Science and Justice 51 (2011) 91–98

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

Measuring the validity and reliability of forensic likelihood-ratio systems ☆

Geoffrey Stewart Morrison

Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, UNSW Sydney, NSW 2052, Australia

ARTICLE INFO

Article history:
Received 8 June 2010
Received in revised form 27 February 2011
Accepted 19 March 2011

Keywords:
Forensic
Likelihood ratio
Validity
Reliability
Accuracy
Precision

ABSTRACT

There has been a great deal of concern recently about validity and reliability in forensic science. This paper reviews for a broad target audience metrics of validity and reliability (accuracy and precision) which have been applied in forensic voice comparison and which are potentially applicable in other branches of forensic science. The metric of validity is the log likelihood-ratio cost (C_{llr}), and the metric of reliability is an empirical estimate of credible intervals. A revised procedure for the calculation of credible intervals is introduced.

Contents

1. Introduction

The present paper describes metrics of the validity and reliability of forensic likelihood-ratio systems which have been adopted or developed by the author and his colleagues. The intent of the current exposition is to present these metrics and the rationale behind them to a broad audience. The descriptions are therefore relatively non-technical, focussing on concepts rather than equations (references to

☆ Variants of this paper have been presented as: G.S. Morrison, Empirically assessing the validity and reliability of forensic-comparison systems, presentation at the International Association for Forensic Phonetics and Acoustics (IAFPA) Conference, Trier, Germany, July 2010; G.S. Morrison, Empirical measures of accuracy and reliability in forensic comparison science, presentation at the Australia and New Zealand Forensic Science Society (ANZFSS) 20th International Symposium on Forensic Science, Sydney, Australia, September 2010; and G.S. Morrison, Measuring validity and reliability in forensic science, keynote presentation at BIT's 1st Annual World Congress of Forensic Science, Dalian, China, October 2010.

1355-0306/$ – see front matter
doi:10.1016/j.scijus.2011.03.002

more technical works are provided throughout). Examples are drawn from forensic voice comparison, but the concepts and metrics are potentially applicable to many other branches of forensic science.

1.1. Concern about validity and reliability

There has been a great deal of concern recently about validity and reliability in forensic science [1–3]. The National Research Council (NRC) report to Congress on Strengthening Forensic Science in the United States [2] urged that procedures be adopted which include "quantifiable measures of the reliability and accuracy of forensic analyses" (p. 23), "the reporting of a measurement with an interval that has a high probability of containing the true value" (p. 121), and "the conducting of validation studies of the performance of a forensic procedure" (p. 121). Quantification of accuracy and precision are standard practice in modern science in general and the NRC report can be summarised in part as "forensic science should be scientific".

1.2. What are validity and reliability?

In statistics and science in general *validity* is synonymous with *accuracy*, and *reliability* is synonymous with *precision*. Fig. 1 illustrates the difference between accuracy and precision. If some system takes a number of measurements on the same object and there is a wide range of values in that set of measurements but their average value is close to the true value, then the system has good accuracy but poor precision (Fig. 1b). If another system takes a number of measurements on the same object and there is a small range of values in that set of measurements but their average value is far from the true value, then the system has good precision but poor accuracy (Fig. 1c). Ideally, one wants a system with both good accuracy and good precision (Fig. 1d), although in practice it might be difficult to obtain good values for both.

The distinction between validity and reliability has not always been maintained in the judicial and forensic-science literature. In *Daubert* [Daubert v Merrell Dow Pharmaceuticals (92–102) 509 US 579, 1993] the US Supreme Court equated "evidentiary reliability" with "scientific validity". Although both validity and reliability are clearly of concern in the NRC report [2] it does not appear to consistently distinguish the two terms according to their conventional scientific meanings. The Law Commission of England & Wales consultation paper [3] uses the term reliability where its concern actually appears to be primarily related to validity. Influenced by the latter, in an earlier paper [4] the author of the present paper used the term reliability in place of validity. Arguments about why it is important in forensic science to consider both validity and reliability will be presented towards the end of the present paper (Section 5) once the required conceptual framework has been built.

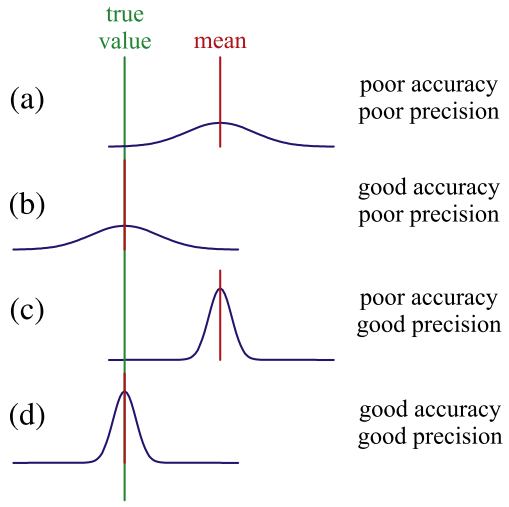

Fig. 1. Examples of accuracy and precision.

1.3. The likelihood-ratio framework

The likelihood-ratio framework for the evaluation of forensic evidence is described in numerous works including Robertson and Vigneaux [5] and Aitken and Taroni [6] for forensic science in general, Balding [7] and Buckleton [8] in relation to forensic-DNA comparison, and Champod and Meuwly [9], Rose [10,11] and Morrison [4,12] in relation to forensic voice comparison. It can be applied to any evidence type for which the ultimate question is whether two samples have the same origin or not.

In the likelihood-ratio framework the task of the forensic scientist is to determine the probability of obtaining the observed properties of the sample of known origin and the sample of questioned origin under the hypothesis that the two samples have the same origin versus under the hypothesis that they have different origins, see Eq. (1).

$$LR = \frac{p(\ E|H_{so}\ )}{p(\ E|H_{do}\ )} \tag{1}$$

where LR is the likelihood ratio, E is the evidence, i.e., the properties of the sample of known origin and the properties of the sample of questioned origin, H_{so} is the same-origin hypothesis and H_{do} is the different-origin hypothesis.

