apply to FVC but also any biometric evidence comparison are rooted in data protection schemes across different laboratories and the fact that conditions under which the crime scene data were collected cannot be replicated [31].

Within the field of FVC, there are publicly available corpora designed for the purpose of speaker comparison in various languages, for example English [64], [65], French [66], Spanish [67]. We advocate researchers and experts utilise these databases for cross-laboratory cooperation in reliability evaluation, establishing *measurement process* and the state of *statistical control* as well as the *tolerable variation*. This would be similar to those 'black-box' studies that are conducted in other areas of forensic evidence comparison [6], [8], [9], [10]. It is important to note that the comparison results obtained from any automatic system are coloured by everything involved in the measurement process, e.g., training and test data, statistical models, experts etc [31]. Although this sort of *statistical control* and *tolerable variation* established using speech corpora cannot fully represent the 'true' conditions under real case scenario, we could at least obtain a knowledge of the baseline validity and reliability of a given measurement process. We also propose researchers and experts using different subset and sample size of speakers for reliability evaluation through sampling of the sort described in the current study. Last but not least, the impact of sampling variability is likely underestimated in the current study because we simulated from test and training scores, thereby missing the effect of sampling variability in the reference data. Ideally, sampling from all three datasets would allow us to fully understand the impact of sampling variability. However, this would require a much larger database.

5. Conclusion

The current study simulated data from previous FVC studies to demonstrate the possible fluctuation in validation caused by sampling variability and sample size. It is worth noting that the motivation of the current paper was not a validation exercise or to compare the absolute validity across the three generations of ASR systems, but to demonstrate the potential fluctuations in comparison results of one system under different conditions. Providing such information under different conditions is critical for the trier-of-fact to evaluate the evidence provided by the expert. There must be a conceptual change in the validation exercise in FVC. The ultimate goal of the expert is to reduce the degree of uncertainty at every stage of the measurement process in a validation exercise, rather than maximising the discrimination. Focusing only on discrimination is likely to increase the probability of miscarriages of justice.

Declaration of competing interesting

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Disclosure instructions

None.

6. References

[1] CPD, “England & Wales Criminal Practice Directions.” 2015. Accessed: Jul. 29, 2021. [Online]. Available: https://www.justice.gov.uk/courts/procedure-rules/criminal/docs/2015/crim-practice-directions-V-evidence-2015.pdf

[2] CPS, “UK Crown Prosecution Service.” 2019. Accessed: Jul. 29, 2021. [Online]. Available: https://www.cps.gov.uk/legal-guidance/expert-evidence

[3] National Research Council, “Strengthening forensic science in the United States: A path forward.,” National Academies Press, 2009.

[4] President’s Council of Advisors on Science and Technology, “Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods,” President’s Council of Advisors on Science and Technology, 2016. [Online]. Available: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

[5] G. S. Morrison and E. Enzinger, “Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Conclusion,” *Speech Commun.*, vol. 112, pp. 37–39, Sep. 2019, doi: 10.1016/j.specom.2019.06.007.

[6] R. A. Hicklin *et al.*, “Accuracy and reproducibility of conclusions by forensic bloodstain pattern analysts,” *Forensic Sci. Int.*, vol. 325, p. 110856, Aug. 2021, doi: 10.1016/j.forsciint.2021.110856.

[7] K. L. Monson, E. D. Smith, and E. M. Peters, “Accuracy of comparison decisions by forensic firearms examiners,” *J. Forensic Sci.*, vol. 68, no. 1, pp. 86–100, Jan. 2023, doi: 10.1111/1556-4029.15152.

[8] R. Austin Hicklin *et al.*, “Accuracy, reproducibility, and repeatability of forensic footwear examiner decisions,” *Forensic Sci. Int.*, vol. 339, p. 111418, Oct. 2022, doi: 10.1016/j.forsciint.2022.111418.

[9] H. Arora, N. Kaplan-Damary, and H. S. Stern, “Combining reproducibility and repeatability studies with applications in forensic science,” *Law Probab. Risk*, p. mgad007, Oct. 2023, doi: 10.1093/lpr/mgad007.

[10] B. T. Ulery, R. A. Hicklin, J. Buscaglia, and M. A. Roberts, “Accuracy and reliability of forensic latent fingerprint decisions,” *Proc. Natl. Acad. Sci.*, vol. 108, no. 19, pp. 7733–7738, May 2011, doi: 10.1073/pnas.1018707108.

