Chapter 14
From Biometric Scores to Forensic Likelihood Ratios

Daniel Ramos, Ram P. Krish, Julian Fierrez and Didier Meuwly

Abstract In this chapter, we describe the issue of the interpretation of forensic evidence from scores computed by a biometric system. This is one of the most important topics into the so-called area of forensic biometrics. We will show the importance of the topic, introducing some of the key concepts of forensic science with respect to the interpretation of results prior to their presentation in court, which is increasingly addressed by the computation of likelihood ratios (LR). We will describe the LR methodology, and will illustrate it with an example of the evaluation of fingerprint evidence in forensic conditions, by means of a fingerprint biometric system.

14.1 Likelihood Ratio Framework for Evidence Evaluation

The evaluation of the relationship between two pieces of evidence at judicial trials has been the subject of discussion in the past years [1]. Here, the problem is to give a value to a comparison of a trace specimen of unknown origin (for instance a fingermark revealed in a crime scene, or a wire tapping involving an incriminating conver-

D. Ramos (✉) · R.P. Krish · J. Fierrez
ATVS - Biometric Recognition Group, Escuela Politecnica Superior,
Universidad Autonoma de Madrid, Calle Francisco Tomas y Valiente 11,
28049 Madrid, Spain
e-mail: daniel.ramos@uam.es

R.P. Krish
e-mail: ram.krish@uam.es

J. Fierrez
e-mail: julian.fierrez@uam.es

D. Meuwly
Netherlands Forensic Institute, Laan van Ypenburg 6, 2497GB
The Hague, The Netherlands
e-mail: d.meuwly@utwente.nl

D. Meuwly
University of Twente, Drienerlolaan 5, 7522NB Enschede, The Netherlands
e-mail: d.meuwly@utwente.nl

M. Tistarelli and C. Champod (eds.), *Handbook of Biometrics for Forensic Science*, Advances in Computer Vision and Pattern Recognition,
DOI 10.1007/978-3-319-50673-9_14

sation) with a reference specimen of known origin (for instance, a fingerprint from a suspect, or some recordings of a known individual). From a formal logical perspective [2], the given value should represent the degree of support of the comparison to any of the *propositions* (also called *hypotheses*) involved in the trial. Examples of simple hypotheses might be "the trace and the reference specimens originated from the same source" or "the trace and the reference specimens originated from different sources", but more complex hypotheses can be considered [2]. In some sense, the value of the evidence represents the strength of the link between the trace and the reference specimen in the context of the propositions considered.

Evidence evaluation using a Bayesian probabilistic framework has been proposed in recent years as a logical and appropriate way to report evidence to a court of law [3]. In Europe, there have been initiatives to foster this approach, some of them in response of notorious reluctance to the use of statistics in courts [4]. These have been the main reason leading to the release of a Guideline [5] for the expression of conclusions in evaluative forensic reports. This Guideline is proposed by the European Network of Forensic Science Institutes (ENFSI), an organization that includes almost all the main forensic laboratories in Europe.[1] According to this Guideline, a Bayesian framework for forensic evaluative reports is recommended for all disciplines and laboratories within ENFSI. Under this Bayesian approach, a likelihood ratio (LR) is computed to represent the value of the evidence, and to be reported to a court of law (mainly in the form of a verbal equivalent). This framework clearly complies with the requirements of evidence-based forensic science [1]: it is scientifically sound (transparent procedures, testability, formal correctness), and clearly separates the responsibilities of the forensic examiner and the court.

The increasing establishment of this Bayesian evaluative framework has motivated the convergence of pattern recognition and machine-learning approaches to yield probabilistic outputs in the form of likelihood ratios. A common architecture for this considers two steps: first, the computation of a discriminating score between two specimens, trace and reference (e.g., a fingermark in the crime scene and an exemplar fingerprint from a known suspect), which can be obtained from a standard biometric system; and second, the transformation of the score into a likelihood ratio [6–9]. This architecture is especially suited for biometric systems, where the output of a conventional biometric system is typically expressed as a score, even though it is used as *black-box* technology. Therefore, the score is most of the times a necessary intermediate step to the likelihood ratio.

14.1.1 Challenges in LR-Based Evidence Evaluation

Despite its advantages, the computation of likelihood ratios still presents important challenges. We enumerate the most important as follows.

[1] http://www.enfsi.eu/.

First, the typical scenario in forensic science involves data presenting diverse and unfavorable conditions, which means that automatic comparisons between the specimens will result in a challenging problem. Efforts to model or compensate the effects of these adverse conditions in likelihood ratio computation should be improved. Some works such as [10] have contributed to evaluate the impact of this problem. Moreover, integration of advanced machine-learning algorithms (like in [11, 12]) for the compensation of adverse conditions into forensic evaluation helps in this direction. However, adverse condition compensation still remains a challenge.

Second, in forensic science the databases are difficult to obtain and to use, even for research purposes. This is because, although there is plenty of forensic data in some disciplines (e.g., large fingerprint databases), there are legal, privacy and interoperability issues that hamper the use of this data by academic and research institutions. This leads to two opposite situations: either the databases are big when there is access to the data, and therefore big-data solutions are a challenge to face; or the databases are highly scarce, and the use of robust models is necessary. Data scarcity has been tackled by different techniques as in [13, 14]. However, to our knowledge, evidence evaluation models have not been adapted to big-data scenarios to handle big databases when possible, which represents a loss of information in these scenarios. Other lines of research have proposed the use of simulated forensic data in order to prevent the problem of data scarcity [15]. The involvement of simulated data is a big improvement against data scarcity situations, but the testing of the validity of simulated databases for the operational use of systems in a real setup is still controversial.

Third, although likelihood ratio computation methods are becoming more and more popular, the validation of those methods for its use in forensic casework is still not standardized. Even if likelihood ratios are computed to evaluate the links between evidential materials, this does not guarantee that they will be able to be integrated into a Bayesian decision framework to ultimately allow a fact finder to do optimal decisions. In this sense, the measurement of the performance characteristics that a likelihood ratio model should manifest is of paramount importance. Recent work has shown that one of the most important characteristic that forensic likelihood ratios should present is the so-called calibration [9]. This is a property of a set of likelihood ratios, by which the LR is itself a measure of evidential weight. This leads to the property that "The LR of the LR is the LR", meaning that the LR is interpreting the evidence with the best possible probabilistic meaning in terms of Bayes decisions [16]. Therefore, computing likelihood ratios is not enough, they should also be the best calibrated as possible. There are current efforts of the forensic community in order to establish formal frameworks for the validation of likelihood ratio models [9, 17, 18], but research is still needed. Also, a framework for the validation of likelihood ratio methods has been recently published [19].

