Between-Source Modelling for Likelihood Ratio Computation in Forensic Biometric Recognition

Daniel Ramos-Castro[1], Joaquin Gonzalez-Rodriguez[1], Christophe Champod[2], Julian Fierrez-Aguilar[1], and Javier Ortega-Garcia[1]

[1] ATVS (Speech and Signal Processing Group), Escuela Politecnica Superior
Universidad Autonoma de Madrid, E-28049 Madrid, Spain
{daniel.ramos,joaquin.gonzalez,javier.ortega}@uam.es
[2] Institut de Police Scientifique, Ecole de Sciences Criminelles
Universite de Lausanne, CH-1015, Lausanne, Switzerland

Abstract. In this paper, the use of biometric systems in forensic applications is reviewed. Main differences between the aim of commercial biometric systems and forensic reporting are highlighted, showing that commercial biometric systems are not suited to directly report results to a court of law. We propose the use of a Bayesian approach for forensic reporting, in which the forensic scientist has to assess a meaningful value, in the form of a likelihood ratio (LR). This value assist the court in their decision making in a clear way, and can be computed using scores coming from any biometric system, with independence of the biometric discipline. LR computation in biometric systems is reviewed, and statistical assumptions regarding estimations involved in the process are addressed. The paper is focused in handling small sample size effects in such estimations, presenting novel experiments using a fingerprint and a voice biometric system.

1 Introduction

The number of commercial applications of biometric systems has significantly increased in the last years. As a consequence, forensic applications of biometric systems arise then in a natural way. Forensic reporting in cases involving anthropomorphical or behavioral patterns can be assisted by using a biometric system. For example, a sample pattern is recovered at the scene of a crime (e. g., a fingermark) and a court of law requests an expert opinion on the comparison of such a *mark* with a template (e. g., a suspect's fingerprint) from a suspect. The aim of a forensic system in such a case is to report a *meaningful value* in order for the court to assess the *strength of the forensic evidence* in this context of identification of sources [1][2]. However, when a biometric system is used, this value cannot be given neither by a decision or a threshold nor directly by a similarity measure [1][3], because it may lead the forensic scientist to usurp the role of the court, responsible of the actual decision [4]. Our point is that commercial, score-based biometric systems are not suited for direct forensic reporting to a court of law as has been stated in previous work [1][3].

To overcome this difficulty, the application of a *likelihood ratio* (LR) paradigm suited for forensic evidence to score-based biometric system has been proposed

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 1080–1089, 2005.

[2][3][5][6]. In this paper we propose to use this LR framework. Following this approach the forensic scientist assesses and reports one meaningful value: the LR, that allows the court to progress to a posterior opinion starting from his prior opinion about the case before the forensic evidence analysis [5][7]. This logical Bayesian framework implies a change of opinion when new information is considered, i. e., when the weight of the evidence has been assessed [1]. LR Computation can be performed using the scores from any biometric system [3][8][9], the process being independent of the biometric discipline. Thus, the LR assessed from the system scores can be used for direct forensic reporting. Our recent work presents examples of LR computation using on-line signature, face and fingerprint biometric systems [9]. In [10], the ATVS forensic voice biometric system is presented and excellent results in NIST Speaker Recognition Evaluation 2004 [11] and NFI-TNO Forensic Evaluation 2003 [12] using robust LR computation algorithms are shown. LRs can be used to compare the strength of the evidence between different biometric systems and expert opinions, and allow the combination of evidence weights coming from different and independent systems [5].

The present paper describes briefly forensic interpretation and reporting using biometric systems by means of LR computation. Then the paper highlights statistical assumptions regarding estimations involved in LR computation [9][10]. The novel contribution is focused on small data set effects [13] using different estimation techniques. The paper is organized as follows: Sect. 2 describes the Bayesian analysis of forensic evidence and its motivation. In Sect. 3, the LR computation process is reviewed, statistical assumptions commonly considered are presented, and main approaches found in the literature in order to cope with them are reviewed. Sect. 4 presents new experiments regarding generalization against small data set effects using different estimation techniques for fingerprint and voice biometric systems. In Sect. 5, conclusions are extracted.

2 Forensic Interpretation of the Evidence

2.1 Score-Based Biometric Systems vs. Forensic Interpretation

The aim of commercial score-based biometric systems is to output a similarity measure (score) between a user of the system, represented by a biometric test pattern, and a claimed identity, represented by a biometric template. Biometric verification is a classification problem involving two classes, namely *target users* of the system and *non-target users* or *impostors*. A decision is made by comparing the output score with a threshold. Assessment of these systems can be done by means of decision theory tools such as ROC or DET curves [3].

