Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Estimating the Precision of the Likelihood-Ratio Output of a Forensic-Voice-Comparison System

Geoffrey Stewart Morrison[1,2], Tharmarajah Thiruvaran[2], and Julien Epps[2,3]

[1]School of Language Studies, Australian National University, Canberra, ACT 0200, Australia
[2]School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia
[3]National ICT Australia (NICTA), Australian Technology Park, Sydney, NSW 1430, Australia
geoff.morrison@anu.edu.au, thiruvaran@student.unsw.edu.au, j.epps@unsw.edu.au

Abstract

The issues of validity and reliability are important in forensic science. Within the likelihood-ratio framework for the evaluation of forensic evidence, the log-likelihood-ratio cost (C_{llr}) has been applied as an appropriate metric for evaluating the accuracy of the output of a forensic-voice-comparison system, but there has been little research on developing a quantitative metric of precision. The present paper describes two procedures for estimating the precision of the output of a forensic-comparison system, a non-parametric estimate and a parametric estimate of its 95% credible interval. The procedures are applied to estimate the precision of a basic automatic forensic-voice-comparison system presented with different amounts of questioned-speaker data. The importance of considering precision is discussed.

1. Introduction

1.1. Concern about accuracy and precision in forensic science

Recently there has been a great deal of concern in forensic science about validity and reliability [1–4]. The National Research Council report to Congress on Strengthening Forensic Science in the United States [3] urged that procedures be adopted which include "the reporting of a measurement with an interval that has a high probability of containing the true value; . . . [and] the conducting of validation studies of the performance of a forensic procedure" (p. 121); the latter requiring the use of "quantifiable measures of the reliability and accuracy of forensic analyses" (p. 23).

1.2. Accuracy

In statistics and scientific literature *validity* is synonymous with *accuracy* and *reliability* with *precision*; however, in judicial and forensic-science literature reliability has often been discussed without explicit definition, or has been defined in terms of a measure of validity: *classification-error rates*, i.e., the proportion of same-origin comparisons in a test set which are classified as different-origin (misses), and the proportion of different-origin comparisons which are classified as same-origin (false alarms).

If one accepts that the *likelihood-ratio framework* is the correct framework for the evaluation of forensic comparison evidence [5–18], then a metric such as classification-error rate, based on a hard-thresholding of posterior probabilities, is not an appropriate measure of accuracy (by extension, this is also true for equal error rate, EER). Rather, an appropriate metric should be based on likelihood ratios (LRs) and should be continuous in nature – an LR which provides greater support for a contrary-to-fact hypothesis should attract a heavier penalty than one which provides more limited support for the contrary-to-fact hypothesis, since the former has a greater potential to contribute to a miscarriage of justice. An appropriate measure of accuracy, developed for use in automatic speaker recognition [19, 20] and subsequently applied in forensic voice comparison, e.g., [14, 21], is the *log-likelihood-ratio cost* (C_{llr}), which, at least in automatic speaker recognition, may now be considered a standard metric of the accuracy of a system which outputs LRs.

1.3. Precision

In addition to accuracy, however, it is also important to consider precision [22, 23]. Imagine two systems that are assessed as having the same accuracy and when tested on a particular pair of objects multiple times give the same average $\log_{10}$(LR) of −2, but the test results on one system have a wide range of LR output values leading to an estimated 95% credible interval, in $\log_{10}$(LR), of ±0.1 whereas the other has an estimated 95% credible interval of ±3. The former system (with a 95% LR credible interval for this pair of objects ranging from 79 to 126 in favor of the different-origin hypothesis) would be preferred over the latter (with a 95% LR credible interval for this pair of objects ranging from 100 000 in favor of the different-origin hypothesis to 10 in favor of the same-origin hypothesis). The former, more precise, system would be much more useful in assisting the trier of fact to weigh the forensic-comparison evidence as part of making their ultimate decision as to the guilt or innocence of the accused (the *trier of fact* is the judge, the panel of judges, or the jury, depending on the legal system).

The present paper describes and provides examples of the use of two procedures for calculating a metric of the precision of the LR output of a forensic comparison system. The metric is an estimate of the *95% credible interval* (CI) [24]. One procedure is non-parametric and the other parametric, and the examples are of their application to an automatic forensic-voice-comparison system. The aim of developing this metric is to allow forensic scientists to compare developmental systems, and to allow them to report the precision, as well as the accuracy, of the final system so that a judge can consider whether testimony based on the system should be admitted in court [1]. Finally, as part of their testimony, it would allow a forensic scientist to make a statement such as the following:

> Based on my evaluation of the evidence, I have calculated that one would be *X* times more likely to obtain the acoustic differences between the voice samples if the questioned-voice sample had been

produced by the accused than if it had been produced by someone other than the accused. Based on my calculations, I am 95% certain that it is at least X_{lower} times more likely and not more than X_{upper} times more likely.

1.4. Precision at the activity level

The underlying idea is that there is a true LR for the comparison of a pair of speakers (at least for the specified speaking styles), and each LR which is calculated on the basis of a pair of samples from the two speakers is an estimate of the true LR. If we take multiple non-overlapping pairs of voice samples from a pair of speakers we can calculate an LR estimate from each pair of samples. According to the central-limit theorem, if we take the mean of all our estimates this is our best estimate of the true LR. We can also look at the variance of our individual estimates around the best estimate.