Fourth, evidence evaluation in complex cases is still problematic. Probabilistic graphical models, particularly Bayesian networks [20], have been proposed to address those situations. However, this emerging field is an active area of research in forensic science. More efforts are needed in order to provide forensic examiners with appropriate tools in operational scenarios, especially if those models are to be learned from data.

14.2 Case Assessment and Interpretation Methodology

A milestone in the use of the LR methodology in Europe was the Case Assessment and Interpretation (CAI) methodology developed by the Forensic Science Service (FSS) in the late 1990s [2]. This was the result of the efforts of the now closed Forensic Science Service of the United Kingdom, in order to homogenize and make more agile the relationship between courts and forensic service providers (e.g., police forces or other public or private forensic laboratories). An ultimate aim is the use of a logical methodology to avoid pitfalls of reasoning and fallacies. The methodology has been described in several papers during the end of the twentieth century, remarkably [2, 21, 22], and serves as the core of likelihood ratio-based evidence interpretation.

There are several characteristic features of the CAI methodology, which we summarize below.

- Full integration of the LR methodology into the forensic evidence evaluation process. In this sense, all the elements typical from LR evidence evaluation are present, namely the evidence, propositions, probabilistic reasoning, etc.
- A particular emphasis is put in the definition of the propositions in a given case, which have to be informed by the circumstances of the case themselves. Thus, the relationship between the court and the forensic science provider should be essential in order to define the propositions. Issues like the definition of the population considered to model the alternative proposition, the specificity of the propositions with respect to the population, the suspect and the trace, or the selection of the most appropriate database to address the propositions [23], are of particular importance.
- A hierarchy of propositions [21] is introduced in order to address the forensic casework in the most appropriate manner with respect to the information in the case. In this sense, there are three basic levels in the hierarchy: *source level*, where the source which originated the trace(s) is considered; *activity level*, where the activities from which originate the traces are under discussion; and *offence level*, that focuses on the question whether the activity from which originate the traces is an infraction. Depending on the question asked by the requester/fact finder and on the information in the case available to the forensic scientist, it is possible to escalate the inference a to higher level, but in most cases the forensic examiner is requested to report at source level, reason for which most of the effort to produce increasingly robust models has been focused on source level. Nevertheless,

nowadays there is a push towards the use of activity-level propositions in casework (even in the ENFSI Guideline for evaluative reports [5]), although LR models for activity or offence levels are mostly in a research stage.

- Case pre-assessment is encouraged by the model. Under this concept, a preliminary LR value is reported prior to the case itself, in order to indicate what would be the expected outcome of the forensic analysis by the examiner. This helps to focus the expectations of the fact finder, and has important implications regarding the efficiency of resources in a case.

14.3 Evidence Evaluation with Likelihood Ratios

The LR framework for interpretation of the evidence represents a mathematical and logical tool in order to aid in the inference process derived from the analysis of the evidence. In this methodology, the objective of the forensic scientist is computing the likelihood ratio (LR) as a degree of support of one proposition versus its alternative [3, 24].

The LR framework is stated as follows. Consider a forensic case. There is a forensic evidence E, which contains the specimens to compare in a forensic case, namely, in a fingerprint case, a *recovered* fingermark of unknown origin and a *reference* fingerprint (namely the *exemplar*) whose origin is known to be a given suspect involved in the case. In this context, the unobserved variable of interest is the true proposition H with values $\left\{H_p, H_d\right\}$, where H_p and H_d are the possible relevant propositions defined in the case, according to the CAI methodology. As mentioned before, the definition of H_p and H_d varies in each case. A possible definition at the source level could be as follows:

H_p: The origin of the fingermark (trace) and the fingerprint (reference) is the same finger of one single donor.

H_d: The origin of the fingermark (trace) and the fingerprint (reference) are fingers from two different donors.

H_p is typically called *prosecution* proposition, whereas H_d is referred to as *defense* proposition. This is due to the fact that alternative, and mutually exclusive propositions arise naturally in an adversarial trial system like in the UK, where the CAI methodology was developed. Other propositions can be addressed, and variation of their statement can lead to a radically different selection of databases for LR computation [23], and even to different likelihood ratio models [25]. Therefore, care should be taken in order to clearly and appropriately define the propositions in a case.

In Bayesian decision theory, decisions are made considering the probability distribution of the variable of interest (in this case, the proposition variable H), given all the observed information. In a forensic case, this can be represented as $P\left(H = H_p \mid E, I\right)$ and $P\left(H = H_d \mid E, I\right)$, or simply $P\left(H_p \mid E, I\right)$ and $P\left(H_d \mid E, I\right)$, where I is the background information available in the case not related to the evidence E, as defined by the CAI methodology. H_p and H_d are in most cases mutually exclusive. Then, Bayes' theorem [3] relates probabilities before and after evidence analysis.

$$P\left(H_p \mid E, I\right) = \frac{P\left(E \mid H_p, I\right) \cdot P\left(H_p \mid I\right)}{P(E \mid I)} \tag{14.1}$$

In terms of interpretation, it is useful to use ratios of probabilities. Then, Eq. 14.1 becomes

$$\frac{P\left(H_p \mid E, I\right)}{P\left(H_d \mid E, I\right)} = LR \cdot \frac{P\left(H_p \mid I\right)}{P\left(H_d \mid I\right)} \tag{14.2}$$

$$LR = \frac{P\left(E \mid H_p, I\right)}{P\left(E \mid H_d, I\right)} \tag{14.3}$$

In Eq. 14.2, we can distinguish the following:

1. The prior probabilities $P\left(H_p \mid I\right)$ and $P\left(H_d \mid I\right)$, which are province of the fact finder and should be stated assuming only the background information (I) in the case [24].
2. The LR (Eq. 14.3), assigned or computed by the forensic practitioner [3].

A critical point in the application of the LR methodology is the selection of proper databases to address the propositions and also the trace material, for example the language of the trace will determine the language of the speech databases used in the case, but also from the definition of the propositions themselves. We will address this issue in the example below, but many works in the literature give recommendations on how to select these databases, both in fingerprints [26] or in forensic science in general [23].

This LR-based framework for interpretation presents many advantages

- It allows forensic practitioners to evaluate and report a meaningful value for the weight of the evidence to the court, with a universal interpretation, allowing for the combination of results across disciplines when the same propositions are considered [5, 24].
- The role of the examiner is clearly defined, leaving to the court the task of using prior judgments or costs in the decision process.
- Probabilities can be interpreted as degrees of belief [27], allowing the incorporation of subjective opinions as probabilities in the inference process in a clear and scientific way.

The LR value has an interpretation as a *support* to a previously stated opinion, due to the analysis of the evidence E. In other words

- If the LR > 1 the evidence will support that $H = H_p$, i.e., the prosecutor proposition.

- If the LR < 1 the evidence will support that $H = H_d$, i.e., the defense proposition.