A likelihood ratio greater than one lends support to the same-origin hypothesis, e.g., if the likelihood ratio is 100, then, whatever the trier of fact's[1] prior belief, after hearing this they should be 100 times more likely than before to believe that the samples have the same origin rather than different origins. Similarly, a likelihood ratio less than one lends support to the different-origin hypothesis, e.g., if the likelihood ratio is 1/100, then, whatever the trier of fact's prior belief, after hearing this they should be 100 times more likely than before to believe that the samples have different origins rather than the same origin. The deviation of the likelihood ratio from one is a quantification of the strength of the evidence with respect to the competing same-origin and different-origin hypotheses.

It is often convenient to convert likelihood ratios to log likelihood ratios such that the strength of evidence is expressed on a scale which is symmetrical about zero. Table 1 provides conversions between selected likelihood-ratio values and their corresponding log-base-ten likelihood-ratio values. Count the number of zeros!

2. Measuring the validity of forensic likelihood-ratio systems

2.1. How to measure validity

Irrespective of the actual metric used to quantify validity, the general procedures for measuring validity are the same: take a large number of pairs of test samples, some of which are known to be same-origin pairs and the remainder of which are known to be different-origin pairs. Run each pair through the system and in each case determine whether the output value is good or bad with respect to whether it accords with the desired value given the same-origin or different-origin status of the input. Calculate overall validity as a value describing the average performance over all of the test pairs. The

[1] The trier of fact is the entity which makes the ultimate decision in a trial. Depending on the legal system the trier of fact could be a judge, a panel of judges, or a lay jury.

Table 1
Conversion of likelihood ratios, *LR*, to log-base-ten likelihood ratios, $\log_{10}(LR)$.

LR	1/1000	1/100	1/10	1	10	100	1000
$\log_{10}(LR)$	−3	−2	−1	0	1	2	3

Table 2
Correct classifications and classification errors.

Truth	System output	
	Same-origin	Different-origin
Same-origin	True positive	False negative
Different-origin	False positive	True negative

details of the determination of goodness and the averaging constitute the metric[2].

An important point to make here is that measurement of validity depends both on the system and the test set. For this reason, if a measurement of validity is to be relevant in the context of a particular case at trial, the samples in the test set should be matched as closely as possible to the conditions in the case at trial. For example, in forensic voice comparison, if in the case at trial the offender recording is of a one-minute long mobile telephone conversation and the suspect recording is five minutes long and consists of answers to police interview questions recorded using a regular microphone, then each pair of recordings in the test set should consist of a one-minute long mobile telephone conversation and a five-minute long recording of answers to interview questions recorded using a microphone.

2.2. Correct-classification rates/classification-error rates as a metric of validity

The NRC report [2, ch. 4] and Koehler [13] recommend the use of correct-classification rates/classification-error rates for measuring the validity of a forensic-comparison system. Decisions are binary, as shown in Table 2. Errors are counted if the system declares two samples to have the same origin when in fact they have a different origin (false positive), or when it declares two samples to have different origins when in fact they have the same origin (false negative). However, as will be explained below, such a metric of validity is not appropriate if one is working within the likelihood-ratio framework.

2.3. Requirements for a metric of the validity of forensic likelihood-ratio systems

Correct-classification rates/classification-error rates are not appropriate for use within the likelihood-ratio framework because they are based on posterior probabilities rather than likelihood ratios, and they are based on a categorical thresholding, error versus not-error, rather than a gradient strength of evidence.

In order to calculate a posterior probability, one would have to combine prior probabilities with the likelihood ratio, as shown in Eq. (2) (odds form of Bayes' Theorem).

$$\underbrace{\frac{p(H_{so}|E)}{p(H_{do}|E)}}_{\substack{\text{posterior}\\ \text{odds}}} = \underbrace{\frac{p(E|H_{so})}{p(E|H_{do})}}_{\substack{\text{likelihood}\\ \text{ratio}}} \times \underbrace{\frac{p(H_{so})}{p(H_{do})}}_{\substack{\text{prior}\\ \text{odds}}} \tag{2}$$

The prior odds, the trier of fact's belief about the relative probability of the competing hypotheses before the evidence is presented, and the posterior odds, the trier of fact's belief about the relative probability of the competing hypotheses after the evidence has been presented, are not within the purview of the forensic scientist. In fact it is a strength of the likelihood-ratio framework that the forensic scientist focuses only on calculating the strength of evidence on the basis of the samples presented to them for analysis, and does not consider the prior odds, thus reducing the likelihood that their analysis could be consciously or unconsciously influenced by other knowledge about the case (such potential sources of human bias were a concern of the NRC report [2]). Since posterior probabilities are not calculated as part of the likelihood-ratio framework, a metric of validity cannot be based on posterior probabilities, rather it should be based on likelihood ratios.

The size of a likelihood ratio indicates the strength of its support for one hypothesis over the other. It would be worse to report a likelihood ratio of one million in favour of a contrary-to-fact hypothesis (a likelihood-ratio which supports the same-origin hypothesis when the objects actually have different origins, or a likelihood-ratio which supports the different-origin hypothesis when the objects actually have the same origin) than to report a likelihood-ratio of ten in favour of a contrary-to-fact hypothesis, because the former provides greater support for the contrary-to-fact hypothesis and would thus have a greater potential to contribute to a miscarriage of justice. A metric of validity should therefore assign a greater penalty to the former than to the latter. Such a gradient metric contrasts with the binary classification-error rate metric which would assign an equal penalty to both.