[11] M. Jessen, “Forensic Phonetics,” *Lang. Linguist. Compass*, vol. 2, no. 4, Art. no. 4, Jul. 2008, doi: 10.1111/j.1749-818X.2008.00066.x.

[12] M. J. Saks and J. J. Koehler, “The Coming Paradigm Shift in Forensic Identification Science,” *Science*, vol. 309, no. 5736, pp. 892–895, Aug. 2005, doi: 10.1126/science.1111565.

[13] E. Enzinger, “Likelihood Ratio Calculation in Acoustic-Phonetic Forensic Voice Comparison: Comparison of Three Statistical Modelling Approaches,” presented at the Interspeech 2016, Sep. 2016, pp. 535–539. doi: 10.21437/Interspeech.2016-1611.

[14] V. Hughes, S. Wood, and P. Foulkes, “Strength of forensic voice comparison evidence from the acoustics of filled pauses,” *Int. J. Speech Lang. Law*, vol. 23, no. 1, pp. 99–132, Jun. 2016, doi: 10.1558/ijsll.v23i1.29874.

[15] P. Rose and X. Wang, “Cantonese forensic voice comparison with higher-level features: likelihood ratio-based validation using F-pattern and tonal F0 trajectories over a disyllabic hexaphone,” presented at the Odyssey 2016, Jun. 2016, pp. 326–333. doi: 10.21437/Odyssey.2016-47.

[16] C. Zhang, G. S. Morrison, and T. Thiruvaran, “FORENSIC VOICE COMPARISON USING CHINESE /iau/,” *ICPHS Hong Kong*, pp. 2280–2283, 2011.

[17] B. X. Wang and V. Hughes, “Reducing uncertainty at the score-to-LR stage in likelihood ratio-based forensic voice comparison using automatic speaker recognition systems,” in *Interspeech 2022*, ISCA, Sep. 2022, pp. 5243–5247. doi: 10.21437/Interspeech.2022-518.

[18] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, and T. Niemi, “Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition,” 2015.

[19] B. Robertson, G. A. Vignaux, and C. E. H. Berger, *Interpreting evidence: evaluating forensic science in the courtroom*, Second edition. Chichester, West Sussex, UK ; Hoboken: John Wiley and Sons, Inc, 2016.

[20] 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.

[21] E. Gold and P. French, “International practices in forensic speaker comparisons: second survey,” *Int. J. Speech Lang. Law*, vol. 26, no. 1, Art. no. 1, Jun. 2019, doi: 10.1558/ijsll.38028.

[22] E. Gold and P. French, “International Practices in Forensic Speaker Comparison,” *Int. J. Speech Lang. Law*, vol. 18, no. 2, Art. no. 2, Nov. 2011, doi: 10.1558/ijsll.v18i2.293.

[23] D. Van Der Vloed, “Data strategies in forensic automatic speaker comparison,” *Forensic Sci. Int.*, vol. 350, p. 111790, Sep. 2023, doi: 10.1016/j.forsciint.2023.111790.

[24] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-Vectors: Robust DNN Embeddings for Speaker Recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB: IEEE, Apr. 2018, pp. 5329–5333. doi: 10.1109/ICASSP.2018.8461375.

[25] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia, “Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition,” *IEEE Trans. AUDIO*, vol. 15, no. 7, Art. no. 7, 2007.

[26] G. Morrison *et al.*, “Consensus on validation of forensic voice comparison,” *Sci. Justice*, vol. 61, no. 3, Art. no. 3, Mar. 2021, doi: 10.1016/j.scijus.2021.02.002.

[27] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digit. Signal Process.*, vol. 10, no. 1–3, pp. 19–41, Jan. 2000, doi: 10.1006/dspr.1999.0361.

[28] J. Franco-Pedroso and J. Gonzalez-Rodriguez, “Linguistically-constrained formant-based i-vectors for automatic speaker recognition,” *Speech Commun.*, vol. 76, pp. 61–81, Feb. 2016, doi: 10.1016/j.specom.2015.11.002.

