Biometric Identification in Forensic Cases According to the Bayesian Approach

J. Gonzalez-Rodriguez[1], J. Fiérrez-Aguilar[1], J. Ortega-Garcia[1], and
J.J. Lucena-Molina[2]

[1]ATVS (Speech and Signal Processing Group)
DIAC-EUITT - Universidad Politecnica de Madrid, Spain
{jgonzalez,jortega}@diac.upm.es www.atvs.diac.upm.es

[2]Acoustics and Image Processing Lab.-DGGC
Ministry of Internal Affairs, Spain

Abstract. On the one hand, commercial biometric systems and forensic identification require different approaches in order to evaluate system outputs. On the other hand, bayesian approach for evidence analysis and forensic reporting perfectly suits the needs of the court and the forensic scientist. Inside this bayesian framework, any biometric system can be adapted to provide its results in the form of *likelihood ratios* (LR) (being so converted in a forensic identification system), and performance of the forensic system can be then assessed according to the bayesian approach. We will focus on a specific biometric characteristic, showing how forensic speaker recognition can be reported by means of bayesian technique. Results including NIST-Ahumada and providing LR scores in the form of Tippet plots (and compared with DET plots) will be finally presented.

1. Introduction

While commercial biometric systems performance, oriented to acceptance or rejection decisions, are widely assessed through different classical decision-based criteria, as type I and II errors or ROC and DET plots, an intense debate among forensic practitioners have taken place during the last decade in order to achieve a common framework for the evaluation of evidence and its interpretation to the court (as shown in Fig. 1), and then how to assess the performance of forensic systems. Nowadays, the Bayesian (or *Likelihood-Ratio*, LR) approach is firmly established as a theoretical framework for any forensic discipline, where systems providing its results according to this approach, from the large experience gained in DNA-based person identification, are assessed through Tippet plots. In this contribution, we will show the different nature of the outputs that automatic recognition systems must provide respectively in commercial and forensic approaches, even if the systems use the same core technology, and subsequently the need for different assessment tools specially suited for their corresponding applications.

M. Tistarelli, J. Bigun, A.K. Jain (Eds.): Biometric Authentication, LNCS 2359, pp. 177-185, 2002.

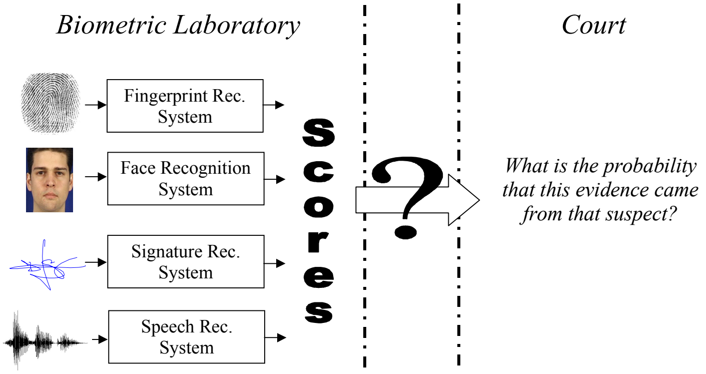

Fig. 1. The problem of biometric scores submission to Court

The paper is organized as follows. In Section 2, generic biometric recognition systems will be presented, and problems regarding forensic evaluation and reporting will be shown. In Section 3, the Bayesian approach to forensic identificaction will be established as a theoretical framework for biometric recognition. In Section 4, assessment of forensic biometrics will be discused, and finally, in Section 5, forensic speaker recognition will be introduced, where comparative results using DET and Tippet plots will be shown.