The aim of forensic interpretation is different. Forensic evidence is defined as the relationship between the suspect material (samples of biometric patterns obtained from the suspect) and the mark (biometric pattern generally left in association with a crime of disputed origin) involved in a case. The role of the forensic scientist is to examine the material available (mark and control material) and to assess the contribution of these findings with regards to competing propositions arising from the circumstances and often the adversarial nature of the criminal trial [6]. In sources attribution issues [2] such as the ones considered here, the prosecutor view will suggest that the suspect left the mark whereas the defense will support that an unknown contributor is the source [1][3][6]. The LR framework we suggest bellow imply that the forensic scientist will only guide as to the degree of support for one proposition versus the other and not comment, probabilistically or otherwise, on the hypotheses themselves [1][6]. This role differs fundamentally from the natural objectives of a commercial biometric system (i. e., making a decision) [3][4].

2.2 Bayesian Analysis of Forensic Evidence

The problems described above are handled elegantly when using the Bayesian analysis of forensic evidences [1][5][6]. Following this approach, the interpretation of the forensic findings is based on two competing hypotheses, namely H_p (*the biometric trace originates from the suspect*, also called *prosecutor hypothesis*) and H_d (*the biometric trace originates from any other unknown individual*, also called *defence hypothesis*). The decision of the judge or jury (in one word the *fact finder*) is based on the probabilities of the two hypotheses given all the information of the case, that can be split into *forensic information* (E), and *background information* (I) (i. e., all other information related to the case). Using the Bayes Theorem [5], we can write in odds form:

$$\frac{\Pr\left(H_p \middle| E, I\right)}{\Pr\left(H_d \middle| E, I\right)} = \frac{\Pr\left(E \middle| H_p, I\right)}{\Pr\left(E \middle| H_d, I\right)} \cdot \frac{\Pr\left(H_p \middle| I\right)}{\Pr\left(H_d \middle| I\right)} \tag{1}$$

In this way, the posterior probabilities needed by the fact finder can be separated into prior probabilities, based only on the background information, and a *likelihood ratio* (LR) that represents the strength of the analysis of the forensic evidence in the inference from prior to posterior odds:

$$LR = \frac{\Pr\left(E \middle| H_p, I\right)}{\Pr\left(E \middle| H_d, I\right)} \tag{2}$$

The role of the forensic scientist lies therefore with the assessment of this LR. The meaning of the LR is in essence *independent of the forensic discipline* [5], and its assessment in a case can involve computation (such as in [8][9][3][14]) or informed judgements expressed as subjective probabilities [15].

Assessment of forensic systems performance can be made using Tippett plots [3][9] (see Fig. 3), which are cumulative distributions of LR for targets (when H_p is true) and non target (when H_d is true) respectively.

3 Likelihood Ratio Computation in Score-Based Biometric Systems

As noted in [1], the numerator of the LR (Eq. 2), is obtained from knowledge of the within-source variability (WS) of the suspect material. This distribution can be estimated using scores obtained by comparing biometric patterns (*controls*) from the suspect to templates originating from the same suspect. On the other hand, the denominator of the LR is obtained from knowledge of the between-source (BS) distribution of the mark, which can be estimated from scores resulting from the comparison of the mark with a set of biometric templates from a relevant population of individuals. The *evidence score* is computed by comparing the mark with the suspect biometric template. Finally, the LR value will be the ratio of the density of the evidence score under respectively WS and BS [3][8], as is shown in Fig. 1. As the LR is conditioned by the prosecutor (H_p) and defence (H_d) hypothesis and background information (I), the forensic scientist has to estimate the WS and BS distributions based on the data available in the case. Evidence scores significantly different from the data set used in distribution estimations will give a non-informative LR value of one.

