Published in IET Biometrics
Received on 7th February 2014
Revised on 8th July 2014
Accepted on 16th July 2014
doi: 10.1049/iet-bmt.2014.0009

ISSN 2047-4938

Biometric evidence evaluation: an empirical assessment of the effect of different training data

Tauseef Ali[1], *Luuk Spreeuwers*[1], *Raymond Veldhuis*[1], *Didier Meuwly*[2]
[1]Biometric Pattern Recognition Group, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands
[2]Netherlands Forensic Institute, Laan van Ypenburg 6, 2497 GB, The Hague, The Netherlands
E-mail: T.Ali@utwente.nl

Abstract: For an automatic comparison of a pair of biometric specimens, a similarity metric called 'score' is computed by the employed biometric recognition system. In forensic evaluation, it is desirable to convert this score into a likelihood ratio. This process is referred to as calibration. A likelihood ratio is the probability of the score given the prosecution hypothesis (which states that the pair of biometric specimens are originated from the suspect) is true divided by the probability of the score given the defence hypothesis (which states that the pair of biometric specimens are not originated from the suspect) is true. In practice, a set of scores (called training scores) obtained from the within-source and between-sources comparison is needed to compute a likelihood ratio value for a score. In likelihood ratio computation, the within-source and between-sources conditions can be anchored to a specific suspect in a forensic case or it can be generic within-source and between-sources comparisons independent of the suspect involved in the case. This results in two likelihood ratio values which differ in the nature of training scores they use and therefore consider slightly different interpretations of the two hypotheses. The goal of this study is to quantify the differences in these two likelihood ratio values in the context of evidence evaluation from a face, a fingerprint and a speaker recognition system. For each biometric modality, a simple forensic case is simulated by randomly selecting a small subset of biometric specimens from a large database. In order to be able to carry out a comparison across the three biometric modalities, the same protocol is followed for training scores set generation. It is observed that there is a significant variation in the two likelihood ratio values.

1 Introduction

For a given pair of biometric specimens, a score computed by a biometric recognition system quantifies the similarity between the input pair of biometric specimens while taking into account their typicality. In biometric applications such as access-control and e-passport gates at some airports, the developer of the system chooses a threshold from the range of the score and consequently any score above the threshold implies a positive decision and vice versa [1, 2]. However, in a criminal case, it is desirable to report a likelihood ratio (LR) instead of a score or a decision based on a selected threshold [3]. This distinction between biometric and forensic applications is addressed in detail recently by Meuwly [2]. Once a forensic scientist has computed the LR, it is the responsibility of the judge or the jury to make a decision which involves other sources of information about the case at hand such as other types of evidences. Use of a LR value to report the output of a biometric comparison is gradually becoming a standard way of evidence evaluation from score-based biometric systems. A LR is a more informative, balanced and useful output in forensic evaluation than simply a score [4]. A general description of the LR concept for evidence evaluation from biometric systems can be found in [4, 5]. It is applied to several biometric modalities including forensic voice [6], speech [7–9] and fingerprint comparison [10]. Preliminary results of evidence evaluation using a LR value in the context of face and handwriting recognition systems are presented in [11–14]. A LR is the probability of the score given the prosecution hypothesis is true divided by the probability of the score given the defence hypothesis is true

$$LR(s) = \frac{P(s|H_p, I)}{P(s|H_d, I)} \tag{1}$$

where s, considered as the evidence, is the score obtained by comparison of the biometric specimen from the suspect with that found at the crime scene. I refers to background information which may or may not be domain specific. H_p and H_d are two mutually exclusive and exhaustive source-level hypotheses defined as follows:

- H_p: The pair of biometric specimens is originated from the suspect.
- H_d: The pair of biometric specimens are not originated from the suspect.

Once a forensic scientist has computed the LR value, one way to interpret it is as a multiplicative factor which updates the prior odds (before observing the evidence from a biometric system) to the posterior odds (after observing the evidence from a biometric system) using the Bayesian probabilistic framework

$$\frac{P(H_p|s, I)}{P(H_d|s, I)} = \frac{P(s|H_p, I)}{P(s|H_d, I)} \times \frac{P(H_p|I)}{P(H_d|I)} \tag{2}$$

In this framework, the judge or the jury is responsible for quantification of the prior beliefs about H_p and H_d while the forensic scientist is responsible for the scientific analysis of the pair of biometric specimens and quantification of its evidential value in the form of a LR.

The hypotheses H_p and H_d can be specifically interpreted in a slightly different way so that the within-source and between-sources conditions are linked with the specific suspect in a forensic case or they are independent of the suspect. These different interpretations correspond to difference in the pairs of biometric specimens used to obtain the distribution of scores under H_p and H_d. In forensic evaluation, these pairs of biometric specimens are called training data (or calibration data). The purpose of this paper is to quantify the evidential value from a face, a fingerprint and a speaker recognition system using the likelihood ratio concept and to study the effect of different training data on resultant likelihood ratio values using a simple simulated forensic LR evaluation scenario. The study is carried out for a single biometric recognition system from each of the three biometric modalities; however, the concepts and procedure described to obtain the training data and compute LR values apply to any biometric system which computes a score for an input pair of biometric specimens.

The paper is organised as follows. In Section 2, we briefly review the procedure of LR computation from scores and discuss the employed score-to-LR computation method. Section 3 discusses the two different approaches for selection of the training data and the differences in the interpretation of the hypotheses they imply. Section 4 reviews existing work which studies the effects of different training data on LR values and presents the comparison procedures followed in this paper. Section 5 explains experimental setup by introducing the three biometric recognition systems, databases of biometric specimens and the way the training scores sets are constructed. Section 6 presents results by mapping score values to Log_{10} LR (LLR) values using the two different sets of the training scores. Finally Section 7 draws conclusions and points toward future research directions.

2 Computation of a likelihood ratio (LR) from a score

2.1 Computation of training scores

Score-based biometric systems output two classes of scores. The first one is the result of the comparison of two biometric specimens produced by a same source. When comparing a set of biometric specimens produced by a same source, there is some variation in the score values output by a biometric system. Each biometric modality has different nature of variations in the biometric specimens produced by a same source, for example, in case of face recognition systems it is caused by lighting condition, facial expressions, partial occlusion of the face and so on. A set of scores is obtained by comparing biometric specimens from a same source represent the within-source variability of the score and is referred to as the within-source scores. Similarly, comparing a set of biometric specimens produced by different sources results in a set of scores that represent the between-source variability of the score and is referred to as the between-source scores (Researchers in different biometric modalities use different terminologies for the within-source and between-source scores such as 'target and non-target scores', 'genuine and impostor scores' and 'same-source and different-sources scores'.) (see Fig. 1). Scores in the within-source and in the between-source sets are collectively called training scores where the pair of biometric specimens to obtain these training scores are referred to as training data.