2. Biometric Systems and Classical Forensic Reporting

In order to assess the identification abilities of any biometric system, the system must be tested with known users and impostors, task which is usually performed using databases of the corresponding input patterns (fingerprints, voices, signatures, faces, ...). Two types of error can occur in a detection system: false rejections (type I error), when a true user is rejected, and false acceptances (type II error), when an impostor is accepted. The probability of any of these two errors depends on the value of the threshold, in a complementary manner. This means that if the threshold is increased, the false acceptances will be reduced but the false rejections will be increased, and vice versa. As the same system or technology could work in different operating conditions, it is usual to show all possible operating points. This has been done classically in detection tasks by means of ROC curves, showing the tradeoff between missed detections (false rejections) and false alarms (false acceptance). In order to have a single value characterizing the performance of the system, the Equal Error Rate (EER) is usually given, which is the point where the probability of a missed detection equals the probability of false alarms.

However, as performance regarding biometric systems increases, comparison of systems have become extremely difficult with this representation, as curves from different systems are extremely close to the lower left corner. This problem was overcome with the introduction of the DET (Detection Error Tradeoff) curve [1], which allows an almost linear representation of system performances, permitting easy observation of system contrasts.

2.1 Is Acceptance/Rejection the Objective of Forensic Recognition?

In the last years, the value of the different types of forensic evidence (even in firmly established areas, as fingerprint matching) have been severely attacked, questioning their scientific status, as is shown in influential works in the field [2, 3], specially *"...after several highly publicized miscarriages of justice in which forensic expertise played a crucial role*" [4].

Classically, there have been two different approaches to forensic reporting in "individualization of the source" areas, which includes areas as fingerprint, voice, face, signature, or DNA, tool marks, paint, glass, fibers, and firearms. The first approach has been to provide just "identification" or "exclusion/elimination" decisions, which results in a very high percentage of non-reporting cases. This approach has two main drawbacks: the first one is related with the use of subjective thresholds, as these techniques does not provide absolute identifications, specially in forensic conditions, and all that the system/technique can provide is a score or a probability. Then, if the forensic scientist takes the (subjective) decision of identification or exclusion/rejection, he will be ignoring the prior probabilities related to the case (independent of the evidence under analysis), usurping the role of the court in taking this decision, as "*... the use of thresholds is in essence a qualification of the acceptable level of reasonable doubt adopted by the expert*" [5]. The second drawback is the large amount of non-reporting cases that this identification/exclusion process induces, when *"... there is no logical reason to suppress probability statements ... because ... any piece of evidence is relevant if it tends to make the matter which requires proof more or less probable than otherwise"* [5]. The second classical approach to forensic reporting in this area consists in the use of a verbal scale of identification probabilities (typically "identification" / "very probable" / "probable" / "not conclusive" / "elimination"). This approach falls in the same errors as has just been noted, as it makes use of several subjective thresholds, but again ignores the prior probabilities (or usurp the judge/jury role if assigns it) relative to every case.

3. Bayesian Analysis of Forensic Evidence

Fortunately, the Bayesian or LR approach is now firmly established as a theoretical framework for any forensic discipline [6, 7, 8]. As an example, there are eight Working Groups (DNA, Fibers, Fingerprint, Firearms, Handwriting, Tool Marks, Paint and Glass, Speech and Audio) in ENFSI (European Network of Forensic Science Institutes) dealing with individualization of the source. All of them, in discussions open also to non-European participants, have dealt or are dealing with the bayesian approach, looking for common standards and procedures.

In this Bayesian framework, the roles of the scientist and the judge/jury are clearly separated, because the court wants to know the odds in favor of the prosecution proposition (C), ("the suspect has committed the crime"), given the circumstances of the case (I) and the observations made by the forensic scientist (E). These odds in favor of C are obtained from (1):

$$O\left(C\middle|E,I\right)=\frac{\Pr\left(E\middle|C,I\right)}{\Pr\left(E\middle|\overline{C},I\right)}\cdot O\left(C\middle|I\right) \tag{1}$$

Expressed in words, *Posterior odds = Likelihood ratio x Prior odds*, where the prior odds concern to the court (background information relative to the case) and the likelihood ratio is provided by the forensic scientist.

