Science and Justice 58 (2018) 47–58

Contents lists available at ScienceDirect

Science and Justice

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

Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality

Geoffrey Stewart Morrison [a,b,c,*], Ewald Enzinger [d]

[a] Forensic Speech Science Laboratory, Centre for Forensic Linguistics, Aston University, Birmingham, England, United Kingdom
[b] Department of Linguistics, University of Alberta, Edmonton, Alberta, Canada
[c] Isaac Newton Institute for Mathematical Sciences, Cambridge, England, United Kingdom
[d] Eduworks, Corvallis, OR, United States

ARTICLE INFO

Article history:
Received 27 September 2016
Received in revised form 13 May 2017
Accepted 18 June 2017

Keywords:
Likelihood ratio
Similarity
Score
Anchored
Conversion
Calibration

ABSTRACT

Score based procedures for the calculation of forensic likelihood ratios are popular across different branches of forensic science. They have two stages, first a function or model which takes measured features from known-source and questioned-source pairs as input and calculates scores as output, then a subsequent model which converts scores to likelihood ratios. We demonstrate that scores which are purely measures of similarity are not appropriate for calculating forensically interpretable likelihood ratios. In addition to taking account of similarity between the questioned-origin specimen and the known-origin sample, scores must also take account of the typicality of the questioned-origin specimen with respect to a sample of the relevant population specified by the defence hypothesis. We use Monte Carlo simulations to compare the output of three score based procedures with reference likelihood ratio values calculated directly from the fully specified Monte Carlo distributions. The three types of scores compared are: 1. non-anchored similarity-only scores; 2. non-anchored similarity and typicality scores; and 3. known-source anchored same-origin scores and questioned-source anchored different-origin scores. We also make a comparison with the performance of a procedure using a dichotomous "match"/"non-match" similarity score, and compare the performance of 1 and 2 on real data.

1. Introduction

Score based procedures for the calculation of forensic likelihood ratios are popular across different branches of forensic science, including forensic comparison of voice recordings, ink, pharmaceutical tablets, digital camera images, handwriting, fingerprints, identity documents, gasoline residues, face images, and smokeless powders, e.g., [1–10].

Scores quantify the degree of similarity or the degree of difference between measurements made on pairs of objects, e.g., a questioned-source specimen and a known-source sample. Since differences and similarities are the inverse of one another we henceforth only refer to similarities (the argumentation would also apply to differences).

Score based procedures have two stages, first a function or model which takes the measured features as input and calculates a score as output, then a subsequent model which converts the score to a likelihood ratio (takes a score as input and calculates a likelihood ratio as output), see Fig. 1. The score-to-likelihood-ratio model is trained using training data consisting of a number of same-origin scores and a number of different-origin scores. These are generated by entering feature data from same-origin pairs and from different-origin pairs into the feature-to-score model.

Score based procedures may be used in situations where only one (univariate or multivariate) measurement is made on each source, e.g., [3,4,7,8]. The output of commercial biometrics systems (or other pattern recognition systems) may also be used as scores, e.g., [5,9,11], in which case the system generating the scores is often treated as a black box – the details of how the score is calculated may be a commercial secret. Score based procedures may also be used when the measurements made in the original feature space have a complex multidimensional and potentially multimodal distribution, e.g., [5,12–14]. In the latter case, scores can be considered a sort of projection of the complex multidimensional distribution in the feature space onto a simple unidimensional distribution in the score space. This necessarily implies a loss of information, but simple models with only a few parameter values to be fitted in the score space tend to produce much better calibrated likelihood ratio results than complex models requiring many parameter values to be fitted in the original feature space.

The present paper argues that scores which are measures of similarity are not appropriate for calculating forensically interpretable likelihood ratios. In addition to taking account of similarity between a questioned-source specimen and a known-source sample, scores must also take account of the typicality of the questioned-source specimen

* Corresponding author at: Forensic Speech Science Laboratory, Centre for Forensic Linguistics, Aston University, Birmingham, England, United Kingdom.
E-mail address: geoff-morrison@forensic-evaluation.net (G.S. Morrison).

http://dx.doi.org/10.1016/j.scijus.2017.06.005

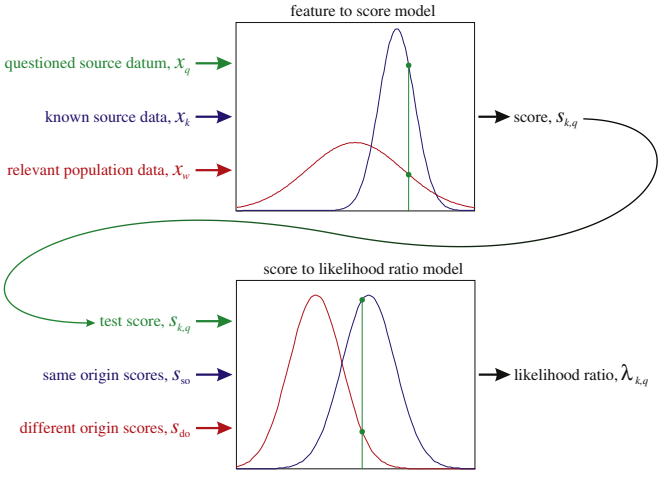

Fig. 1. Schematic of a score based procedure consisting of a feature-to-score model followed by a score-to-likelihood-ratio model. The procedure in this example uses scores which take account of both similarity and typicality. (The images in the boxes are illustrative only, and are not intended to represent feature and score distributions based on actual data.)

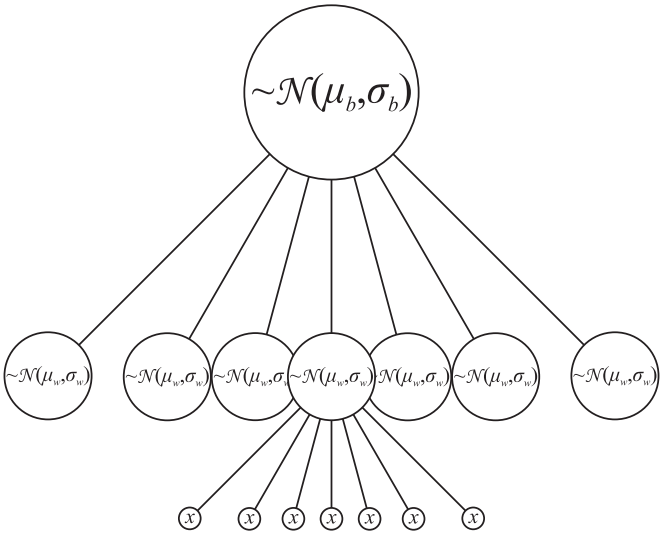

Fig. 2. Structure of the specified Monte Carlo distribution for the relevant population.

with respect to a sample of the relevant population specified by the defence hypothesis. A forensic likelihood ratio is the answer to a specific question specified by the prosecution and defence hypotheses. Usually the prosecution hypothesis is that the questioned-source specimen has the same origin as the known-source sample, and usually the defence hypothesis is that the questioned-source specimen originated not from the known source but from some other source selected at random from the relevant population.

The present paper demonstrates that likelihood ratios calculated via scores will not appropriately reflect typicality with respect to the relevant population if the score is a similarity-only score. Appropriate accounting for typicality cannot be introduced at the score to likelihood ratio conversion stage. The score itself must take account of typicality with respect to the relevant population.[1]

We provide an empirical demonstration based on specified Monte Carlo distributions.[2] The distributions of the feature values for the relevant population and for the known sources are completely specified. This allows us to calculate reference values for likelihood ratios over a range of questioned-source feature values. Since in the Monte Carlo simulation these reference values are based on completely known distributions, they can be considered "true" likelihood ratio values. In contrast, apart from trivial cases, in real life one cannot know the true distributions and hence one cannot know what the true values of likelihood ratios would be.

We draw simulated samples from the specified Monte Carlo distributions, and use these to calculate likelihood ratios via different procedures. We then compare the likelihood ratio values calculated by each procedure with the reference likelihood ratio values. This is similar to the analysis presented in Appendix A of [5], although the latter used a closed-form analytical approach rather than Monte Carlo simulation and also differed in other details.[3]

We begin with a direct calculation from feature values to likelihood ratios using generative models which correspond to the known structure of the data. This provides a baseline against which to compare score-based procedures. We explore the performance of multiple score-based procedures, each using a different type of score, but which are otherwise identical. For simplicity, all the score-based procedures use the same score to likelihood ratio model (the non-parametric pool adjacent violators, PAV, algorithm).

We use simple Monte Carlo distributions for clarity and tractability. Actual forensic data usually have more complex distributions to which the results in the present paper may not be immediately applicable. Any problems observed when procedures are applied to the simple distributions may, however, be cause for concern regarding the applicability of those procedures to more complex distributions.

The different types of scores tested in the present papers are:

- similarity-only scores (non-anchored)
- similarity-and-typicality scores (non-anchored)
- known-source anchored same-origin scores and questioned-source anchored different-origin scores

We will describe each in detail in Section 7.[4]

Appendix A presents a comparison of the application of similarity-only scores and similarity-and-typicality scores to real (i.e., not Monte Carlo simulated) data reflecting the conditions of an actual forensic voice comparison case. Since the reference values for likelihood ratios are unknown for the real data, performance is assessed by comparing likelihood ratio values with knowledge about whether each test pair was a same-origin or a different-origin pair. Results are presented as Tippett plots and log likelihood ratio cost (C_{llr}).