2.4. The log-likelihood-ratio cost as a metric of validity

An appropriate metric of validity for use within the likelihood-ratio framework, a gradient metric based on likelihood ratios, is the log-likelihood-ratio cost (C_{llr}). The log-likelihood-ratio cost was developed for use in automatic speaker recognition (Brümmer and du Preez [14], van Leeuwen and Brümmer [15]) and has subsequently been applied in forensic voice comparison (e.g., Ramos-Castro et al. [16], González-Rodríguez et al. [17], Morrison and Kinoshita [18], Thiruvaran, Ambikairajah, and Epps [19], Morrison [20], Enzinger [21]). Its increasing popularity within the latter branch of forensic science is attested by the fact that it was mentioned in the abstracts of 6 of the 18 papers originally accepted for presentation in the special session on forensic voice comparison and forensic acoustics at the 2nd Pan-American/Iberian Meeting on Acoustics in November 2010[3].

[2] Note that the type of accuracy of concern in the present paper is with respect to the ground truth of whether a pair of samples have the same origin or different origins. This type of accuracy can be assessed irrespective of what kind of measurements are made on the objects and is even applicable if the output consists of subjective judgments. There is another type of accuracy which could be assessed but it is specific to a particular kind of measurement made on the object. For example, if the true distribution of the fundamental frequencies of speakers voices were known for a population, then the true likelihood ratio for a comparison of any pair of samples could be calculated. It could be the case that for a particular same-origin comparison the true likelihood ratio is actually less than one, and for a particular different-origin comparison the true likelihood ratio could be greater than one. This type of validity can be explored using a statistical technique known as Monte Carlo simulation, but in practice for any non-trivial situation it is impossible to known the true distribution of any particular kind of measurement. Also, the latter type of validity is specific to the particular kind of measurement and a system using that kind of measurement cannot be compared with a system using a different kind of measurement, e.g, a forensic-voice-comparison system using mel-frequency cepstral coefficients could not be compared with a system using vowel-formant measurements.

[3] http://cancun2010.forensic-voice-comparison.net/.

In order to calculate C_{llr} one must have a test database from which one can draw a large number of pairs of samples known to have the same origin and a large number of pairs of samples known to have different origins. The pairs of test samples are presented to the system and the knowledge about the same-origin or different-origin status of the test pairs is compared with the output of the system. As mentioned above (Section 2.1) as is the case for measurement of validity using any metric, the value of C_{llr} is dependent on the test samples as well as the forensic-comparison system. For the test of system validity to be meaningful in any particular case the test database must therefore be representative of the relevant population and reflect the conditions of the known and questions samples in the case.

The log-likelihood-ratio cost is calculated using Eq. (3).

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{so}}\sum_{i=1}^{N_{so}} log_2\left(1+\frac{1}{LR_{so_i}}\right) + \frac{1}{N_{do}}\sum_{j=1}^{N_{do}} log_2\left(1+LR_{do_j}\right)\right) \quad (3)$$

where N_{so} and N_{do} are the number of same-origin and different-origin comparisons respectively, and LR_{so} and LR_{do} are the likelihood ratios derived from test pairs known to be same-origin and different origin comparisons respectively. Note that part of the equation is the mean of the output of a function applied to all the likelihood ratios derived from same-origin comparisons (left side within the outer brackets), another part is the mean of the output of a function applied to all the likelihood ratios derived from different-origin comparisons (right side within the outer brackets), and C_{llr} is the mean of these two means. A plot of the same-origin and different-origin penalty functions is shown in Fig. 2. Ideally, a same-origin comparison should result in a large positive log likelihood ratio and would contribute a very small penalty value to C_{llr}. For a same-speaker comparison, a positive log likelihood ratio close to zero would not provide as much support for the same-speaker hypothesis as a large positive log likelihood ratio, and it would contribute a somewhat larger penalty value to C_{llr}. For a same-speaker comparison, a negative log likelihood ratio would contrary-to-fact lend support to the different-speaker hypothesis and would contribute a larger penalty value to C_{llr}, with that penalty value increasing as the magnitude of the negative log likelihood ratio increases and lends greater support to the contrary-to-fact different-origin hypothesis. Mutatis mutandis for a different origin comparison.

If a forensic-comparison system were always to return a likelihood ratio of one, i.e., equal likelihood of obtaining the evidence under either the same-origin or different-origin hypothesis, then its C_{llr} would be one. The extent to which the C_{llr} is less than one is a measure of the validity of the system, the lower the C_{llr} the better the validity of the system (it is possible for a system to have a C_{llr} greater than one, although this can be fixed via a process known as calibration [14,15]).

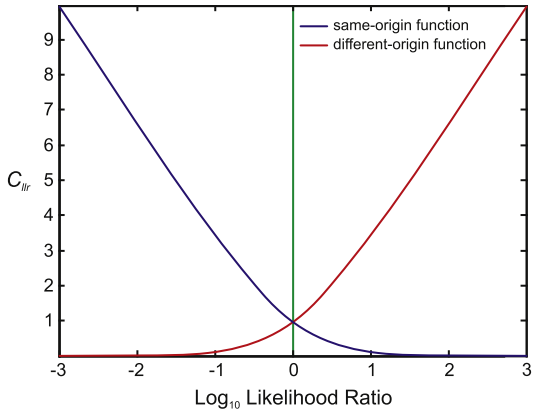

Fig. 2. Plot of C_{llr} functions for same-origin and different-origin comparisons.