The use of the Bayesian approach is recommended because "*... assists scientists to assess the value of scientific evidence, help jurists to interpret scientific evidence, and clarify the respective roles of scientists and of members of the court*" [5]. In this way, the scientist alone cannot infer the identity of the speaker from the analysis of the scientific evidence, but gives the court the likelihood ratio of the two competing hypothesis (usually - *C*, the questioned pattern *was made* by the suspect, and $\overline{C}$, the questioned pattern *was not made* by the suspect).

This LR, or Bayes factor, must be determined by the forensic scientist. In order to compute these numerator and denominator probabilities, population data need to be available in order to determine objective probabilities. For score-based systems, as all biometric techniques, data are needed in order to model the distribution of measurements, both within and between sources, as this LR is in this case a ratio of probability density functions, rather than a ratio of probabilities. Moreover, the bayesian approach allows to combine different types of evidence present in the process (blood type, fingerprint, ...) and even the incorporation of subjective probabilities related to uncertain events, as shown in [8].

4. Assessment of Forensic Biometric Systems

In order to test the abilities of systems providing their results in the form of LR values, some system calibration experiments have to be performed. In [9] and [10], a useful representation for between-source comparisons in any forensic discipline, the so-called Tippet plots, is provided, representing *proportion of cases with "LR values greater than..."*. Then, we will draw in Tippet plots simultaneously two curves, one for the C hypothesis (the pattern belongs to the suspect – target), where the system must provide high LR values (LR>>1), and another one for the $\overline{C}$ hypothesis (the pattern does not belong to the suspect – non-target), where the system must provide low LR values (LR<<1). In this way, for any *x*-axis value each curve shows *proportion of cases* with *LR greater than x*. Then, the greater the separation between curves, the higher the discriminating capability and the better the system.

5. Specific Application in Speaker Recognition

Regarding a particular biometric field, like forensic speaker recognition, the most common question is: *What is the probability that this evidence (voice) came from that person?* In [5], the roles of classical commercial techniques as speaker verification (discrimination task), speaker identification (classification task) and type I and II error reporting have been properly criticized as alternatives to provide conclusions to the court, basically because these techniques usurp the role of the judge or the jury in the process, as happens also in the assignment of prior probabilities if type I and II error reporting [11] is the selected alternative.

In the following subsections we will show how any speaker recognition system can be turned into a forensic system, and assessed as such, according to the bayesian approach for analysis of the speech evidence. Basic knowledge on automatic speaker recognition [12] and Gaussian Mixture Models [13] is assumed from now on.

5.1 Computation of Likelihood Ratios in Forensic Speaker Recognition

However, there is no closed solution to the problem of LR computation, and an agreement must be achieved in every identification area, especially in the process of selection of the involved populations, and what characteristics to be used from this population. While it is assumed that the numerator of the LR calls for an assessment of the intra-variability of the system, and the denominator is the random match probability, they can be obtained from objective or subjective measures over relative frequencies in the relevant population. One of the main problems arises from the estimation of these probabilities; specially, in open populations as in the case of fibers or tool marks.

In [14], a solution to this problem for forensic speaker recognition is proposed using automatic speaker recognition techniques. In this proposal, we have first to select the adequate population (usually from linguistic analysis or background knowledge), building speaker models (GMMs) with the selected individuals. We have also to record speech from the suspect, building a suspect speaker model (GMM) with a part of it, and obtaining some reference utterances (SC: speech controls) that will be used to estimate the statistical distribution standing for the speaker intravariability. The key issue here is the computation of the probability distributions (*pdf*, probability density functions) of inter- and intra-variability, where the speech evidence, that is, the likelihood of the questioned recording with the suspect model, will be referenced.