2.2 Mapping a score to a LR using the training scores

Score-based LR computation can be considered as a mapping function from score to LR. Given a set of training scores, there are several methods to map the score-axis to LLR-axis. LR values in logarithmic scale are preferred for

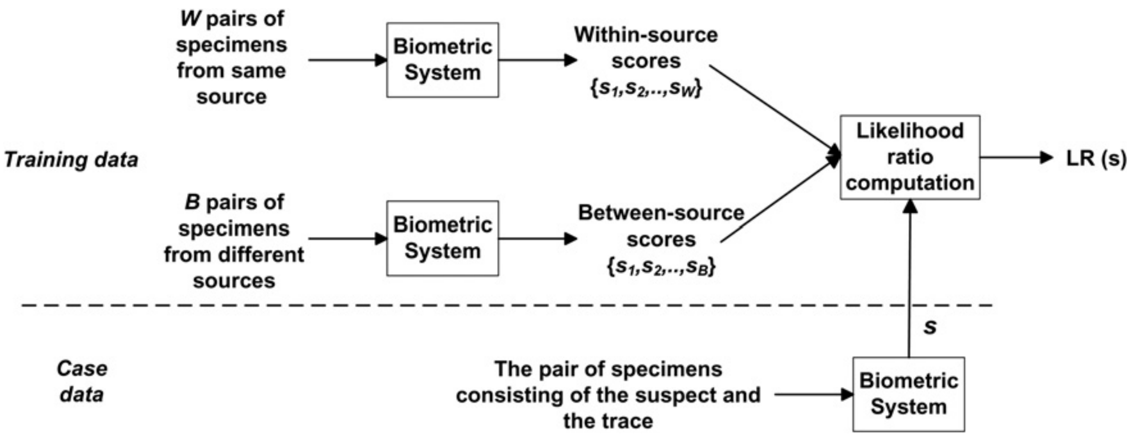
W pairs of specimens from same source
Biometric System
Within-source scores {s1,s2,..,sW}
Training data
B pairs of specimens from different sources
Biometric System
Between-source scores {s1,s2,..,sB}
Likelihood ratio computation
LR (s)
s
Case data
The pair of specimens consisting of the suspect and the trace
Biometric System

Fig. 1 *Computation of a score-based LR for a given pair of biometric specimens consisting of the trace biometric specimen and the suspect biometric specimen*

Same biometric system must be used to compute the within-source scores, the between-source scores and the evidence score *s*

doi: 10.1049/iet-bmt.2014.0009

plotting purposes as well as it has intuitive appeal for forensic practitioners. As an example, a LLR value of 1 can be interpreted as 'It is 10 times more likely that the two biometric specimens are originated from a same source than if they were originated from different sources'. Similarly, a LLR value of −1 can be interpreted as 'It is ten times more likely that the two biometric specimens are originated from different sources than if they were originated from a same source'. Several methods of conversion of biometric scores to LR values are described in [9, 15] and can be classified as parametric or non-parametric. When the distributions of the scores in the within-source and in the between-source sets are similar to available probability density functions (PDFs) such as Gaussian, Exponential and Weibull, the PDFs of the scores under H_p and H_d can be estimated by selection of a pair of PDFs from the available PDFs and finding the specific parameters of those PDFs using maximum likelihood estimation [10]. Once the two PDFs are estimated, a LR is computed by dividing the PDF of scores under H_p by the PDF of scores under H_d (see Fig. 2). Another possible parametric approach is to estimate the ratio of the PDF under H_p and the PDF under H_d using logistic regression [16]. In non-parametric category, there are histogram binning, kernel density estimation (KDE) and finding slope of the receiver operating characteristic convex hull (ROCCH) [15, 17]. Logistic regression and ROCCH approaches have a desirable property: both of them produce a monotonically increasing function from score to LR values. For the purpose of this study, we propose the use of ROCCH procedure because it can ensure to a greater extent that the resultant variation in LR values are because of the difference in the training scores set and not because of the poor fitting of the PDFs or the logistic regression model to the training score sets.

Readers are referred to [17] for the algorithm to construct the ROCCH from a given set of training scores. Once the ROCCH is constructed, LR value for a given score s is the slope of the corresponding segment of the ROCCH on which score s lies and can be computed as follows

$$LR(s) = \frac{w_s}{b_s} \times \frac{W}{B} \tag{3}$$

where w_s and b_s are the number of the within-source and the between-source scores, respectively, in the corresponding segment of the ROCCH on which score s lies. The value W and B are number of scores in the complete sets of the within-source and the between-source in the training scores set. It is interesting to note that computing ROCCH is equivalent to computing receiver operating characteristic (ROC) curve of the posterior probabilities obtained by pool adjacent violators (PAV) algorithm [17]. This argument leads to an alternative way of implementation; computing posterior probabilities using PAV and then plugging it into the Bayesian formula along with W and B to compute LR values. PAV algorithm (or equivalently, the ROCCH approach) is extensively used in forensic speech recognition for computation of LR values [18].

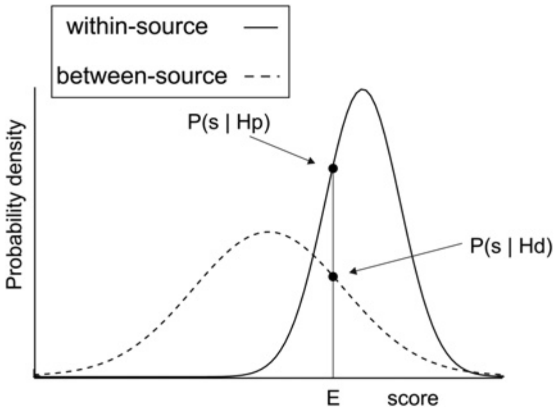

Fig. 2 *Example of LR computation from the estimated PDFs*

E is the evidence score for which the LR is computed. In this example the LR is approximately 2

Once the ROCCH procedure is applied to compute LR values, there is a group of scores for which the LR value is either zero or infinity. The logarithm of these LR values results in minus infinities and plus infinities, respectively. To avoid this problem, a procedure similar to [18] is followed. We insert a score in the between-source set which is equal to the maximum score in the within-source set and a score in the within-source which is equal to the minimum score of the between-source set. These inserted scores can be considered to represent scores which were not encountered in the training scores set because there is not enough training data, but which could have occurred. The resultant LR values replacing zero and infinity are quite intuitive for the use in forensic evaluation and reporting. These values depend on the size of the training scores set. In cases where small training scores set is available, the absolute values of the log-LRs which replace zero and infinity are small. The absolute values of these LLRs increase as the size of the training scores set increase. It simply shows that when small training set of training data is available, only a weak strength of evidence (in terms of LR) is supported empirically by the data. On the other hand, a data-driven approach exploiting large set of training scores offers support for a larger range of likelihood ratios and therefore stronger strengths of evidence can be supported.

3 Choice of the training data

Based on the available number of biometric specimens from the suspect, the within-source and the between-sources conditions can either be anchored to the suspect or it can be general within-source and between-source comparisons using all persons from the potential population defined in a given forensic case. Potential population refers to the set of possible alternate sources of the trace biometric specimen and its size and nature is dependent on the case as well as the analysis of the forensic expert. In extreme situations like in case of a crime in an immerged submarine, an assumption of close-set can be made and the size of the relevant population can be defined precisely. In the vast majority of the forensic cases the relevant population is an open set and its size can only be estimated in terms of ranges.

3.1 *Suspect-specific training data*

To compute the suspect-specific within-source scores, a set of biometric specimens from the suspect can be compared with another set of biometric specimens from the suspect [9]. The two sets of biometric specimens are referred to as 'reference' and 'test' data sets. For better calibration, the biometric specimens in the test data set should be as close as possible to the trace and the biometric specimens in the reference data set should be as close as possible to the database of biometric specimens from the potential population. Cross-comparison of all the biometric specimens in the reference and the test data set results in a set of scores that can be used to model the distribution of scores under the prosecution hypothesis. Similarly, for modelling the distribution of scores under the defence hypothesis, biometric specimens in the test data set are compared with the reference biometric specimens of the potential population database [19]. The suspect-specific approach implies considering the following interpretations of the prosecution and defence hypotheses

- H_p: The pair of biometric specimens is originated from the suspect.
- H_d: The pair of biometric specimens is not originated from the suspect (or alternatively, the pair of biometric specimens is originated from someone else in the potential population).