[29] A. Silnova *et al.*, “Analyzing speaker verification embedding extractors and back-ends under language and channel mismatch,” Mar. 19, 2022, *arXiv*: arXiv:2203.10300. Accessed: Aug. 15, 2024. [Online]. Available: http://arxiv.org/abs/2203.10300

[30] C. Eisenhart, “Realistic evaluation of the precision and accuracy of instrument calibration systems,” *J. Res. Natl. Bur. Stand.*, no. 67, pp. 161–187, 1963.

[31] J. L. Wayman, A. Possolo, and A. J. Mansfield, “Modern statistical and philosophical framework for uncertainty assessment in biometric performance testing,” *IET Biom.*, vol. 2, no. 3, pp. 85–96, 2013, doi: 10.1049/iet-bmt.2013.0009.

[32] W. A. Shewhart, *Statistical method from the viewpoint of quality control*. 1939.

[33] 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, Art. no. 6, Dec. 2015, doi: 10.1016/j.scijus.2015.05.003.

[34] P. Rose, *Forensic Speaker Identification*. London: Taylor & Francis, 2002.

[35] 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, Art. no. 1, Aug. 2019, doi: 10.1558/ijsll.38046.

[36] J.-A. Bright, K. E. Stevenson, J. M. Curran, and J. S. Buckleton, “The variability in likelihood ratios due to different mechanisms,” *Forensic Sci. Int. Genet.*, vol. 14, pp. 187–190, Jan. 2015, doi: 10.1016/j.fsigen.2014.10.013.

[37] 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.

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

[39] M. I. Mandasari, M. McLaren, and D. A. Van Leeuwen, “The effect of noise on modern automatic speaker recognition systems,” in *2012 IEEE International Conference on Acoustics,*

Speech and Signal Processing (ICASSP), Kyoto, Japan: IEEE, Mar. 2012, pp. 4249–4252. doi: 10.1109/ICASSP.2012.6288857.

[40] X. B. Wang, "The effect of sampling variability on overall performance and individual speakers' behaviour in likelihood ratio-based forensic voice comparison", Doctoral Dissertation. University of York, UK, 2021.

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

[42] M. Jessen, J. Bortlík, P. Schwarz, and Y. A. Solewicz, "Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)," *Speech Commun.*, vol. 111, pp. 22–28, Aug. 2019, doi: 10.1016/j.specom.2019.05.002.

[43] 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.

[44] D. G. Da Silva and C. A. Medina, "Evaluation of MSR Identity Toolbox under conditions reflecting those of a real forensic case (forensic_eval_01)," *Speech Commun.*, vol. 94, pp. 42–49, Nov. 2017, doi: 10.1016/j.specom.2017.09.001.

[45] G. S. Morrison, "Bi-Gaussianized calibration of likelihood ratios," *Law Probab. Risk*, vol. 23, no. 1, p. mgae004, Jan. 2024, doi: 10.1093/lpr/mgae004.

[46] P. Weber *et al.*, "Validations of an alpha version of the E3 Forensic Speech Science System (E3FS3) core software tools," *Forensic Sci. Int. Synergy*, vol. 4, p. 100223, Jan. 2022, doi: 10.1016/j.fsisyn.2022.100223.

[47] 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, Art. no. 1, Jan. 2016, doi: 10.1016/j.scijus.2015.06.005.

[48] G. S. Morrison, E. Enzinger, D. Ramos, J. González-Rodríguez, and A. Lozano-Díez, "Statistical models in forensic voice comparison," in *Handbook of Forensic Statistics*, Boca Raton, FL: CRC., 2019, p. 78.

[49] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB: IEEE, Apr. 2018, pp. 5329–5333. doi: 10.1109/ICASSP.2018.8461375.

[50] 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.

[51] R Core Team, *R: A language and environment for statistical computing*. (2023). [Online]. Available: https://www.R-project.org/

[52] A. Azzalini, *The R package "sn": The Skew-Normal and Related Distributions such as the Skew-t*. (2020).

[53] 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.

[54] 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.

[55] 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.