The 2016 forensic science report by President Obama's Council of Advisors on Science and Technology (PCAST) advocated the use of a procedure in which the first stage is "match"/"non-match" [16]. This can be considered a dichotomous similarity-only score, which can be the basis for the calculation of a likelihood ratio. In Appendix B, the output of such a procedure is compared with reference likelihood ratio values.

Appendix C compares the performance of the different procedures in terms of C_{llr}.

[1] [15] ch 4 demonstrates that if the model used to calculate scores takes account of typicality with respect to some population other than the relevant population specified by the defence hypothesis (i.e., is trained using data sampled from some other population), the performance is much worse than if this model takes account of typicality with respect to the relevant population (i.e., is trained using data sampled from the relevant population). In both cases, pairs of samples/specimens sampled from the relevant population were used to train the score to likelihood ratio conversion model.

[2] The Matlab code for this demonstration is provided at http://geoff-morrison.net/#scorebased2016.

[3] One difference is that, unlike in the present study, [5] always used known-source-anchored scores to calculate the numerator of their score-based likelihood ratios.

[4] For simplicity we only demonstrate the use of three types of scores in the present paper. [17] demonstrated the use of additional types of scores.

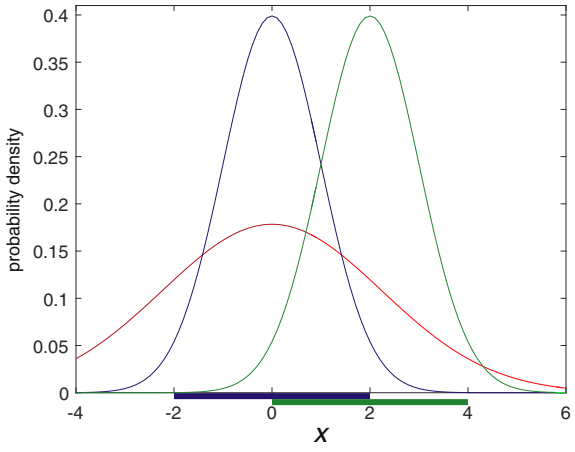

Fig. 3. The specified relevant population's feature distribution (red curve) and the two known sources' specified feature distributions (blue and green curves). The ranges of questioned specimen values compared against each known source are also indicated (blue and green horizontal bars).

2. Monte Carlo distributions

The Monte Carlo distribution for the relevant population is specified as follows, see also Fig. 2. The relevant population has a univariate Gaussian distribution with a grand mean, μ_b, of 0, and a between source standard deviation, σ_b, of 2. Each source has a mean, μ_w, inherited from the between source distribution and a within source standard deviation, σ_w, of 1.[5] Assuming independent sources of variance, the features, x, therefore have a Gaussian distribution which is the sum of the between source and within source distributions, i.e., a mean of 0 and a standard deviation, σ_{b+w}, as in Eq. (1):

$$\sigma_{b+w} = \sqrt{\sigma_b^2 + \sigma_w^2} = \sqrt{2^2 + 1^2} = \sqrt{5} \approx 2.24 \tag{1}$$

The distributions of two ($K=2$) known sources are specified as follows. One has a mean, $\mu_{k=1}$, of 0 and the other has a mean, $\mu_{k=2}$, of 2. Each has a within-source standard deviation, σ_k, of 1.

The relevant population's feature distribution and the two known sources' feature distributions are represented graphically in Fig. 3. Note that the first known source is very typical, its mean corresponds with the population mean, whereas the second known source is relatively atypical, its mean is offset from the population mean.

3. Test probes

A series of questioned source feature values, x_q,[6] are specified as test probes. These range from -2 to $+2$ for comparison with the first known source, and 0 to $+4$ for comparison with the second known source. Each consists of a set of 41 point values ($Q=41$) spaced 0.1 units apart.

4. Reference likelihood ratio values

Reference likelihood ratio values, $\Lambda_{k,q}$, were calculated as follows. For each combination of a questioned-source feature value, x_q, and a known-source model with a specified mean and standard deviation, μ_k and σ_k, the likelihood of the model was calculated. This was used as the numerator of the likelihood ratio, see Eq. (2). For each questioned-source feature value, x_q, and the relevant population model with specified mean and standard deviation, μ_b and σ_{b+w}, the likelihood of the model was calculated. This was used as the denominator of the likelihood ratio, see Eq. (2). The likelihood ratio value was calculated by dividing the former likelihood by the latter, see Eq. (2).

$$\Lambda_{k,q} = \frac{f(x_q|\mu_k, \sigma_k)}{f(x_q|\mu_b, \sigma_{b+w})} \tag{2}$$

where $f(x|\mu,\sigma)$ is the likelihood of a univariate Gaussian distribution with mean μ and standard deviation σ evaluated at x.

Fig. 4 shows the reference likelihood ratio values. The short vertical lines represent the means of the known sources ($\mu_{k=1}=0$ and $\mu_{k=2}=2$). The curves over $x\in\{-2,\ldots,+2\}$ and $x\in\{0,\ldots,+4\}$ represent the $\log_{10}$ likelihood ratios corresponding to the first and second known sources and their corresponding questioned-source values, x_q. Note that since the first source has the same mean as the relevant population mean ($\mu_k=\mu_b$) its curve is symmetrical about its mean, with the highest likelihood ratio value being at its mean (leftmost curve in Fig. 4). In contrast, since the mean of the second source is off centre compared to the mean of the relevant population mean ($\mu_k>\mu_b$), its curve is not symmetrical about its mean (rightmost curve in Fig. 4). The highest likelihood ratio value is to the right of the second source's mean. Except under specific circumstances (e.g., in the first example where $\mu_k=\mu_b$), maximum similarity, i.e., maximum likelihood in the numerator of the likelihood ratio (when $x_q=\mu_k$), does not imply maximum likelihood ratio.

5. Training data

Monte Carlo samples from univariate Gaussian distributions were simulated using a pseudo-random number generator. For each sample set, 300 source means, μ_w, $w\in\{1,\ldots,300\}$, were generated using the mean, μ_b, and standard deviation, σ_b, specified for the between source distribution. For each source within the sample set, 90 tokens, $x_{w,r}$, $r\in\{1,\ldots,90\}$, were generated using its mean, μ_w, and the standard deviation, σ_w, specified for the within source distribution. The choice of 300 sources and 90 tokens per source was arbitrary. The simulations can easily be repeated with different sample sizes. Reducing sample sizes leads to less reliable output, however, the relative performance of the different procedures is unchanged.

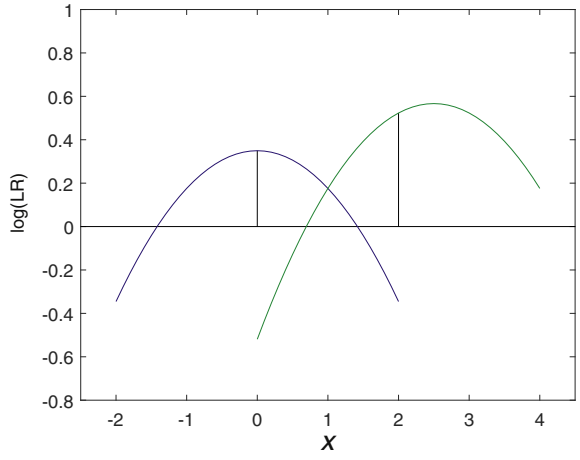

Fig. 4. Reference likelihood ratio values. The blue and green curves represent the reference $\log_{10}$ likelihood ratio values corresponding to the first and second known sources and their respective ranges of questioned origin test values. The black vertical lines represent the means of the two known sources.

[5] Adding a layer of randomization to induce a different within source standard deviation for each source resulted in poorer performance for all procedures, but did not alter the pattern of relative performance of the different procedures. For simplicity, in the present paper we only report the results from a Monte Carlo distribution in which each and every source has the same standard deviation.

[6] We will use the notation x_q to refer to any questioned-source feature value irrespective of its source. We will use the notation to x_{k_q} to refer to a questioned-source feature value whose source is the known source k.

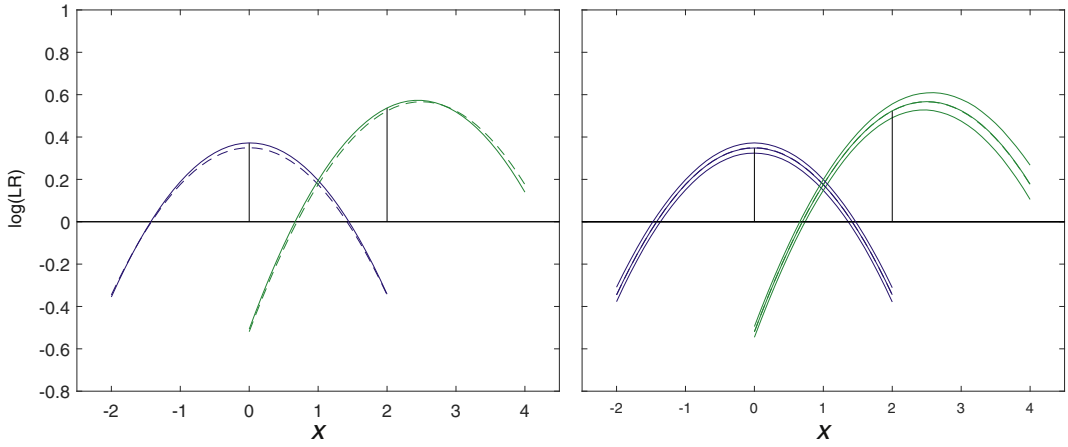

Fig. 5. Likelihood ratio values calculated on Monte Carlo sample data using the baseline direct feature to likelihood ratio procedure. The left panel shows the results from one sample set, and the right panel shows the 5th, 50th, and 95th percentiles across all 1000 sample sets. The solid lines represent the $\log_{10}$ likelihood ratio values calculated on the sample data, and the dashed lines represent the reference $\log_{10}$ likelihood ratio values.