The difficulty in following the suspect-specific approach is that in most cases it may not be possible to obtain a large set of biometric specimens from the suspect. This leads to fewer scores in the training scores set, particularly the within-source scores set.

3.2 Suspect-independent training data

Certain specific solutions have been proposed as how to increase the number of the within-source scores when following the suspect-specific approach [20, 21]. A general solution is to construct the within-source and between-sources scores sets by combining the suspect-specific sets of multiple persons from the potential population database. To compute the suspect-independent within-source scores, a set of reference and test biometric specimens from the potential population are compared where the two biometric specimens in each pair are obtained from a same source. Similarly, for the suspect-independent between-source scores, a set of reference and test biometric specimens from the potential population are compared where the two biometric specimens in each pair are obtained from different sources. Using the suspect-independent approach to LR computation implies the following interpretations of the prosecution and defence hypotheses:

- H_p: The pair of biometric specimens is obtained from a same source.
- H_d: The pair of biometric specimens is obtained from different sources.

For the between-source scores, besides the suspect-specific and suspect-independent approaches, another commonly used approach is to compute trace-anchored scores. In this approach, the trace biometric specimen is compared with all the reference biometric specimens of the potential population to compute the between-source scores [22].

4 Comparing the resultant LR values

4.1 Motivation

When sufficient training data is available, it is preferred to compute a suspect-specific LR because it takes into account more relevant information about the case at hand. Therefore there is some research on how to compute a suspect-specific LR for a biometric comparison when there is limited training data available from the suspect. Ramos [21] proposed a strategy which is based on the adaptation of the suspect-independent within-source scores distribution to the suspect-specific scores via maximum a posteriori (MAP) estimation. Similarly, in forensic handwriting recognition, Davis [20] generated a large set of simulated writing specimens from a small set of suspect specimens to form a data set for computation of the suspect-specific within-source scores. These specific approaches do not generalise in most cases and usually a suspect-independent approach is considered as a last resort to compute a reliable LR for a given pair of biometric specimens [23]. In [23], suspect-independent approach is proposed as a feasible alternative when a single specimen is available from the suspect. Given the common use of the suspect-independent approach as an alternative to the suspect-specific approach to compute a LR value, it is important to study and analyse the differences in the LR values produced by these two approaches.

4.2 Existing work

Quantifying the difference between the LR values using the suspect-specific and suspect-independent training data is still under investigation in most biometric modalities. In [20], authors describe the effect(s) of different training data used to construct the between-source scores set in the context of handwriting recognition. For fingermark evidence, Alberink *et al.* [24] recently discussed different theoretical possibilities of conditioning such as conditioning on specified fingers, fingerprints and fingermarks in order to compute the training scores set. They also studied the asymmetric conditioning in LR computation which is, however, subject to further debate. Similarly, Ramos *et al.* [9] studied the effect of using suspect-independent within-source scores instead of suspect-specific on the resultant LR values in the context of forensic speaker recognition. In [12], we investigated the effect of the two different approaches of LR computation considering a face recognition system. The work presented in the current paper is an extension of [12]. One of the main focal points of the current paper is to investigate how much each of the three biometric modalities show variation in the resultant likelihood ratios when the training data is changed. The work in [12] compares only the two approaches of likelihood ratio computation. The current paper extends the comparison across three biometric modalities relevant in forensics and is, therefore, of interest to a broader domain of audience, including in general forensic biometrics and in specific to face, speech and fingerprint recognition.

4.3 Comparison approach

A common approach to compare systems producing forensic LR values is to compute a set of test LR values for a set of pairs of biometric specimens whose origin is known. Then the criterion is that a better system should result in a larger value of LR for a within-source pair of biometric specimens and a smaller value of LR for a between-sources pair of biometric specimens. Two common tools that compare systems (more precisely, sets of test LR values produced by systems) based on this criterion are Tippett plot [25] and 'Cost of Log LR (C_{llr})' [18]. Such a comparison approach is very useful in practice; however, the focus of this work is to study how close the two LR values are instead of which LR value is preferred. Therefore in this work, instead of

doi: 10.1049/iet-bmt.2014.0009

following the traditional approach to compare forensic LR computation systems, we propose to study the whole mapping function from score to LLR values and for a given random score, observe how much the LR values differ. We compute the functions from score to LLR using the suspect-specific and suspect-independent sets of training scores and sample them uniformly for a quantitative analysis of the differences in the two LR values. The behaviour of the two score-to-LLR functions is studied in different regions of LLR values. Furthermore, using random subsampling, the effect of the different sizes of the training scores set used in each approach is also investigated.

5 Experimental setup

Fig. 3 illustrates computation of the within-source and the between-source scores in the suspect-specific and suspect-independent approach for a single specimen per person case assuming person 1 is the suspect.

To simulate a forensic case, we randomly select five persons from a large database of biometric specimens in each of the biometric modality. For this purpose, face recognition grand challenge (FRGC) [26] database is used for face data, Dutch National Police database is used for fingerprint data and National Institute of Standards and Technology (NIST) 2010 Speaker Recognition Evaluation (SRE) extended database is used for speech data [27]. The important condition for setting up a rational experimental protocol for such a comparative study across the three different biometric modalities is to use equal number of scores per person in the within-source and between-source sets.

For face data, a set of 36 biometric specimens are chosen randomly for each of the five persons in the selected subset. Then half of the biometric specimens are used to generate the test data set while the remaining half are used as a reference data set. Cross-comparison of the 18 biometric specimens in the test data set and reference data set results in 324 scores in the suspect-specific within-source scores set. For face data, images used as test data set are degraded by adding motion blur of 15 pixels with zero angle and downsampling them by half of the original resolution. The goal of the degradation process is to make the images in the test data set similar to a trace image because the original face images in FRGC database are of very-high resolution (231 × 251). The comparison is performed using a commercial face recognition system developed by Cognitec [28]. The specific values for the blurring and resizing were chosen empirically where the objective was to reduce the quality with the constraint that the resultant images could be compared by Cognitec face recognition system [28]. It should be mentioned that selection of different values will have no significant effect on the analysis. This is because it will mainly shift, along the score-axis, the within-source, the between-source and the evidence score values. Since the LR is the ratio of the two PDFs, horizontal shifting of the functions should have no effect. However, it is possible to obtain slightly different distributions of the within-source and between-source scores using different values of the blur and resolution of the images.