Moreover, the value of the LR represents the *degree of support* of the evidence to one value of H against the other. For instance, LR $= 3$ means that "the evidence supports $H = H_p$ against $H = H_d$ with a degree of 3 versus 1". Therefore, a single LR value has a *meaning* by itself, as opposed to a biometric score, that may have only meaning if compared to a reference threshold or another set of scores.

It is important to note that the LR *supports* an opinion about H, but the LR *is not* an opinion about H. Opinions about H are represented as probabilities of propositions, or in our binary case, their ratios. Therefore, it is not possible to make a decision about the value of H based solely on the value of the LR, because decisions will be taken from posterior probabilities, not only from degrees of support.

14.4 Interpreting Biometric System Scores with Likelihood Ratios

According to [6, 7], in biometrics all the information that the systems yield about the propositions after observing E is in most cases condensed into a so-called *score*, a single number which is an observation of a random variable S, and must contain as much information as possible about H. Therefore, the interpretation of the evidence using biometric systems requires that the score will be first computed by the system, yielding the particular value $S = s$, and then the score is interpreted using a likelihood ratio. This leads to the following expression:

$$LR = \frac{P\left(E|\,H_p, I\right)}{P\left(E|\,H_d, I\right)} = \frac{P\left(s|\,H_p, I\right)}{P\left(s|\,H_d, I\right)} \tag{14.4}$$

Moreover, most biometric scores are continuous, and in that case the ratio of probabilities becomes a ratio of probability density functions, yielding

$$LR = \frac{P\left(s|\,H_p, I\right)}{P\left(s|\,H_d, I\right)} = \frac{p\left(s|\,H_p, I\right)}{p\left(s|\,H_d, I\right)} \tag{14.5}$$

Thus, the LR value (Eq. 14.5) is the quotient of two probability densities. On the one hand, the probability density function (pdf) $p\left(S|\,H_p, I\right)$ in the numerator in Eq. 14.3 is known as the intra-variability distribution. Its evaluation in the particular value of the score $S = s$ gives a measure of the probability density of observing the evidence under H_p. On the other hand, the pdf $p\left(S|\,H_d, I\right)$ in the denominator is known as the inter-variability distribution, and its evaluation in the particular value of the score $S = s$ gives a measure of the probability density of observing the evidence under H_d.[2]

The aim of LR methods with biometric score-based systems is to provide a model that transforms scores into LR values in a case. Moreover, the resulting LR values should present adequate performance in order to correctly aid the decisions of fact finders.

14.5 LR Computation Methods from Biometric Scores

In this section, some of the most common algorithms for LR computation from biometric scores are described.

14.5.1 Generating Training Scores

The main commonality of all the methods described in this section is that they need two proposition-dependent sets of *training* scores, namely $\mathbf{S_p}$ and $\mathbf{S_d}$. These sets and some of the issues associated to them are described as follows.

- The set $\mathbf{S_p} = \left\{ s_p^{(1)}, \ldots, s_p^{(N_{pt})} \right\}$ consists of N_{pt} scores computed assuming that $H = H_p$. Therefore, the selection of data to compute the scores in $\mathbf{S_p}$ has to be done accordingly to the definition of the propositions. As H_p proposition typically assumes that the trace and reference specimens in the case come from the same person, the $\mathbf{S_p}$ consists of *same-source* scores.[3] However, the rest of information in H_p can be determinant in order to select the database to generate those same-source scores. For instance, if the particular suspect in the case is included in the propositions (e.g., "the trace was left by Mr. Dean Keaton"), then the propositions will be *person-specific* or *source-specific*, and the database to generate $\mathbf{S_p}$ should include specimens coming from the particular donor, because in many biometric traits each person has a particular behavior regarding their score distribution [28]. On the other hand, *person-generic* or *source-generic* propositions (e.g., "the trace and the reference samples come from the same source") would allow the use of any same-source score from other people, since there is not knowledge of a particular subject. Another example of the influence of propositions in the model for LR computation is related to the definition of suspect-based or finger-based proposition for fingerprint interpretation [25].

[2] The background information about the case I will be eliminated from the notation for the sake of simplicity hereafter. It will be assumed that all the probabilities defined are conditioned to I.

[3] Here we work at the source level, and therefore *same-source* scores refer to scores generated from two biometric specimens coming from the same source. They are what in biometric authentication terminology are called *genuine* scores.

- The set $\mathbf{S_d} = \left\{ s_d^{(1)}, \dots, s_d^{(N_{dt})} \right\}$ consists of N_{dt} scores computed assuming that $H = H_{dt}$. For the computation of these scores, several things should be taken into account. First, H_d typically assumes that the questioned materials were not generated by the suspect in the case, but other person. Therefore, the scores in $\mathbf{S_d}$ will essentially be *different-source* scores,[4] since the case always considers the donor reference specimens as part of the evidence. Second, the way in which these scores are generated is critical, since the selection of different strategies to obtain $\mathbf{S_d}$ might lead to different LR values. Also, theoretical issues should be taken into account regarding these strategies (for a discussion about this, see [29]). Last, but not least, the determination of the population of sources of the test specimen must be handled with care. The key point is that the population must be seen as the set of potential donors of the test specimen, considering the definition of the proposition and the information about the case that is relevant and available to the forensic examiner.

An important remark is in order here. The aim of the $\mathbf{S_p}$ and $\mathbf{S_d}$ score sets is to represent the variation of S conditioned to the propositions. As S is the variable representing the score to be obtained from the evidence in the case, the conditions in the forensic case should be preserved for all comparisons in $\mathbf{S_p}$ and $\mathbf{S_d}$. For instance, if the evidence consists of a degraded, partial fingermark and a fingerprint from a ten-print card of a known suspect, all the scores in $\mathbf{S_p}$ and $\mathbf{S_d}$ should be generated from comparisons of degraded fingermarks and fingerprints from ten-print cards, in the conditions as similar as possible to those in the case. An exception would be if the conditions do not affect the behavior of the scores at all, but this rarely happens in real forensic scenarios.

Moreover, the scores in $\mathbf{S_p}$ and $\mathbf{S_d}$ should represent all possible sources of variability in S. Therefore, the use of models of variability is essential in order compute better likelihood ratios. Good examples exist in the literature of the use of variability models to compute the LR [26], or to compensate this variability at the level of the biometric score [11, 12].

14.5.2 Common Methods for Score-Based LR Computation

14.5.2.1 Generative Assignment of Probability Densities

LR computation in forensic biometrics has been classically performed by the use of generative techniques modeling the hypotheses-conditional distribution of the scores variable S. This is the approach already presented in [30], and has been followed in subsequent works in the literature. Under this approach, the objective is assigning

[4] Here we work at the source level, and therefore *different-source* scores refer to scores generated from two biometric specimens coming from different sources. They are what in biometric authentication terminology are called *impostor* scores.

the likelihoods $p\left(S|\,H_p\right)$ to the training scores $\mathbf{S_p}$, and $p\left(S|\,H_d\right)$ to $\mathbf{S_d}$. Then, the ratio of the particular value of these densities at $S = s$ will be the LR value.

