Likelihood Ratio Calibration in a Transparent and Testable Forensic Speaker Recognition Framework

Daniel Ramos-Castro, Joaquin Gonzalez-Rodriguez and Javier Ortega-Garcia

ATVS (Speech and Signal Processing Group)
Escuela Politecnica Superior, Avda. Tomas y Valiente 11, Universidad Autonoma de Madrid
E-28049 Madrid, Spain
{daniel.ramos, joaquin.gonzalez, javier.ortega}@uam.es

Abstract

A recently reopened debate about the infallibility of some classical forensic disciplines is leading to new requirements in forensic science. Standardization of procedures, proficiency testing, transparency in the scientific evaluation of the evidence and testability of the system and protocols are emphasized in order to guarantee the scientific objectivity of the procedures. Those ideas will be exploited in this paper in order to walk towards an appropriate framework for the use of forensic speaker recognition in courts. Evidence is interpreted using the Bayesian approach for the analysis of the evidence, as a scientific and logical methodology, in a two-stage approach based in the similarity-typicality pair, which facilitates the transparency in the process. The concept of calibration as a way of reporting reliable and accurate opinions is also deeply addressed, presenting experimental results which illustrate its effects. The testability of the system is then accomplished by the use of the NIST SRE 2005 evaluation protocol. Recently proposed application-independent evaluation techniques (C_{llr} and APE curves) are finally addressed as a proper way for presenting results of proficiency testing in courts, as these evaluation metrics clearly show the influence of calibration errors in the accuracy of the inferential decision process.

1. Introduction

The debate about the presentation of forensic evidences in a court of law, including forensic speaker recognition, is currently a hot topic in many scientific and legal forums [1, 2, 3]. One of the main reasons of this discussion arises from the American Daubert rules for the admissibility of the scientific evidence in trials [4]. According to these rules, the U.S. Supreme Court suggests that scientifically sounding techniques presenting standard procedures and demonstrating their testability, accuracy and acceptance in the scientific community are likely to be admitted in a U.S. federal court of law. On the other hand, non-scientific statements, such as expert testimonies lacking of scientific foundations, are likely to be rejected. The implications of these rules are in accordance to many opinions of forensic experts worldwide [1, 5, 6, 7, 8], demanding more transparent procedures and a scientific framework for a logical and testable interpretation of the forensic evidence. The debate also considers that existing techniques which have been assumed by the court as error-free are starting to be questioned (see, for example, [9] for a complete study regarding latent fingerprint identification). This has been partly due to some critical errors in positive identification reports, highlighted by the mass media (like the Mayfield case in Madrid terrorist attacks in 11 March 2004 [10]). Also, forensic case data is not sufficiently integrated into the police investigative processes as it should be, and the use of standard models for crime analysis making use of evidence interpretation are being more and more demanded [11]. All these ideas should be considered in order to use automatic speaker recognition systems for forensic purposes.

In order to cope with this emerging requirements, the speaker recognition community should investigate ways of converging to this new paradigm in forensic science. Standard procedures and protocols for testing and assessing forensic speaker recognition systems may be helpful for their admissibility in courts. Also, proficiency testing using clear protocols in controlled situations should be used in order to clearly determine the capabilities of the system [1, 12, 13]. The procedures in use should be easily tested in order to clarify the accuracy of the systems used and to be conscious of the error rates present in the methodology at hand. In this sense, state of the art automatic speaker recognition is not as accurate as other classical techniques such as fingerprints or DNA. Therefore, caution should be taken in order to use it in courts [14]. Thus, the improvement of the performance of score-based automatic speaker recognition systems constitutes a main challenge and is a task in constant progress, successfully impelled by the yearly NIST Speaker Recognition Evaluations (SRE) [15]. Due to this periodic evaluation process, the methodologies and protocols for the assessment of speaker recognition systems are converging to a common framework. However, it is still needed to stimulate this convergence regarding forensic interpretation of the evidence using speaker recognition systems. In this sense, the Bayesian approach for evidence analysis [6, 7] has been proposed as a common framework for forensic interpretation of the evidence, and recent works demonstrates the adequacy of this technique for forensic speaker recognition [16], both using automatic [12, 13], phonetic-acoustic [17] or semi-automatic approaches [17]. Under such a framework, the fact finder is able to infer posterior probabilities (also known as *confidences* [12]) about the considered hypotheses in a logical and transparent way [16]. However, several problems in automatic forensic speaker recognition still need to be addressed, namely session variability and data scarcity [3, 18, 13]. This two problems may not only affect performance of automatic speaker recognition technology, but also introduce errors in the estimations needed for accurate Bayesian interpretation [13].