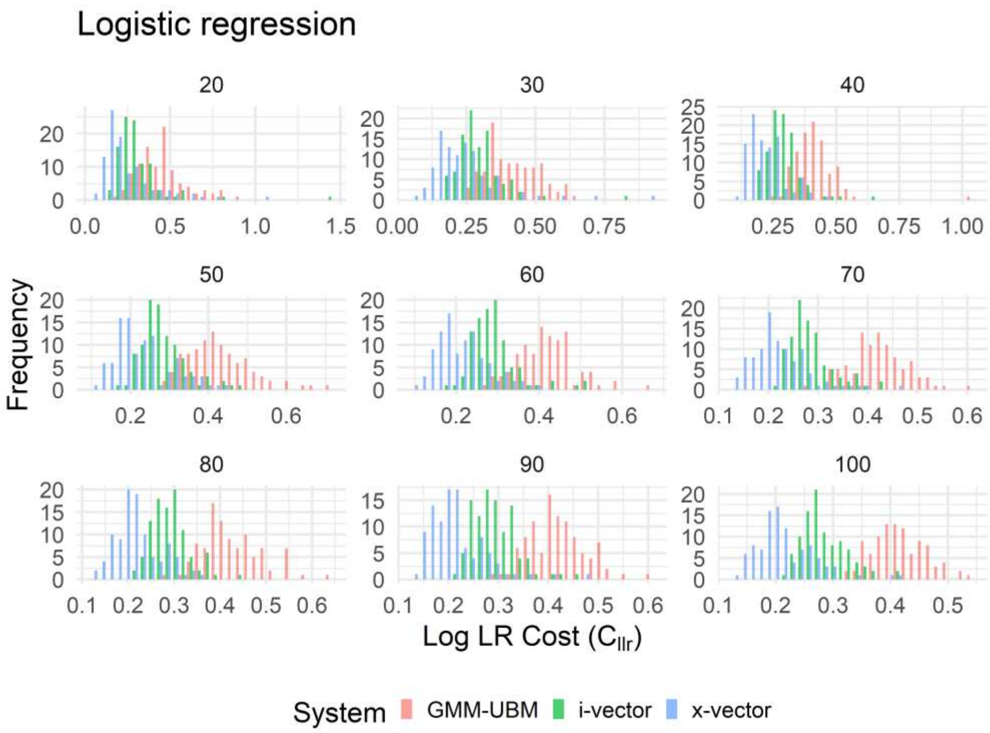
[56] P. Rose and C. Zhang, "Conversational Style Mismatch: its Effect on the Evidential Strength of Long- term F0 in Forensic Voice Comparison," in *Proc. 17th Australasian Int'l conf. on Speech Science and Technology*, Sydney, 2018, pp. 157–160.