2.5. Example

This example is taken from Morrison [22], which is in turn a partial reanalysis of Morrison [23], further details can be found in the latter two publications. The research question concerns whether a new procedure for extracting information from voice recordings is better than an old procedure. From audio recordings of vowels it is possible to extract measurements of the acoustic resonances of the vocal tract (called formants). In certain vowels, called diphthongs, the formant values change over time. The old procedure measured the formant values near the beginning of the diphthong and again near the end of the diphthong. The new procedure involves fitting a parametric curve to the whole trajectory of the change over time of the formant values, and using the coefficient values from the parametric curve to calculate the likelihood ratio. In this example the diphthong measured was /aɪ/, the vowel sound in the words "high", "bye", "cry" etc., and the parametric curve was a cubic polynomial, resulting in four coefficient values per formant trajectory. The test database was small, consisting of recordings of only 25 speakers, but was a database on which the old procedure had previously been tested. (Because of the small database, this experiment should only be considered an initial proof of concept.) The same test pairs were analysed using the same acoustic-phonetic forensic-voice-comparison system with the only difference being the procedures for extracting numeric data from the audio recordings. Likelihood ratios were calculated using Aitken and Lucy's [24] multivariate kernel density formula, followed by logistic-regression calibration (Brümmer and du Preez [14], van Leeuwen and Brümmer [15]). Because of the small amount of data available, a cross-validation procedure was adopted. The old procedure resulted in a C_{llr} of 0.43 and the new procedure resulted in a C_{llr} of 0.10, less than a quarter that of the old procedure. The validity of the new procedure is therefore much better than the validity of the old procedure.

3. Reliability within the likelihood-ratio framework

3.1. Type of variability of concern in the present paper

Curran et al. [25] and Curran [26] discussed and compared metrics of the reliability of likelihood ratios resulting from forensic-DNA-profile comparison. The focus was on the effect of changes in the size of the background database on the precision of likelihood ratios (the background database is the sample of the population used to estimate the typicality of the known- and questioned-origin samples). The type of variability of interest in the present paper, however, is not that due to background-database selection, but rather that due to intrinsic variability at the source of the samples. To a first approximation DNA can be considered invariant, i.e., the DNA of an organism does not change from occasion to occasion. DNA profiles are also quantified using discrete values, e.g., integer values such as 1, 2, 3, 4, but no intermediate values such as 1.5, 2.33, 2.99999 or 3.00001. In contrast human voices are highly variable and from occasion to occasion there will almost inevitably be measurable differences in their acoustic properties, even if a speaker were to try to repeat exactly the same thing the same way. Acoustic measurements are also continuously valued, as opposed to discrete. For discussion of the difference in the data structure of DNA profiles and acoustic measurements of voice recordings, see Morrison [4 §99.190, 12] and Rose and Morrison [27]. Although variability at the source or in the measurement of an object's properties may be extreme in voice data, this type of variability is a problem which potentially affects all branches of forensic science based on continuously-valued measurements[4].

[4] Although the idea that the reliability of likelihood ratios should be estimated has been relatively well received in forensic-science venues, it has not been well received by the automatic-speaker-recognition community. This appears to be due to a lack of understanding about forensic science and the likelihood-ratio framework within that community. The issues are discussed in Morrison [32 Appendix B].

In passing, it should be noted that Alexander [28 §3.4–3.5] proposed a procedure for estimating the reliability of forensic likelihood ratios accounting for both variability in the composition of the background database and source/measurement variability for the known-origin sample. The proposed procedure used as statistical technique known as bootstrapping, and was designed for use in forensic voice comparison, but its use was not demonstrated.

3.2. Credible intervals as a metric of reliability

Credible intervals are the Bayesian analogue of frequentist confidence intervals (see Curran [26] and Bolstad [29]). In the present paper (and others mentioned below) uniform priors are used, hence the numeric calculation of credible intervals and confidence intervals would be equivalent; however, only Bayesian credible intervals are philosophically consistent with the likelihood-ratio framework. Credible intervals also have a more natural interpretation: we are 95% certain that the true value of the parameter we wish to estimate lies within the 95% credible interval.

Morrison, Zhang, and Rose [30] introduced a procedure for calculating credible intervals for likelihood ratios derived from tests of forensic-comparison systems. This procedure was generalised in Morrison, Thiruvaran, and Epps [31] and a second related procedure introduced. These procedures have also been applied in Morrison [32], and Morrison, Thiruvaran, and Epps [33]. The first part of the two procedures is identical but the final part in one is non parametric, applicable when homoscedasticity cannot be assumed, and in the other it is parametric, applicable when homoscedasticity is a reasonable assumption. Homoscedastic data have the same amount of variability irrespective of the baseline values, e.g., at a $\log_{10}(LR)$ of 1 the 95% credible interval is ± 0.5, at a $\log_{10}(LR)$ of 2 it is ± 0.5, and at a $\log_{10}(LR)$ of 3 it is ± 0.5. Heteroscedastic data have different amounts of variability depending on the baseline values, e.g., at a $\log_{10}(LR)$ of 1 the 95% credible interval is ± 0.5, at a $\log_{10}(LR)$ of 2 it is ± 1.0, and at a $\log_{10}(LR)$ of 3 it is ± 1.5.

The present paper introduces a revision of the first part of the procedures described in Morrison, Thiruvaran, and Epps [31]. The revision represents a change in philosophy about exactly what should be measured. For each procedure, the second part (the more mathematical-calculation heavy part) remains unchanged. For clarity of exposition, the new first part of the procedures is described as is and discussion of how it differs from the earlier version is relegated to a footnote.

Assume that in the test database there are three samples of each object, e.g., three noncontemporaneous recordings of each speaker in a database of recordings of speakers from the relevant population (there could be more than three, and more would lead to a better estimate of the degree of reliability of the system, but three is the minimum number necessary). One of these samples (hereafter sample A) has the same conditions as the questioned-origin sample in the case at trial, e.g., same language and dialect, same gender, same speaking style, same recording and transmission channel (if known), and same duration. The other two samples (hereafter B and C) have the same conditions as the known-origin sample in the case at trial. For simplicity, it is assumed that there is only one known-origin recording in the case at trial, if there were say two known-origin recordings, then each of B and C would designate a set of two recordings from the test database. No recording should be reused, e.g., recording A should not reappear in B or C, and a recording from B should not reappear in C. Each same-speaker comparison is now represented by two pairs of recordings, and each different-speaker comparison by four pairs of recordings, see Table (3a) and (3b). We can therefore generate two likelihood-ratio estimates for each same-speaker (same-object) comparison and four likelihood-ratio estimates for each different-speaker (different-object) comparison[5].