One of the main advantages of Bayesian methods is their testability. As opinions about the hypotheses are expressed in the form of posterior probabilities, there is a need of measuring not only the discrimination capabilities of the system, but the reliability of such confidences. Highly discriminant (or *refined*

[19]) systems may lead to wrong posterior probabilities if they do not elicit reliable (or *calibrated*) confidences [19, 20]. A significant work has been developed in the past in order to measure the calibration and refinement of elicited confidences (see [20, 21] and references therein). In this paper, we explore the concept of calibration in forensic speaker recognition systems, emphasizing its effects. The problem is addressed in an experimental way, presenting results of LR-based systems using different interpretation and assessment techniques recently proposed in the literature. A methodology for the interpretation of the evidence according to the new requirements in forensic science has been used for the presentation and evaluation of such results. This work is organized as follows. Section 2 proposes some guidelines as steps towards the "coming paradigm shift" [1] in forensic speaker recognition. In order to obtain accuracy in the systems following the suggested Bayesian framework, Section 3 describes the problem of calibration in Bayesian forensic speaker recognition, addressing a methodology for the assessment of its effects and presenting some examples which clarify its importance. Experimental results are reported in Section 4, where the effect of calibration is highlighted by testing and comparing several robust approaches proposed in the literature for Bayesian forensic speaker recognition. In Section 5 a brief discussion is included for clarity. Finally, conclusions are drawn in Section 6.

2. Towards a new paradigm in forensic speaker recognition

The Daubert rules [4] define a set of requirements for scientific evidence to be accepted in a U.S. federal court of law. Briefly, in order to be admitted in court, any technique must satisfy the following conditions: *i)* it has been or can be tested. *ii)* it has been subjected to peer review or publication, *iii)* there exist standards controlling its use, *iv)* it is generally accepted in the scientific community, and *v)* it has a known or potential (and acceptable) error rate. These rules, added to the evidence of errors in some well-established forensic areas, have lead to reconsider the procedures used for forensic interpretation and reporting [1, 5]. A need of transparency and testability in the techniques used is demanded in order to submit proficiency test results to the court for the assessment of the methodology in use. This is in accordance to the ideas expressed by several forensic experts worldwide [6, 7, 8]. Moreover, it has been demonstrated that no forensic discipline is really error-free, even considering some well established disciplines which were viewed as error-free in the past (e. g., fingerprints [9]). These demonstrations have come either from the scientific community [8, 14] or from mistakes in real trials [9, 10]. In this sense, the idea of "discernible uniqueness" [1] of a given sample should not have validity anymore, as positive identification as a result of forensic analysis constitutes a "leap of faith" [22] adopted by the experts in a subjective way, usually justified by their experience in the field [5, 23]. This obscurity and arbitrariness in positive identification statements leads not only to usurp the judge's role in the decision making process [16], but also to a hardly testable framework.

In [1], DNA analysis is proposed as a model in order to avoid these difficulties. The main characteristics of forensic DNA analysis, highlighted in [1, 5] may be summarized in: *i)* it is scientifically based, avoiding expert opinions based on experience [5]; *ii)* it is clear and standard in their procedures, allowing scrutinizing and inspection by fact finders and forensic scientists [1]; and *iii)* it is probabilistic, avoiding hard *match* or *non-match* statements [1, 8, 22]. This forensic discipline, much newer than fingerprint analysis, has been characterized by the use of a two-stage approach in order to assess the weight of the evidence [24, 6, 5] based on: *i)* a similarity factor which supports that the questioned sample was left by a given suspect, and *ii)* a typicality factor which supports that the questioned sample was left by anyone else in a relevant population. In order to implement this procedure in a scientific way, in DNA analysis the Bayesian methodology for evidence analysis has been used as a model of a clear, standard and probabilistic framework [6, 7] suited to any forensic discipline. In this sense, during the last years recent work in the speaker recognition area has demonstrated that any score-based speaker recognition system can be adapted to work following the Bayesian methodology [25, 12, 13].