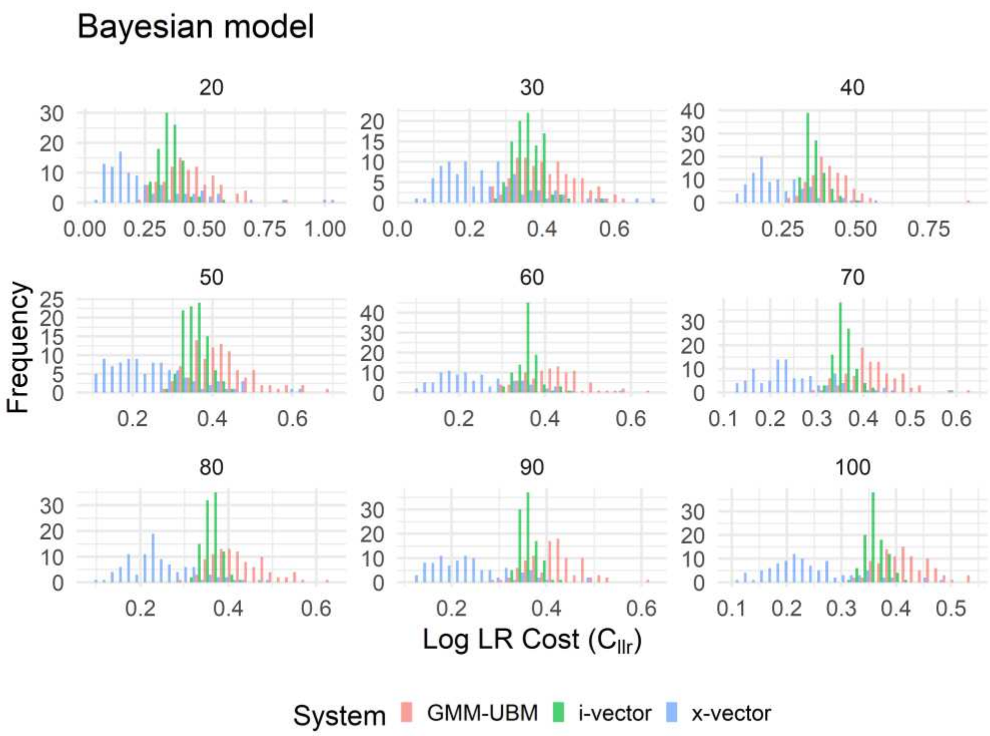
[57] N. Brümmer and A. Swart, “Bayesian Calibration for Forensic Evidence Reporting,” in *Interspeech*, Singapore, 2014, pp. 388–392.
[58] C. Watson, C. Wilson, M. Indovina, and B. Cochran, “Two finger matching with vendor SDK matchers,” National Institute of Standards and Technology, Gaithersburg, MD, NIST IR 7249, 2005. doi: 10.6028/NIST.IR.7249.
[59] G. S. Morrison, “Measuring the validity and reliability of forensic likelihood-ratio systems,” *Sci. Justice*, vol. 51, no. 3, Art. no. 3, Sep. 2011, doi: 10.1016/j.scijus.2011.03.002.
[60] G. S. Morrison, “Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate,” *Sci. Justice*, vol. 56, no. 5, pp. 371–373, Sep. 2016, doi: 10.1016/j.scijus.2016.05.002.
[61] J. M. Curran, J. S. Buckleton, C. M. Triggs, and B. S. Weir, “Assessing uncertainty in DNA evidence caused by sampling effects,” *Sci. Justice*, vol. 42, no. 1, pp. 29–37, Jan. 2002, doi: 10.1016/S1355-0306(02)71794-2.
[62] J. M. Curran, “An introduction to Bayesian credible intervals for sampling error in DNA profiles,” *Law Probab. Risk*, vol. 4, no. 1–2, Art. no. 1–2, Mar. 2005, doi: 10.1093/lpr/mgi009.
[63] 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.
[64] E. Gold, S. Ross, and K. Earnshaw, “The ‘West Yorkshire Regional English Database’: Investigations into the Generalizability of Reference Populations for Forensic Speaker Comparison Casework,” in *Interspeech 2018*, ISCA, Sep. 2018, pp. 2748–2752. doi: 10.21437/Interspeech.2018-65.
[65] F. Nolan, K. McDougall, G. De Jong, and T. Hudson, “The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research,” *Int. J. Speech Lang. Law*, vol. 16, no. 1, Art. no. 1, Sep. 2009, doi: 10.1558/ijsll.v16i1.31.
[66] M. Ajili, J.-F. Bonastre, J. Kahn, S. Rossato, and G. Bernard, “FABIOLE, a Speech Database for Forensic Speaker Comparison,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 726–733. Accessed: Jan. 22, 2024. [Online]. Available: https://aclanthology.org/L16-1115
[67] E. S. Segundo, H. Alves, and M. F. Trinidad, “CIVIL Corpus: Voice Quality for Speaker Forensic Comparison,” *Procedia - Soc. Behav. Sci.*, vol. 95, pp. 587–593, Oct. 2013, doi: 10.1016/j.sbspro.2013.10.686.
[68] 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.

Appendix

C_{llr} distribution across different systems and number of speakers

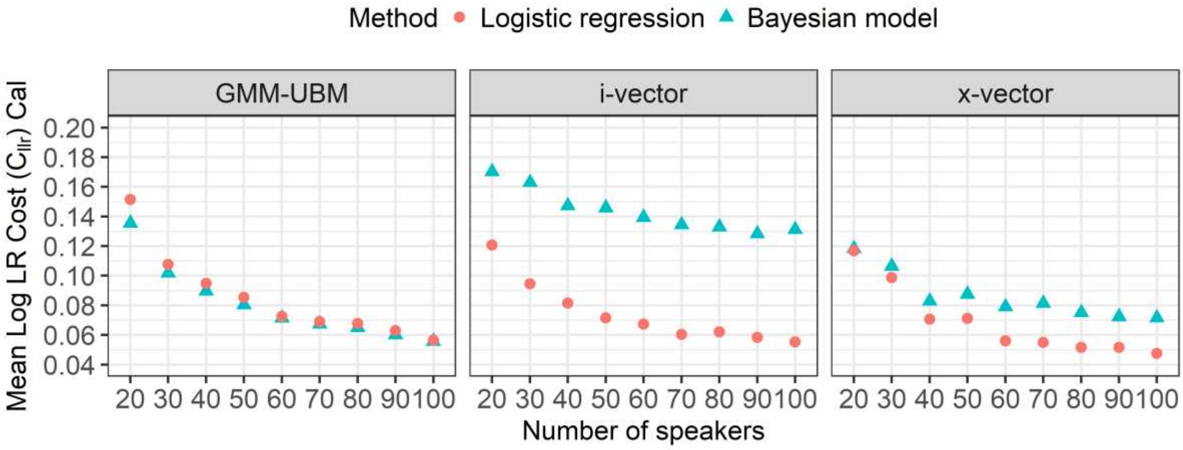
C_{llr} distribution across different systems and number of speakers using logistic regression and Bayesian model for calibration respectively. Numbers on top of each panel (e.g., 20, 30, 40) indicate the number of training and test speakers.