Some additional data are needed for the score-based approaches: For non-anchored score procedures (Sections 7.1, 7.2), one additional sample, x_{w_r}, was generated for each source in the same way as for the first 90 samples. For the anchored score procedure (Section 7.3), an additional set of 90 samples, $x_{k'_r}$, $r \in \{1, \ldots, 90\}$, was generated for each known source. The use of these additional data will be described in Sections 7.1, 7.2, and 7.3 below.

One thousand sample sets were generated.

6. Direct likelihood ratio calculation procedure – baseline performance

6.1. Procedures

Likelihood ratio values, $\lambda_{k,q}$, were calculated from the sample (training) data using generative models which had the same structure as the specified Monte Carlo distributions. These will be used to establish a baseline level of performance.

The calculations, see Eq. (3), were the same as for the calculation of the reference values, except that the value of the known-source standard deviation was the pooled within-source standard deviation, $\hat{\sigma}_{\text{pooled}}$,[7] calculated using sample data, x_{w_r}, from all 300 sources in the training set, and the values of the relevant-population mean and standard deviation, $\hat{\mu}_b$ and $\hat{\sigma}_{b+w}$, were calculated using all the sample data, x_{w_r}, without reference to source.

$$\lambda_{k,q} = \frac{f\left(x_q|\mu_k, \hat{\sigma}_{\text{pooled}}\right)}{f\left(x_q|\hat{\mu}_b, \hat{\sigma}_{b+w}\right)} \tag{3}$$

6.2. Results

Fig. 5 graphically represents the resulting likelihood ratio values. The left panel shows the results from one sample set, and the right panel shows the 5th, 50th, and 95th percentiles across all 1000 sample sets (these were calculated pointwise, i.e., independently for each combination of μ_k and x_q). The closeness of the 50th percentile curve to the reference likelihood ratio curve and the distance between the 5th and 95th percentile curves provide a baseline for the performance (accuracy and precision respectively) that can be expected from these sample data.

Root mean squared (RMS) errors, ε_{RMS}, were calculated based on the distance of each calculated likelihood ratio from its corresponding reference value, see Eq. (4).[8]

$$\varepsilon_{\text{RMS}} = \sqrt{\frac{1}{K}\sum_k \frac{1}{Q}\sum_{k_q}\left(\log_{10}\left(\Lambda_{k,k_q}\right) - \log_{10}\left(\lambda_{k,k_q}\right)\right)^2} \tag{4}$$

A separate RMS error was calculated for each sample set. Fig. 6 graphically represents the distribution of the RMS errors over all 1000 sample sets, both for the baseline procedure described in the present section and for the score-based procedures described below.

7. Score based procedures

7.1. Similarity-only scores (non-anchored)

Similarity-only scores include Manhattan distance, Euclidian distance, Pearson correlation, and Kullback–Leibler divergence, e.g., [2–5,7,18–21].

In the present paper we demonstrate the performance of one type of similarity-only score.[9] In this example, a score, $s_{k,q}$, is calculated as the likelihood of a known-source model, k, evaluated at a specified questioned-source value, x_q. The known-source model is a univariate Gaussian distribution with mean, μ_k, and a standard deviation which is the pooled within source standard deviation, $\hat{\sigma}_{\text{pooled}}$, calculated using sample data, x_{w_r}, from all 300 sources in the sample set. The similarity-only score (Eq. (5)) is therefore the same as the numerator of the likelihood ratio calculated using the direct procedure that was described in Section 6.1.

$$s_{k,q} = f\left(x_q|\mu_k, \hat{\sigma}_{pooled}\right) \tag{5}$$

Scores for training the score to likelihood ratio conversion model were generated as follows. The pooled within source standard deviation, $\hat{\sigma}_{\text{pooled}}$, was calculated using sample data, x_{w_r}, from all 300 sources in the training set. For each sample source, w, the x_{w_r} sample data were

[7] For all procedures for which there was the option to use either a pooled calculation of standard deviation or a separate standard deviation for each source, the latter resulted in poorer performance (even when a Monte Carlo distribution that had a different standard deviation for each source was used), but did not alter the pattern of relative performance of the different procedures. For simplicity, in the present paper, whenever it is an option, we only report the results from the version of each procedure which uses a pooled calculation of standard deviation.

[8] A different set of questioned-origin specimens is associated with each known-source sample, hence Eq. (4) includes a subscript on the subscript: $_{k_q}$. The likelihood ratio values corresponding to when $x_{k_q} = \mu_k$ were excluded from the RMS error calculation. The reason for this is explained in Section 8.

[9] Logically, the same pattern of results would be observed for other types of similarity-only scores. [17] includes results from a distance score.

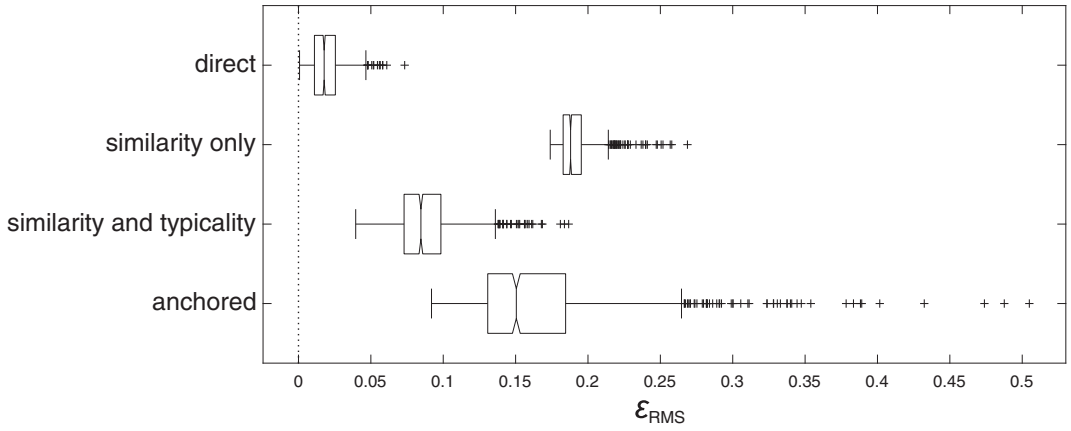

Fig. 6. Boxplot representations of distributions of RMS errors over all 1000 sample sets, for the baseline direct procedure and for each score-based procedure.

used to calculate a mean, $\hat{\mu}_w$. A same-origin score, s_{so}, was calculated for each sample source w by evaluating the likelihood of a univariate Gaussian distribution with mean $\hat{\mu}_w$ and standard deviation $\hat{\sigma}_{\text{pooled}}$ at the value of the additional sample from the same source x_{w_r}, see Eq. (6)a. A different-origin score, s_{do}, was calculated for each possible pair of different-origin sources in the training set by evaluating the likelihood of each source model at the additional sample value, x_{w_r}, from every other source in the sample set, see Eq. (6b).

$$s_{\text{so}} = f\left(x_{w_{1r}}|\hat{\mu}_{w_2}, \hat{\sigma}_{pooled}\right)\Big|_{w_1=w_2} \tag{6a}$$

$$s_{\text{do}} = f\left(x_{w_{1r}}|\hat{\mu}_{w_2}, \hat{\sigma}_{pooled}\right)\Big|_{w_1\neq w_2} \tag{6b}$$

7.2. Similarity-and-typicality scores (non-anchored)

A similarity-and-typicality score, as the name implies, takes account of both similarity and typicality. Use of this type of score is common in forensic comparison of voice recordings, e.g., [1,12,13] [15] ch. 4, [22], and it has also been used in forensic comparison of face images, e.g., [23]. A similarity-and-typicality score can be considered an attempt to calculate a likelihood ratio, but the resulting value is not treated as an interpretable forensic likelihood ratio answering the question posed by the prosecution and defence hypotheses. The attempt to calculate a likelihood ratio may have used a small amount of data to estimate a large number of parameter values in a complex (e.g., multimodal and multi-dimensional) model, or may have violated modelling assumptions, thus there is reason not to trust the calculated likelihood ratio value.[10] Such values are therefore treated as scores and subjected to a score to likelihood ratio conversion procedure.[11]

Scores were calculated as in Eq. (7) (which is the same as for calculating likelihood ratios using the direct procedure).

$$s_{k,q} = \frac{f\left(x_q|\mu_k, \hat{\sigma}_{pooled}\right)}{f\left(x_q|\hat{\mu}_b, \hat{\sigma}_{b+w}\right)} \tag{7}$$