Fig. 3a shows a schematic of two log-likelihood-ratio values derived from the two pairs of samples from a single same-object comparison and four log-likelihood-ratio values derived from the four pairs of samples from a single different-object comparison. In practice the test set would include many same-origin and many different-origin comparisons. To simplify description these sets of two or four values will henceforth be referred to as groups. The mean log-likelihood-ratio value is calculated for each group, Fig. 3b. Each log likelihood ratio within each group is then converted from its original value to a deviation-from-the-mean value, Fig. 3c. Note that, relative to their location in Fig. 3b, the points in Fig. 3c are simply rotated by 90° about their within-group means. The calculation of a credible interval, say a 95% credible interval, is then essentially a matter of finding the boundaries between the 95% of points which have the least deviation from the mean and the 5% of points which have the greatest deviation from the mean, Fig. 3d.

Fig. 3d shows a heteroscedastic credible interval, i.e., the width of the interval is not fixed but depends on the underlying likelihood ratio value. In this example the credible interval gets narrower as the log likelihood ratio increases. Morrison, Thiruvaran, and Epps [31] describe a non-parametric procedure for calculating a heteroscedastic credible interval. The procedure is simplified by assuming that the deviation-from-mean distribution is symmetrical in a log-likelihood-ratio space (this assumption was already implied in the use of the mean value), in which case one can ignore the sign of the deviation-from-mean values and only work with their absolute values, i.e., flip all of the negative values over to positive values, Fig. 3e. This simplifies the calculation because it is now only necessary to find one boundary between the 95% lowest magnitude deviation-from-mean values and the 5% highest magnitude deviation-from-mean values. The procedure for finding this boundary is based on local linear regression, see Morrison, Thiruvaran, and Epps [31] for details.

If homoscedasticity can be assumed then a simpler and more robust parametric procedure can be applied. In the parametric

Table (3a)
Same-object comparisons.

Known origin		Questioned origin	
Object	Sample	Object	Sample
1	B	1	A
1	C	1	A
2	B	2	A
2	C	2	A
:	:	:	:

[5] In the earlier version of the procedures, multiple questioned-origin test samples were used as well as multiple known-origin test samples. No samples were reused and it was argued that because the pairs within each group were nonoverlapping they were statistically independent. In the revised version only one questioned-origin test sample is used. The logic is that in a case at trial only one questioned-origin sample can be tested at a time and the same-origin–different-origin likelihood-ratio question asked is with respect to this one particular questioned-origin sample not other samples which could have come from the same origin. The questioned-origin sample is therefore treated as a fixed point (as in Alexander [28]). Because it is a fixed point, although sample A is used in multiple pairs in each group (e.g., B–A, C–A), the likelihood ratios from these pairs can still be treated as statistically independent. The strength-of-evidence question is still concerned with the speaker level, as opposed to the recording level, with respect to the known-origin recording — if more known-origin recordings were available a better estimate of the within-speaker variability for the known speaker could be obtained, but given the number of known-origin recordings available how might the likelihood ratio vary if different sets of the same number of recordings could be obtained from the known speaker? The estimate of the imprecision of the system using the revised procedure is actually a quantification of the extent to which the system has failed to adequately model the within-speaker variability.

Table (3b)
Different-object comparisons.

Known origin		Questioned origin	
Object	Sample	Object	Sample
2	B	1	A
2	C	1	A
3	B	1	A
3	C	1	A
:	:	:	:
1	B	2	A
1	C	2	A
3	B	2	A
3	C	2	A
:	:	:	:

procedure, one simply ignores the underlying log-likelihood-ratio values and calculates the pooled within-group sample variance, i.e., the sample variance of all the deviation-from-mean values irrespective of their group membership, Fig. 3f. The sample variance is then used with a *t* distribution to calculate the credible interval. Uniform priors are adopted and hence the credible interval can be calculated using only the sample variance and the degrees of freedom, see Morrison, Thiruvaran, and Epps [31] for details.

The description above assumes a research scenario using multiple simulated questioned-origin samples. In cases of likelihood ratios greater than one, the estimate of the credible interval could be made more specific to the case at trial by having the different-origin comparison pairs each be a comparison between the sample of questioned origin from the case at trial and a sample from the test database (assuming the test database does not contain any samples from the same origin as the questioned-origin sample). Under this scenario, only different-origin comparisons could be used to calculate reliability, and the number of sample pairs per comparison group would be equal to the number samples from each object on the test set.

3.3. Validity revisited

If C_{llr} were calculated using all sample pairs in a group (e.g., B–A, C–A) it would include the contribution of imprecision of the system. To remove as much as possible the contribution of imprecision, C_{llr} should be calculated using the mean value from each group. Via the central-limit theorem the mean gives a better estimate of the true value than any individual value. The log-likelihood-ratio cost can then be presented as a measure of the accuracy and a credible interval as a measure of the precision of the likelihood-ratio results from the same test set.

3.4. Example

This example is an application of the new procedure using a system and data otherwise identical to an example previously presented in Morrison, Thiruvaran, and Epps [31]. The system is a generic automatic forensic-voice-comparison system. The system characterises the entire spectrum of the speech-active portion of each recording using mel-frequency cepstral coefficients (MFCCs) with a measurement taken every 10 ms. Likelihood ratios are then calculated using a Gaussian mixture model–universal background model (GMM-UBM, Reynolds, Quatieri, and Dunn [34]), followed by logistic-regression calibration (Brümmer and du Preez [14], van Leeuwen and Brümmer [15]). The test database consisted of three audio recordings of each of 100 speakers and the system was tested under two conditions: in the first condition the test samples simulating the questioned-origin sample consisted of 20 ms net speech, and in the second condition these test samples consisted of 40 ms net speech