When it comes to the actual suspect and offender data we only have two voice samples and the LR we calculate for this pair of samples is all we have. We could run the same system several times on this pair and measure the variability due to any imprecision in measurement and statistical modeling; however, in addition to the latter, what we are interested in and trying to estimate here is what would our estimate of the variability be if we could obtain multiple additional estimates of the LR for the suspect and offender using additional voice samples from this pair of speakers (we cannot in practice obtain multiple samples from the offender because we don't know who the offender is). In calculating and presenting such an estimate of precision, we have shifted from addressing the same-speaker versus different-speaker propositions purely at the *source level* and are now addressing the *activity level* [25]. We are doing this because what the court cares about is what this evidence and our forensic expertise can tell them about the speakers. Our forensic expertise, based on tests of our forensic-voice-comparison system, tells us that multiple tests on the same pair of speakers will result in a range of LR estimates, therefore when we only have one LR estimate there is a degree of uncertainty as to how representative it is of this pair of speakers. It is therefore our duty as forensic scientists to inform the court of this degree of uncertainty. To take an extreme example, imagine that tests of our system on one pair of voice samples from a given pair of speakers we obtained an LR of one million in favor of the same-speaker hypothesis, but on another pair of voice samples from the same pair of speakers we obtained an LR of one million in favor of the different-speaker hypothesis, then in a court case the comparison of the suspect and offender samples resulted in an LR of one million in favor of the same-speaker hypothesis. It would be very misleading to the court if we were only to report the latter result and not also report the tests of the reliability of our system.

1.5. Black-box approach

Forensic-voice-comparison systems typically have several stages involving selection of the portion of the speech signal to measure, the measurement of acoustic properties, score calculation, and calibration and fusion. It may be possible to apply precision analyses to each stage, but combining these to form an analytic solution for the whole system would seem to be an intractable problem, and such a solution developed for one system would not be immediately transferrable to a different system based on different acoustic measurements or different modeling techniques, etc. We therefore treat the system as a black box and in a test situation simply compare the output of the black box with what we know about the input. This allows us to apply the same precision-measurement procedure to systems with very different architectures, e.g. an acoustic-phonetic system and an automatic system [27].

2. Calculation of Precision

2.1. Calculation of sets of independent likelihood ratios

Assume that one has a test database containing a large number of speakers and four non-contemporaneous recordings of the voice of each speaker, labeled *A*, *B*, *C*, and *D*. A larger number of recordings per speaker could be used, but four recordings per speaker is the minimum necessary for the CI estimate to be based on LRs calculated from both same-speaker and different-speaker comparisons; two is the minimum necessary to estimate the CI from different-speaker comparisons only.

For each possible same-speaker comparison in the test database, a suspect (known-speaker) model can be constructed using data from recording *A*, and data from recording *B* can be used as offender (questioned-speaker) data (probe data) to generate an LR. A second LR for the same same-speaker comparison can be calculated using *C* to create a suspect model with *D* used as offender data (see Table 1). This results in two LR estimates of the strength of evidence for each same-speaker comparison calculated using *independent*, i.e., *non-overlapping*, pairs of test data.

Similarly, for each possible different-speaker comparison in the test database, each speaker's *A* recording is used to create a suspect model and the other speaker's *B* data is used as offender data, and each speaker's *C* is used to create a suspect model with the other speaker's *D* used as offender data (see Table 2). This results in four LR estimates for each different-speaker comparison calculated using *independent*, i.e., *non-overlapping*, pairs of test data.

Table 1: Same-speaker comparison pairs

Suspect model	Recording	Offender data	Recording
001	*A*	001	*B*
001	*C*	001	*D*
002	*A*	002	*B*
002	*C*	002	*D*
:	:	:	:

Table 2: Different-speaker comparison pairs

Suspect model	Recording	Offender data	Recording
001	*A*	002	*B*
001	*C*	002	*D*
001	*A*	003	*B*
001	*C*	003	*D*
:	:	:	:
002	*A*	001	*B*
002	*C*	001	*D*
002	*A*	003	*B*
002	*C*	003	*D*
:	:	:	:

Given two independent estimates for each same-speaker comparison and four independent estimates for each different-speaker comparison, a pooled within-group (within-comparison) sample variance can be estimated.

For a database with N speakers, given two independent estimates for each same-speaker comparison and four independent estimates for each different-speaker comparison, there are a total of $2N$ same-speaker comparisons and $4((N^2-N)/2)$, i.e., $2N(N-1)$, different-speaker comparisons.

2.2. Non-parametric procedure for the calculation of a credible interval

Earlier experimental results indicate that the distribution of different-speaker log LRs generated by acoustic-phonetic forensic-voice-comparison systems may be non-normal and heteroscedastic [26, 27]. Rather than estimating the CI via a parametric estimate of the sample variance, we therefore first adopt a non-parametric procedure which finds the boundary between the most outlying α data points and the $1-\alpha$ least outlying. The procedure is as described below, and Matlab® scripts and functions are provided on the first author's website <http://geoff-morrison.net>. All calculations are carried out using log-LR values.

For each same-speaker and different-speaker comparison, i, first calculate the within-comparison mean, $\bar{x}_i$, of the individual log-LR estimates, x_{ij}:

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \quad (1)$$

where n_i is the number of log-LR estimates calculated for comparison i (herein two for same-speaker comparisons, e.g., 001*A*–001*B* and 001*C*–001*D* , and four for different-speaker comparisons, e.g., 001*A*–002*B*, 001*C*–002*D*, 002*A*–001*B*, and 002*C*–001*D*), and x_{ij} is the jth LR estimate of comparison i.