Assigning $p\left(S|\,H_p\right)$ and $p\left(S|\,H_d\right)$ implies the selection of a proper model. The most straightforward choice for biometric scores could be the Gaussian distribution, obtained via Maximum Likelihood from the training set of scores. However, this requires the distributions involved to present a good fitting with Gaussian probability density functions, which is not typically the case. Fortunately, some score normalization techniques such as T-Norm tend to generate Gaussian distributions for scores when H_d is true [31]. Other approaches for generative ML fitting includes the use of Kernel Density Functions [30, 32], Gaussian Mixture Models [32] and other parametric distributions [18, 42].

14.5.2.2 Logistic Regression

Logistic regression is a well-known pattern recognition technique widely used for many problems including fusion [33, 34] and more recently likelihood ratio computation [7, 35]. The aim of logistic regression is obtaining an affine transformation (i.e., shifting and scaling) of an input dataset in order to optimize an objective function. Let $\mathbf{S_f} = \left\{s_f^{(1)}, s_f^{(2)}, \ldots, s_f^{(K)}\right\}$ be a set of scores from K different biometric systems. The affine transformation performed by the logistic regression model can be defined as

$$f_{lr} = \log\left(\frac{P\left(H_p\middle|\mathbf{S_f}, I\right)}{P\left(H_d\middle|\mathbf{S_f}, I\right)}\right) = a_0 + a_1 \cdot s_f^{(1)} + a_2 \cdot s_f^{(2)} + \cdots + a_K \cdot s_f^{(K)} \qquad (14.6)$$

This leads to the following *logistic regression model*:

$$P\left(H_p\middle|\mathbf{S_f}, I\right) = \frac{1}{1 + e^{-f_{lr}}} = \frac{1}{1 + e^{-\log(LR) - \log\left(O(H_p)\right)}} \qquad (14.7)$$

where $O(H_p)$ determines the prior odds in favor of H_p.

The weighting terms $\left\{a_0, a_1, a_2, \ldots, a_K\right\}$ can be obtained from a set of training data with optimization procedures found in the literature.[5] Moreover, by training the weights for some given simulated value of the prior odds, and removing the influence of that value of the prior odds after computing f_{lr}, likelihood ratios are obtained.

Notice that logistic regression can be used for computing likelihood ratios from a single biometric score ($K = 1$), but also to perform fusion and LR computation simultaneously (when $K > 1$) [34]. This fact, joined to the good behavior that logistic

[5] Typical implementations used in biometrics include toolkits like FoCal or BOSARIS, which can be found in http://niko.brummer.googlepages.com.

regression presents in most situations, have made this LR computation algorithm one of the most popular ones.

14.5.2.3 Pool Adjacent Violators (PAV)

Another approach to score-to-LR transformation has been proposed by the use of the Pool Adjacent Violators (PAV) algorithm [7]. The PAV algorithm transforms a set of scores into a set of LR values presenting optimal calibration. However, it is only possible to apply an optimal PAV transformation if the ground-truth labels of the propositions for each score in the set are known. As suggested in [8], a PAV transformation can be trained on the set of training scores $\mathbf{S_p}$ and $\mathbf{S_d}$, and then apply the trained transformation to a score in a forensic case. Although a straightforward use of PAV leads to a non-invertible transformation, several *smoothing* techniques can be applied to PAV in order to keep it monotonically increasing. For instance, adding a small slope to the function defining the PAV transformation will lead to an invertible transformation. Interpolating with linear, quadratic or splines approaches are also possible smoothing schemes.

14.6 Performance Measurement of LR Methods

As it was previously stated, the issue of performance measurement of LR methods is paramount in order to achieve validation of forensic interpretation prior to its use in casework [18]. In this section, we describe some of the performance metrics adequate for LR-based forensic interpretation.

At the source level, performance measurement is typically carried out in an empirical way. In order to measure the performance of a LR method, a test set of LR values should be generated by comparisons of specimens from a biometric database using that LR method. These comparisons should be in fact simulated cases, where the conditions of specimens should be similar to the conditions of the case scenario whose performance is to be measured. This would lead to N_p LR values computed when H_p is true and N_d LR values computed when H_p is true.

A solution to measure the performance of likelihood ratio values has been proposed in [7] for speaker recognition, and has been dubbed *log-likelihood-ratio cost* (C_{llr}). Later, it has been used in many other fields in forensic sciences [14, 17, 36, 37]. C_{llr} is defined as follows:

$$C_{llr} = \frac{1}{2 \cdot N_p} \sum_{i_p} \log_2 \left(1 + \frac{1}{LR_{i_p}} \right) + \frac{1}{2 \cdot N_d} \sum_{j_d} \log_2 \left(1 + LR_{j_d} \right) \tag{14.8}$$

The indices i_p and j_d respectively denote summing over the LR values of the simulated cases where each proposition is respectively true.

An important result is derived in [7], where it is demonstrated that minimizing the value of C_{llr} also encourages to obtain reduced Bayes decision costs for all possible decision costs and prior probabilities [38]. This property has been highlighted as extremely important in forensic science [9]. Moreover, in [7], the Pool Adjacent Violators algorithm is used in order to decompose C_{llr} as follows:

$$C_{llr} = C_{llr}^{min} + C_{llr}^{cal} \tag{14.9}$$

where

- C_{llr}^{min} represents the *discrimination cost* of the LR method, and it is due to non-perfect discriminating power.
- C_{llr}^{cal} represents the *calibration cost* of the system.

C_{llr} is a scalar measure of performance of LR values, the lower its value the better the performance. Another useful measure of performance, with interpretation in terms of information loss, is the Empirical Cross-Entropy (ECE), which is a generalization of C_{llr}, as follows:

$$\begin{aligned} ECE = &- \frac{P\left(H_p \mid I\right)}{N_p} \sum_{i_p} log_2 P\left(H_p | s_{i_p}, I\right) \\ &- \frac{P\left(H_d \mid I\right)}{N_d} \sum_{j_d} log_2 P\left(H_d | s_{j_d}, I\right), \end{aligned} \tag{14.10}$$

where s_{i_p} and s_{j_d} denote the scores from trace and reference specimens in each of the simulated cases, where either H_p or H_d is respectively true.

As it happens with C_{llr}, ECE can be additively decomposed also using the PAV algorithm into ECE = ECE^{min} + ECE^{cal}. This leads to ECE^{min} measuring information loss due to bad discriminating power, and ECE^{cal} measuring information loss due to miscalibration.