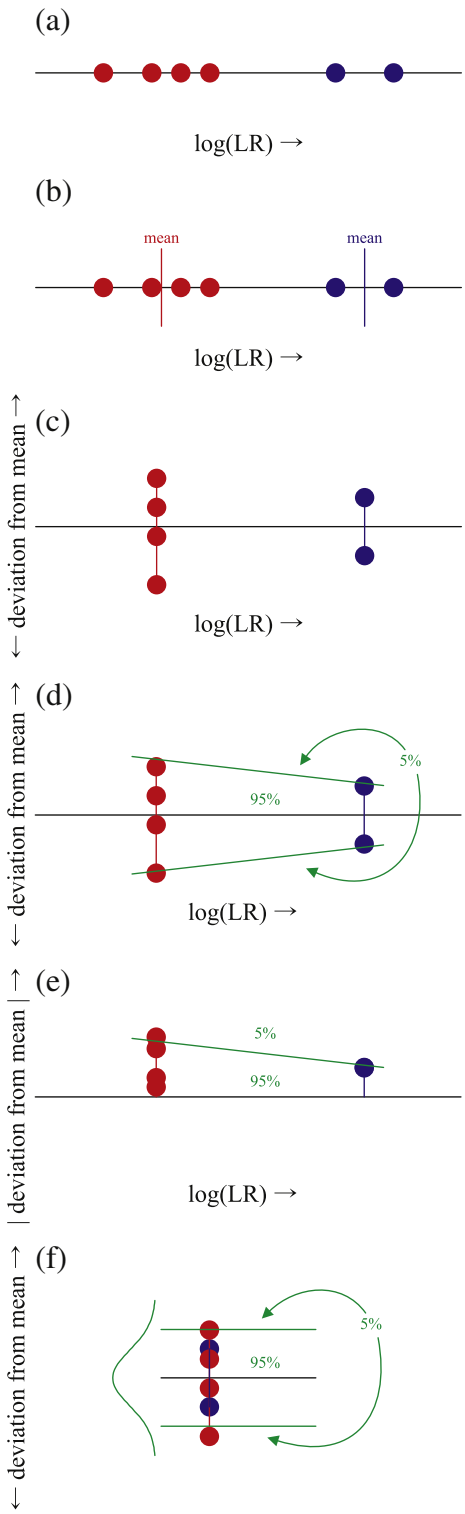

Fig. 3. (a) Two groups of likelihood ratios, one group derived from a same-object comparison and the other from a different-object comparison. (b) The means of each group. (c) The deviations from the mean of each group. (d) Heteroscedastic 95% credible interval. (e) Non-parametric calculation of heteroscedastic 95% credible interval using absolute deviation-from-mean values. (f) Parametric calculation of homoscedastic 95% credible interval using deviation-from-mean values and a *t* distribution.

(the 20 ms and 40 ms were non-overlapping portions taken from the same original recordings). It is not uncommon in forensic voice comparison for the length of the audio recordings to be relatively

short. 95% credible intervals were estimated using the parametric procedure.

Results are given in Table 4. Somewhat unexpectedly, the validity of the 20 ms condition was slightly better than for the 40 ms condition, but, as might be expected, the reliability for the 40 ms condition was slightly better than that for the 20 ms condition. Given the small size of the differences in C_{llr} and the 95% credible interval, one could conclude that this system is relatively robust to differences in the length of the questioned-voice recording in the range 20–40 ms.

3.5. The percentile of the credible interval

In the example above 95% credible intervals were calculated; however, calculating a credible interval at a single percentile could be misleading — there is nothing special about 95%, 99%, or any other percentile and a cliff-edge interpretation should be avoided, e.g., there is unlikely to be any meaningful difference between a value which is at the 94.99th percentile and the 95.01th percentile. A statistically naïve trier of fact could mistakenly think that the belief about the true likelihood-ratio value is zero outside the 95% credible interval and equal for all values within the interval, when in fact the belief about the true value is high close to the actual likelihood-ratio value which was calculated and tails off as one moves away from that value. The likelihood-ratio value calculated is still the best single-value estimate for the strength of evidence. One way to address this could be to calculate estimates of the credible interval over a range of percentile values, e.g., 50, 60, 70, 80, 90, 95, 99, 99.5, 99.9. A histogram of the deviation-from-mean values could also be plotted, especially in the case where the data can be assumed to be homoscedastic. Fig. 4 provides a histogram of the results from the example above for the 40 ms condition, with the 95% credible interval marked (for a non-technical introduction to histograms, see Morrison [12 §99.220]).

4. Probability of observing equally or more misleading evidence

In addition to the validity and reliability metrics described above, it is also possible to calculate for a specific likelihood-ratio value the probability of observing misleading evidence of equal or greater strength. For example, if the specific likelihood ratio obtained is 100, and there were 1000 test-pairs in the test set which are known to be different-origin comparisons and 5 of these had likelihood ratios equal to or greater than 100, then it could be reported that the probability of observing a likelihood ratio equal to or greater than 100 when the comparison pair has a different origin is 0.5%. In cases of likelihood ratios greater than one, this probability could be made more specific to the case at trial by having the different-origin comparison pairs each be a comparison between the sample of questioned origin form the case at trial and a sample from the test database (assuming the test database does not contain any samples from the same origin as the questioned-origin sample). For a fully-elaborated statistical approach to calculating the probability of observing misleading evidence, see Royall [35]. Arguably, the probability of observing equally or more misleading evidence is an error metric which can be explained to a statistically naïve trier of fact more easily than the metrics of validity and reliability described above.

Table 4
Measures of validity (log-likelihood-ratio cost) and reliability (95% credible interval in $\log_{10}(LR)$) of a forensic-voice-comparison system under two different conditions: 20 ms v 40 ms for questioned recording.