Next, calculate the deviation-from-mean value, y_{ij}, of each log-LR estimate, x_{ij}:

$$y_{ij} = x_{ij} - \bar{x}_i \quad (2)$$

Then estimate the credible interval using a procedure based on *local linear regression* with a *nearest-neighbor kernel* [28, §6.1.1] (to calculate a 95% credible interval, set $\alpha = 0.05$):

1. For the value x_0 at which one wishes to estimate the credible interval, find its k nearest neighbors among the $\bar{x}_i$. Designate this group of i values as K, and the number of y_{ij} $\{i \in K\}$ data points as n_k.

2. Set $m = n_k$. Set $M = K$.

3. For all $i \in M$, fit a linear regression of $|y_{ij}|$ on $\bar{x}_i$. (The use of within-comparison means, $\bar{x}_i$, and the absolute deviation-from-mean values, $|y_{ij}|$, implies an assumption that the distribution is symmetrical.)

4. If $m = \lfloor 2n_k\alpha \rceil$, go to step 8.

5. For all $i \in M$, calculate the signed residuals, ε_{ij}, between each observed value, $|y_{ij}|$, and its corresponding value, $\hat{y}_{ij}$, estimated from the linear regression.

6. If $\lfloor 3n_k\alpha \rceil > m > \lfloor 2n_k\alpha \rceil$, discard the $m - \lfloor 2n_k\alpha \rceil$ data points with the most negative ε_{ij} values, leaving m data points M. Else discard the $\lfloor n_k\alpha \rceil$ data points with the most negative ε_{ij} values, leaving m data points M.

7. Repeat steps 3 through 6.

8. Use the estimated coefficient values from the linear regression at the last iteration to calculate the estimated value $\hat{y}_0$ at x_0, and use this to calculate the estimated value of the CI at x_0: $\text{CI} = x_0 \pm \hat{y}_0$.

Figure 1 provides a graphical representation of the non-parametric procedure for the calculation of the 95% CI for a $\log_{10}$(LR) value of x_0 = +2 (LR of 100 in favor of the same-speaker hypothesis). The example makes use of the system described below (§3) using 40 s of questioned-speaker data. The dots in Figure 1 show the 500 nearest neighbors to x_0. The straight green lines show the fits of the successive linear regressions (the lowest line is from the first iteration and the highest line from the last). The blue dots are the 90% of data points discarded and the red dots are the 10% of data points remaining at the last iteration – the last linear regression is fitted to this last 10% of the data. The triangle shows the estimated y_0 value of 1.166. The estimated 95% $\log_{10}$ LR CI at this point is therefore 2±1.166, or a 95% LR CI ranging from 6.82 to 1 452 in favor of the same-speaker hypothesis.

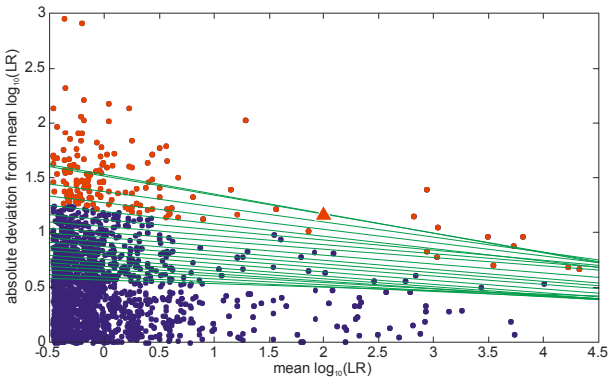

Figure 1: Graphical representation of the non-parametric procedure for the estimation of a 95% CI.

2.3. Parametric procedure for the calculation of a credible interval

If homoscedasticity and normality can be assumed, then the CI can be estimated using the t distribution of the pooled-within-group posterior standard deviation of the x values (σ') using degrees of freedom (df) equal to the total number of LR estimates minus the total number of speaker-comparisons [24]:

$$\text{CI} = \pm t_{1-\frac{\alpha}{2},df}\sigma' \quad (3)$$

$$df = \sum_i (n_i - 1) \quad (4)$$

In principle, the posterior standard deviation (σ') is calculated using the prior standard deviation (σ) and the sample standard deviation ($\hat{\sigma}$). In practice we will use flat priors, hence in Equation 3 we simply substitute $\hat{\sigma}$ for σ' and our estimate of the CI will be based only on the sample variance:

$$\hat{\sigma}^2 = \frac{1}{df}\sum_i \left(\sum_{j=1}^{n_i}(\bar{x}_i - x_{ij})^2\right) \quad (5)$$

Note that we have used the unbiased least-squares estimate of $\hat{\sigma}^2$. If the biased maximum-likelihood estimate were used, i.e., using $\sum_i n_i$ in place of df in Equation 5, the estimated CI would be narrower.

3. Experimental Methodology

3.1. Databases

Usually in forensic voice comparison, the language and dialect spoken in the questioned-voice recording can be determined without being disputed by either the prosecution or defense. The universal background model (UBM) should be representative of the potential population of offenders, and should therefore match the language and dialect of the questioned-voice recordings. For expedience, the present study made use of recordings which were labeled in their source databases as US English. In constructing a real forensic voice comparison system one would have to be more specific about the dialect in question and would have to verify that each recording was a match for that dialect.