For fingerprint data, there is only one fingermark available which is used as the test data set. However, the number of reference fingerprints are very large. In order to have equal number of scores in the training scores set for each

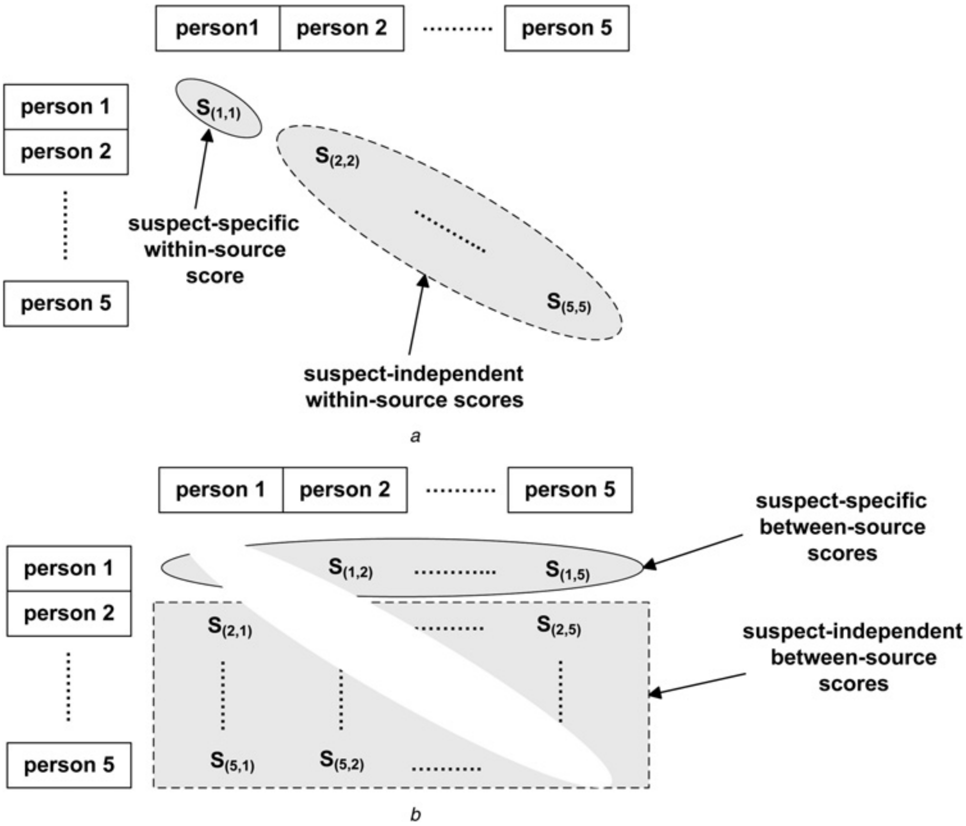
person1
person 2
person 5
person 1
person 2
person 5
S(1,1)
S(2,2)
S(5,5)
suspect-specific within-source score
suspect-independent within-source scores
a
person 1
person 2
person 5
person 1
person 2
person 5
S(1,2)
S(1,5)
S(2,1)
S(2,5)
S(5,1)
S(5,2)
suspect-specific between-source scores
suspect-independent between-source scores
b

Fig. 3 *Within-source and the between-source scores sets assuming the first person as the suspect and 1 biometric specimen per person*

a Computation of the within-source scores sets
b Computation of the between-source scores sets

biometric modality, the size of the suspect-reference data set is increased to 324. Comparison of the one fingermark in the test data set with the 324 fingerprints in the reference data set results in 324 scores in the suspect-specific within-source scores. Motorola Biometric Identification System (BIS) software (version 9.1) is used for comparison of the fingermark to the reference fingerprints. In case of the fingerprint, a fingermark (which simulates a trace biometric specimen) is compared with a high-quality fingerprint and therefore no degradation in the quality of the biometric specimens is required.

For speech data, similar to the face data, 18 specimens are used as the test data set and 18 specimens are used as the reference data set. The recognition algorithm is based on probabilistic linear discriminant analysis approach [29] which models the distribution of *i*-vectors as a multivariate Gaussian. The system is described in [29, Section 2.5] in detail.

Table 1 Number of scores in the set of the within-source and the between-source scores

within-source scores	
suspect-specific	324
suspect-independent	$4 \times 324 = 1296$
between-source scores	
suspect-specific	$4 \times 324 = 1296$
suspect-independent	$4 \times 1296 = 5184$

Table 1 shows the number of unique comparisons (and hence the number of scores) in each approach of the within-source and the between-source scores sets computation given there are five persons in the selected subset.

Beside studying the overall score-to-LLR functions for comparison, for a more quantitative analysis of the differences in LR values, we define score-axis as starting from the minimum value of the score in the suspect-independent between-source scores set and ending at the maximum value of the scores in the suspect-independent within-source scores set. Then we generate 100 evidence scores by uniformly sampling the score-axis and compute the number of cases in which the two LR values agree and disagree on a given range of LR. A disagreement is reported when one approach produces a LR that falls into a different range. These ranges correspond to different verbal equivalents of the numerical LR values which can be used in certain situations to report the forensic evaluation of the evidence. These ranges along with their corresponding verbal equivalents are shown in the left two columns of Table 2 [3].

6 Results

The score-to-LLR functions are computed using the suspect-specific as well as the suspect-independent training

Table 2 Number of times in which the LR values computed by the two approaches fall into a same range considering all of the five persons (P1, P2, P3, P4 and P5) in the selected subset

Ranges	Verbal equivalents		Number of agreements					
			P1	P2	P3	P4	P5	Total
$4 < \mathrm{LLR}$	very-strong evidence to support H_p	face	0	0	0	0	0	0
		fingerprint	0	0	0	0	0	0
		speech	0	0	0	0	0	0
$3 < \mathrm{LLR} \le 4$	strong evidence to support H_p	face	54	54	54	0	0	162
		fingerprint	0	40	0	0	44	84
		speech	0	36	0	38	38	112
$2 < \mathrm{LLR} \le 3$	moderately strong evidence to support H_p	face	0	0	0	0	0	0
		fingerprint	0	0	0	0	0	0
		speech	2	0	2	0	0	4
$1 < \mathrm{LLR} \le 2$	moderate evidence to support H_p	face	9	8	10	8	9	44
		fingerprint	5	0	8	5	4	22
		speech	1	0	2	0	4	7
$0 < \mathrm{LLR} \le 1$	limited evidence to support H_p	face	9	10	6	13	7	45
		fingerprint	8	0	5	6	8	27
		speech	4	4	1	2	3	14
$-1 < \mathrm{LLR} \le 0$	limited evidence to support H_d	face	2	6	0	4	3	15
		fingerprint	5	0	7	4	7	23
		speech	7	4	3	2	3	19
$-2 < \mathrm{LLR} \le -1$	moderate evidence to support H_d	face	0	2	0	5	0	7
		fingerprint	21	0	23	0	23	67
		speech	0	0	0	0	0	0
$-3 < \mathrm{LLR} \le -2$	moderately strong evidence to support H_d	face	4	4	4	2	4	18
		fingerprint	5	5	0	0	3	13
		speech	37	28	0	31	37	133
$-4 < \mathrm{LLR} \le -3$	strong evidence to support H_d	face	0	0	0	0	0	0
		fingerprint	0	0	0	0	0	0
		speech	0	0	0	0	0	0
$\mathrm{LLR} < -4$	very-strong evidence to support H_d	face	1	1	1	1	1	5
		fingerprint	1	1	1	1	1	5
		speech	1	1	1	1	1	5
	total number of agreements	face	79	85	75	33	24	296
		fingerprint	45	46	44	16	90	241
		speech	52	73	9	74	86	294

For each person considered as a suspect, there are 100 values of *s* generated by uniformly sampling the score-axis. Out of a total of 500 LR values computed by the two approaches, 296, 241 and 294 times the two LR values agree on one range for face, fingerprint and speaker recognition systems, respectively

doi: 10.1049/iet-bmt.2014.0009

scores set in order to compare the general behaviour of these functions. Figs. 4–6 show the frequency histograms of the scores in the within-source and in the between-source sets, the ROC curves of the training scores in the suspect-specific and suspect-independent approach and the score-to-LLR functions computed by the ROCCH procedure as described in Section 2.2.