A set of same-origin and different-origin scores for training the score to likelihood ratio conversion model were calculated in the same manner as described in Section 7.1, see Eq. (8).

$$s_{\text{so}} = \frac{f\left(x_{w_{1r}}|\hat{\mu}_{w_2}, \hat{\sigma}_{pooled}\right)}{f\left(x_{w_{1r}}|\hat{\mu}_b, \hat{\sigma}_{b+w}\right)}\Bigg|_{w_1=w_2} \tag{8a}$$

$$s_{\text{do}} = \frac{f\left(x_{w_{1r}}|\hat{\mu}_{w_2}, \hat{\sigma}_{pooled}\right)}{f\left(x_{w_{1r}}|\hat{\mu}_b, \hat{\sigma}_{b+w}\right)}\Bigg|_{w_1\neq w_2} \tag{8b}$$

7.3. Known-source anchored same-origin scores and questioned-source anchored different-origin scores

With respect to the similarity-only scores and the similarity-and-typicality scores described above, the scores for training the score to likelihood ratio conversion model were non-anchored. These scores were calculated using pairs of samples taken from the relevant population. Neither the known-source data nor the questioned-source data played any part in the calculation of these training scores. We now describe a procedure in which the same-origin scores are anchored on the known source and the different-origin scores are anchored on the questioned source (the former can be considered similarity scores and the latter typicality scores).[12] Versions of the procedure are described in [5,26–28].

Same-origin scores for training the score to likelihood ratio conversion model were calculated as follows. For a known-source test sample, a univariate Gaussian distribution model was calculated using its specified mean μ_k and the standard deviation $\hat{\sigma}_{\text{pooled}}$ calculated using the training data, x_{w_r}. The likelihood of this model was evaluated at each of the 90 training samples, $x_{k'_r}$, associated with that test source, see Eq. (9). This resulted in a set of 90 same-origin scores anchored on the known source.

$$s_{\text{so},k,r} = f\left(x_{k'_r}|\mu_k, \hat{\sigma}_{pooled}\right) \tag{9}$$

Different-origin scores for training the score to likelihood ratio conversion model were calculated as follows. For each of the 300 sources in the sample of the relevant population, a univariate Gaussian distribution model with mean $\hat{\mu}_w$ (calculated from the x_{w_r} values for that source) and standard deviation $\hat{\sigma}_{\text{pooled}}$ was calculated. The likelihood of each of these models was evaluated at a questioned source value, x_q, see Eq.

[10] For example, in forensic voice comparison common procedures for calculating scores [24,25] use Gaussian mixture models with usually 1024 Gaussian components fitted to feature vectors which can have 32 or more dimensions. Feature vectors are typically extracted every 10 ms from voice recordings which may be tens of seconds or several minutes long, thus there are usually tens of thousands of vectors, but the vectors are not random samples and there is correlation between feature vectors which are proximal in time.

[11] The output of the attempt to calculate a likelihood ratio can alternatively be considered an uncalibrated likelihood ratio, and the second level of modelling considered a calibration procedure.

[12] Other anchoring schemes are attested in the literature, [5,17,28].

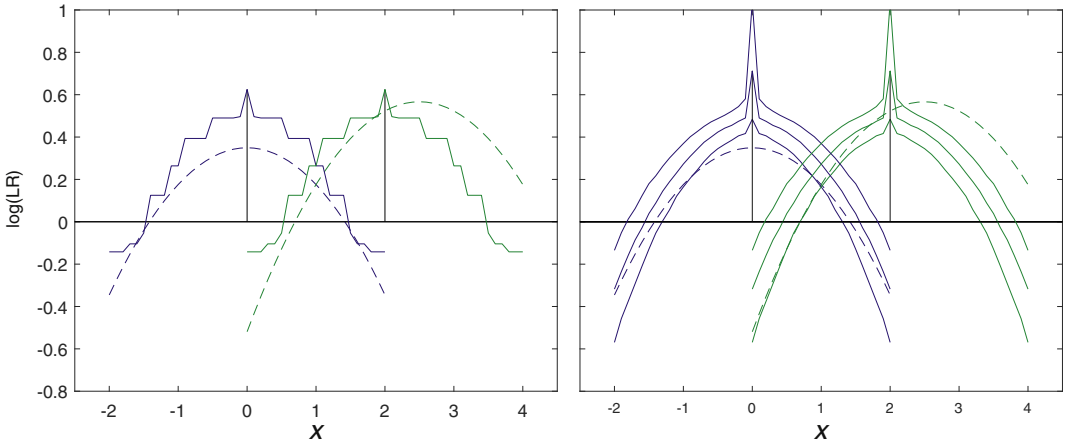

Fig. 7. Likelihood ratio values calculated on Monte Carlo sample data using the similarity-only score procedure.

(10). This resulted in a set of 300 different-origin scores anchored on the questioned source.

$$s_{\mathrm{do},w,q} = f\left(x_q|\hat{\mu}_w, \hat{\sigma}_{pooled}\right) \tag{10}$$

A score, $s_{k,q}$, for the combination of the known source and questioned source was calculated as in Eq. (11).

$$s_{k,q} = f\left(x_q|\mu_k, \hat{\sigma}_{pooled}\right) \tag{11}$$

The distributions of the same-origin scores $s_{\mathrm{so},k,r}$ and of the different-origin scores $s_{\mathrm{do},w,q}$ coincide at only one point: the score for the comparison of the known source and the questioned source, $s_{k,q}$ (compare μ_k in Eqs. (9), (11), and x_q in Eqs. (10), (11)). A separate set of 90 same-origin scores and 300 different-origin scores for training the score to likelihood ratio conversion model had to be calculated for each combination of known source and questioned source.

7.4. Score to likelihood ratio conversion

Several different models have been proposed for score to likelihood ratio conversion, including kernel density models, logistic regression, Gaussian models with equal variances, Gaussian models with separate variances, and the pool adjacent violators (PAV) algorithm, e.g., [12,13,17,29–31]. Perhaps the simplest model to understand is one which fits two generative distributions, one to the same-origin scores and another to the different-origin scores, then evaluates the likelihoods of these two Gaussians at the score value derived from the comparison of the known and questioned sources, $s_{k,q}$, see Fig. 1. Since types of score are the focus of the present paper, not models for converting from scores to likelihood ratios, we use only one model. The model we use is PAV, which is a non-parametric model that can be applied irrespective of the actual distribution of the scores (this is not the model in the illustrations in Fig. 1). As a non-parametric model it is susceptible to overfitting on the training data and not generalising well to new data. We therefore do not recommend it for calculation of likelihood ratios in real applications, but in the present demonstration it has the advantage of being distribution neutral and thus not a priori favouring some types of scores over others. The PAV algorithm is described in [32–36]. We used the implementation in Brümmer's FoCal Toolkit.[13]

For each type of score, the same-origin and different-origin scores, s_{so} and s_{do}, were used to train the score to likelihood ratio conversion model. This model was then used to convert each score from the comparison of a known source and a questioned source, $s_{k,q}$, to a likelihood ratio, $\lambda_{k,q}$. The PAV algorithm provides a mapping from each score value in the training set to a corresponding likelihood ratio value. To convert an $s_{k,q}$ value from the test set to a $\lambda_{k,q}$ value, first the training set $s_{k,q}$ value closest to the test set $s_{k,q}$ value was found (closeness was assessed using absolute distance in a natural log score space), then the $\lambda_{k,q}$ value corresponding to that training set $s_{k,q}$ value was used.

8. Results and discussion

The likelihood ratio values, $\lambda_{k,q}$, calculated using the score based procedures were compared to the reference likelihood ratio values, $\Lambda_{k,q}$, in the same way as was done for the baseline likelihood ratio values calculated using the direct procedure (see Section 6.2). Graphical representations of the likelihood ratio value results are given in Figs. 7, 8, 9. The distributions of RMS error results are shown in Fig. 6.

Note, in Fig. 7, that the likelihood ratios calculated using similarity-only scores are in general far from the reference likelihood ratio values. The curves associated with the first and second known sources are simply shifted versions of each other, shifted on the x axis. The curve associated with the second known source does not display the same asymmetry as seen for the reference values. The procedure based on similarity-only scores has not taken account of the fact that the second known source and its associated questioned-source values are more atypical than the first known source and its associated questioned-source values.

In contrast, in Fig. 8, the likelihood ratio values calculated using non-anchored similarity-and-typicality scores are closer to the reference likelihood ratio values, and the curve associated with the second known source shows the same asymmetry as seen in the reference values. The procedure based on non-anchored similarity-and-typicality scores has taken account of the fact that the second known source and its associated questioned-source values are more atypical than the first known source and its associated-questioned source values.[14]

As seen in Fig. 9, the likelihood ratio values calculated using the known-source anchored same-origin scores and questioned-source anchored different-origin scores deviate more from the reference likelihood ratio values than do those calculated using the non-anchored similarity-and-typicality scores. Performance could potentially be improved by increasing the number of tokens in the additional sample from the known source ($x_{k'_r}$), but the fact that an additional known-source sample is needed at all is already a practical disadvantage and

[13] http://www.dsp.sun.ac.za/nbrummer/focal.