All training, calibration, and test data were from telephone recordings of speakers labeled in their respective source databases as adult male US English speakers (although the data used to train the UBM may have contained some recordings of people speaking other languages.)

The *training* database, for training the UBM, consisted of the 800 longest recordings from the National Institute of Standards and Technology (NIST) 2004 speaker recognition evaluation (SRE) database [29].

The calibration and test databases were compiled from the telephone subset of the 8conv condition from the NIST 2008 SRE database [30], within which there were 132 US English speakers. The *calibration* database consisted of two non-contemporaneous recordings (*A* and *B*) from each of 32 speakers, each of voice-active duration greater than one minute. The *evaluation* database consisted of four non-contemporaneous recordings (*A*, *B*, *C*, and *D*) from each of the remaining 100 speakers, each of voice-active duration greater than one minute.

The decision as to which recording from each speaker to assign to *A* and *B* or *A*, *B*, *C*, and *D*, was arbitrary.

3.2. Forensic-voice-comparison system

The forensic-voice-comparison system tested was a basic automatic system. The front-end extracted 16 mel-frequency cepstral coefficient (MFCC) values from 20 ms frames overlapped by 10 ms, which were appended with delta coefficients [31]. Cumulative density mapping was used for feature normalization [32]. The back-end was based on a 512-mixture Gaussian mixture model – UBM (GMM-UBM) system [33]. For simplicity, no additional channel compensation procedures were applied.

The UBM was trained using expectation maximization (EM), and suspect-speaker models were created using five-iteration mean-only maximum a posteriori (MAP) adaptation from the UBM [33]. The system was calibrated using linear logistic regression as per [14, 21], using the FoCal Toolkit [34]. Calibration weights were calculated using same-speaker and different-speaker (lower numbered speaker *A* as suspect model and higher numbered speaker *B* as offender data) scores derived from the calibration data (all the calibration data were used), and these weights were then used to calibrate the LRs derived from the test data.

3.3. Procedures

The procedures described in §2 were used to estimate the 95% CI. For the non-parametric system the CI was estimated at each $\bar{x}_i$ value, with k set to 500. The C_{llr} value for the $\bar{x}_i$ values was also calculated.

Usually in forensic casework, a relatively large amount of suspect data is available, but the amount of offender data is relatively small. We therefore conducted tests of two conditions using all the available suspect data, but simulating having different amounts of offender data.

The suspect models were built using all the data available in the *A* and the *C* recordings (range 84 to 131 s, median 110 s). Two sets of suspect test data were analyzed. The first set consisted of the first 20 s of speech from each of the *B* and *D* recordings in the evaluation database, and the second set consisted of the next 40 s of speech from each of the *B* and *D* recordings (there was no overlap between these two sets).

4. Results

4.1. Raw results

C_{llr} values for test sets *AB* and *CD* for the 20 s and 40 s tests are given in Table 3 (only lower-numbered suspect to higher-numbered offender comparisons are included for different-speaker comparisons). Tippett plots are provided in Figures 2 and 3.

Although the C_{llr} values for the tests using 40 s of offender data were slightly less than for those using 20 s of offender data, the differences were not substantial – the differences between the *AB* and the *CD* pairs were greater than the differences between the 20 s and 40 s pairs, and the division of the former was arbitrary.

Table 3: C_{llr} values for test sets *AB* and *CD*

	Test set	
Duration	*AB*	*CD*
20 s	0.282	0.250
40 s	0.279	0.226

4.2. Accuracy results

C_{llr} values for the within-comparison mean LRs, $\bar{x}_i$, for the 20 s and 40 s tests were, to three figures, both 0.150.

4.3. Precision results (non-parametric procedure)

Figures 4 and 5 provide scatter plots of the deviation-from-mean, y_{ij}, values (y axis) against the within-comparison mean, $\bar{x}_i$, values (x axis), for the 20 s tests and 40 s tests respectively (red dots represent different-speaker comparisons, and blue dots same-speaker comparisons). The plots also include the 95% CI estimated at each $\bar{x}_i$ value (green lines). The means of these estimates are given in Table 4. The estimated 95% CI for the tests using more data was generally narrower than for the tests using less data.

Figures 6 and 7 provide Tippett plots of the mean within-comparison LRs (solid lines) and their corresponding 95% CIs (dashed lines to the left and right of the solid lines) for the 20 s and 40 s tests respectively. As might be expected given the small differences in the accuracy and precision results reported above, the two Tippett plots are visually almost identical.

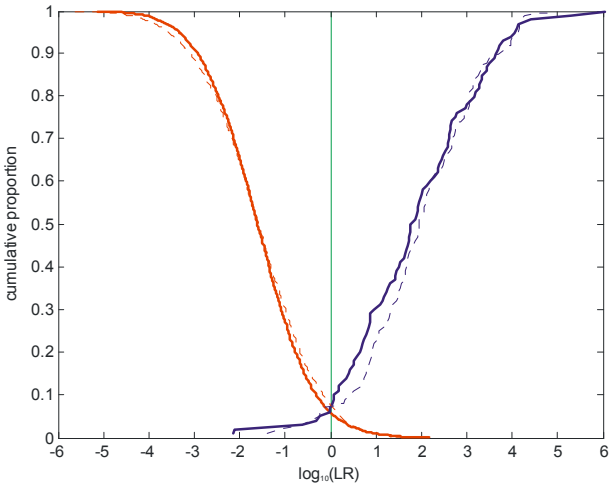

Figure 2: Tippett plot of LR values from test sets AB (sold lines) and CD (dashed lines) for 20 s of questioned-voice data.