Metric	Condition	
	20 ms	40 ms
C_{llr}	0.200	0.212
95% CI	±1.58	±1.51

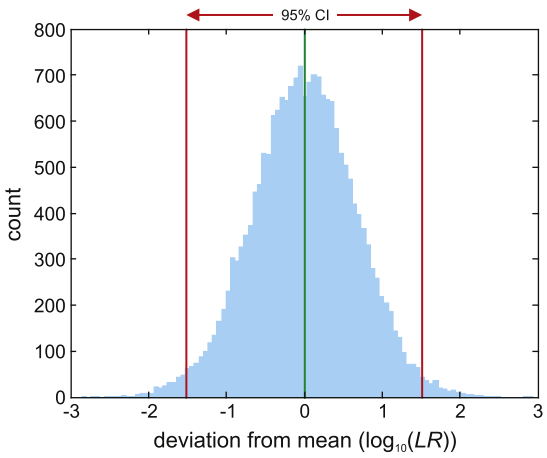

Fig. 4. Histogram of the deviation-from-mean values for the 40 ms condition, with the 95% credible interval marked.

5. The importance of metrics of validity and reliability

As the examples above illustrate, measurements of validity and reliability are important in research on developing forensic-comparison systems: a parametric-curve procedure outperforms a two-point measurement procedure in terms of validity and the former should therefore be employed rather than the latter; using longer voice recordings results in better reliability than shorter voice recordings and therefore longer recordings should be employed if reliability is the primary concern.

As a scientist, a forensic scientist should calculate estimates of the degree of validity and reliability of their forensic-comparison system under conditions as closely matched as possible to the conditions of the case at trial. These estimates of validity and reliability should them be presented to the judge to allow the judge to make a rational decision as to whether the scientific methods and procedures have been demonstrated to be of sufficient validity and reliability that evidence based on them should be admitted in court (see *Daubert* requirements). Even if the judicial system in a particular jurisdiction does not require forensic scientist to present measures of validity and reliability, this would still represent best scientific practice. If the courts are confused about the conventional scientific meanings of the terms *validity* and *reliability* (*accuracy* and *precision*), then, as a scientific expert, the forensic scientist should do their best to educate the courts on these matters.

It is important to present a measure of the reliability of the system as well as its validity, and reliability should be presented to the trier of fact during the trial, not just during an admissibility hearing. Not presenting the reliability of the forensic-comparison system under the conditions of the case at trial could seriously mislead the trier of fact. For example, if the likelihood ratio derived for the comparison of the known and questioned samples in a particular case were 1000 ($\log_{10}(LR)$ of +3) and tests of the system under conditions similar to those of the case indicated that the 95% credible interval was ±2 in $\log_{10}(LR)$, then the 95% credible interval in this case would be between likelihood ratios of 10 and 10 000 in favour of the same-origin hypothesis. It would be very misleading if the forensic scientist were simply to state that the likelihood of obtaining the evidence is 1000 times more likely under the same-origin hypothesis than under the different-origin hypothesis, since with no other information the trier of fact would probably take this as a precise estimate of the strength of evidence. In contrast, if the trier of fact is informed that the likelihood

ratio presented is a relatively imprecise estimate of the strength of evidence, the trier of fact may at their discretion decide to work with a more neutral value, i.e., closer to a likelihood ratio of one, and in this example could choose to use a value of say 10 from the bottom of the 95% credible interval. The trier of fact's choice of using 10 rather than 1000 for the estimate of the strength of evidence could have a substantial effect on their ultimate decision and might make the difference between them ultimately deciding guilty or not guilty, but it is a choice for the trier of fact to make not for the forensic scientist to make. Rather the forensic scientist should present all relevant information to assist the trier of fact in making their ultimate decision, including presenting the forensic scientist's best estimate of the strength of evidence and an estimate of the validity and reliability of that value given the measured accuracy and precision of their forensic-comparison system.

One of the reviewers pointed out that because of the work of Curran et al. [25,26] credible intervals for likelihood ratios derived from DNA-profile comparison are presented in court in New Zealand, but that in the United Kingdom it is thought they lead to an additional layer of complexity but little in the way of a difference in order of magnitude. Likelihood ratios derived from DNA profile comparisons can be very large and it may be that in practice even a two-order-of-magnitude difference has little effect on the trier of fact's ultimate decision, e.g., whether the trier of fact uses ten billion or a more conservative one-hundred million, this likelihood ratio may still be so large as to outweigh any other evidence. Given the data structure of voice data (and probably data in many other branches of forensic science), however, such large-magnitude likelihood ratios cannot be expected. Expectations of likelihood ratios on the order of hundreds or thousands are more reasonable. A likelihood ratio of 100 in favour of the same-origin hypothesis may, in combination with other evidence presented during a trial, be very helpful to the trier of fact when making their decision; however, if that likelihood ratio of 100 has a 95% credible interval of plus or minus two orders of magnitude then the lower bound of this interval is a likelihood ratio of 10 in favour of the different-origin hypothesis. The trier of fact may decide that the reliability of the likelihood ratio presented is so poor that they will not consider this evidence when making their final decision (the may decide that this likelihood ratio is not meaningfully different from one). Even if the situation is not so extreme, knowing about the precision of the likelihood-ratio estimate presented may substantially affect the trier of fact's ultimate decision. In forensic voice comparison, and probably many other branches of forensic science, it is therefore essential that measurements of the reliability of likelihood ratios be presented to the trier of fact.

6. Conclusion

Procedures for estimating the validity and reliability of forensic-comparison systems have been reviewed, and the importance of presenting measurements of validity and reliability discussed. These procedures have been applied in forensic voice comparison, but are also potentially applicable in other branches of forensic science.