[14] The similarity-and-typicality scores used to train and test this score-to-likelihood-ratio conversion model (aka calibration model) were actually well calibrated likelihood ratios to begin with. Post calibration performance is worse than pre calibration performance because when the original test scores are actually well calibrated and different data are used for training and testing, sampling variability between the training and test samples leads to worse results on the particular test data.

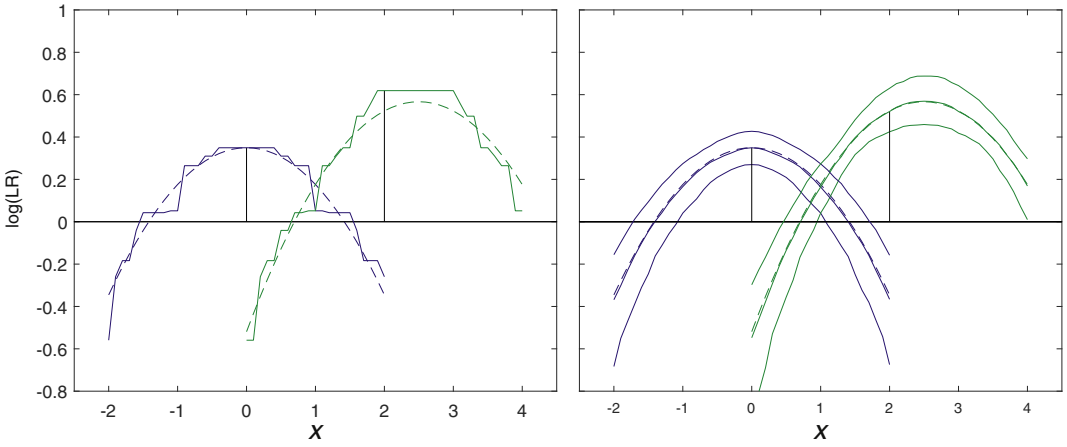

Fig. 8. Likelihood ratio values calculated on Monte Carlo sample data using the similarity-and-typicality score procedure.

further increasing the size of that sample may be even more impractical. In real life, such data may simply not be available.[15] [37] also criticised the known-source anchored same-origin scores and questioned-source anchored different-origin scores procedure as not theoretically justified.

Comparison of the RMS error distributions in Fig. 6 shows that only the output of the procedure based on non-anchored similarity-and-typicality scores comes anywhere near the output of the direct baseline procedure.[16]

[5] set out to demonstrate that score-based likelihood ratios are not the same as likelihood ratios calculated directly on the original features, and that they cannot be interpreted as such. They did not, however, examine exactly the same type of similarity-and-typicality score as examined here – a score which is essentially an attempt to calculate a likelihood ratio. Although the demonstration presented here is limited, we interpret the results as indicating that likelihood ratios calculated using non-anchored similarity-and-typicality scores may reasonably be considered sufficiently close to the reference likelihood ratio values that they may be interpreted as meaningful answers to the original question posed by the prosecution and defence hypotheses, the question which is answered directly by the baseline procedure, i.e.:

> What is the probability of obtaining the measured properties of the questioned-source specimen if it came from the known source?
> versus
> What is the probability of obtaining the measured properties of the questioned-source specimen if it came from a source selected at random from the relevant population?

A procedure based on similarity-only scores clearly asks a different question to that above, hence it would be expected to give a different answer. The non-anchored similarity-only score procedure asks:

> What is the probability of obtaining the degree of similarity between the measured properties of the questioned-source specimen and the measured properties of the known-source sample if the questioned-source specimen and known-source sample came from the same source, a source selected at random from the relevant population?
> versus
> What is the probability of obtaining the degree of similarity between the measured properties of the questioned-source specimen and the measured properties of the known-source sample if the questioned-source specimen and the known-source sample each came from a different source each of which was selected at random from the relevant population?

It could therefore be argued that it is not appropriate to compare the output of a similarity-only score procedure with reference likelihood ratio values as we have done here. It could be argued that similarity-only scores are a feature and will produce different likelihood ratio values compared to when other features are used, in the same way that if the widths instead of the lengths of objects were measured, different likelihood ratios values would result (unless there were 100% correlation between the widths and the lengths). This is, however, a false analogy. The decision to use similarity-only scores is not a choice of feature, a similarity-only score is a derivative of a feature that has already been measured (or features that have already been measured). Procedures based on similarity-only scores do not appropriately account for typicality with respect to the relevant population. In our opinion, this is a fundamental problem that should not be swept under the carpet by reformulating the question.

The focus in the present paper has been on comparison of calculated likelihood ratio values with "true" reference values. Procedures for calculating likelihood ratios which account for both similarity and typicality have, however, also been shown to outperform procedures based only on similarity when performance is assessed in relation to whether each test trial is a same-origin or a different-origin pair. Such results were found for classification-error rates reported in [14],[17] and, in a non-forensic context, area under receiver operating curve reported in [38] ch. 3. Appendix A of the present paper reports log likelihood ratio cost (C_{llr}) results for similarity-only scores and similarity-and-typicality scores derived from real data. Appendix C of the present paper reports C_{llr} results for the direct procedure and all three score-based procedures (plus a dichotomous similarity-only score procedure, see Appendix B) using the same simulated data as used above.

9. Conclusion

In a Monte Carlo simulation, a score based procedure for likelihood ratio calculation based on similarity-only scores produced likelihood ratio values which were far from reference likelihood ratio values calculated on the basis of complete knowledge of relevant population and known-source distributions. A procedure based on known-source

[15] A potential alternative would be to assume within-source homogeneity and use non-anchored same-origin scores versus questioned-specimen anchored different-origin scores.

[16] In Fig. 6, for the similarity-only procedure and for the known-source anchored and questioned-source anchored procedure, there are spikes in the likelihood ratio values when $x_{q_k} = \mu_k$. These are likely artefacts of the combination of the PAV algorithm with these procedures. The $s_{k,q}$ value at this point is often greater than any value encountered in the training data, and the PAV algorithm returns its maximum value. So as not to bias the RMS error results, these likelihood ratio values were excluded from the RMS error calculations for all procedures.

[17] For continuously-valued data, [14] compared the performance of a procedure based on distance-only scores with a procedure that directly calculated likelihood ratios. Both synthetic and real data were tested.

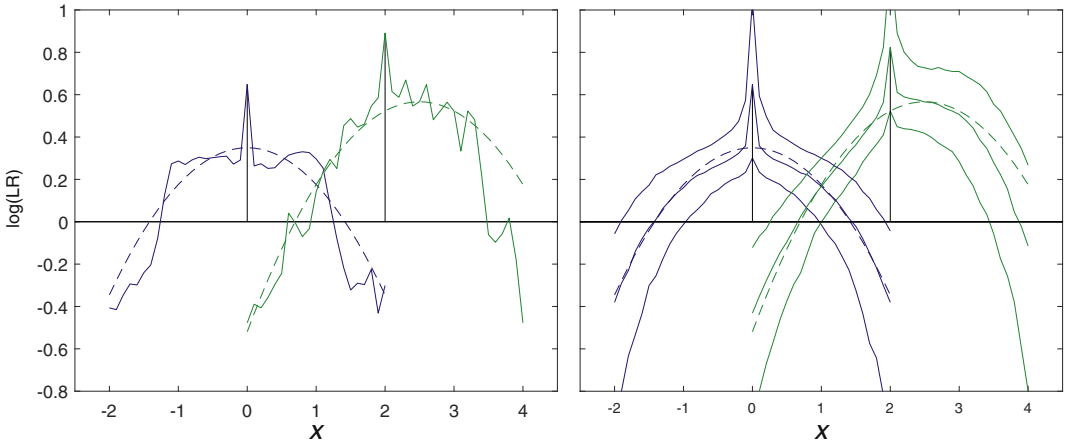

Fig. 9. Likelihood ratio values calculated on Monte Carlo sample data using the known-source anchored same-origin scores and questioned-source anchored different-origin scores procedure.

anchored same-origin scores and questioned-source anchored different-origin scores has practical disadvantages and produced results that were further from reference likelihood ratio values than a procedure based on non-anchored similarity-and-typicality scores. When using a score-based procedure for calculating likelihood ratios, we therefore recommend that the scores themselves take account of both similarity and typicality, and that non-anchored similarity-and-typicality scores be used.

Acknowledgement

This work was funded in-part by a fellowship from the Simons Foundation awarded to the first author. The first author would like to thank the Isaac Newton Institute for Mathematical Sciences for its hospitality during the programme Probability and Statistics in Forensic Science which was supported by EPSRC Grant Number EP/K032208/1.

Appendix A. Comparison of score-based procedures applied to real data

This appendix compares the performance of similarity-only scores and similarity-and-typicality scores as part of procedures applied to real data, i.e., not Monte Carlo simulated data (there were insufficient data to be able to apply a procedure involving known-source anchoring). The data are from a forensic voice comparison analysis performed under conditions reflecting those of an actual forensic case. The questioned-source recording was of a telephone call. It included background office noise and signal compression for storage. The known-source recording was of a police interview. It included room reverberation and background ventilation system noise. The analyses reported here are adapted from data and analytical procedures previously reported in [22]. Additional details can be found in that reference, and the description below is deliberately terse – the intent is to focus on the question at hand, and minimise details which are tangential to that question.