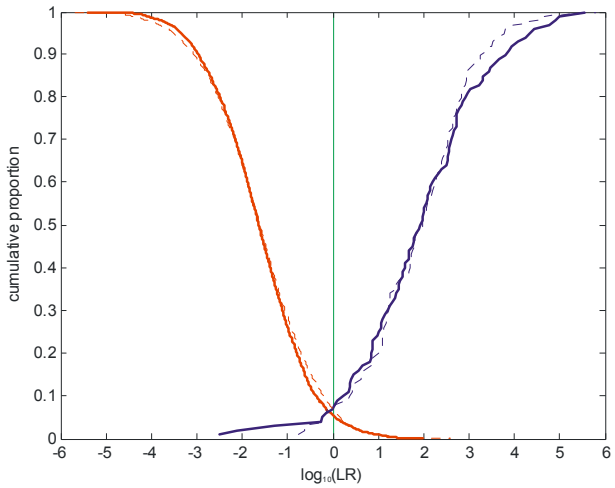

Figure 3: Tippett plot of LR values from test sets *AB* (sold lines) and *CD* (dashed lines) for 40 s of questioned-voice data.

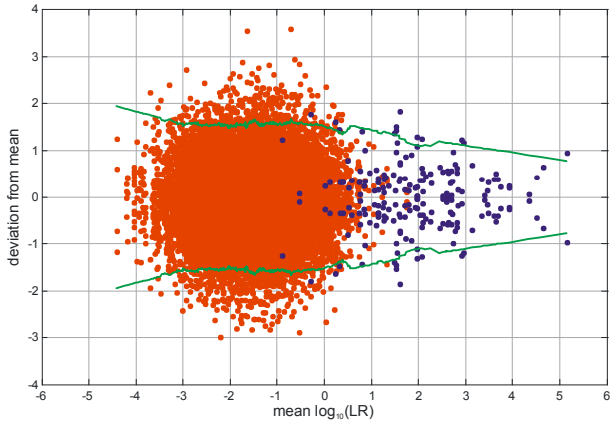

Figure 4: Scatter plot of deviation-from-mean values against within-comparison mean values for 20 s of questioned-voice data.

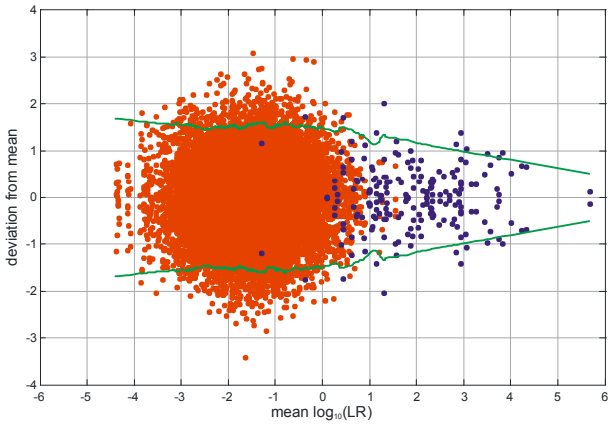

Figure 5: Scatter plot of deviation-from-mean values against within-comparison mean values for 40 s of questioned-voice data.

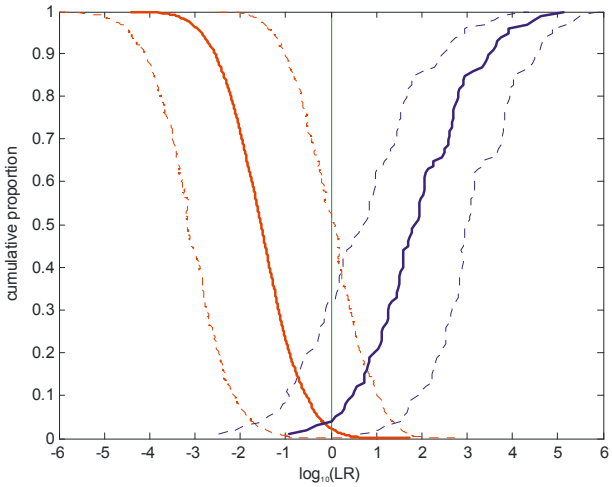

Figure 6: Tippett plot of within-comparison mean log-LR values from 20 s of questioned-voice data.

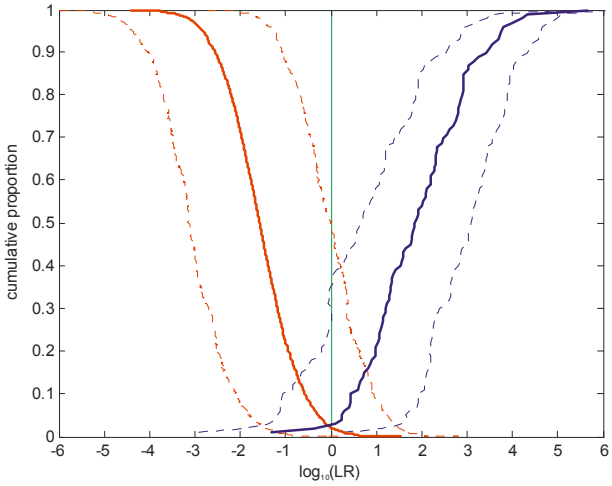

Figure 7: Tippett plot of within-comparison mean log-LR values from 40 s of questioned-voice data.

