Information-theoretical comparison of evidence evaluation methods for score-based biometric systems

Daniel Ramos, Joaquin Gonzalez-Rodriguez and Julian Fierrez

ATVS - Biometric Recognition Group, Escuela Politecnica Superior
C./ Francisco Toms y Valiente 11, Universidad Autonoma de Madrid, E-28049 Madrid, Spain
{daniel.ramos,joaquin.gonzalez,julian.fierrez}@uam.es

Abstract

Biometric systems are a powerful tool in many forensic disciplines in order to aid scientists to evaluate the weight of the evidence. However, uprising requirements of admissibility in forensic science demand scientific methods in order to test the accuracy of the forensic evidence evaluation process. In this work we analyze and compare several evidence analysis methods for score-based biometric systems. For all of them, the score given by the system is transformed into a likelihood ratio (LR) which expresses the weight of the evidence. The accuracy of each LR computation method will be assessed by classical Tippett plots- We also propose measuring accuracy in terms of average information given by the evidence evaluation process, by means of Empirical Cross-Entropy (ECE) plots. Preliminary results are presented using a voice biometric system and the NIST SRE 2006 experimental protocol.

1. Introduction

Biometric systems aim at automatically recognizing an individual from any biometric trait, such as their voice, written signature, face, fingerprint, etc [1]. However, the use of biometric systems for forensic evidence evaluation is not straightforward [2]. In particular, the typical architecture of a biometric system is score-based, meaning that a matching score is computed from a biometric sample and a previously enrolled template. Although this matching score can be useful for ranking purposes or intelligence applications, it is useless as a direct measure of the weight of the evidence, as it has been previously discussed in the literature [2, 3, 4, 5]. In this sense, the likelihood ratio (LR) methodology for evidence evaluation [6], proposed in other forensic disciplines such as DNA or glass analysis, has been successfully applied to forensic reporting using biometric systems [3, 7]. A typical $LR-$ based approach for biometric systems can be as follows: first, from recovered and control biometric samples involved in a given forensic case, a biometric system computes a matching score; then, using databases of biometric material, the score is transformed into a LR value, which represents the weight of the evidence in such given comparison. This approach has the advantage of keeping the score-based architecture of the system unchanged, and therefore it can be applied to any biometric system yielding matching scores.

In order to be successfully used for evidence evaluation, biometric systems should consider the emerging requirements in forensic identification [8]. In particular, the scientific assessment of the accuracy of forensic disciplines is important regarding the admissibility of a given procedure in court. In this work, we assess the accuracy of several LR computation methods for score-based biometric systems using an experimental and scientific procedure. Preliminary results in this short paper present two techniques, namely suspect-independent and suspect-adapted LR computation from matching scores. The accuracy of the techniques will be measured in the form of classical Tippett plots. We also propose an information-theoretical framework for accuracy assessment, namely Empirical Cross-Entropy (ECE) plots. These ECE plots show the amount of information lost in the inferential process about the identity of the source. This loss may be due either to inaccuracies in the LR computation process or to the intrinsic non-perfect discriminating power of the matching scores. An ATVS-UAM voice biometric system has been used in order to generate preliminary results, using the NIST Speaker Recognition Evaluation 2006 (NIST SRE 2006) protocol. For the final presentation, results using other LR computation techniques such as logistic regression will be presented, as well as results for other biometric traits such as fingerprint or on-line written signature.

2. LR computation methods for score-based biometric systems

The starting point for LR computation for score-based biometric systems is the matching score. The transformation of the score into the LR value has been classically performed by generative models [3, 7], where within-source and between-source variations are modelled from same-source and different-source matching scores respectively. For preliminary results in this short paper we will use two of such techniques, namely suspect-independent and suspect-adapted LR computation. The suspect-independent method proposes a framework where an accurate model of the within-source distribution for a given suspect can be obtained using same-source scores from different individuals in the same conditions. On the other hand, the suspect-adapted method obtains the within-source distribution from same-source scores coming from the suspect involved in the case, adapting it from a global, suspect-independent distribution. Therefore, an *adapted* within-source pdf is obtained. See [9, 5] for details.

2.1. Accuracy assessment and representation

In this work, we use two methods for the assessment of the accuracy of LR values, namely Tippett and ECE plots. Tippett plots have been classically used for empirical performance assessment of evidence evaluation [10]. They represent the proportion of same-source and different-source comparisons whose

Julian Fierrez's research is supported by a Marie Curie Fellowship from the European Commission.

LR exceed a given value. Important values shown by these curves are the actual distributions of the LR values for the given experimental set-up, and the rates of misleading evidence, defined as the proportion of LR values giving support to the wrong hypotheses. Details can be found in [10]. Example of Tippett plots can be found in Figure 1.