A.1. Data and methodology

The feature-level data consisted of 28 dimensional vectors (mel frequency cepstral coefficients, MFCCs, plus deltas, to which feature warping and probabilistic feature mapping had been applied). The statistical models used were Gaussian mixture models (GMMs) with 512 Gaussian components, each with a diagonal-only covariance matrix.

A relevant-population GMM was trained using data pooled from recordings of 105 speakers, one recording per speaker, 10,453 vectors per recording.

Data from an additional 61 speakers, one questioned-source-condition recording per speaker, 4137 feature vectors per recording, and two or three known-source-condition recordings per speaker, 10,453 feature vectors per recording (total 61 questioned-source-condition and 162 known-source-condition recordings), were used to generate training scores. The multiple known-source-condition recordings from a speaker came from separate recording sessions separated by approximately a week. The questioned-source-condition recordings came from the first recording session. For each known-source-condition recording from each speaker, a known-source GMM was trained via

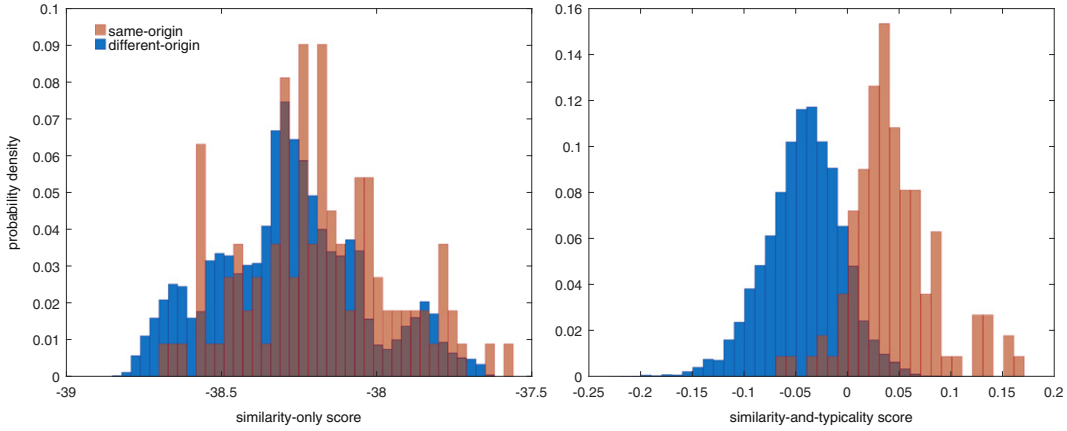

Fig. 10. Histograms of distributions of same-origin and different-origin test scores from the forensic voice comparison analysis. Left panel: similarity-only scores. Right panel: similarity-and-typicality scores.

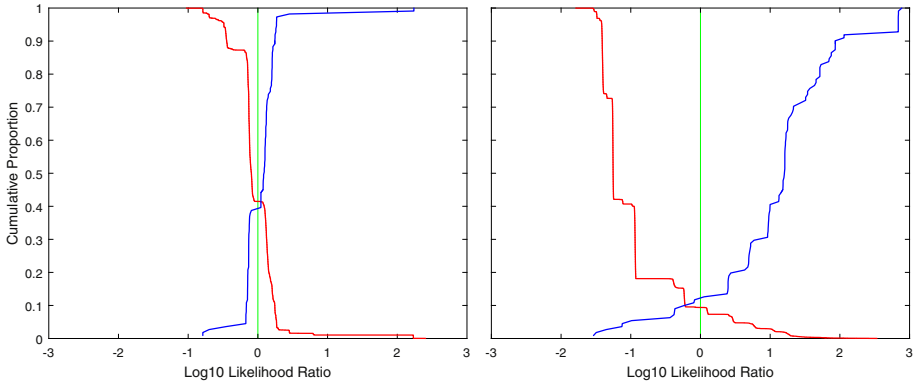

Fig. 11. Tippett plots for test results from the forensic voice comparison analysis. Left panel: similarity-only score procedure. Right panel: similarity-and-typicality score procedure.

mean-only maximum a posteriori adaptation from the relevant-population GMM.

The likelihoods of known-source models and the relevant-population model were evaluated at the values of each vector from a questioned-source-condition recording. Since there were 4137 feature vectors per questioned-source-condition recording, this resulted in 4137 likelihoods (for the similarity-only score procedure) or likelihood ratios (for the similarity-and-typicality score procedure) for each comparison of a known-source-condition and a questioned-source-condition recording. These were converted to a single score by taking the geometric mean of the 4137 values, i.e., the means of the natural-log likelihood values or the means of the natural-log likelihood-ratio values as applicable. Score values were calculated for all possible same-origin comparisons (excluding comparison of the session one known-source-condition recording with the session one questioned-source-condition recording) and all possible different-origin comparisons. These scores were then used to train PAV models to be used later to convert scores to likelihood ratios.[18]

The test set consisted of the same 223 recordings of 61 speakers as were used for calculating training scores, but a cross-validation procedure was adopted to prevent training and testing on the same data. For a same-origin comparison all the recordings of the speaker being tested were excluded from the calculation of the training scores. For a different-origin comparison all the recordings of the two speakers being tested were excluded from the calculation of the training scores. Score-based likelihood ratios were calculated for all possible same-origin comparisons (excluding comparison of the session one known-source-condition recording with the session one questioned-source-condition recording) and all possible different-origin comparisons.

A.2. Results and discussion

Fig. 10 shows histograms of the distributions of the test scores prior to conversion to likelihood ratios. The same-origin and different-origin similarity-and-typicality scores have distributions which appear to be approximately Gaussian with a reasonable degree of separation between the two distributions. In contrast, the distributions of the same-origin and different-origin similarity-only scores have greater overlap, and the distribution of the different-origin similarity-only scores appears to be multimodal.

For simplicity, to represent the performance of each system we present only one numeric summary (C_{llr}) and one graphical summary (Tippett plot).

The procedure for calculating the log likelihood ratio cost (C_{llr}, [33,39]) assigns a continuously valued cost to the test results. For same-origin test pairs, for which high likelihood ratio values are the desired result, the procedure assigns low cost values to likelihood ratio values much greater than 1, moderate cost values to likelihood ratio values around 1, and high cost values to likelihood ratio values much less than 1. Mutatis mutandis for different origin test pairs, for which low likelihood ratio values are the desired result. A cost is assigned to each likelihood ratio value in the test results and C_{llr} is averaged over these – mean of the same-origin set, mean of the different-origin set, then the mean of the latter two means. A system which always outputs a likelihood ratio of 1, and thus provides no information (posterior odds would always equal prior odds), would have a C_{llr} value of 1. The better the performance of the system, the lower the C_{llr} value.

The C_{llr} values for the similarity-only and for the similarity-and-typicality score based procedures were 1.018 and 0.415 respectively. The performance of the similarity-and-typicality score procedure is substantially better than that of the similarity-only score procedure. On this test set, the similarity-only score procedure results had a C_{llr} value of approximately 1, i.e., on average there is no benefit from using this system.

A Tippett plot [40,41] shows log likelihood ratios on the *x* axis; and, on the *y* axis, cumulative proportion of same-origin test pairs with likelihood ratios less than or equal to the value on the *x* axis, and cumulative proportion of different-origin test pairs with likelihood ratios equal to or greater than the value on the *x* axis. In general, the further to the right the same-origin curve, and the further to the left the different-origin curve, the better the performance of the system.

The Tippett plots in Fig. 11 show that for the similarity-only score procedure almost all the likelihood ratio values were close to 1 (log likelihood ratios close to 0),[19] whereas for the similarity-and-typicality score procedure there was a greater proportion of relatively large likelihood ratios from same-origin comparisons and a greater proportion of relatively small likelihood ratios from different-origin comparisons. The procedure based on similarity-and-typicality scores clearly outperformed that based on similarity-only scores.[20]

[18] We actually got better performance for both systems when we used logistic regression to convert from scores to likelihood ratios, but the relative difference between the two different types of scores was about the same. [31] also got better performance from logistic regression than from PAV when applied to scores derived from fingerprint-fingermark comparisons.

[19] The extreme large positive log likelihood ratio results are likely artefacts of the non-parametric PAV procedure. They do not occur if logistic regression is used instead.

[20] This result will not be of surprise to anyone familiar with automatic speaker recognition. In that field, use of a model trained on a large diverse sample has been a standard way of normalising scores for many years. The sample is generally not selected to represent a carefully defined relevant population, as would be required for a forensic analysis, and the aim is generally not to produce a likelihood ratio value that will be directly interpretable as a probabilistic answer to a particular question. Instead, the aim is generally to make a decision by comparing the score with a threshold. Normalising scores in this manner allows a single threshold value to be used, rather than having to determine a separate threshold value for each known speaker.