As it can be seen, ECE is dependent of the prior probabilities both explicitly and through $P\left(H_p | s_{i_p}, I\right)$ and $P\left(H_d | s_{i_d}, I\right)$. Thus, ECE can be represented in a prior-dependent way. This has been proposed to be done by a so-called ECE plot [17], which shows three comparative performance curves together (Fig. 14.1)

- solid, red curve: accuracy. This curve is the ECE of the LR values in the validation set, as a function of the prior log-odds. The lower this curve, the more accurate the method. This curve shows the overall performance of the LR method;
- dashed, blue curve: ECE^{min}. This curve is the ECE of the validation set of LR values after the application of the PAV algorithm. This shows the best possible ECE in terms of calibration, and it is a measure of discriminating power;

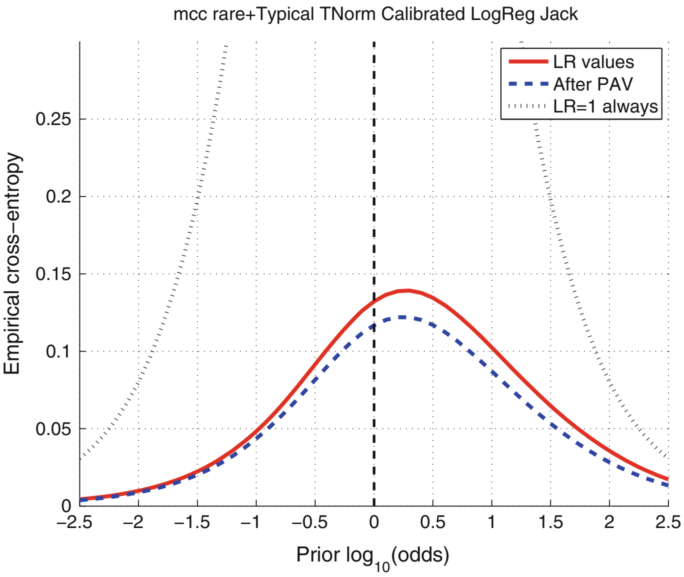


Fig. 14.1 Example of ECE plot

- dotted curve: neutral reference. It represents the comparative performance of a so-called *neutral LR method*, defined as the one which always delivers LR = 1 for each forensic case simulated in the set of LR values. This neutral method is taken as a *floor of performance*: the accuracy should always be better than the neutral reference. Therefore, the solid curve in an ECE plot should be always lower than the dotted curve, for all represented values of the prior log-odds (the names *floor* and *ceiling* are the opposite of the usual physical connotations but are chosen to represent the lowest and highest levels of performance).

A free Matlab$^{\text{TM}}$ software package to draw ECE plots can be found in the following webpage: http://arantxa.ii.uam.es/~dramos/software.html.

14.7 Computing LR Values from Biometric Scores: An Example with Forensic Fingerprint Recognition

In this section, we illustrate the process of computing forensic LR values from biometric fingerprint scores. We use a database collected from real cases, and in collaboration with Spanish Guardia Civil. Also, we use a state-of-the-art fingerprint system based on Minutiae Cylinder Code [39–41].

Evidence evaluation in fingerprints by the use of LR has been recently proposed in remarkable works like in [26] for minutiae configurations extracted manually from forensic examiners. However, other models based on the use of AFIS scores to compute likelihood ratio values can be found in [42], and more recently [18]. The reasons of modeling AFIS scores are diverse. On the other hand, it may give a complementary information to other methods more based on the direct statistical modeling of the minutiae extracted by the examiners. On the other hand, it allows the use of powerful systems to extract the information of identity, after which a likelihood ratio model performs the interpretation with the less loss of information possible [6].

14.7.1 Database and Statistics

The database used in this work was obtained from Guardia Civil, the Spanish law enforcement agency. The Guardia Civil database (GCDB) is a realistic forensic fingerprint casework database. Apart from having typical minutiae feature types (*ridge-endings, bifurcations*), GCDB also comprises rare minutiae types like *fragments*, *enclosures*, *dots*, *interruptions, etc.* [43]. A comprehensive list of rare minutia features used by Guardia Civil are shown in Fig. 14.2 and the corresponding minutiae type names are listed in Table 14.1.

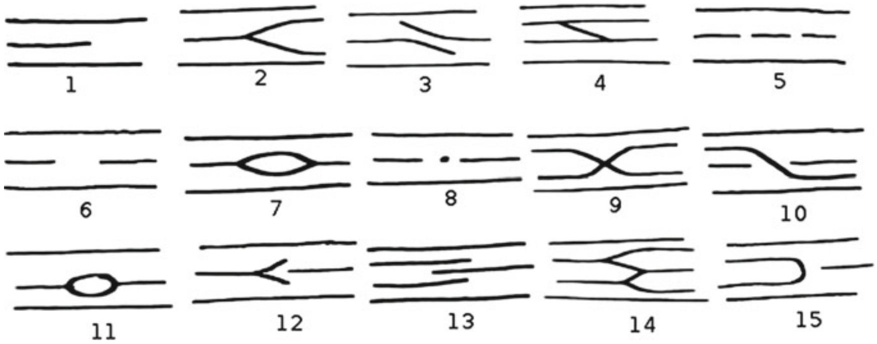


Fig. 14.2 Minutia types used by Guardia Civil. Names corresponding to individual minutia type numbers can be found in Table 14.1

Table 14.1 List of minutia types used by Guardia Civil. Numbering with respect to Fig. 14.2

No	Minutiae type	No	Minutiae type	No	Minutiae type
1	Ridge ending	6	Interruption	11	Circle
2	Bifurcation	7	Enclosure	12	Delta
3	Deviation	8	Point	13	Assemble
4	Bridge	9	Ridge crossing	14	M-structure
5	Fragment	10	Transversal	15	Return

Table 14.2 The probability of occurrence and the entropy-based weights for the minutia types present in the 268 fingermarks of GCDB. The numbers correspond to minutia types in Fig. 14.2

No	Minutiae type	Probability (p_i)	Weight ($w_i = -\log_{10} p_i$)
1	Ridge-ending	0.5634	0.2492
2	Bifurcation	0.3620	0.4413
3	Deviation	0.0015	2.8294
4	Bridge	0.0024	2.6253
5	Fragment	0.0444	1.3523
6	Interruption	0.0021	2.6833
7	Enclosure	0.0204	1.6896
8	Point	0.0036	2.4492
10	Transversal	0.0003	3.5284

GCDB used in this work consists of 268 fingermark and reference fingerprint images and minutia sets. All the minutiae in the fingermark images were manually extracted by forensic examiners of Guardia Civil. The corresponding mated minutiae in the reference fingerprints were also manually established. This includes the typical (ridge-endings and bifurcations) minutiae and the rare minutiae. The minutiae in the reference fingerprints were combined with the minutiae extracted by Neurotechnology Verifinger algorithm, in order to generate a complete set of minutiae for the reference fingerprint. For the fingermark, the minutiae will be the ones marked by the examiners. The average number of minutiae in the fingermarks was 13 and that of tenprints was 125.