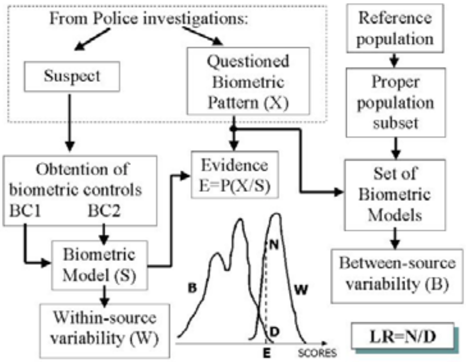

From Police investigations:
Suspect
Questioned Biometric Pattern (X)
Reference population
Proper population subset
Obtention of biometric controls
BC1
BC2
Evidence E=P(X/S)
Set of Biometric Models
Biometric Model (S)
Between-source variability (B)
Within-source variability (W)
B
N
W
D
E
SCORES
LR=N/D

Fig. 1. LR Computation Steps

3.1 Statistical Assumptions

In the estimation of WS and BS distributions for LR computation, some assumptions have to be made. In order to estimate WS distribution, matching conditions between the suspect biometric template and controls (see Fig. 1) is needed [16][17]. However obtaining matching controls in real forensic casework can be a very difficult task, especially in some biometric disciplines, leading to a paucity of data. Therefore, generalization is desirable to avoid small sample size effects [13]. Approaches based on modelling WS distributions using databases can be found in [16]. More robust techniques based on additional knowledge about the system behavior are shown in [10], in which they can be also found procedures to optimize the use of the suspect data.

BS estimation problems related to mismatch between the considered relevant population and the conditions of the mark have been explored in [10] and [17] for voice biometric systems. In [18], corpus-based techniques are applied to reduce the mismatch between the population and suspect templates. Also, the nature of the population is conditioned to the circumstances of the case (I). The relevant population can then be reduced, either according to I, or because of the lack of databases matching the conditions of the case under study. If the population size is small, non-matching conditions between population templates and questioned patterns can seriously degrade system performance.

The novel contribution of this work focuses on small sample size effects, not related to forensic issues of the partiality, poor quality or degradation of marks. Thus, the scenarios explored will postulate marks of quality comparable with the control material. In that sense, we have a symmetry in term of amount of information between the mark and the suspect material.

3.2 Between-Source Distribution Modelling Techniques

In this paper, we concentrate on BS modelling. We propose to assess two different estimation techniques, one parametric and one non-parametric, to model BS distributions. The parametric approach, proposed in [3], consist in modelling BS with a one-dimensional mixture of gaussian components:

$$p\left(x\right) = \sum_{m=1}^{M} p_m \cdot b_m\left(x\right) \tag{3}$$

where M is the number of mixtures used, and p_m are restricted to:

$$\sum_{m=1}^{M} p_m = 1 \tag{4}$$

Maximum Likelihood (ML) estimation using this parametric model is carried out by the Expectation-Maximization (EM) algorithm [19].

On the other hand, Kernel Density Functions (KDF) [19] are used. In this non-parametric technique the score-axis is divided in regions (*bins*) of length h. If N samples are available, and k_N of these samples fall in a bin, the probability estimated for that bin will be k_N/N. So the corresponding density will be:

$$\hat{p}\left(x\right) \equiv \hat{p}\left(x_0\right) \approx \frac{1}{h}\frac{k_N}{N},\ \left|x - x_0\right| \leq \frac{h}{2} \tag{5}$$

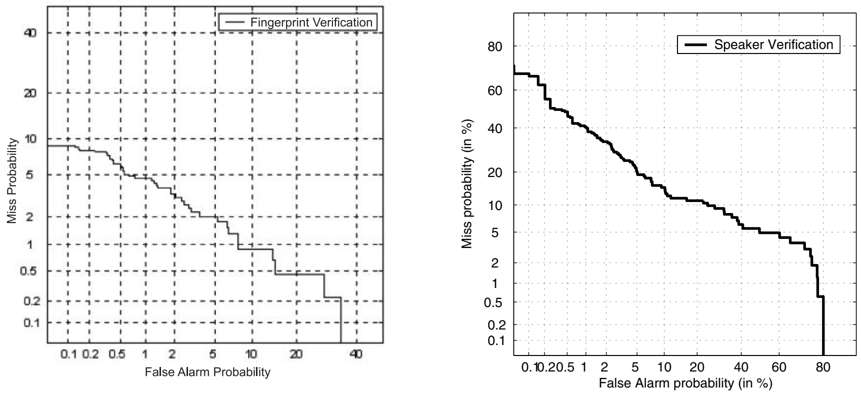
Using smooth functions ϕ, known as *kernels*, where $\phi \geq 0$ and:

$$\int_x \phi\left(x\right) \cdot dx = 1 \tag{6}$$

then the resulting estimated function is a legitimate pdf.