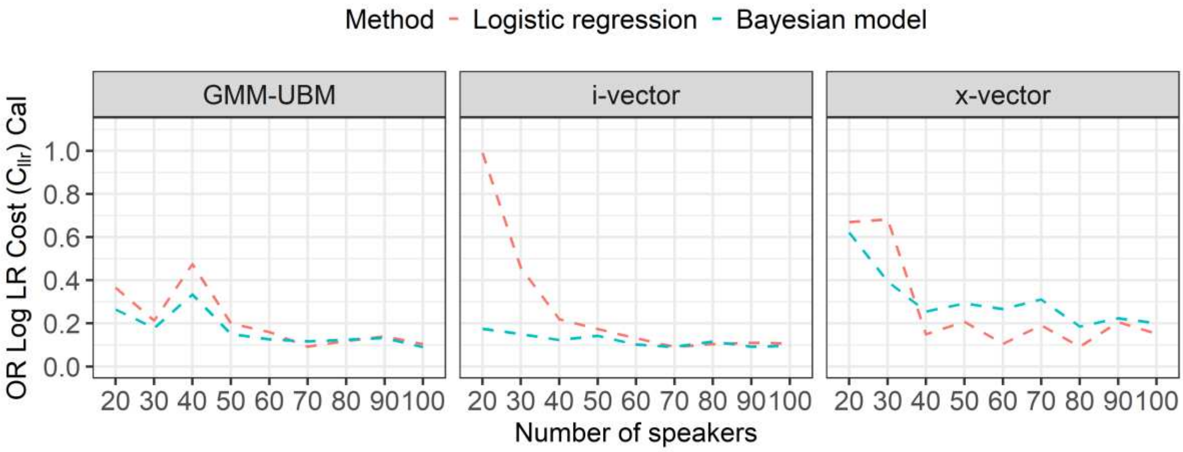




Mean and OR of C_{llrCal}

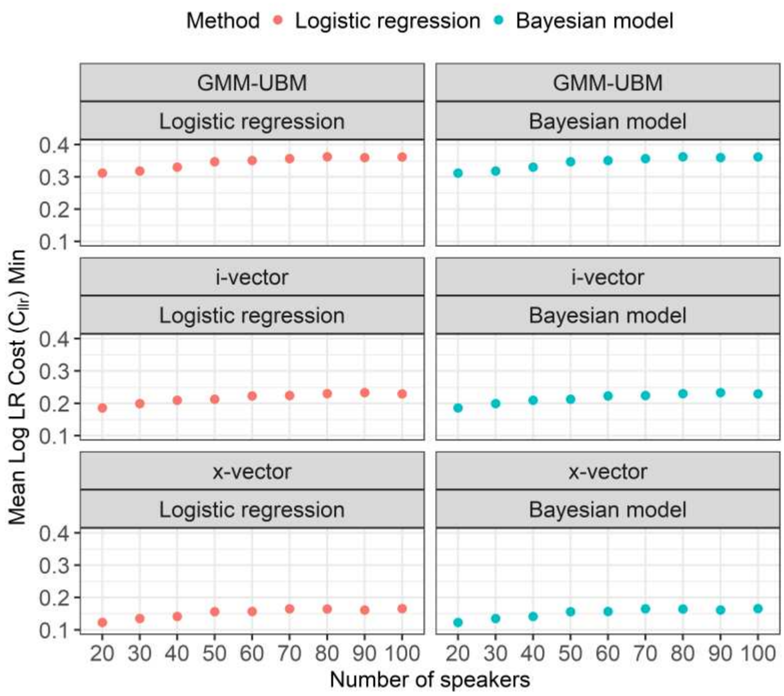
C_{llrCal} calculates the calibration loss, representing how well the likelihood ratios are calculated to reflect the true probabilities. Mean and OR C_{llrCal} plots show Bayesian model has lower mean C_{llrCal} than logistic regression for GMM-UBM system (top panel), but higher for i-vector and x-vector systems across all sample sizes (i.e., number of speakers). Meanwhile, the OR of C_{llrCal} using Bayesian model calibration is lower than those using logistic regression when sample size is small (lower panel), i.e., between 20 and 40 speakers in the training and test data respectively.