Note the large variations in the histograms of the within-source scores for the suspect-specific approach. Within-source biometric specimens of each person are selected in such a way so that the variations are as close as possible across the five persons. However, still we observe considerable variation in the suspect-specific frequency histograms of scores. These variations are caused by either the slight variation in the specimen acquisition process or because of the fact that some people are easy to be recognised or differentiated from others [30]. As can be observed from the histograms of scores, besides the slight difference in the within-source specimens from person to person, identity itself has a considerable effect on the suspect-specific within-source scores distribution. A biometric recognition system may perform differently for different persons when it is used to match a set of pairs of within-source and between-sources specimens.

The area under the ROC curve is used as a summary metric to assess the discrimination power of a set of within-source and between-sources scores. The motivation behind plotting ROC curves in this context is to demonstrate that there is no general conclusion about the discrimination power when comparing the suspect-specific and suspect-independent training scores.

There is a significant difference in the LR values computed using the suspect-specific and suspect-independent approach. For example, for person 1 in face recognition case, at the score location of 0.44, the suspect-specific and the suspect-independent LR values are 1052 and 78, respectively. The uppermost and the lowermost horizontal lines in the mapping functions are because of the proposed strategy to avoid infinite LLR values. We will refer to LR values along these horizontal lines as ‘saturated LR values’. The magnitude of these LR values is directly proportional to the size of the training scores set and therefore the suspect-independent approach results in saturated LR values of larger magnitude than the suspect-specific approach. In general, for all of the three biometric systems, it can be stated that there are significant differences in the suspect-specific and suspect-independent LR values. Therefore it can be argued that anchoring plays a crucial role in computation of a LR for a given pair of biometric specimens.

In most cases, the exact numerical value of a likelihood ratio is of less importance than the range in which it lies. This fact should be taken into consideration when performing such a comparative study. To this end, the score-axis is uniformly sampled to simulate 100 values of evidence score s. These scores are converted to LLR values using both suspect-specific and suspect-independent training data. For five persons, this implies computation of 500 LLR values using suspect-specific as well as suspect-independent approach in each biometric modality. Table 2 shows the number of cases in which the two approaches compute LLR values which fall into a same range. As seen from Table 2, in 296 cases out of 500 for face recognition, in 241 cases out of 500 for fingerprint recognition and in 294 cases out of 500 for speaker recognition system, the two LR values agree on a same verbal equivalents resulting in 59.2, 48.2 and 58.8% agreement rates for face, fingerprint and speaker recognition systems, respectively.

In case of a disagreement, the differences between the two LRs are mostly to the adjacent classes. However, bigger errors, where the two LR translate to verbal equivalents (classes) which are not adjacent, are also observed. For example, in case of fingerprint recognition system, the two LRs of the score value of 1044 are 2.5 and 0.004. Their verbal equivalents are ‘Moderately strong evidence to support H_d’ and ‘limited evidence to support H_p’. The actual differences can be larger even for adjacent classes. For example, for person 1 in face recognition system, at score location of 0.44, the suspect-specific and the suspect-independent LR values are 1052 and 78, respectively. Although this difference is only to the adjacent class, still it can have significant effect on the conclusion of the forensic analysis. Whether the difference can change the decision or not also depends on the priors which are calculated from the background information about the case at hand.

Note that using a different method to map from score values to LR values may lead to completely different results. Similarly, a different database of biometric specimens and a different biometric recognition system to compute scores can also slightly influence the difference between the suspect-specific and suspect-independent approach of LR computation. Authors are currently investigating the effect on the results when other methods of score-to-LR conversion such as KDE and logistic regression are used.

In the experiments not reported in the paper, a third possibility, ‘generic approach’, is also considered. This is a similar approach to suspect-independent; however, in the generic approach, the suspect data was also included in the suspect-independent sets. The results using the generic approach were very similar to the suspect-independent approach and therefore are not shown in order to avoid cluttered graphs. Furthermore, the suspect-independent approach, unlike the generic approach, is more useful since it provides a way for computation of a LR without obtaining biometric specimens from the suspect.

An obvious difference between the two approaches is the use of different sizes of the training scores sets. One way to study the effect of the difference in the sizes of the training sets between the suspect-specific and suspect-independent approach is to randomly sample a number of scores equal to the size of the suspect-specific sets from the suspect-independent sets. Given the size of the within-source and between-source sets in the two approach is the same, the variation in the LR values is only caused by the nature of the distributions of the scores. Fig. 7 shows the mapping functions computed by the two approaches when the within-source and the between-source sets are equally sized by random subsampling the suspect-independent within-source and between-source sets so that the sizes of these sets in the suspect-independent approach are equal to those in the suspect-specific approach. Note that reduction in the size of the training scores reduces the range of LR values that can be computed. Besides the saturated region of LR values, the difference in the sizes of the training sets has very small effect on the resultant mapping function from score to LLR values. For example, in case of the face recognition system, for $\text{LLR} \leq 2$, the number of agreements considering original sizes of the training sets is 129 whereas it is 134 considering equal sizes of the training sets.