4 Experiments

In order to test the performance of the BS estimation techniques proposed, we present experiments using a fingerprint and a voice biometric system respectively.



(a)

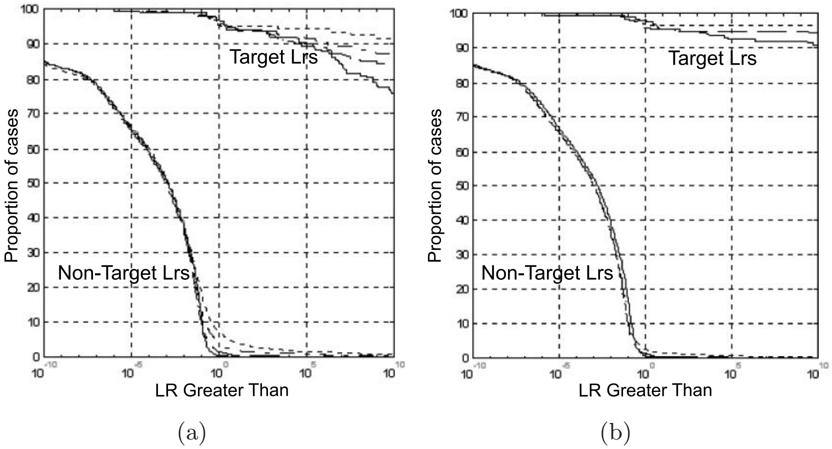
(b)

Fig. 2. DET curve of: (a) ATVS reference fingerprint system with MCYT corpus, and (b) ATVS reference voice biometric system with NIST SRE 2004 sub-corpus

4.1 Databases, Experimental Protocol and Biometric Systems

For fingerprint experiments, the ATVS fingerprint recognition system based on minutiae comparison [20] has been used. A sub-corpus from the MCYT fingerprint database [21] has been selected, consisting of 50 users each one having 10 fingerprint samples. One sample per user will be used as reference biometric template. For score-based system performance assessment via DET plots, the 9 remaining samples will be used as test patterns (marks), so a total of $50 \times 9 = 450$ target trials and $50 \times 49 \times 9 = 22050$ non-target trials have been considered. For the forensic interpretation system, 5 (out of 9) fingerprint patterns have been used as biometric controls in order to obtain WS scores, and the remaining 4 will be used as marks. No technique will be used to predict degradation in WS distribution, as fingerprint biometric patterns are all acquired in the same conditions. Therefore, a total of $50 \times 4 = 250$ target trials and $50 \times 49 \times 4 = 9800$ non-target trials will be used for Tippett plot computation. Population data has been taken from the same corpus too.

For voice biometric system experiments, the ATVS UBM-MAP-GMM system [10] has been used. The scores used in the LR computation experiments are extracted from the ATVS results in the NIST Speaker Recognition Evaluation 2004 [11], using only a male subcorpus of 50 users, and all the trials defined in the evaluation for these users in the core condition, i. e., one conversation side (5 minutes) for training and one for testing. Strong mismatch on channel and language conditions is present in this data set, and it exists variability in the amount of speech per conversation side (as silence removal has not been performed). As only one speech segment is used as suspect biometric material, jackknife and prediction techniques described in [10] are used to perform robust WS estimation. In summary, a total of 163 target trials and 1969 non-target trials are performed. Population data consists of a channel-balanced English set of GMM models obtained from development corpora from past NIST SRE.



(a) (b)

Fig. 3. Tippett plots for fingerprint system estimating BS distribution with different techniques. (a): ML with M gaussian mixtures: M=1 (solid), M=3 (dashed), M=10 (dash-dot) and M=30 (dotted). (b): KDF with bin size h=10 (solid), h=3 (dashed), h=1 (dotted)