The original fingermark minutia sets provided by Guardia Civil and the post-processed VeriFinger generated minutia sets are used in all our experiments. To represent some rare minutiae, multiple points were needed. For example, to represent a *deviation* two points are needed (see type 3 in Fig. 14.2), and to represent an *assemble* three points are needed (see type 13 in Fig. 14.2). Whenever multiple points are needed to represent a rare minutia, we mapped them to a single point representation by taking the average of locations and orientations of all points.

From the 268 fingermark minutia sets, we estimated the probability of occurrence (p_i) of various minutia types. The probability (p_i) and the entropy-based weights ($w_i = -\log_{10} p_i$) for each minutia type present in GCDB are listed in Table 14.2. In the 268 fingermarks of GCDB, we noticed only seven types of rare minutia features. They are listed in Table 14.2. Other rare minutia types are not found in the current database used in this study, because they did not appear in the whole database.

14.7.2 Biometric System

The system used to compare the minutiae was based on Minutiae Cylinder Code (MCC) representation, also extensively presented in another chapter of this book, deeply described in [39–41].[6] It is not the aim to deeply describe the score computation system in this chapter, because we aim at the likelihood ratio computation process. Details about the algorithm can be found in the relevant references [39–41].

In order to exploit the information in the rare minutiae features in the GCDB, the minutiae included in those rare points are part of the features directly added to the ridge endings and bifurcations. Thus, everything together has been used to feed the MCC system. Therefore, the scores obtained by the system include information from both typical and rare minutiae.

Finally, a T-Norm stage has been performed in order to align and normalize the output different-source scores of the system. Test-Normalization, or T-Norm [44] has been used to perform score normalization. In order to do that, a set of different-source fingerprints, namely a *cohort* of different sources, is needed. From those so-called T-Norm scores, the mean and the standard deviation μ_{Tnorm} and σ_{Tnorm} are computed. The T-Norm technique is then applied to a particular score computed form a given fingermark query as follows:

$$s_{Tnorm} = \frac{s_{raw} - \mu_{Tnorm}}{\sigma_{Tnorm}} \tag{14.11}$$

Thus, T-Norm performs query-dependent score normalization, and the result is the alignment of the query-dependent different-source score distributions for all comparisons in the particular set of scores.

Thus, this normalization technique compensates variability in the scores due to the recovered fingermark. The T-Norm cohort in this experiment has been selected from the same Guardia Civil database that has been used to simulate real forensic fingermark-reference fingerprint comparisons, and therefore the results may be overstated in terms of performance. However, for the sake of illustration on the computation of likelihood ratios, this is not a problem.

It has been reported that the different-source scores of a biometric system tend to be more Gaussian after the application of T-Norm [31]. Therefore, we will assume that a Gaussian model will be appropriate for the MCC scores after T-Norm is applied.

[6] We have used the implementation of this score computation system provided by the authors.

14.7.3 Methodology and Proposed LR Methods

This section proposes several methods for likelihood ratio computation using scores from the MCC algorithm with the Guardia Civil database (GCDB) described in Sect. 14.7.1.

14.7.3.1 Definition of Propositions

According to the methodology of CAI, the first step to compute likelihood ratios is to establish the propositions considering the information present in the case. The *simulated cases* that we are going to conduct here consist of the comparison of one fingermark and one reference fingerprint. Both fingermark and reference fingerprint come from GCDB. The scores used to train the models for the LR computation are the rest of scores in the GCDB generated from individuals different from the donors of the fingermark and the reference fingerprint. In this way, the models are trained with scores not used in the case, and the data handling is honest in the sense of the performance measurement.

According to this setup, there are several observation that are in order:

- The information in the case is almost non-existent. We only have the images of the fingermark and the reference fingerprint, and therefore no assumption can be done about the donors of fingermark and reference (e.g. ethnicity, gender, etc.). This only allows generic propositions about the populations involved.
- The trace and reference specimens are pseudonymised because the metadata of the donors are not necessary.
- We only have a single same-source comparison for each subject in the database. Therefore, it is impossible for us to focus in source-specific models, because there are no additional data available to model the particular behavior of their scores in comparison to the whole population of scores.
- There is no information whatsoever about the relevance of the donor of fingermark and reference fingerprint with respect to the action in the crime scene, or even more with respect to any offense. Therefore, only propositions at source level can be addressed.
- Because of the way it was built, we assume that all fingermarks in the GCDB dubbed as different in the ground-truth labels are generated by different people. It is assumed also in the corresponding reference fingerprint. Therefore, in this database it will be equivalent to talk about donors as about fingers, since different fingerprints will definitely belong to different donors (and not to different fingers of the same donor).

Under these premises, we decide to state source level, person-generic and general-population propositions for this case. Therefore, we have the following propositions:

H_p: The origin of the fingermark (trace) and the fingerprint (reference) is the same finger of the donor.

H_d: The origin of the fingermark (trace) and the fingerprint (reference) are fingers from two different donors.

This definition of the propositions implies that, for a forensic case involving the comparison of a fingermark and its corresponding reference fingerprint, the scores needed to train the LR model should be generated with comparisons of fingermarks and reference fingerprints without the constrain of belonging to a particular individual. This implies that more scores will be typically available to train the models, therefore improving their statistical robustness. On the other hand, the use of person-generic propositions inevitably implies an important loss of information in cases where the identity of the individual is known, as it is typical in court. However, for this example we will consider this person-specific scenario because of the limitations of the GCDB, as explained above.

14.7.4 Likelihood Ratio Models

As example in this chapter, we will compare the performance of the following common models for likelihood ratio computation.

- Pool Adjacent Violators.
- Gaussian-ML.
- Logistic regression.

14.7.5 Experimental Results

14.7.5.1 Experimental Protocol

The experimental protocol has been designed in order to simulate a real forensic scenario where fingermarks are compared with reference fingerprints using typical minutia features and also rare minutia features.

In our experiments, we used the Guardia Civil database (as described in Sect. 14.7.1), because it is the only forensic fingerprint database which contains rare minutiae, as it has been previously described. Since the GCDB is limited in size, a cross-validation strategy has been followed in order to optimally use the data without using the same dataset to train and test the LR models proposed. This cross-validation strategy is described as follows: for each same-source comparison of a fingermark and a reference fingerprint, the scores to train the LR model for that particular comparison will consist of all the scores generated with the GCDB, except those generated with either the fingermark or the reference fingerprint involved in the case. Therefore, the separation between the fingermark and reference fingerprint and the individuals in the training database is guaranteed.