The speaker intravariability is computed as the distribution, assumed to be gaussian, of the likelihoods of the speech controls (reference recordings from the suspect) with the suspect model. The intervariability is obtained as an statistical model of the likelihoods of the questioned recording with the models of the (selected) reference population. This is performed in [14] using kernel density estimation. In our proposal [15], this is performed with a multigaussian estimate (where the number of gaussians involved is relative to the size of the population – $M=2$ for $51<N<100$, $M=3$ for $101<N<1000$, $M=4$ for $1001<N<10000$) in order to avoid excessive details in the distribution, as the selected population (usually hundreds or thousands of speakers) is representing all possible speakers relative to the case (language, dialect, sex,..). Finally, the LR value is obtained as the quotient of the amplitudes of both distributions at the evidence likelihood.

All this computation is carried out through IdentiVox© forensic tool [16], based in state-of-the-art UBM-MAP-adapted Gaussian-Mixture-Models (GMM) text-independent speaker recognition, developed to solve the needs of forensic speech scientist. The system, perfectly suited to the bayesian approach for Forensic Speaker Recognition, accomplishes speaker modeling, population management and likelihood ratio (LR) computation. The system also includes classical speaker identification, threshold establishment, speaker verification, channel normalization, likelihood normalization with a Universal Background Model, and optimal parameterization (MFCC). The basic technology in IdentiVox© is a proprietary implementation of UBM-MAP-Gaussian Mixture Models and other well-known speech-related techniques, which have been shown as the present best solutions, regarding the problem of text-independent speaker recognition, as have been shown in last NIST evaluations (www.nist.gov/speech/spkrinfo.htm).

5.2 DET and Tippet Plots with NIST-Ahumada Data

In this subsection, we will show the close relations and significant differences in the assessment of biometric systems when used in commercial or forensic applications.

An interesting example is presented herewith, where the suitability and the role played by both DET and Tippet curves in different environments is compared, yet using the same evaluation data. NIST'2001 Ahumada [17] eval data is used, in order to assess respectively our technology (ATVS-UPM), -as to be used in any commercial/decision application-, and the forensic system developed, according to the bayesian approach, based in this technology.

Our GMM implementation uses UBM MAP-adapted GMM system with Tnorm, with a basic coefficient vector of 8 MFCC+delta+deltadelta. Since last evaluation, the system has been improved, by suppressing deltadelta coefficients, and increasing the basic vector size to 12 or 19 MFCC. As it can be seen in Fig. 2 (left), the 19 MFCC-based system outperforms the 12 MFCC-based one, assessed from a DET curve closer to the origin of coordinates (note that the best NISTeval'01 system with these data was just 1~2% better in EER). However, if we want to use any of these two systems in a forensic application, apart from the theoretical problems exposed previously (subjective thresholds and suppression of prior probabilities), the operating point of the system must have a very low (or even null) false acceptance rate, which would mean a miss detection rate much greater than 40% of the cases. Does it mean that we cannot use automatic speaker recognition technology in forensic cases?

In this experiment, the same eval'2001 raw scores have been used to compute LR values, in order to show the performance of a GMM-based forensic system. As we have just available in this dataset one speech file per speaker to build a model, and two test files per speaker, we will always use one of the files as test file, and the other one will be used as speech control, that is, the information needed to estimate the intravariability distribution (as just one likelihood is available, it will be used as mean value of a single-gaussian distribution with variance that of all speakers with his own test files). The intervariability is obtained as the distribution of the likelihoods of every test file with respect to all non-target models. Once we have the two distributions available for every test file, we compute the LR values and summarize them in the following Tippet plots (Fig. 2, right) with both systems (12/19 MFCC). Every Tippet plot is composed of two curves, one for target speakers (103*x*2=206 trials) and other for non-target speakers (103*x*2*x*102= 21,012 trials).