We also propose an information-theoretical representation of accuracy, namely Empirical Cross-Entropy (ECE) plots, which computes and represents the value of ECE for each prior probability in a partition of the $(0, 1)$ range. ECE can be interpreted as the mean information over different comparisons which is needed in order to know whether the control and recovered biometric samples actually come from the same source or not [11]. Figure 2 shows examples of ECE plots. The solid curve is the ECE (average information loss) of the LR values. For a given prior probability, the higher this ECE curve, the higher the information needed on average in order to know the true hypothesis, and therefore the worse the evidence evaluation process. As the prior probability is province of the court and may be even unknown by the forensic scientist, ECE is represented for a wide range of possible values of the prior. Also, the dashed curve represents the accuracy of the experimental set of LR values after a calibration algorithm, namely Pool Adjacent Violators (PAV) [12]. The PAV transformation obtains a calibrated set of LR values [13], but it is needed to know whether the recovered and control material come from the same source or not for each comparison. Therefore, calibrated LR values can only be obtained by PAV for an experimental database where the answers of the hypothesis are known, but not in real casework. Details about ECE plots and their interpretation can be found in [11, 5]. ECE is in essence the average value over different comparisons of the logarithmic *strictly proper scoring rule*. Such a value has been proposed in the statistics literature in order to evaluate opinions expressed in the form of posterior probabilities (forecasts) [13].

3. Preliminary results

The scores needed for LR computation have been obtained using the ATVS GMM-UBM-MAP system, which is based on modelling the likelihoods of the speaker data using an adapted Gaussian Mixture Model (GMM) from an Universal Background Model (UBM). Details can be found in [14]. Experiments have been performed using the evaluation protocol proposed in NIST 2006 SRE for the 1 conversation side training and 1 conversation side testing task (1c-1c, see [15] for details). More than comparisons have been performed in this condition.

3.1. Results

Figure 1 shows Tippett plots illustrating the distribution of the matching scores (as given by the biometric system) and the LR values for both compared methods. In Figure 1a, it can be seen that the magnitude of the strength of misleading evidence (LR values supporting the wrong hypothesis) is high for the scores for different-source LR values, which suggests a bad accuracy if the scores are going to be used directly as $\log(LR)$ values. On the other hand, after LR computation the rate of misleading evidence is quite limited for both presented techniques (Figures 1b-c). However the Tippett plots do not allow us to easily conclude which LR computation method is more accurate.

Figure 2 shows the performance for the presented LR computation techniques in the form of ECE plots. It can be seen that for values of the prior log-odds between -1 and $+1$, the ECE of the scores (Figure 2a, solid line) is higher than for all LR values computed using the presented techniques. It is also observed that such ECE value is far from the calibrated system after PAV. That means that if the matching scores are directly used as $\log(LR)$ values they will lead to a loss of information due to miscalibration, on average over different comparisons. However, ECE (solid curve) dramatically reduces after LR computation, and therefore the fact finder will need less average information in order to know whether the recovered and the control biometric samples come from the same source. Finally, for both LR computation schemes presented, the ECE after PAV calibration (dashed curve) is near the ECE of the LR values computed (solid curve), which demonstrates the good calibration performance of these methods.

4. Conclusions

This short paper has presented preliminary results comparing the accuracy of several methods for the evaluation of the evidence using score-based biometric systems. The assessment of the accuracy has been presented in the form of Tippett plots, and also by means of information-theoretical measures based on empirical cross-entropy (ECE). Both assessment techniques, added to a clear and standard protocol such as those developed by NIST in their yearly SREs, give a method to perform the accuracy assessment of the evidence evaluation process in a scientific way, according to the recent demands in forensic science. Results show that suspect-adapted LR computation outperforms a suspect-independent approach in voice biometrics. For the final presentation, results will be extended with additional methods for LR computation from scores, such as logistic regression. Moreover, results using other biometric traits such as fingerprint or on-line written signature will be also presented.

5. References

[1] Anil K. Jain, Patrick Flynn, and Arun Ross, *Handbook of Biometrics*, Springer, 2008.

[2] D. Dessimoz and C. Champod, "Linkages between biometrics and forensic science," in *Handbook of Biometrics*, Anil K. Jain, Patrick Flynn, and Arun A. Ross, Eds. 2007, Springer.