This cross-validation strategy has many advantages in the sense of the optimal usage of the available database. However, it also presents the disadvantage that the conditions of the training scores matches the conditions of the fingermark and reference fingerprint under comparison to a higher degree than in a potential real case. Thus, the results presented here could be overstated in terms of performance. However, due to the limitation of the database, and also because the aim of the work is to show how to apply the methodology, we consider it appropriate to use this protocol.

Notice that this cross-validation strategy does not only guarantee that the data used to train and test the models are different. Moreover, it also guarantees that the T-Norm scores generated with the cohort are not present in the training database. This is because the T-Norm cohort scores must be generated with the scores of the query fingermark, which will be not present in the training database. Therefore, the situation is realistic in the sense of the data handling to normalize the scores and also to train the LR models.

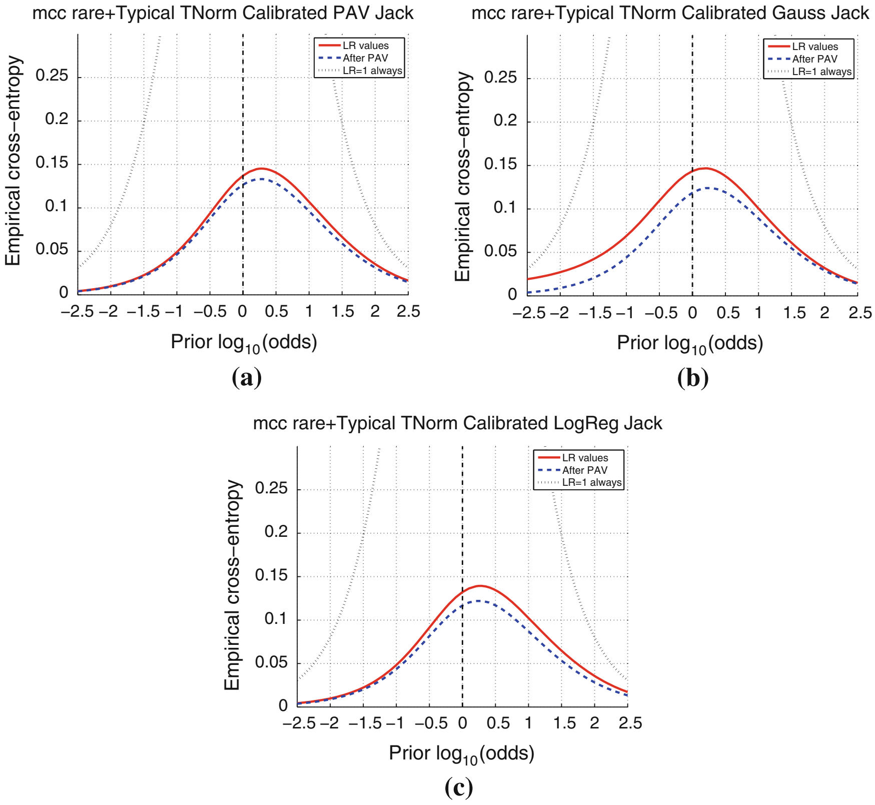
14.7.6 Results on the Comparison of LR Computation Methods

In this section, we compare all the proposed LR computation methods not only from the perspective of the discriminating power, but also with respect to the calibration loss. Thus, accuracy as the sum of both performance measures will allow us to select the best choice for LR computation. From Fig. 14.3, it is seen that the logistic regression model presents the best accuracy (red solid curve), and therefore is the best of the three methods proposed.

We now analyze calibration (separation between red and blue curves) more deeply. It is generally seen in Fig. 14.3 that the calibration loss is better for PAV and logistic regression methods rather than for the Gaussian method. Thus, both methods are apparent good options for LR computation. Regarding the Gaussian-ML method, it is seen that the calibration performance is worse than for PAV or logistic regression, especially in the region of low prior odds. As a possible explanation, although T-Norm different-source scores tend to be Gaussian when they are pooled among all queries, it is not the case for the same-source scores, and this makes the same-source distribution to seriously diverge from Gaussianity even after T-Norm is applied.

An additional warning is in order here. The cross-validation procedure to train LR models and to select T-Norm scores implies a scenario with low dataset shift between training and testing data. In a forensically realistic setup, where dataset shift between training and testing data can be severe, the performance of LR methods that excessively fits the training data can seriously degrade. On the other hand, it is known in pattern recognition that models with lower complexity are more robust to this effect to avoid overfitting. Therefore, the much lower complexity of logistic regression with respect to PAV indicates that the former can be potentially more robust to overfitting and dataset shift than the latter in forensically realistic conditions. Due to this reason, logistic regression is more preferable to PAV computation method in this scenario.

As a conclusion of this section, the calibration loss represents a low percentage of all the loss of accuracy for logistic regression and PAV LR computation methods, in this order. This makes the overall performance of logistic regression superior, which among other reasons makes it the best choice. On the other hand, Gaussian-ML presents poorer calibration loss, sometimes presenting worse performance than the neutral reference, which makes it less recommendable for the score computation systems proposed in this chapter.



(a) **(b)** **(c)**

Fig. 14.3 ECE plots showing performance of LR methods with T-Normed MCC scores. The different LR methods are PAV (**a**), Guassian-ML (**b**) and Logistic regression (**c**)

14.8 Conclusions

In this chapter, we have described the methodology for computing likelihood ratios from biometric scores, with an example of fingerprint recognition in forensic conditions using rare minutiae. This has allowed the interpretation of the evidence from the fingermark-to-reference-fingerprint comparisons simulating a real forensic case by the use of a cross-validation strategy with the GCDB. Several LR methods have been proposed and compared in terms of discriminating power and calibration performance. These results clearly show that the proposed methods present far better performance than the neutral reference and therefore are useful for forensic interpretation.