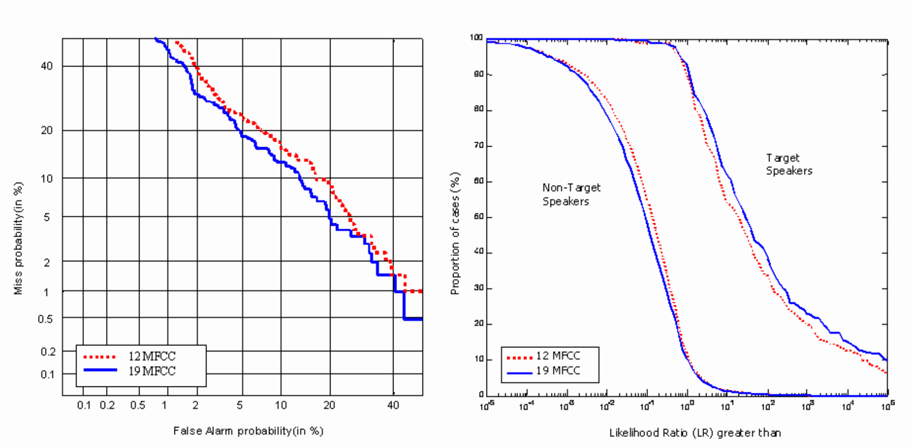

Fig. 2. *Left:* DET-plots for two versions (12/19 MFCC as base vector) of our system with NIST-Ahumada eval'2001 data. *Right:* Tippet plots for the same two versions of ATVS system with same NIST data

As it can be seen, the better the system the greater the separation between target and non-target curves for each system. But the most important fact here is that observing results in Fig. 2, and independently of the system used (12/19 MFCC), we can provide a meaningful LR value for every single file, strengthening clearly the prosecution hypothesis for the case of target-speakers, and attenuating it in the case of non-target speakers, so an excellent forensic performance of the system can be derived. Moreover, the system is not assuming any prior probability nor taking any decision (which corresponds to the court) and just limits its role to reinforce or not the prosecution hypothesis.

Addittional improvements have been developed in our lab to our basic technology presented to NIST'2001 evalulation. As can be seen in figure 3, we have obtained similar results to the best reported system for this task in last 2001 eval through the use os specific UBMs of differents sizes (with our best previous parameterization, as shown in fig. 2). Unfortunatly, we have not finished by the time of printing this paper the computation of the Tippet plots for the technology improvements shown in figure 3. However, as have been shown in figure 2, a direct improvement in the form of greater separation between the curves for target and non-target speakers is expected.

6. Conclusions

We have shown in this contribution how any biometric system can be adapted to work in the forensic environment according to the bayesian approach. In addition, the roles of ROC/DET and Tippet plots in commercial and forensic applications have been clarified. While ROC/DET curves assess system/technology performance, they cannot be used to provide conclusions to the court as acceptance or rejection of speakers is not the goal of forensic speaker recognition, as has been shown. An interesting example is presented with NIST-Ahumada eval'2001 data, showing how easily a

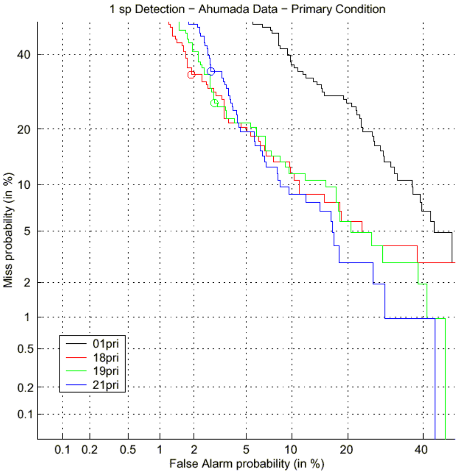

Fig. 3. ATVS results at NIST2001 eval with a generic UBM and 8 MFCC+delta+deltadelta as coefficient vector (01-pri), and present ATVS system with 19 MFCC+ delta and specific UBMs of differents sizes (M=1024 in 18pri, M=512 in 19pri, and M=2048 in 21 pri)

GMM-based system can be adapted to provide LR values according to the bayesian approach, firmly established in any forensic discipline, comparing DET and Tippet plots, the latter strongly recommended to assess LR-based systems, appropriate for forensic speaker recognition.