2.1. The Bayesian methodology: a two-stage approach

The Bayesian framework for interpretation of the evidence represents a mathematical an logical tool in order to implement the two-stage approach in the evidence analysis process. This Bayesian framework presents many advantages in the forensic context. First, it allows the forensic scientists to estimate and report a meaningful value to the court [16]. Second, the role of the scientist is clearly defined, leaving to the court the task of using prior judgements or costs in the decision process [26]. Third, probabilities can be interpreted as degrees of belief [27], allowing the incorporation of subjective opinions as probabilities in the inference process in a clear and scientific way.

Classically, Bayesian interpretation of the forensic evidence using automatic systems has been performed by generative statistical models [28, 6, 25, 13], whereas discriminative techniques have been also recently applied to this task [12]. In both cases, the objective is to compute the likelihood ratio (LR) as a degree of support of one hypothesis versus its opposite. This LR can be estimated from similarity scores computed by an automatic system [12, 13]. We assume that the evidence E is the information extracted from the questioned mark (e. g., a wire-tapping) and the suspect material (e. g., a recording from the suspect in controlled situations). Typically, using automatic systems this E will be a similarity score between the mark and the suspect material. However, other kind of meta-information (such as signal to noise ratio, transmission channels, subjective quality of the speech signal, etc.) may be also used in order to compute this LR value [12]. Therefore:

$$LR = \frac{f\left(E|\, H_p, I\right)}{f\left(E|\, H_d, I\right)} \tag{1}$$

where H_p (a given suspect is the author of the questioned recording involved in the crime) and H_d (another individual is the author of the questioned recording involved in the crime) are the relevant hypothesis and I is the background information available in the case. The hypothesis are defined in the court from I, the prosecutor and defense propositions and often because of the adversarial nature of the criminal system.

Equation 1 represents the two-stage approach in LR computation. The likelihood $f\left(e|\, H_p, I\right)$ in the numerator in Equation 1 is known as the within-source distribution, and models the variability of the speaker between sessions evaluated in $e = E$. The evaluation of this function in $e = E$ gives a measure of the similarity between the questioned material and the suspect. On the other hand, the likelihood $f\left(e|\, H_d, I\right)$ in the denominator is known as the between-source distribution, and its evaluation

in $e = E$ can be seen as a measure of the typicality or rarity of the suspect in a relevant population of individuals. Both values, similarity and typicality, are computed in a transparent way by the speaker recognition system or expert, and it is the duty of the forensic scientist, following the background information of the case (I), to select the population of individuals which will be proper for the case at hand. This two-stage approach can be easily documented by the forensic scientist and understood by fact finders [6, 5].

2.2. Testability

Proficiency testing is being seen as a key issue for the admissibility of forensic systems in courts [1]. According to Daubert, the knowledge about the error rates of the technique in use demands unified protocols for system evaluation in an scientific way. We identify two main factors as critical for the achievement of this goal in forensic speaker recognition. First, an effort for generating common protocols and databases for proficiency testing should be done. In this sense, the work by NIST and NFI/TNO in their respective SREs has been fundamental in the last years [15]. Second, the use of a common methodology for presenting results in court will measure and clarify the reliability of the system to be used for forensic analysis. In this paper we use the NIST 2005 SRE protocol for testing system performance, and we also use several evaluation methods for a clear presentation of results in a Bayesian framework, which are described in Section 3.1.

3. Calibration in Bayesian forensic speaker recognition

The concept of calibration was introduced in [19] in the context of weather forecasting. There, posterior probabilities (or confidences) were used as degrees of belief about a given hypothesis (tomorrow it will rain) against its opposite (tomorrow it will not rain). The accuracy of the forecaster was then assessed by means of *strictly proper scoring rules*, which may be viewed as cost functions which assign a penalty to a given confidence depending on: *i)* the probabilistic value of the forecast, and *ii)* the true hypothesis which actually occurred (see [21] for details).