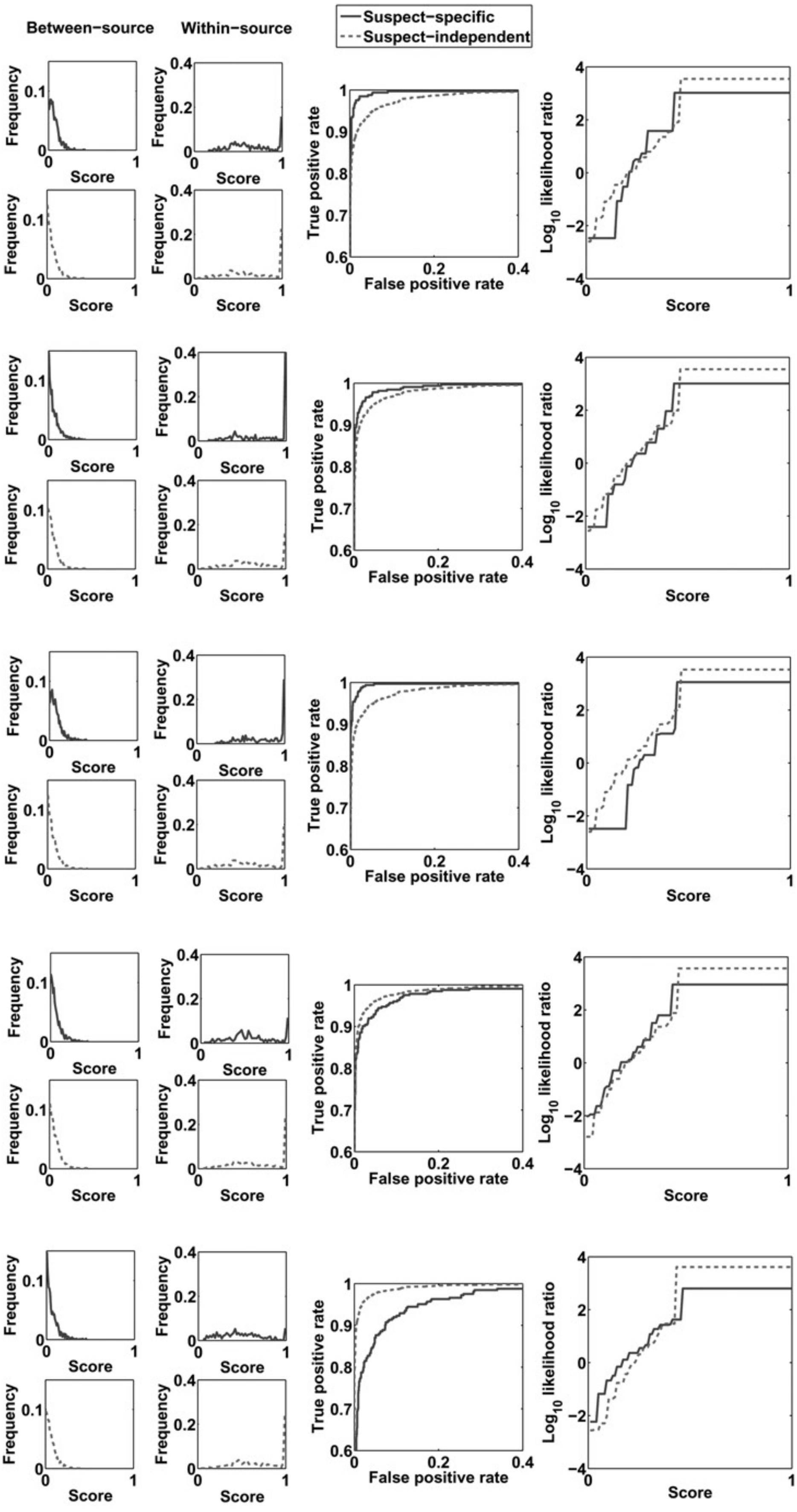

Fig. 4 *Frequency histograms of scores, ROC curves and score-to-LLR functions for the five persons in the selected subset of FRGC face images database*

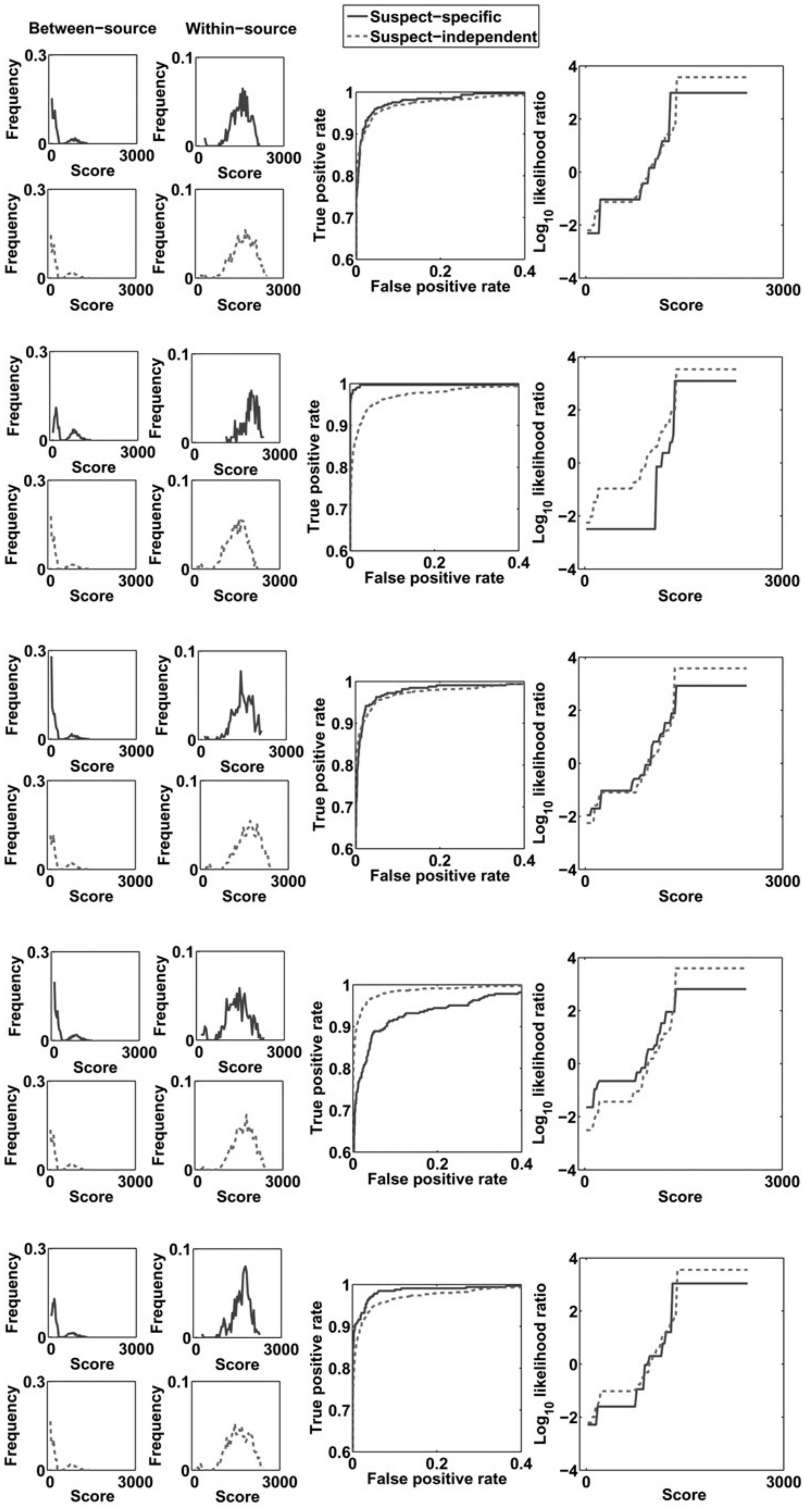

Fig. 5 *Frequency histograms of scores, ROC curves and score-to-LLR functions for the five persons in the selected subset of the fingerprints database*