4.4. Precision results (parametric procedure)

Visual inspection of the scatter plots in Figures 4 and 5, and of running histograms (not shown), suggest that the assumptions of homoscedasticity and normality are not unreasonable for the deviation-from-mean log-LR values, y_{ij}, output by this system. The parametric estimates of the 95% CIs are given in Table 4. The estimated 95% CI for the tests using 40 s of data was narrower than for the tests using 20 s of data.

Table 4: Estimates of the 95% CIs (for the non-parametric procedure the value reported is the mean of the CI values estimated at each of the $\bar{x}_i$ values).

Test data duration	Procedure	
	non-parametric	parametric
20 s	±1.56	±1.69
40 s	±1.52	±1.63

5. Discussion

5.1. Comparison of using 40 s versus 20 s of questioned-voice data

When 40 s as opposed to 20 s of questioned-voice data were used there was no substantial difference in the accuracy of the results (§4.2); however, there was a slight increase in precision (decrease in the CI, §4.3–4.4). The primary purpose of the present paper is to explain procedures for calculating precision rather than examine the effect of using different amounts of questioned-voice data. To properly explore the latter, additional testing would be necessary using more data and/or randomization tests or bootstrapping.

5.2. The importance of presenting information about precision

To illustrate the importance of presenting information about precision, imagine that in casework, for which the 40 s system described in the present paper is appropriate, an LR for the comparison of the known- and questioned-voice recordings of 100 in favor of the same-speaker hypothesis is obtained.

Without having calculated an estimate of precision, the forensic scientist would simply report that one would be 100 times more likely to observe the measured acoustic differences between the known- and questioned-voice recordings under the hypothesis that the speaker on the questioned-voice recordings was the accused, than under the hypothesis that it was someone other than the accused.

When an estimate of precision is available, the forensic scientist can report that one would be 100 times more likely to observe the measured acoustic differences between the known- and questioned-voice recordings under the hypothesis that the speaker on the questioned-voice recordings was the accused, than under the hypothesis that it was someone other than the accused. In addition, they can report that, based on tests of the system, they are 95% certain that one would be at least 6.82 times (approximately 7 times) more likely and not more than 1,452 times (approximately 1,450 times) more likely to observe these acoustic differences given the same-speaker hypothesis than given the different-speaker hypothesis ($\log_{10}$(LR) of 2±1.166 taken from the results of the non-parametric procedure, see §2.2 and Figure 1).

Given this information about the precision of the forensic-comparison-system, the trier of fact may, for example, decide to use a conservative value (i.e., a value closer to an LR of one), and use a value of say 10, near the bottom of the 95% CI, rather than the raw calculated value of 100. Whereas the trier of fact is permitted to make a decision of this sort, it would be inappropriate for the forensic scientist to do so and, say, only report an LR value of 10 to the trier of fact. This would be stepping beyond an objective-as-possible scientific evaluation of the evidence, and usurping part of the rôle of the trier of fact.

Although the LR of 100 would still be the forensic scientist's best single-valued estimate of the strength of evidence, having estimated the precision of the system, it would also be inappropriate for them to simply report the raw LR value of 100. Rather, the rôle of the forensic scientist should be to provide the trier of fact with all the relevant information about the results of the analysis of the voice recordings, and the performance of the forensic-comparison system, including its precision at the activity level, so as to assist the trier of fact in coming to a maximally informed decision.

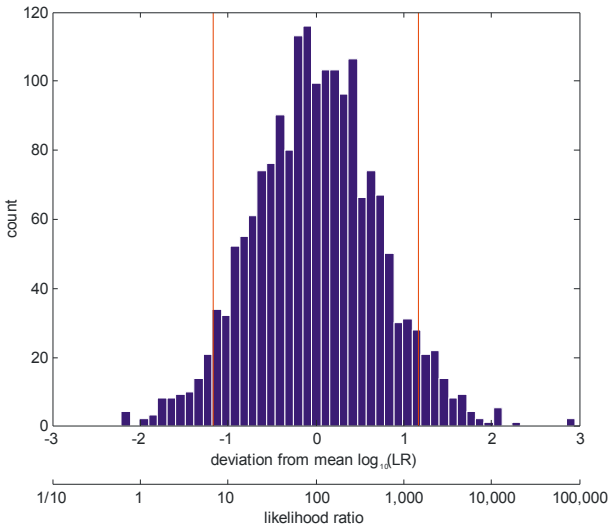

Figure 8: Histogram of deviation-from-mean $\log_{10}$(LR) values for the 500 nearest neighbors to $\log_{10}$(LR) = 2 using 40 s of questioned-voice data (blue bars). Estimated 95% CI from non-parametric procedure (red lines).

To help the trier of fact understand the 95% CI, the forensic scientist could present a histogram of the results from the *k* nearest neighbors, see Figure 8. Rather than an *x* axis labeled in log LRs, an *x* axis labeled in LRs would likely be more easily understood by the trier of fact, given that the trier of fact cannot be assumed to have a background in statistics etc. Whereas, if only supplied with a CI, a statistically sophisticated person may mistakenly assume a normal distribution, a statistically naïve person may mistakenly assume a uniform distribution.