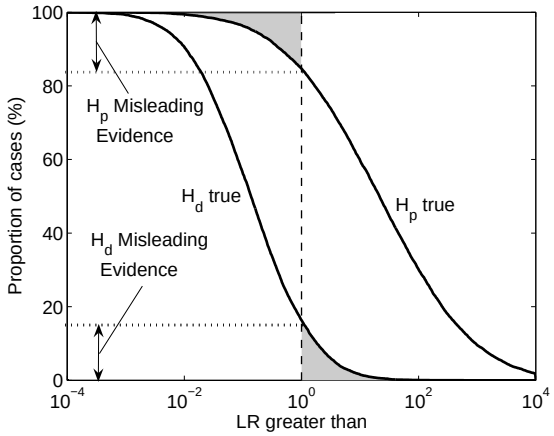

Figure 1: Example of Tippett plots showing the actual LR distributions (with its meaningful values) and the rates of misleading evidence when H_p and H_d are respectively true.

For example, if a probabilistic forecast gives a high probability of rain for tomorrow (value of the forecast) and tomorrow it does not rain (true hypothesis), a proper scoring rule will assign a high penalty to the forecast, and vice-versa. Strictly proper scoring rules have interesting properties. First, the *only* confidence value which optimize a strictly proper scoring rule is the *actual* probability of occurrence of the hypothesis [19]. Thus, any opinion expressed by the forecaster which deviates from the actual probability of occurrence of the hypothesis will lead to a higher penalty. Second, in [19] it is demonstrated that any proper scoring rule can be split into a *refinement* component, measuring the discrimination capabilities of the confidence values elicited, and a *calibration* component, which measures the deviation of such confidence values from the actual probabilities of occurrence of the hypothesis.

The use of proper scoring rules in order to assess speaker recognition systems delivering LR values has been recently proposed in the literature [21, 12]. In a speaker recognition context, each *forecast* is represented with the confidence on the hypothesis "the speaker is the author of the test utterance" or its opposite, which may be inferred from the LR computed by the speaker recognition system and the prior probabilities (not necessarily estimated by the system). This assessment framework is perfectly suited for the methodology proposed in Section 2.1 for forensic speaker recognition considering: *i)* the hypotheses used are H_p and H_d as defined in Section 2.1, *ii)* the prior judgements are province of the court, and *iii)* the LR is computed by the forensic speaker recognition system.

3.1. Assessing calibration in forensic speaker recognition

In NIST SREs, DET plots have been used to measure the discrimination performance of speaker detection technology. However, LR values are not only used as a discrimination score, but as a measure of the degree of support to a hypothesis against its opposite. Using the LR and the prior odds (province of the court [16]) we obtain a posterior probability or *confidence* for each hypothesis. Thus, the accuracy of the LR values does not only depend on their discrimination power for trials where H_p or H_d is true (measured by the refinement of the LR values), but in their actual values (calibrated LR values will lead to reliable confidences). Therefore, in Bayesian analysis of forensic evidences, Tippett plots have been classically used for performance evaluation [24, 13], as in NFI/TNO forensic SRE [29]. In this representation, the distribution of the LR values being H_p or H_d respectively true are plotted together. Important values shown by these curves (and not by DET plots) are the actual distributions of the LR values and the rates of misleading evidence. The rate of misleading evidence is defined as the proportion of LR values giving support to the wrong hypotheses ($LR > 1$ when H_d is true and $LR < 1$ when H_p is true). In Figure 1 an example of Tippett plots is shown, highlighting the rate of misleading evidence values (the H_p and H_d rates of misleading evidence are different in general).

Recent approaches for speaker recognition evaluation have proposed the use of application-independent metrics such as C_{llr} [21], where *application*, as defined in [21], is the set of prior probabilities and decision costs involved in the inferential process [26]. C_{llr} is a single scalar value defined as:

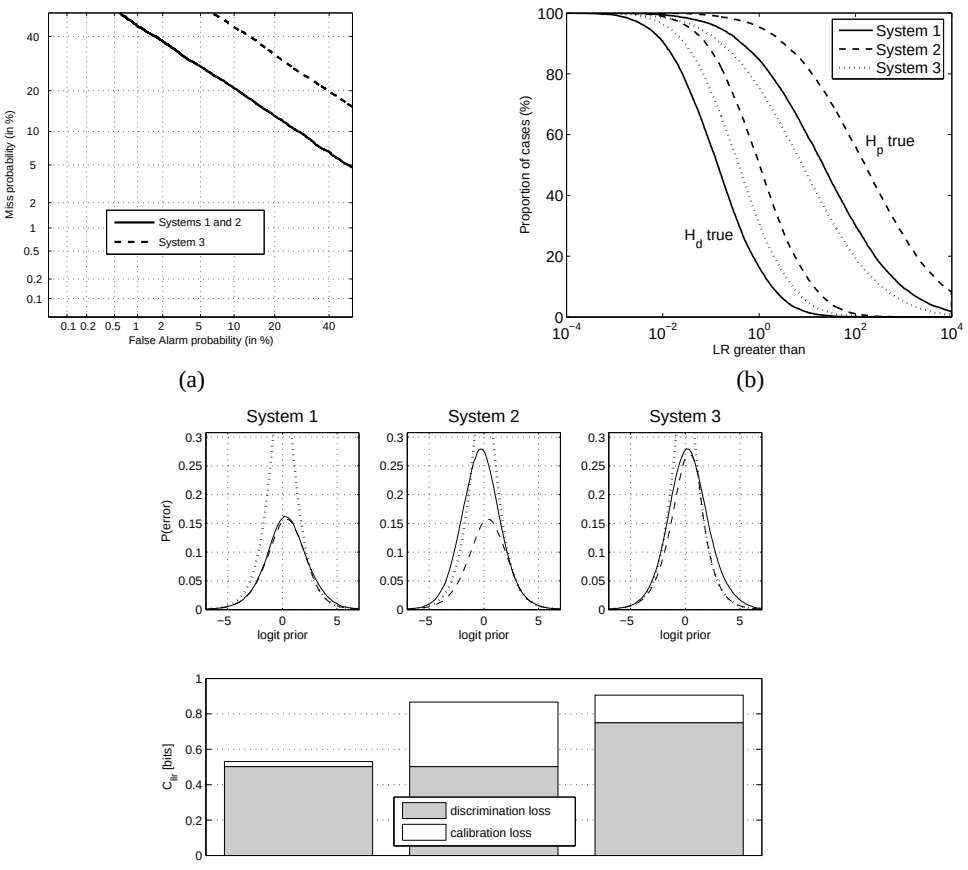

Figure 2: DET curves (a), Tippett plots (b) and APE curves (c) for three simulated systems (System 1, System 2 and System 3). LR values have been randomly generated in order to plot these curves.

$$C_{llr} = \frac{1}{N_{H_p}} \sum_{i for H_p=true} log_2 \left(1 + \frac{1}{LR_i}\right) + \frac{1}{N_{H_d}} \sum_{j for H_d=true} log_2(1 + LR_j) \quad (2)$$

where N_{H_p} and N_{H_d} are respectively the number of LR values in the evaluation set for H_p or H_d true. As it can be seen in Equation 2, hypothesis-dependent logarithmic cost functions are applied to the LR values being evaluated, and thus they are assessed depending on their numerical value: highly misleading LR values will have a strong penalty (high C_{llr}) and viceversa.

C_{llr} presents several interesting properties. First, the LR values are evaluated in an application-independent way, which in forensics means *case-independent*, where different costs and priors may be involved in the decision process of each different case [26]. Second, as a single scalar value, C_{llr} is very useful in order to easyly compare and rank systems. Third, C_{llr} has also an information-theoretical interpretation: given a system delivering LR values, $1 - C_{llr}$ measures the amount of information that is delivered from the system to the user (in our case, the fact finder) assuming a maximum entropy prior (in our binary case, $P(H_p) = P(H_d) = 1/2$). So, the lower the C_{llr} value, the higher the information delivered from the system to the fact finder. Finally, C_{llr} is a strictly proper scoring rule [21], and it can be split into discrimination loss (C_{llr}^{min}) and calibration loss ($C_{llr} - C_{llr}^{min}$). The C_{llr}^{min} value is obtained by optimal calibration via a monotonic transformation of the LR values, knowing the actual hypothesis occurred for each LR value. Details may be found in [21].