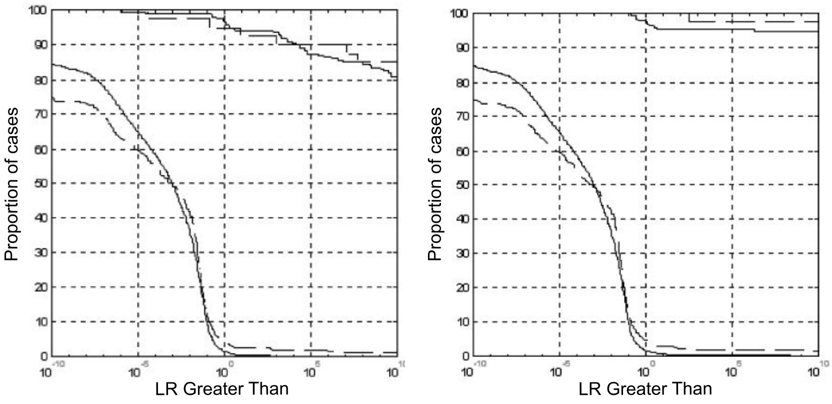
4.2 Results

Fig. 2 show the performance of the score-based biometric systems in the scenarios described. The performance of the forensic fingerprint system using the two techniques described is shown in Fig. 3. As can be seen in Fig. 3(a), performance of the forensic system in non-target trials degrades as the number of mixtures (M) increases. For M=30 mixtures, a small but not negligible proportion of non-target trials have values of LR greater than 100.000, which is alarming because the rate of misleading evidence for the non-target curve is critical in forensic systems [10]. The same conclusion can be extracted for KDF in Fig. 3(b), when the bin size h is small. This effect is due to an over-fitting effect of the BS model on the available data set.

Generalization against small sample size effects is inferred from Fig. 4. Two populations of $L = 50$ and $L = 10$ (obtained by sampling) biometric templates has been used. It can be seen that KDF and ML estimation presents very similar performance when the data set size is reduced. However, performance of targets for KDF estimation is better when population size decreases, which means over-estimation of target LRs due to over-fitting in BS distribution estimation.

In the experiments presented using voice biometrics system, ML estimation is performed to model BS distribution. In Fig. 5, the same effect noticed in Fig. 3(a) can be observed, i. e., the proportion of non-target trials having LR values greater than one grows as M increases.

Generalization for the voice biometric system is shown in Fig. 6. ML estimation of BS distribution with M=1 and M=8 has been used. It can be seen that as population sample size decreases, over-fitting in the data (M grows) implies degradation on system performance (i.e., bigger proportion of non-target $LRs > 1$, and over-estimated target LR).



(a)

(b)

Fig. 4. Analysis of generalization effects with small sample-size data for the fingerprint biometric system. Population size: L=50 (solid) and L=10 (dotted). (a): ML with M=3 gaussian mixtures, (b): KDF with bin size h=3

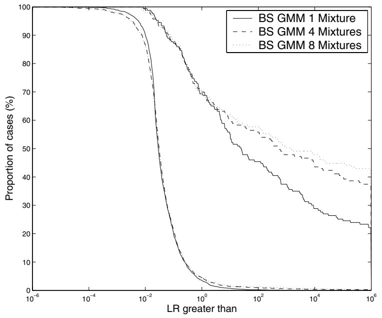
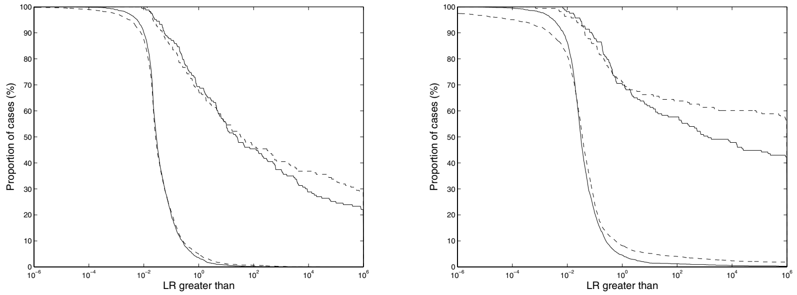


Fig. 5. Tippett plots for voice biometric system estimating BS distribution with ML and different number of gaussian mixtures



(a)

(b)

Fig. 6. Analysis of generalization effects using ML for voice biometric system. Population size L=60 (solid) and L=12 (dotted) (a): M=1 gaussian; (b): M=8 gaussian

5 Conclusions

In this paper, we have shown how biometric systems can be used in forensic applications, using a *LR* as a measure of the strength of the evidence computed from the scores. The need of proper forensic reporting has been highlighted, as the fact finder needs a meaningful value to assist his decision making. Direct reporting using score-based biometric systems has been shown in the literature to be misleading, and we promote a *LR* based reporting system. Bayesian analysis of forensic evidence has been referred as the logical way for evaluating forensic findings. *LR* computation process has been reviewed, highlighting that it can be performed using any score-based biometric system, regardless of the biometric discipline. Statistical assumptions regarding estimations involved in the *LR* computation process have been discussed. The main contribution of the paper are the experiments regarding small sample size effects in *BS* estimation, which can appear in forensic casework when the relevant population is reduced, either because of the background information on the case (I) or the availability of databases matching the suspect biometric template conditions. It has been shown that the performance of the system degrades when *BS* distribution overfits the data set when its size is small, and misleading evidence in non-target trials can increase, which is a highly undesirable effect in forensic systems. *LRs* for target trials might also be over-estimated in these conditions.