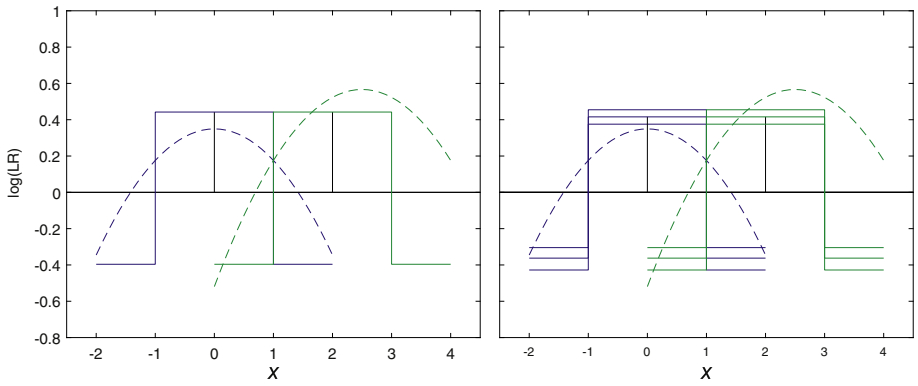

Fig. 12. Likelihood ratio values calculated on Monte Carlo sample data using the dichotomous "match"/"non-match" scoring procedure with a threshold of 1.0.

Appendix B. "match"/"non-match" scores

The 2016 report on "Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods" [16] by President Obama's Council of Advisors on Science and Technology (PCAST) advocated the use of a two-stage procedure in which the first stage is to compare a known-source sample and questioned-source specimen and declare a "match" or a "non-match" based on a predetermined threshold. This can be considered a dichotomous similarity-only score. If a "match" is declared, the second stage is to estimate and report the probability of declaring a "match" if the two objects came from the same source (correct acceptance rate) and the probability of declaring a "match" if the two objects came from different sources (false acceptance rate). Mutatis mutandis if a "non-match" is declared (correct rejection and false rejection rates). The PCAST report itself does not take the next step, but dividing the correct acceptance rate by the false acceptance rate provides a likelihood ratio for when a "match" has been declared, and dividing the false rejection rate by the correct rejection rate provides a likelihood ratio for when a "non-match" has been declared, Eq. (11)

$$\lambda_{\text{"match"}} = \frac{p(\text{"match"}|H_{same-origin})}{p(\text{"match"}|H_{different-origin})} \tag{11a}$$

$$\lambda_{\text{"non-match"}} = \frac{p(\text{"non-match"}|H_{same-origin})}{p(\text{"non-match"}|H_{different-origin})} \tag{11b}$$

In [42], this approach was criticised for dichotomising continuously-valued data, and thus discarding information that could be exploited by more appropriate statistical procedures. Here, we empirically compare likelihood ratio values calculated using the dichotomous "match"/"non-match" scoring procedure with reference likelihood ratio values. We calculated the distance between a known-source sample mean value and a questioned-source value, and determined whether it was less than a threshold value. We used a range of threshold values, from 0.1 to 2.0. The same training data as described in Section 5 were used, as were the same general procedures as described in Section 7 for generating score-based likelihood ratios and comparing them with reference likelihood-ratio values (note that the second stage is that described in the present appendix, it is not PAV).

A graphical representation of the likelihood ratio value results for a threshold value of 1.0 is given in Fig. 12. Fig. 13 shows RMS error over a range of threshold values. The scale on the *x* axis of Fig. 13 is the same as that in Fig. 6. It is clear that, irrespective of the threshold chosen, with respect to deviation from reference likelihood ratio values, the performance of the dichotomous "match"/"non-match" scoring procedure is much worse than that of any of the other scoring procedures examined in the present paper.

Appendix C. Performance in terms of C_{llr}

In order to assess the performance of the different scoring procedures with respect to whether the test data were same-origin or different-origin pairs, an additional set of test data was simulated. Instead of evenly spaced probes, as were used for the RMS error calculation in the main text, these test data were generated using the specified means and within-source standard deviations for the two specified sources. A fresh set of 90 Monte Carlo samples was generated for each source. The likelihood ratio at each test probe was calculated, and C_{llr} assessed according to whether each test probe was generated from the first or the second source: a likelihood ratio calculated using the first-source model and a test probe generated from the first source was a same-origin comparison, a likelihood ratio calculated using the first-source model and a test probe generated from the second source was a different-origin comparison, etc.

Fig. 14 shows the results for the direct procedure and the three scoring procedures described in the main text. Fig. 15 shows the results for the dichotomous "match"/"non-match" scoring procedure described in Appendix B. The relative pattern of performance was the same as for RMS error. Of the scoring procedures, the scoring procedure using non-anchored similarity-and-typicality scores had the best performance. The procedure using dichotomous "match"/"non-match" scores performed particularly poorly.

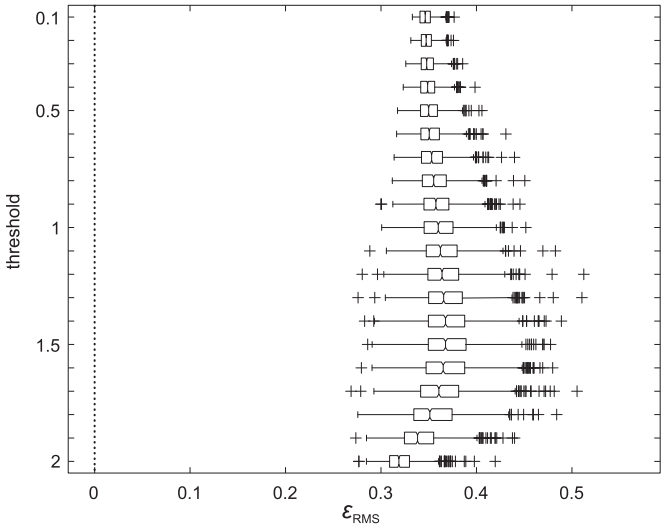

Fig. 13. Boxplot representations of distributions of RMS errors over all 1000 sample sets, for the dichotomous "match"/"non-match" scoring procedure with a range of threshold values. The scale on the *x* axis is the same as the scale for Fig. 6.

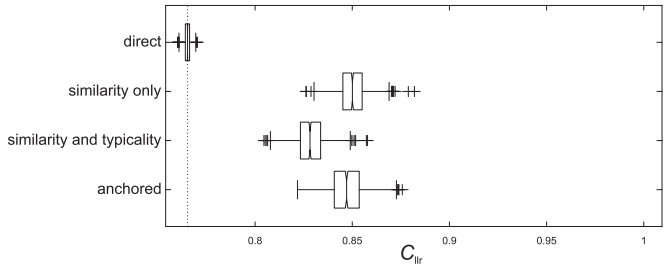

Fig. 14. Boxplot representations of distributions of C_{llr} values over all 1000 sample sets, for the baseline direct procedure and for each score-based procedure. The vertical dotted line indicates the performance of the reference likelihood ratio values. Note that the scale on the *x* axis does not start at 0.

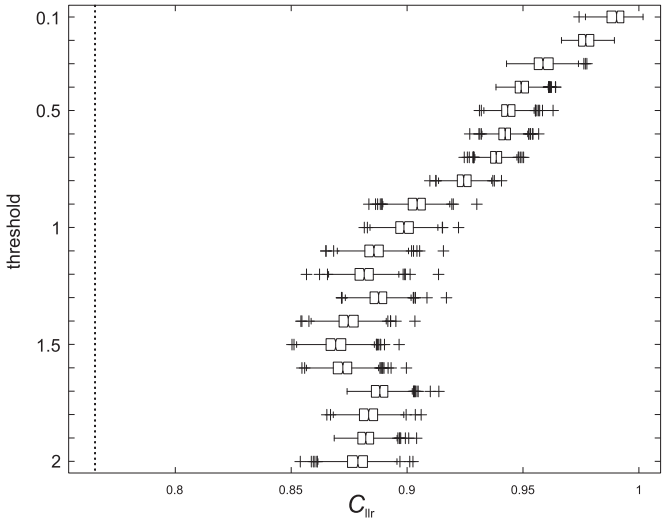

Fig. 15. Boxplot representations of distributions of C_{llr} values over all 1000 sample sets, for the dichotomous "match"/"non-match" scoring procedure with a range of threshold values. The vertical dotted line indicates the performance of the reference likelihood ratio values. Note that the scale on the *x* axis does not start at 0, and that it is the same as the scale for Fig. 14.

References

[1] J. González-Rodríguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-García J, Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Trans. Audio Speech Lang. Process. 15 (2007) 2104–2115, http://dx.doi.org/10.1109/TASL.2007.902747.

[2] C. Neumann, P. Margot, New perspectives in the use of ink evidence in forensic science. Part III. Operational applications and evaluation, Forensic Sci. Int. 192 (2009) 29–42, http://dx.doi.org/10.1016/j.forsciint.2009.07.013.

[3] A. Bolck, C. Weyermann, L. Dujourdy, P. Esseiva, J. Berg, Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons, Forensic Sci. Int. 191 (2009) 42–51, http://dx.doi.org/10.1016/j.forsciint.2009.06.006.

[4] A. Nordgaard, T. Höglund, Assessment of approximate likelihood ratios from continuous distributions: a case study of digital camera identification, J. Forensic Sci. 56 (2011) 390–402, http://dx.doi.org/10.1111/j.1556-4029.2010.01665.x.

[5] A.B. Hepler, C.P. Saunders, L.J. Davis, J. Buscaglia, Score-based likelihood ratios for handwriting evidence, Forensic Sci. Int. 219 (2012) 129–140, http://dx.doi.org/10.1016/j.forsciint.2011.12.009.