Note that, unlike the local-linear-regression procedure used to calculate the non-parametric estimate of the CI, the histogram does not take account of heteroscedasticity over the range of values covered by the *k* nearest neighbors (compare

Figure 8 with Figure 1 – in Figure 8 the *x* dimension of Figure 1 has been collapsed and the folding of the *y* dimension about $\log_{10}(\text{LR}) = 0$ has been undone). The proportion of the area of the histogram bars beyond the *CI* lines may therefore not be equal to α, and the histogram should only be used as a rough guide to the shape of the distribution of the deviation-from-mean values in the vicinity of x_0. This would not be a concern if the parametric procedure were deemed appropriate and applied.

5.3. Separating accuracy and precision

At the beginning of the present paper (§1.2), C_{llr} was described as a measure of accuracy. Theoretically this is an appropriate characterization, and also practically when it is based on $\bar{x}_i$ values, i.e., on the within-group means calculated using multiple independent measures comparing the same pair of speakers / comparing the same speaker with themselves, as reported in §4.2. However, when C_{llr} is calculated using a single set of comparison values, as reported in §4.1, it is actually a goodness metric which combines accuracy and precision, with the relative contribution of each being unknown. When multiple samples are available for each comparison pair, it is possible to distinguish the contribution of accuracy and the contribution of precision. As the number of samples increases, so does the ability to separately estimate each of accuracy and precision.

5.4. Technical issues related to the non-parametric procedure

The non-parametric estimate of the CI will itself be more accurate and reliable in more densely populated regions of the $\bar{x}_i$ by y_{ij} space than in less densely populated regions (see Figures 4 and 5). Since there are many fewer same-speaker comparisons than different-speaker comparisons and only two independent test pairs for each same-speaker comparison, as compared to four for each different-speaker comparison, and the positive log-LR region is dominated by same-speaker results, the positive log-LR region is relatively sparsely populated. Where sampling is sparse, the size of the CI is likely to be underestimated because few extreme values are likely to be generated. Unlike the parametric system which uses a *t* distribution which is wider for smaller degrees of freedom, the non-parametric procedure does not take account of this phenomenon.

Changing the value of *k*, the number of nearest neighbors, also affects the accuracy and precision of the estimate of the CI. A small value of *k* fits the sample data more closely (lower bias) and results in a more jagged line in the scatter and Tippett plots (higher variance). For use in casework, it would be important to choose and fix the value of *k* before calculating the LR and the CI estimate for the real suspect and offender samples.

It may be desirable to find a procedure which produces smoother, less jagged, results. One possibility could be to find the most extreme 2α points at each $\bar{x}_i$ value over the entire $\bar{x}_i$ range, then fit a spline to the superset of these points.

5.5. Desirability of using the parametric procedure

If the assumptions of normality and homoscedasticity are reasonable, then it would be advantageous to use the parametric procedure rather than the non-parametric procedure. The latter is a more standard statistical procedure and is also easier to calculate. Also, when there are only two recordings per speaker available, only different-speaker results can be used to estimate a CI, but, if the assumptions for the parametric procedure hold, the parametric estimate of the CI will also be applicable to same-speaker results.

Assumptions of homoscedasticity and normality appear to be reasonable for the output of the GMM-UBM system used in the present paper, and also appear to be reasonable for the output of the GMM-UBM system, but not the multivariate-kernel-density system, reported in [27]. If the assumptions hold for GMM-UBM systems in general, then this will greatly simplify calculating and reporting the precision of GMM-UBM forensic-comparison systems.

6. Conclusion

In addition to accuracy, precision is an important aspect of the performance of a forensic-comparison system. Not reporting the estimated precision for a likelihood ratio calculated from known and questioned samples could mislead the trier of fact into giving a different weighting to the evidence than would be the case if they were aware of the activity-level precision limitations of the forensic-comparison system.

Results of the experiment reported in the present paper suggest that even if an increase in the length of questioned-voice samples does not lead to an improvement in system accuracy, it could lead to an improvement in precision.

7. Acknowledgments

This research was funded in part by Australian Research Council Discovery Grant No. DP0774115. Thanks to James M. Curran (Department of Statistics, University of Auckland), Daniel Ramos (ATVS - Biometric Recognition Group, Universidad Autónoma de Madrid), and two anonymous reviewers for comments on earlier versions of this paper.

8. References

[1] Daubert v Merrell Dow Pharmaceuticals (92–102) 509 US 579, 1993.

[2] Law Commission of England & Wales, *The Admissibility of Expert Evidence in Criminal Proceedings in England and Wales: A New Approach to the Determination of Evidentiary Reliability*, Law Commission, London, UK, 2009. <http://www.lawcom.gov.uk/expert_evidence.htm>

[3] National Research Council, *Strengthening Forensic Science in the United States: A Path Forward*, National Academies Press, Washington, DC, 2009.

[4] M.J. Saks, J.J. Koehler, "The coming paradigm shift in forensic identification science". *Science,* vol. 309, 2005, pp. 892–895. doi:10.1126/science.1111565

[5] C.G.G. Aitken, F. Taroni, Statistics and the Evaluation of Forensic Evidence for Forensic Scientist, 2nd ed, Wiley, Chichester, UK, 2004.