Acknowledgements

This work has been partially supported by the Spanish MCYT projects TIC03 09068-C02-01 and TIC2000 1669-C04-01. D. R.-C. and J. F.-A. are also supported by a FPI scholarship from Comunidad de Madrid.

References

1. Champod, C., Meuwly, D.: The inference of identity in forensic speaker recognition. Speech Communication, **31** (2000) 193-203
2. Champod, C.: Identification/Individualization: Overview and Meaning of ID. Encyclopedia of Forensic Science, J. Siegel, P. Saukko and G. Knupfer, Editors. Academic Press: London (2000) 1077-1083
3. Gonzalez-Rodriguez, J., Fierrez-Aguilar, J., Ortega-Garcia, J.: Forensic Identification Reporting Using Automatic Speaker Recognition Systems. Proc. ICASSP (2003)
4. Taroni, F., Aitken, C.: Forensic Science at Trial. Jurimetrics Journal **37** (1997) 327-337
5. Aitken, C., Taroni, F.: Statistics and the Evaluation of Evidence for Forensic Scientists. John Wiley & Sons, Chichester (2004)
6. Evett, I.: Towards a uniform framework for reporting opinions in forensic science casework. Science and Justice **38(3)** (1998) 198-202
7. Kwan, Q.: Inference of Identity of Source. Department of Forensic Science, Berkeley University, CA (1977)

8. Meuwly, D.: Reconaissance de Locuteurs en Sciences Forensiques: L'apport d'une Approache Automatique. Ph.D. thesis, IPSC-Université de Lausanne (2001)
9. Gonzalez-Rodriguez, J., Fierrez-Aguilar, J., Ramos-Castro, D., Ortega-Garcia, J.: Bayesian Analysis of Fingerprint, Face and Signature Evidences with Automatic Biometric Systems Forensic Science International (2005) (accepted)
10. Gonzalez-Rodriguez, J., et al.: Robust Estimation, Interpretation and Assessment of Likelihood Ratios in Forensic Speaker Recognition. Computer, Speech and Language (2005) (submitted)
11. Home page of NIST Speech Group: http://www.nist.gov/speech
12. Van-Leeuwen, D., Bouten, J.: Results of the 2003 NFI-TNO Forensic Speaker Recognition Evaluation. Proc. of ODYSSEY (2004) 75-82
13. Raudys, S., Jain, A.: Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. IEEE Trans. on PAMI **13(3)** (1991) 252-264
14. Curran, J.: Forensic Applications of Bayesian Inference to Glass Evidence. Ph.D. thesis, Statistics Department, University of Waikato, New Zealand (1997)
15. Taroni, F., et al.: De Finetti's Subjectivism, the Assessment of Probabilities and the Evaluation of Evidence: A Commentary for Forensic Scientists. Science and Justice **41(3)** (2001) 145-150
16. Botti, F., et al.: An Interpretation Framework for the Evaluation of Evidence in Forensic Automatic Speaker Recognition with Limited Suspect Data. Proc. ODYSSEY (2004) 63-68
17. Gonzalez-Rodriguez, J., et al.: Robust Likelihood Ratio Estimation in Bayesian Forensic Speaker Recognition. Proc. Eurospeech (2003) 693-696
18. Alexander, A., et al.: Handling Mismatch in Corpus-Based Forensic Speaker Recognition. Proc. ODYSSEY (2004) 69-74.
19. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley (2001)
20. Simon-Zorita, D., et al.: Quality and Position Variability Assessment in Minutiae-Based Fingerprint Verification. IEE Proc. Vision, Image and Signal Processing **150(6)** (2003) 402-408
21. Ortega-Garcia, J., et al.: MCYT Baseline Corpus: A Bimodal Biometric Database. IEE Proc. Vision, Image and Signal Processing **150(6)** (2003) 395-401