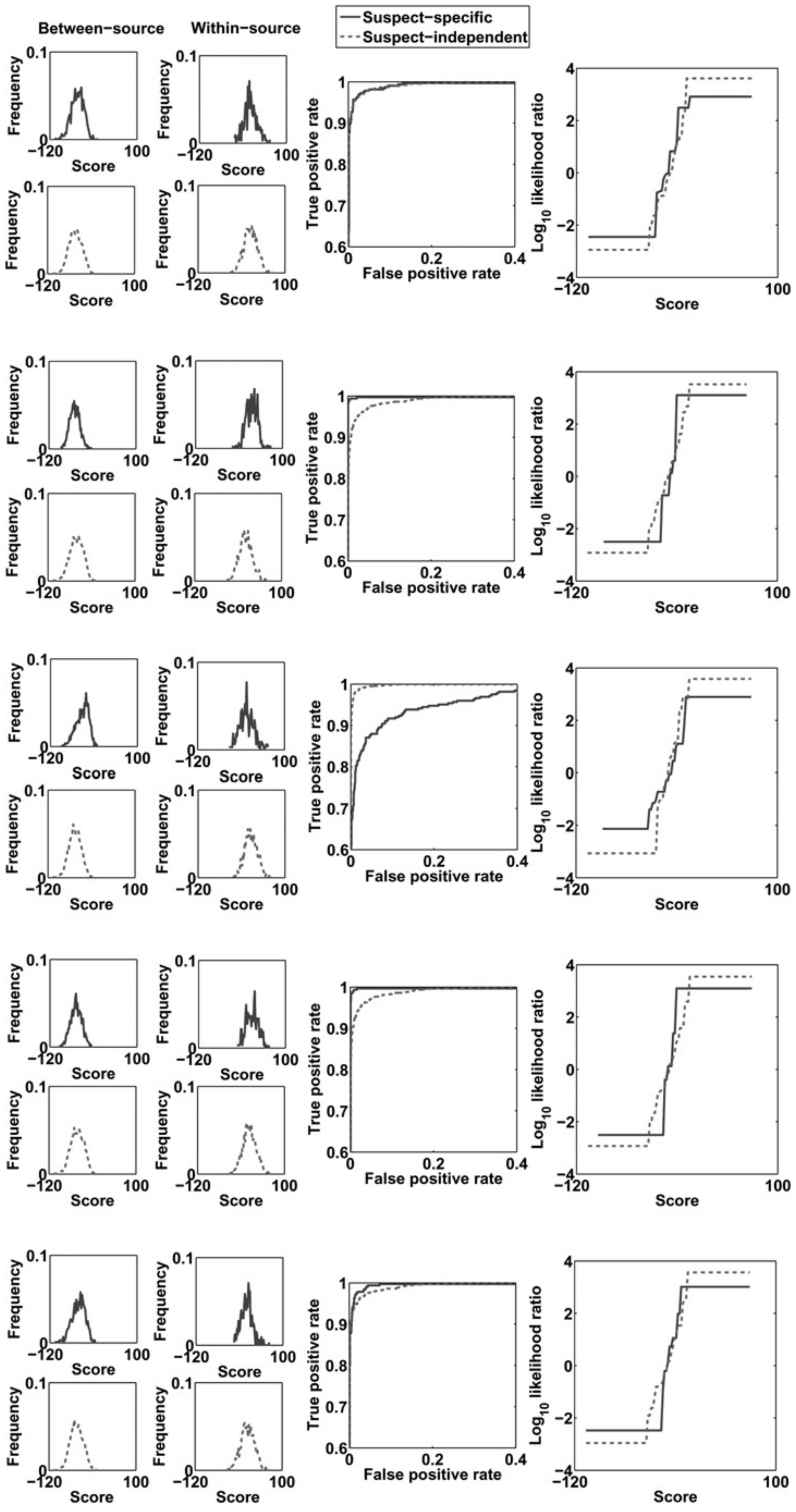

Fig. 6 *Frequency histograms of scores, ROC curves and score-to-LLR functions for the five persons in the selected subset of NIST SRE speech recordings database*

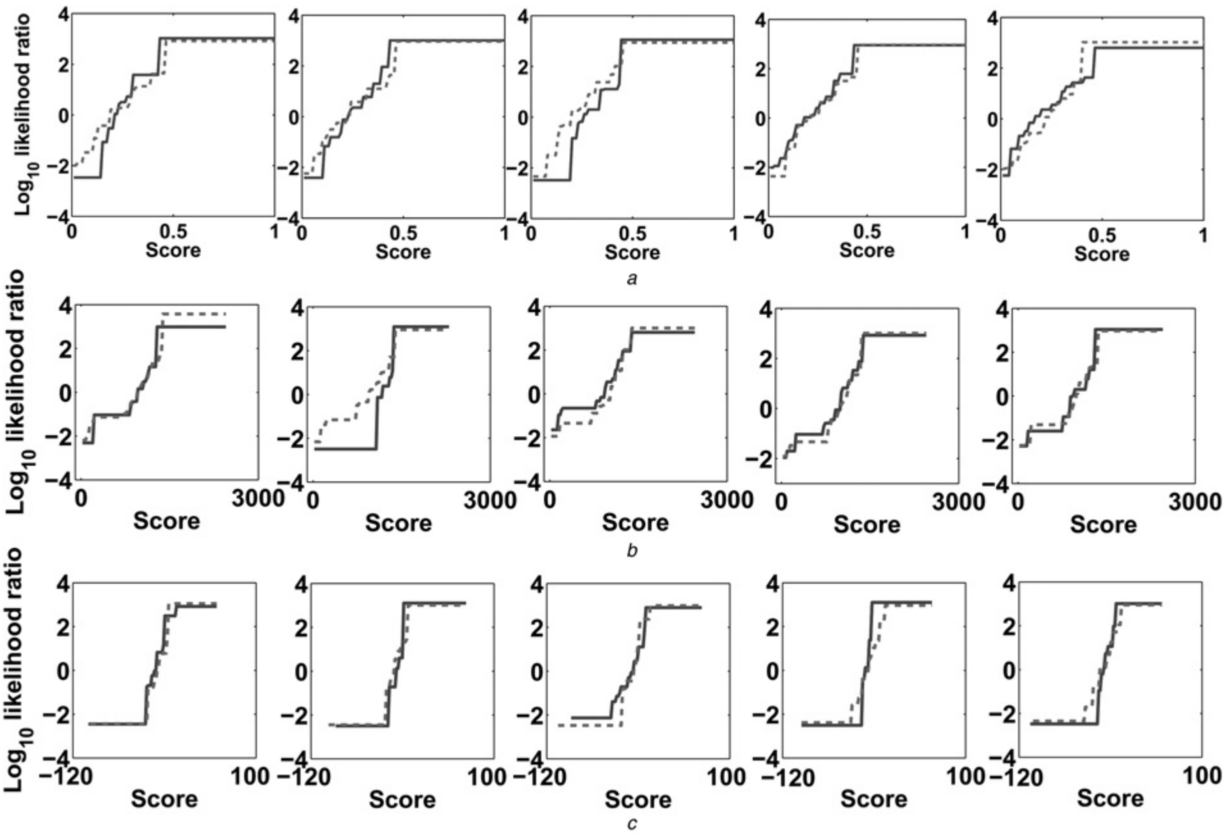

Fig. 7 *Score-to-LLR functions using equal number of specimens in the within-source and between-source sets of the suspect-specific and suspect-independent approach*

Suspect-independent within-source and between-source sets are randomly subsampled so that there are equal number of scores in these sets for the two approaches
a Face recognition system
b Fingerprint recognition system
c Speaker recognition system

It should be mentioned that the suspect-independent approach to compute LRs is more reliable since it uses more training data compared to the suspect-specific approach. The suspect-independent approach is preferred in cases where small datasets of biometric specimens are available from the suspect. The advantage of the suspect-specific approach is that it addresses a more specific set of hypotheses that the suspect-independent approach and therefore provides a more relevant answer. However, a LR computed using small training sets can be very sensitive to random variations in the training sets and the LR computation methods. Since reliability is very important in forensic science therefore it can be concluded that when enough training data for the suspect-specific approach is not available, the suspect-independent approach is more reliable and should be preferred. However, the question of the relevance of the hypotheses tested should be explicitly mentioned while reporting the LR.

7 Discussion and conclusions

We discussed the effect of the different training data on the resultant LR values in the context of face, fingerprint and speaker recognition systems. The process of conversion of a score, computed from the comparison of the crime scene biometric specimen with the suspect biometric specimen, to a forensic LR is described. It is observed that there is a significant variation between the LR values computed using the suspect-specific and the suspect-independent approach. The differences are more prominent in the higher ranges of LR values and therefore more caution should be taken if one approach is used as an alternative to the other.