References

1. Saks MJ, Koehler JJ (2005) The coming paradigm shift in forensic identification science. Science 309(5736):892–895
2. Cook R, Evett IW, Jackson G, Jones PJ, Lambert JA (1998) A model for case assessment and interpretation. Sci Justice 38:151–156
3. Aitken CGG, Taroni F (2004) Statistics and the evaluation of evidence for forensic scientists. Wiley, Chichester
4. Berger CA, Champod JS, Curran C, Dawid J, Kloosterman AP (2011) Expressing evaluative opinions: a position statement. Sci Justice 51:1–2. Several signatories
5. Willis S (2015) ENFSI guideline for the formulation of evaluative reports in forensic science. Monopoly Project MP2010: the development and implementation of an ENFSI standard for reporting evaluative forensic evidence. Technical report, European Network of Forensic Science Institutes
6. Ramos D (2007) Forensic evaluation of the evidence using automatic speaker recognition systems. PhD thesis, Depto. de Ingenieria Informatica, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid, Spain. http://atvs.ii.uam.es
7. Brümmer N, du Preez J (2006) Application independent evaluation of speaker detection. Comput Speech Lang 20(2–3):230–275
8. van Leeuwen D, Brümmer N (2007) An introduction to application-independent evaluation of speaker recognition systems. In: Müller C (ed) Speaker classification. Lecture notes in computer science/Artificial intelligence, vol 4343. Springer, Heidelberg, Berlin, New York
9. Ramos D, Gonzalez-Rodriguez J (2013) Reliable support: measuring calibration of likelihood ratios. Forensic Sci Int 230:156–169
10. Zadora G, Ramos D (2010) Evaluation of glass samples for forensic purposes–an application of likelihood ratio model and information-theoretical approach. Chemometr Intell Lab Syst 102:62–63
11. Li P, Fu Y, Mohammed U, Elder J, Prince SJD (2010) Probabilistic models for inference about identity. IEEE Trans Pattern Anal Mach Intell (PAMI) 34(1):144–157
12. Dehak N, Kenny P, Dehak R, Dumouchel P, Ouellet P (2010) Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Lang Process 19(4):788–798
13. Villalba J, Brümmer N (2011) Towards fully Bayesian speaker recognition: integrating out the between-speaker covariance. Proceedings of the 12th annual conference of the international speech communication association, Interspeech 2011. Florence, Italy, pp 505–508
14. Zadora G, Martyna A, Ramos D, Aitken C (2014) Statistical analysis in forensic science: evidential values of multivariate physicochemical data. Wiley

15. Rodriguez CM, de Jongh A, Meuwly D (2013) Introducing a semiautomatic method to simulate large numbers of forensic fingermarks for research on fingerprint identification. J Forensic Sci 57(2):334–342
16. van Leeuwen DA, Brümmer N (2013) The distribution of calibrated likelihood-ratios in speaker recognition. arXiv preprint
17. Ramos D, Gonzalez-Rodriguez J, Zadora G, Aitken C (2013) Information-theoretical assessment of the performance of likelihood ratio models. J Forensic Sci 58:1503–1518
18. Haraksim R, Ramos D, Meuwly D, Berger CE (2015) Measuring coherence of computer-assisted likelihood ratio methods. Forensic Sci Int 249:123–132
19. Meuwly D, Ramos D, Haraksim R (in press) A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. Forensic Sci Int. 10.1016/j.forsciint.2016.03.048
20. Taroni F, Aitken C, Garbolino P, Biedermann A (2006) Bayesian networks and probabilistic inference in forensic science. Wiley
21. Cook R, Evett IW, Jackson G, Jones PJ, Lambert JA (1998) A hierarchy of propositions: deciding which level to address in casework. Sci Justice 38(4):231–239
22. Evett IW, Jackson G, Lambert JA (2000) More on the hierarchy of propositions: exploring the distinction between explanations and propositions. Sci Justice 401(1):3–10
23. Champod C, Evett IW, Jackson G (2004) Establishing the most appropriate databases for addressing source level propositions. Sci Justice 44(3):153–164
24. Evett IW (1998) Towards a uniform framework for reporting opinions in forensic science casework. Sci Justice 38(3):198–202
25. Neumann C, Evett IW, Skerrett JE, Mateos-Garcia I (2011) Quantitative assessment of evidential weight for a fingerprint comparison I: generalisation to the comparison of a mark with set of ten prints from a suspect. Forensic Sci Int 207:101–105
26. Neumann C, Evett I, Skerret JE (2012) Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm. J R Stat Soc Ser A: Stat Soc 175(2):371–415
27. Taroni F, Aitken CGG, Garbolino P (2001) De Finetti's subjectivism, the assessment of probabilities and the evaluation of evidence: a commentary for forensic scientists. Sci Justice 41(3):145–150
28. Doddington G, Liggett W, Martin A, Przybocki M, Reynolds DA (1998) Sheeps, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In: Proceedings of ICSLP
29. Hepler AB, Saunders CP, Davis LJ, Buscaglia J (2011) Score-based likelihood ratios for handwriting evidence. Foresic Sci Int 219(1–3):129–140
30. Meuwly D (2001) Reconaissance de Locuteurs en Sciences Forensiques: L'apport d'une Approache Automatique. PhD thesis, IPSC-Universite de Lausanne
31. Navratil J, Ramaswamy G (2003) The awe and mystery of T-Norm. In: Proceedings of ESCA European conference on speech, communication and technology, EuroSpeech, pp 2009–2012
32. Gonzalez-Rodriguez J, Fierrez-Aguilar J, Ramos-Castro D, Ortega-Garcia J (2005) Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems. Forensic Sci Int 155(2–3):126–140
33. Pigeon S, Druyts P, Verlinde P (2000) Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. Digit Signal Process 10(1):237–248
34. Brümmer N, Burget L, Cernocky J, Glembek O, Grezl F, Karafiat M, van Leeuwen DA, Matejka P, Scwartz P, Strasheim A (2007) Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. IEEE Trans Audio Speech Signal Process 15(7):2072–2084
35. Gonzalez-Rodriguez J, Rose P, Ramos D, Toledano DT, Ortega-Garcia J (2007) Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. IEEE Trans Audio Speech Signal Process 15(7):2072–2084
36. Vergeer P, Bolck A, Peschier LJ, Berger CE, Hendriks JN (2014) Likelihood ratio methods for forensic comparison of evaporated gasoline residues. Sci Justice 56(6):401–411

37. Morrison GS (2009) Likelihood-ratio-based forensic speaker comparison using parametric representations of vowel formant trajectories. J Acoust Soc Am 125:2387–2397
38. Duda RO, Hart PE, Stork DG (2001) Pattern classification. Wiley
39. Cappelli R, Ferrara M, Maltoni D (2010) Minutia cylinder-code: a new representation and matching technique for fingerprint recognition. IEEE Trans Pattern Anal Mach Intell 32:2128–2141
40. Cappelli R, Ferrara M, Maltoni D (2010) Fingerprint indexing based on minutia cylinder code. IEEE Trans Pattern Anal Mach Intell 33:1051–1057
41. Ferrara M, Maltoni D, Cappelli R (2012) Noninvertible minutia cylinder-code representation. IEEE Trans Inf Forensics Secur 7:1727–1737
42. Egli N (2009) Interpretation of partial fingermarks using an automated fingerprint identification system. PhD thesis, Institute de Police Scientifique, Ecole de Sciences Criminelles
43. Santamaria F (1955) A new method of evaluating ridge characteristics. Fingerprint Ident Mag
44. Auckenthaler R, Carey M, Lloyd-Tomas H (2000) Score normalization for text-independent speaker verification systems. Digit Signal Process 10:42–54