[6] Association of Forensic Science Providers, "Standards for the formulation of evaluative forensic science expert opinion", *Sci. Justice,* vol. 49, 2009, pp. 161–164. doi:10.1016/j.scijus.2009.07.004

[7] D.J. Balding, *Weight-of-evidence for Forensic DNA Profiles*, Wiley, Chichester, UK, 2005.

[8] J. Buckleton, "A framework for interpreting evidence", in J. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), *Forensic DNA Evidence Interpretation*, CRC, Boca Raton, FL, 2005, pp. 27–63.

[9] I.W. Evett, “Interpretation: A personal odyssey”, in C.G.G. Aitken, D.A. Stoney (Eds.), The Use of Statistics in Forensic Science, Ellis Horwood, Chichester, UK, 1991, pp. 9–22.

[10] I.W. Evett, “Towards a uniform framework for reporting opinions in forensic science case-work”, *Sci. Justice*, vol. 38, 1998, pp. 198–202. doi:10.1016/S1355-0306(98)72105-7

[11] B. Robertson, G.A. Vignaux, *Interpreting Evidence*, Wiley, Chichester, UK, 1995

[12] C. Champod, D. Meuwly, “The inference of identity in forensic speaker recognition”, *Speech Commun.*, vol. 31, 2000, pp. 193-203. doi:10.1016/S0167-6393(99)00078-3

[13] J. González-Rodríguez, A. Drygajlo, D. Ramos-Castro, M. García-Gomar, J. Ortega-García, “Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition”, *Computer Speech and Language*, vol. 20, 2006, pp. 331–355. doi:10.1016/j.csl.2005.08.005

[14] J. González-Rodríguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-García, “Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition”, *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 15, 2007, pp. 2104–2115. doi:10.1109/TASL.2007.902747

[15] G.S. Morrison, “Forensic voice comparison and the paradigm shift”, *Sci. Justice*, vol. 49, no. 4, 2009, pp. 298–308. doi:10.1016/j.scijus.2009.09.002

[16] P. Rose, *Forensic Speaker Identification*, Taylor and Francis, London, UK, 2002.

[17] P. Rose, “Technical forensic speaker recognition”, *Comp. Speech Lang.*, vol. 20, 2006, pp. 159–191. doi:10.1016/j.csl.2005.07.003

[18] P. Rose, G.S. Morrison, “A response to the UK position statement on forensic speaker comparison”, *Int. J. Speech, Lang. Law*, vol. 16, 2009, pp. 139–163. doi:10.1558/ijsll.v16i1.139

[19] N. Brümmer, J. du Preez, “Application independent evaluation of speaker detection”, *Comp. Speech Lang*., vol. 20, 2006, pp. 230–275. doi:10.1016/j.csl.2005.08.001

[20] D.A. van Leeuwen, N. Brümmer, “An introduction to application-independent evaluation of speaker recognition systems”, in C. Müller (Ed.), *Speaker Classification I: Fundamentals, Features, and Methods*, Springer-Verlag, Heidelberg, Germany, 2007, pp. 330–353. doi:10.1007/978-3-540-74200-5_19

[21] G.S. Morrison, “Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs”, *J. Acoust. Soc. Americ.,* vol. 125, 2009, pp. 2387–2397. doi:10.1121/1.3081384

[22] J.M. Curran, “An introduction to Bayesian credible intervals for sampling error in DNA profiles”, *Law, Prob. Risk*, vol. 4, 2005, 115–126. doi:10.1093/lpr/mgi009

[23] J.M. Curran, J.S. Buckleton, C.M. Triggs, B.S. Weir, “Assessing uncertainty in DNA evidence caused by sampling effects”, *Sci. Justice*, vol. 42, 2002, 29–37. doi:10.1016/S1355-0306(02)71794-2

[24] W.M. Bolstad, *Introduction to Bayesian Statistics*, 2nd Ed., Wiley, Hoboken, NJ, 2007.

[25] R. Cook, I.W. Evett, G. Jackson, P.J. Jones, J.A. Lambert, “A hierarchy of propositions: Deciding which level to address in casework”, *Sci. Justice*, vol. 38, 1998, 231–239. doi:10.1016/S1355-0306(98)72099-4

[26] G.S. Morrison, C. Zhang, P. Rose, “An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system”, 2010. Manuscript submitted for publication.

[27] G.S. Morrison, “A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multvariate kernel density (MVKD) versus Gaussian mixture model – universal background model (GMM-UBM)”, 2010. Manuscript submitted for publication.

[28] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, NY, 2009.

[29] The NIST Year 2004 Speaker Recognition Evaluation Plan. <http://www.itl.nist.gov/iad/mig/tests/spk/2004/SRE-04_evalplan-v1a.pdf>

[30] The NIST Year 2008 Speaker Recognition Evaluation Plan. <http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf>

[31] S. Furui, “Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum”, *IEEE Trans. Acoust., Speech and Sig. Proc.*, vol. 34 no.1, 1986, pp. 52–59.

[32] J. Pelecanos, S. Sridharan, “Feature warping for robust speaker verification”, *Odyssey Workshop*, 2001, pp. 213–218.

[33] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, “Speaker verification using adapted Gaussian mixture models,”, *Digital Signal Processing*, vol.10, 2000, no. 1/2/3, pp.19-41. doi:10.1006/dspr.1999.0361

[34] N. Brümmer, FoCal Toolkit, July 2005. <http://niko.brummer.googlepages.com/focal/>