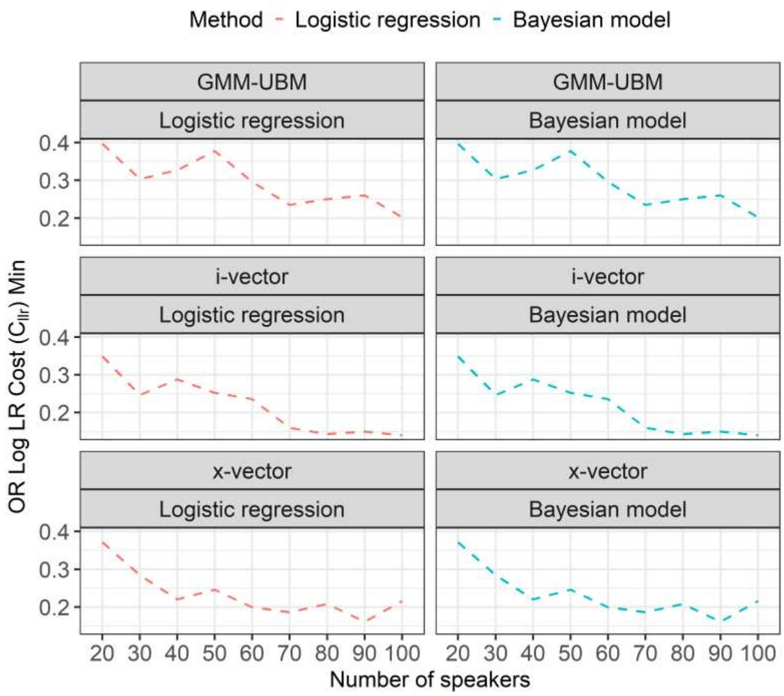




Mean and OR of C_{llrMin}

Mean and OR C_{llrMin} values for different systems using logistic regression and Bayesian model calibration. C_{llrMin} represents the discrimination performance. Note that the mean and OR C_{llrMin} values are identical for the same system regardless of calibration method. This is because calibration does not make a difference in the C_{llrMin} values.





Bayesian model calibration

The Bayesian calibration model employs hyperparameters to shrink LRs when uncertainty is high [57], [68]. The model is estimated using training scores, and the likelihood of obtaining this model is evaluated using test scores [57]. Further, the prior belief and the strength of the belief for the mean and variance of the training scores need to be specified, and the uninformative priors (Jeffreys prior) are often used in FVC. A simplified Bayesian model estimation formula is shown in Equation 1.

Bayesian model (with Jeffreys reference) :

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

Where *t* is the *t* distribution, *n is the* number of speakers, *x* is the test score, $\hat{\mu}$ and $\hat{\sigma}^2$ are the mean and variance of the training score. The calculation of Bayes factors is the ratio between the likelihood of the Bayesian models evaluated using test scores is shown in Equation 2.

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

(Equation 2)

To reduce the extent of non-monotonicity, we followed [68] using the pooled sample variance ($\hat{\sigma}^2$), rather than the variance of same-speaker ($\hat{\sigma}_{ss}^2$) and different-speaker ($\hat{\sigma}_{ds}^2$) comparisons individually . Meanwhile, the degrees of freedom ($n_{ss}+n_{ds}-2$) need to be adjusted to take the pooled variance calculation into consideration. Equation 2 can be modified to Equation 3,

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

(Equation 3)