There has been a lot of research in forensic biometric community and the LR is now considered as one of the most appropriate metrics to be used for reporting the biometric evidence in court. The aspects of the LR computation such as the variability because of different nature of the training data; however, need to be considered. For likelihood ratio computation, ideally the probability density functions of the two sets of training scores are required. However, this is not the case in practice and in best-case scenario, a forensic scientist can use a large number of score in the training sets to model the two scores distributions. This is usually feasible when a suspect-independent approach is followed. However, this approach answers a less relevant question in forensic evidence evaluation than the suspect-specific approach. To enhance the usability of biometric evidence, further works in at least two directions are needed: improvement in the technology for automatic comparison of two biometric specimens and more research on the different aspects of score-to-LR conversion procedure; particularly studies are needed which point out issues so that practitioners are aware of them and take them into account when reporting a LR. A possible implication of the presented study could be the investigation of the robustness of each biometric modality towards the two different kinds of training data. Furthermore, the knowledge of how much the two LRs can be different in worst-case is also useful for forensic practitioners. Further directions for future work include quantifying the influence of biometric specimens from other databases, different biometric recognition systems and other score-to-LR computation methods.

8 Acknowledgments

This research is funded by the European commission as Marie-Curie ITN-project (FP7-PEOPLE-ITN-2008) ‘Bayesian Biometrics for Forensics’.

We would like to thank Cognitec Systems GmbH. for supporting our research by providing the FaceVACS software. Results obtained for FaceVACS were produced in experiments conducted by the University of Twente, and should therefore not be construed as a vendor’s maximum effort full capability result.

Many thanks to Dr. Rahim Saeidi from University of Eastern Finland and Rudolf Haraksim from Netherlands Forensic Institute for providing data for experiments.

9 References

1 Jain, A.K., Flynn, P., Ross, A.: ‘Handbook of biometrics’ (Springer-Verlag, 2007)

2 Meuwly, D., Veldhuis, R.N.J.: ‘Forensic biometrics: from two communities to one discipline’. Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2012, pp. 1–12

3 Lucy, D.: ‘Introduction to statistics for forensic scientists’ (John Wiley & Sons, Inc., 2005)

4 Aitken, C.G.G., Taroni, F.: ‘Statistics and the evaluation of forensic evidence for forensic scientist’ (Wiley, Chichester, UK, 2004, 2nd edn)

5 Robertson, B., Vignaux, G.A.: ‘Interpreting evidence’ (Wiley, Chichester, UK, 1995)

6 Morrison, G.S.: ‘Forensic voice comparison’, in Freckelton, I., Selby, H. (Eds.): ‘Expert evidence’ (Thomson Reuters, Sydney, Australia, 2010), ch. 99

7 Champod, C., Meuwly, D.: ‘The inference of identity in forensic speaker recognition’, *Speech Commun.*, 2000, **31**, (2), pp. 193–203, doi:10.1016/S0167-6393(99)00078-3

8 Rose, P.: ‘Technical forensic speaker recognition’, *Comput. Speech Lang.*, 2006, **20**, pp. 159–191, doi:10.1016/j.csl.2005.07.003

9 Ramos, D.: ‘Forensic evaluation of the evidence using automatic speaker recognition systems’. PhD dissertation, Universidad Autonoma de Madrid, 2007

10 Neumann, C., Evett, I.W., Skerrett, J.: ‘Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm’, *J. R. Stat. Soc. Ser. A*, 2012, **175**, pp. 1–26

11 Peacock, C., Goode, A., Brett, A.: ‘Automatic forensic face recognition from digital images’, *Sci. Justice*, 2004, **44**, pp. 29–34

12 Ali, T., Spreeuwers, L.J., Veldhuis, R.N.J., *et al.*: ‘Effect of calibration data on forensic likelihood ratio from a face recognition system’. IEEE Sixth Int. Conf. Biometrics: Theory, Applications and Systems (BTAS), Washington DC, USA, 2013

13 Ali, T., Spreeuwers, L.J., Veldhuis, R.N.J.: ‘Towards automatic forensic face recognition’. Int. Conf. Informatics Engineering and Information Science (ICIEIS), Communications in Computer and Information Science, Springer Verlag, Kuala Lumpur, Malaysia, 2011, vol. 252, pp. 47–55

14 Hepler, A.B., Saunders, C.P., Davis, L.J., *et al.*: ‘Score-based likelihood ratios for handwriting evidence’, *Forensic Sci. Int.*, 2012, **219**, pp. 129–140, doi:10.1016/j.forsciint.2011.12.009

15 Ali, T., Spreeuwers, L.J., Veldhuis, R.N.J.: ‘A review of calibration methods for biometric systems in forensic applications’. 33rd WIC Symp. on Information Theory in the Benelux, Boekelo, Netherlands, 2012, pp. 126–133

16 Morrison, G.S.: ‘Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio’, *Aust. J. Forensic Sci.*, 2012, **45**, pp. 173–197

17 Fawcett, T., Niculescu-Mizil, A.: ‘PAV and the ROC convex hull’, *Mach. Learn.*, 2007, **68**, pp. 97–106

18 Brummer, N., Preez, J.: ‘Application-independent evaluation of speaker detection’, *Comput. Speech Lang.*, 2006, **20**, pp. 230–275

19 Nordgaard, A., Hoglund, T.: ‘Assessment of approximate likelihood ratios from continuous distributions: a case study of digital camera identification’, *J. Forensic Sci.*, 2011, **56**, pp. 390–402

20 Davis, L.J., Saunders, C.P., Hepler, A.B., *et al.*: ‘Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios’, *Forensic Sci. Int.*, 2012, **219**, pp. 129–140

21 Ramos-Castro, D., González-Rodrguez, J., Montero-Asenjo, A., *et al.*: ‘Suspect-adapted map estimation of within-source distributions in generative likelihood ratio estimation’. Proc. IEEE Odyssey Speaker and Language Recognition Workshop, 2012, doi:10.1109/ODYSSEY.2006.248090

22 Meuwly, D.: ‘Forensic individualization from biometric data’, *Sci. Justice*, 2006, **46**, pp. 205–213

23 Botti, F., Alexander, A., Drygajlo, A.: ‘An interpretation framework for the evaluation of evidence in forensic automatic speaker recognition with limited suspect data’. Proc. Odyssey the Speaker and Language Recognition Workshop, Toledo, Spain, 2004, pp. 63–68

24 Alberink, I., Jongh, A., Rodriguez, C.M.: ‘Fingermark evidence evaluation based on AFIS matching scores: sensitivity of likelihood ratios to different types of conditioning’, *J Forensic Sci*, 2014, **59**, (1), pp. 70–81, doi:10.1111/1556-4029

25 Evett, I.W., Buckleton, J.S.: ‘Statistical analysis of STR data’, in Carracedo, A., Brinkmann, B., Bär, W. (eds): ‘Advances in forensic haemogenetics’ (Springer-Verlag, New York, 1996)

26 Phillips, P.J., Flynn, P.J., Scruggs, T., *et al.*: ‘Overview of the face recognition grand challenge’. Int. Conf. Computer Vision Pattern Recognition, 2005

27 NIST Speaker Recognition Evaluation 2010, http://www.itl.nist.gov/iad/mig/tests/sre/2010/

28 Cognitec FaceVACS SDK version 8.4.0, 2010, http://www.cognitec.com/

29 Mandasari, M.I., McLaren, M., Van Leeuwen, D.: ‘The effect of noise on modern automatic speaker recognition systems’. Proc. ICASSP, Kyoto, 2012

30 Doddington, G., Liggett, W., Martin, A., *et al.*: ‘Sheep, Goats, Lambs and Wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation’. Proc. Int. Conf. Spoken Language Processing, 1998

doi: 10.1049/iet-bmt.2014.0009