Acknowledgements

Authors wish to thank Marta Garcia-Gomar and Oscar Garcia-Ledesma from ATVS-UPM, and Juan-Jesus Diaz-Gomez from DGGC for extensive testing and improvement suggestions.

References

1. Martin, A. *et al.*, “The DET curve in assessment of detection task performance”, Proc. EuroSpeech’97, pp. 1895-1898, Rhodes (Greece), 1997
2. Robertson, B., Vignaux, G.A., *Interpreting Evidence – Evaluating Forensic Science in the Courtroom*, Wiley, Chichester (UK), 1995
3. Foster, K.R., and P.W. Huber, *Judging Science: Scientific Knowledge and the Federal Courts*, MIT Press, Cambridge MA (USA), 1997
4. Broeders, A.P.A., “Forensic Speech and Audio Analysis: the State of the Art in 2000 AD”, Proc. of SEAF-2000 (1st National Conference of the Spanish Forensic Acoustics Society), Ed. J. Ortega-Garcia, Madrid (Spain), 2000
5. Champod, C., Meuwly, D., “The Inference of Identity in Forensic Speaker Recognition”, Speech Communication, vol. 31, pp. 193-203, June 2000
6. Evett, I.W., “Towards a Uniform Framework for Reporting Opinions in Forensic Science Casework”, Science & Justice 1998: 38(3), pp. 198-202
7. Champod, C., “Overview and Meaning of Identification”, Encyclopedia of Forensic Sciences, pp. 1077-1084, Academic Press, 2000
8. Aitken, C.G.C., “Statistical Interpretation of Evidence/Bayesian Analysis”, Encyclopedia of Forensic Sciences, pp. 717-724, Academic Press, 2000

9. Tippet C.F. *et al*., “The evidential value of the comparison of paint flakes from sources other than vehicles”, Journal of the Forensic Science Society, vol. 8, pp. 61-65, 1968
10. Evett, I.W. and Buckleton, J.S., “Statistical Analysis of STR (short tandem repeat) data”, Advances in Forensic Haemogenetics, A. Carracedo, B. Brickmann, and W. Bär, Editors. Springer-Verlag: Heidelberg, pp. 79-86, 1996
11. Taroni, F., Aitken, C.G.C., “Forensic Science at Trial”, Jurimetrics J. 37, 327-337, 1997
12. André-Obrecht, R. (ed.), *Special Issue on Speaker Recognition and its Commercial and Forensic Applications*, Speech Communication, pp. 87-270, Elsevier, 2000
13. Reynolds, D.A., “Speaker Identification and Verification Using Gaussian Mixture Models”, Speech Communication, vol. 17, pp. 91-108, Elsevier, 1995
14. Meuwly, D., Drygajlo, A., “Forensic Speaker Recognition based on a Bayesian Framework and Gaussian Mixture Modeling”, Proc. of Odyssey’2001 ISCA Speaker Recognition Workshop, Crete (Greece), 2001
15. Gonzalez-Rodriguez, J., Ortega-Garcia, J., Lucena-Molina, J.J., “On the Application of the Bayesian Framework to Real Forensic Conditions with GMM-based Systems”, Proc. of Odyssey’2001 ISCA Speaker Recognition Workshop, pp. 135-138, Crete (Greece), 2001
16. Gonzalez-Rodriguez, J., Ortega-Garcia, J., and Lucena-Molina, J.J., “Bayesian Evaluation of Speech Evidences with IdentiVox Automatic Speaker Recognition System”, Joint ENFSI-IAFP (European Network of Forensic Science Institutes – International Association of Forensic Phonetics) Meeting, Paris (France), July 2001
17. Ortega-Garcia, J., Gonzalez-Rodriguez, J., Marrero-Aguiar, V., “AHUMADA: a Large Speech Corpus in Spanish for Speaker Characterization and Identification”, Speech Communication, vol. 31, pp. 255-264, June 2000