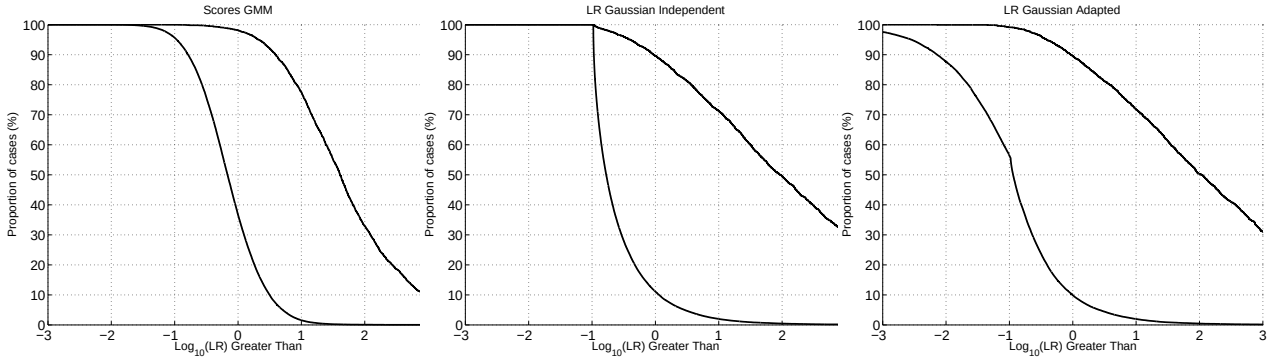
[3] D. Meuwly, *Reconaissance de Locuteurs en Sciences Forensiques: L'apport d'une Approache Automatique*, Ph.D. thesis, IPSC-Universite de Lausanne, 2001.

[4] A. Alexander, *Forensic automatic speaker reconition using Bayesian interpretation and statistical compensation for mismatched conditions*, Ph.D. thesis, École Politechnique Fédérale de Lausanne, Lausanne, Switzerland, 2005.

[5] D. Ramos, *Forensic evaluation of the evidence using automatic speaker recognition systems*, Ph.D. thesis, Depto. de Ingenieria Informatica, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid, Spain, 2007, Available at http://atvs.ii.uam.es.

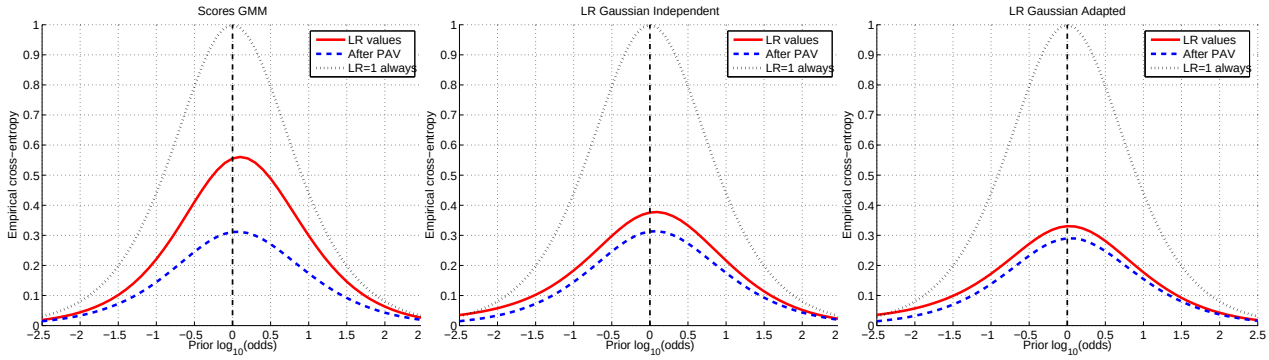
[6] C. G. G. Aitken and F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, John Wiley & Sons, Chichester, 2004.

[7] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro, and J. Ortega-Garcia, "Bayesian analysis of fingerprint, face and signature evidences with automatic bio-



(a) (b) (c)

Figure 1: Tippett plots showing distribution of matching scores and $\log_{10}(LR)$ values for the ATVS-UAM GMM-UBM system for different LR computation methods.



(a) (b) (c)

Figure 2: ECE plots showing the discrimination and calibration performance of the ATVS-UAM GMM-UBM system using information-theoretical magnitudes, for the matching scores directly given by the system and the different compared LR computation methods.

metric systems," *Forensic Science International*, vol. 155, no. 2-3, pp. 126–140, 2005.

[8] M. J. Saks and J. J. Koehler, "The coming paradigm shift in forensic identification science," *Science*, vol. 309, no. 5736, pp. 892–895, 2005.

[9] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Likelihood ratio calibration in transparent and testable forensic speaker recognition," in *Proc. of Odyssey*, 2006.

[10] I. W. Evett and J. Buckleton, "Statistical analysis of STR data," *Advances in Forensic Haemogenetics, Springer-Verlag, Heilderberg*, vol. 6, pp. 79–86, 1996.

[11] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, J. Zieba-Palus, and C. G. G. Aitken, "Information-theoretical comparison of likelihood ratio methods of forensic evidence evaluation," in *Proceedings of International Workshop on Computational Forensics (in IAS 2007)*, 2007, pp. 411–416.

[12] N. Brümmer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[13] M. H. deGroot and S. E. Fienberg, "The comparison and evaluation of forecasters," *The Statistician*, vol. 32, pp. 12–22, 1982.

[14] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[15] "NIST speech group website: http://www.nist.gov/speech," .