References

[1] M.J. Saks, J.J. Koehler, The coming paradigm shift in forensic identification science, Science 309 (2005) 892–895, doi:10.1126/science.1111565.
[2] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, National Academies Press, Washington, DC, 2009.
[3] 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, Retrieved April 2009 from:.
[4] G.S. Morrison, Forensic voice comparison and the paradigm shift, Science & Justice 49 (2009) 298–308, doi:10.1016/j.scijus.2009.09.002.
[5] B. Robertson, G.A. Vignaux, Interpreting Evidence, Wiley, Chichester, UK, 1995.
[6] C.G.G. Aitken, F. Taroni, Statistics and the Evaluation of Forensic Evidence for Forensic Scientist, 2nd ed, Wiley, Chichester, UK, 2004.
[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] C. Champod, D. Meuwly, The inference of identity in forensic speaker recognition, Speech Communication 31 (2000) 193–203, doi:10.1016/S0167-6393(99)00078-3.
[10] P. Rose, Forensic Speaker Identification, Taylor and Francis London, UK, 2002.
[11] P. Rose, Technical forensic speaker recognition, Computer Speech and Language 20 (2006) 159–191, doi:10.1016/j.csl.2005.07.003.
[12] G.S. Morrison, Forensic voice comparison, in: I. Freckelton, H. Selby (Eds.), Expert Evidence, Thomson Reuters, Sydney, Australia, 2010, ch. 99.
[13] J.J. Koehler, Fingerprint error rates and proficiency tests: what are they and why do they matter, Hastings Law Journal 59 (2008) 1077–1100.
[14] N. Brümmer, J. du Preez, Application independent evaluation of speaker detection, Computer Speech and Language 20 (2006) 230–275, doi:10.1016/j.csl.2005.08.001.
[15] D.A. van Leeuwen, N. Brümmer, An introduction to application-independent evaluation of speaker recognition systems, in: C. Müller (Ed.), Speaker Classification I: Selected Projects, Springer-Verlag, Heidelberg, Germany, 2007, pp. 330–353.
[16] D. Ramos-Castro, J. González-Rodríguez, A. Montero-Asenjo, J. Ortega-García, Suspect-adapted map estimation of within-source distributions in generative likelihood ratio estimation. Proceedings of the IEEE Odyssey 2006 Speaker and Language Recognition Workshop, doi:10.1109/ODYSSEY.2006.248090.
[17] 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 Transactions on Audio, Speech, and Language Processing 15 (2007) 2104–2115, doi:10.1109/TASL.2007.902747.
[18] G.S. Morrison, Y. Kinoshita, Automatic-type calibration of traditionally derived likelihood ratios: forensic analysis of Australian English/o/formant trajectories, Proceedings of Interspeech 2008 Incorporating SST 2008, International Speech Communication Association, 2008, pp. 1501–1504.
[19] T. Thiruvaran, E. Ambikairajah, J. Epps, FM features for automatic forensic speaker recognition, Proceedings of Interspeech 2008 Incorporating SST 2008, International Speech Communication Association, 2008, pp. 1497–1500.
[20] G.S. Morrison, Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs, The Journal of the Acoustical Society of America 125 (2009) 2387–2397, doi:10.1121/1.3081384.
[21] E. Enzinger, Characterising formant tracks in Viennese diphthongs for forensic speaker comparison. Proceedings of the 39th Audio Engineering Society Conference - Audio Forensics: Practices and Challenges, Hillerød, Denmark.
[22] G.S. Morrison, Vowel inherent spectral change in forensic voice comparison, in G.S. Morrison, P. Assmann (Eds.), Vowel Inherent Spectral Change, Springer-Verlag, Heidelberg, Germany, in press.
[23] G.S. Morrison, Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English/aɪ/, International Journal of Speech, Language and the Law 15 (2008) 247–264, doi:10.1558/ijsll.v15i2.249.
[24] C.G.G. Aitken, D. Lucy, Evaluation of trace evidence in the form of multivariate data, Applied Statistics 54 (2004) 109–122, doi:10.1046/j.0035-9254.2003.05271.x.
[25] J.M. Curran, J.S. Buckleton, C.M. Triggs, B.S. Weir, Assessing uncertainty in DNA evidence caused by sampling effects, Science & Justice 42 (2002) 29–37, doi:10.1016/S1355-0306(02)71794-2.
[26] J.M. Curran, An introduction to Bayesian credible intervals for sampling error in DNA profiles, Law, Probability and Risk 4 (2005) 115–126, doi:10.1093/lpr/mgi009.
[27] P. Rose, G.S. Morrison, A response to the UK position statement on forensic speaker comparison, International Journal of Speech, Language and the Law 16 (2009) 139–163, doi:10.1558/ijsll.v16i1.139.
[28] A. Alexander, Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions. PhD Dissertation, École polytechnique fédérale de Lausanne, Lausanne, Switzerland, 2005.
[29] W.M. Bolstad, Introduction to Bayesian Statistics, 2nd Ed.Hoboken, NJ, Wiley, 2007.
[30] G.S. Morrison, C. Zhang, P. Rose, An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. Forensic Science International (published online 4 December 2010), doi:10.1016/j.forsciint.2010.11.001.
[31] G.S. Morrison, T. Thiruvaran, J. Epps, Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system, in: H. Cernocký, L. Burget (Eds.), Proceedings of Odyssey 2010: The Language and Speaker Recognition Workshop, International Speech Communication Association, Brno, Czech Republic, 2010, pp. 63–70.
[32] 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), Speech Communication 53 (2011) 242–256, doi:10.1016/j.specom.2010.09.005.
[33] G.S. Morrison, T. Thiruvaran, J. Epps, An issue in the calculation of logistic-regression calibration and fusion weights for forensic voice comparison, in: M. Tabain, J. Fletcher, D. Grayden, J. Hajek, A. Butcher (Eds.), Proceedings of the 13th Australasian International Conference on Speech Science and Technology, Australasian Speech Science and Technology Association, Melbourne, 2010, pp. 74–77.
[34] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models, Digital Signal Processing 10 (2000) 19–41, doi:10.1006/dspr.1999.0361.
[35] R. Royall, On the probability of observing misleading evidence, Journal of the American Statistical Association 95 (2000) 760–768. Stable URL:http://www.jstor.org/stable/2669456.