[6] J. Abraham, C. Champod, C. Lennard, C. Roux C, Modern statistical models for forensic fingerprint examinations: a critical review, Forensic Sci. Int. 232 (2013) 131–150, http://dx.doi.org/10.1016/j.forsciint.2013.07.005.

[7] S. Baechler, V. Terrasse, J.-P. Pujol, T. Fritz, O. Ribaux, P. Margot, The systematic profiling of false identity documents: method validation and performance evaluation using seizures known to originate from common and different sources, Forensic Sci. Int. 232 (2013) 180–190, http://dx.doi.org/10.1016/j.forsciint.2013.07.022.

[8] P. Vergeer, A. Bolck, L.J.C. Peschier, C.E.H. Berger, J.N. Hendrikse, Likelihood ratio methods for forensic comparison of evaporated gasoline residues, Sci. Justice 54 (2014) 401–411, http://dx.doi.org/10.1016/j.scijus.2014.04.008.

[9] T. Ali, L.J. Spreeuwers, R.N.J. Veldhuis, D. Meuwly, Effect of calibration data on forensic likelihood ratio from a face recognition system, IET Biom. 3 (2014) 335–346, http://dx.doi.org/10.1049/iet-bmt.2014.0009.

[10] D.-M. Dennis, M.R. Williams, M.E. Sigman, Assessing the evidentiary value of smokeless powder comparisons, Forensic Sci. Int. 259 (2016) 179–187, http://dx.doi.org/10.1016/j.forsciint.2015.12.034.

[11] A.J. Leegwater, D. Meuwly, M. Sjerps, P. Vergeer, I. Alberink, Performance study of a score-based likelihood ratio system for forensic fingermark comparison, J. Forensic Sci. 62 (2017) 626–640, http://dx.doi.org/10.1111/1556-4029.13339.

[12] D. Ramos-Castro, Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems, Universidad Autónoma de Madrid, Madrid, Spain, 2007 http://atvs.ii.uam.es/files/2007_11_28_thesis_daniel_ramos_searchable_v1.pdf (PhD dissertation).

[13] G.S. Morrison, Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio, Aust. J. Forensic Sci. 45 (2013) 173–197, http://dx.doi.org/10.1080/00450618.2012.733025.

[14] Y. Tang, S.N. Srihari, Likelihood ratio estimation in forensic identification using similarity and rarity, Pattern Recogn. 47 (2014) 945–958, http://dx.doi.org/10.1016/j.patcog.2013.07.014.

[15] E. Enzinger, Implementation of Forensic Voice Comparison within the new Paradigm for the Evaluation of Forensic Evidence, University of New South Wales, Sydney, New South Wales, Australia, 2016 http://handle.unsw.edu.au/1959.4/55772 (PhD dissertation).

[16] President's Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods, Executive Office of The President's Council of Advisors on Science and Technology, Washington DC, 2016 https://www.whitehouse.gov/administration/eop/ostp/pcast/docsreports (2016).

[17] G.S. Morrison, Calculation of forensic likelihood ratios: Use of Monte Carlo simulations to compare the output of score-based approaches with true likelihood-ratio values, Research Report, http://geoff-morrison.net/#ICFIS2014 2015http://arxiv.org/abs/1612.08165.

[18] C.E.H. Berger, Objective ink color comparison through image processing and machine learning, Sci. Justice 53 (2013) 55–59, http://dx.doi.org/10.1016/j.scijus.2012.09.003.

[19] C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, A. Bromage-Griffiths, Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae, J. Forensic Sci. 52 (2007) 54–64, http://dx.doi.org/10.1016/10.1111/j.1556-4029.2006.00327.x.

[20] W. van Houten, I. Alberink, Z. Geradts, Implementation of the likelihood ratio framework for camera identification based on sensor noise patterns, Law Probab. Risk 10 (2011) 149–159, http://dx.doi.org/10.1093/lpr/mgr006.

[21] A. Bolk, H. Ni, M. Lopatka, Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison, Law Probab. Risk 14 (2015) 243–266, http://dx.doi.org/10.1093/lpr/mgv009.

[22] E. Enzinger, G.S. Morrison, F. Ochoa, A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case, Sci. Justice 56 (2016) 42–57, http://dx.doi.org/10.1016/j.scijus.2015.06.005.

[23] M.I. Mandasari, M. Günther, R. Wallace, R. Saeidi, S. Marcel, D.A. van Leeuwen, Score calibration in face recognition, IET Biom. 3 (2014) 246–256, http://dx.doi.org/10.1049/iet-bmt.2013.0066.

[24] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models, Digit. Signal Process. 10 (2000) 19–41, http://dx.doi.org/10.1006/dspr.1999.0361.

[25] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Trans. Audio Speech Lang. Process. 19 (2011) 788–798http://dox.doi.org/10.1109/TASL.2010.2064307.

[26] A. Alexander, A. Drygajlo, Scoring and direct methods for the interpretation of evidence in forensic speaker recognition, Proceedings of Interspeech 2004, Jeju, Korea 2004, pp. 2397–2400http://www.isca-speech.org/archive/interspeech_2004/i04_2397.html.

[27] R. Haraksim, D. Meuwly, Influence of the datasets size on the stability of the LR in the lower region of the within source distribution, Proceedings of Biometric Technologies in Forensic Science, BTFS 2013, Nijmegen, The Netherlands 2013, pp. 34–38http://cls.ru.nl/staff/dvleeuwen/btfs-2013/harakasim-btfs2013.pdf.

[28] I. Alberink, A. de Jongh, C. Rodríguez, Fingermark evidence evaluation based on automated fingerprint identification system matching scores: the effect of different types of conditioning on likelihood ratios, J. Forensic Sci. 59 (2014) 70–81, http://dx.doi.org/10.1111/1556-4029.12105.

[29] N. Brümmer, A. Swart, D. van Leeuwen, A comparison of linear and non-linear calibrations for speaker recognition, Proceedings of Odyssey 2014 The Speaker and Language Recognition Workshop, Joensuu, Finland 2014, pp. 14–18http://www.isca-speech.org/archive/odyssey_2014/pdfs/31.pdf.

[30] T. Ali, L.J. Spreeuwers, R.N.J. Veldhuis, D. Meuwly, Sampling variability in forensic likelihood-ratio computation: a simulation study, Sci. Justice 55 (2015) 499–508, http://dx.doi.org/10.1016/j.scijus.2015.05.003.

[31] D. Ramos, R.P. Krish, J. Fierrez, D. Meuwly, D., From biometric scores to forensic likelihood ratios, in: M. Tistarelli, C. Champod (Eds.), Handbook of Biometrics for Forensic Science, Springer, Cham, Switzerland 2017, pp. 305–327, http://dx.doi.org/10.1007/978-3-319-50673-9_14 (ch. 14).

[32] B. Zadrozny, C. Elkan, Transforming classifier scores into accurate multiclass probability estimates, Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining 2002, pp. 694–699, http://dx.doi.org/10.1145/775047.775151.

[33] N. Brümmer, J. du Preez, Application-independent evaluation of speaker detection, Comput. Speech Lang. 20 (2006) 230–275, http://dx.doi.org/10.1016/j.csl.2005.08.001.

[34] D. Van Leeuwen, N. Brümmer, An introduction to application-independent evaluation of speaker recognition systems, Speaker Classification I, Springer, Berlin Heidelberg 2007, pp. 330–353, http://dx.doi.org/10.1007/978-3-540-74200-5_19.

[35] T. Fawcett, A. Niculescu-Mizil, PAV and the ROC convex hull, Mach. Learn. 68 (2007) 97–106, http://dx.doi.org/10.1007/s10994-007-5011-0.
[36] G. Zadora, A. Martyna, D. Ramos, C.G.G. Aitken, Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data, Wiley, Chichester, UK, 2014http://dx.doi.org/10.1002/9781118763155.
[37] C. Neumann, C. Saunders, et al., Fingermark: the effect of different types of conditioning on likelihood ratios, J. Forensic Sci. 60 (2015) (2014) 252–256, http://dx.doi.org/10.1111/1556-4029. 12634 (Commentary on Alberink).
[38] L.D. Friedland, Detecting Anomalously Similar Entities in Unlabeled Data, http://scholarworks.umass.edu/dissertations_2/845/ 2016 (doctoral dissertations).
[39] G.S. Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems, Sci. Justice 51 (2011) 91–98, http://dx.doi.org/10.1016/j.scijus.2011.03.002.
[40] D. Meuwly, Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique, University of Lausanne, 2001 (PhD dissertation).
[41] G.S. Morrison, Forensic voice comparison, in: I. Freckelton, H. Selby (Eds.), Expert Evidence, Thomson Reuters, Sydney, Australia, 2010 http://expert-evidence.forensic-voice-comparison.net/ (ch. 99).
[42] G.S. Morrison, D.H. Kaye, D.J. Balding, D. Taylor, P. Dawid, C.G.G. Aitken, S. Gittelson, G. Zadora, B. Robertson, S. Willis, S. Pope, M. Neil, K.A. Martire, A. Hepler, R.D. Gill, A. Jamieson, J. de Zoete, R.B. Ostrum, A. Caliebe, A comment on the PCAST report: skip the "match"/"non-match" stage, Forensic Sci. Int. 272 (2017) e7–e9, http://dx.doi.org/10.1016/j.forsciint.2016.10.018.