Based on this C_{llr} value, the APE-curve (Applied Probability of Error) [21] has been also proposed as a way of measuring the probability of error of the LR values computed by the forensic system in a wide range of applications (different costs and priors). This probability of error is represented for the actual LR values computed by the speaker recognition system and also for optimally calibrated LR values obtained as cited above for C_{llr}. Therefore, this representation clearly illustrates the effects of a lack of calibration: highly discriminant LR values may lead to a high probability of erroneous decisions if they are not properly calibrated. Because of their interesting properties, APE curves and C_{llr} will be used as an evaluation metric in coming NIST 2006 SRE [30]. In this paper, we have used the evaluation tools for C_{llr} and APE curve computation included in the toolkit FoCal [31].

3.2. Effects of calibration

The effects of calibration in forensic speaker recognition are illustrated in this section with an example using synthetic data. Here, three sets of LR values have been synthetically generated for each of the H_p and H_d hypotheses simulating three different forensic speaker recognition systems, two of them with the same discrimination ability. Figure 2 shows the performance of these *synthetic systems* (namely *System 1*, *System 2* and *System 3*) in terms of DET curves, Tippett plots, C_{llr} values and APE curves. DET curves in Figure 2(a) show that the discrimination power of *System 1* and *System 2* are the same in all operating points, outperforming *System 3*. However, Tippett plots in Figure 2(b) show that, although the separation between H_p and H_d curves is similar in *System 1* and *System 2*, the latter presents a significantly higher rate of misleading evidence for H_d. Also, confidences inferred from LR values computed by *System 2* and *System 3* will lead to important errors because of the high proportion of misleading LR values.

These results are clearly observed in Figure 2(c), which presents the same results in the form of C_{llr} values and APE curves. Overall performance is given by C_{llr}, split into discrimination loss (C_{llr}^{min}) and calibration loss ($C_{llr} - C_{llr}^{min}$). It is observed that *System 1* and *System 2* present the same discrimination performance (same discrimination loss), clearly outperforming System 3. However, C_{llr} values for *System 2* and System 3 are quite simillar, because of the high calibration loss presented by *System 2*. On the other hand, the calibration performance of *System 1* is the best for all systems.

In order to complete the analysis, APE curves in Figure 2(c) show the probability of error for all possible values of prior probabilities and decision costs (horizontal axis)[1]. The dashed line shows the performance of optimally calibrated LR values obtained by monotonic transformation from LR values given by the system [21]. The solid line shows the actual probability of error of the LR values computed. It is observed that the probability of error dramatically increases when the system is not properly calibrated. Due to this lack of calibration, posteriors inferred using *System 2* and *System 3* will have a similar probability of error, even when *System 2* has a much higher discrimination performance.

4. Experiments

In order to confirm the effects presented in Section 3 using actual speaker recognition systems, we present some experimental results using the techniques described below in the field of forensic speaker recognition.

4.1. System description

We carry out our experiments using the ATVS GMM-MAP-UBM system submitted to NIST 2005 SRE, which includes KL-Tnorm, an efficient and adaptive speaker- and test-dependent score normalization technique [32]. The comparative results presented here consider three techniques for the evaluation of the forensic evidence recenly proposed in the literature, namely: *i)* suspect-independent within-source computation [33], *ii)* suspect-adapted Maximum A Posteriori (MAP) estimation of within-source distributions [34] and *iii)* Within-source Degradation Prediction (WDP) [13]. We briefly describe each interpretation technique below.

In suspect-independent within-source estimation a framework is proposed assuming that an accurate model of the within-source distribution for a given suspect can be obtained using target scores from different individuals in the same conditions. Thus, we define $X_G = \{x_{G1}, \ldots, x_{GN}\}$ as a set of *global* target scores computed using speech from speakers other than the suspect. In this paper, we assume single Gaussian distributions for all estimations involved in within-source computation, and therefore we estimate the *global* within-source distribution $f(e|H_p, I) \equiv f_G(e) = N(\mu_G, \sigma_G)$ via Maximum Likelihood estimation from the X_G set.

On the other hand, suspect-adapted MAP estimation of within-source distributions is based on the fact that, even in the same conditions, the target scores coming from different speakers may present different distributions [35]. Therefore, accuracy in within-source estimation may be improved by exploiting suspect-specific scores, because the H_p condition claims that the suspect *and no other individual* is the author of the questioned recording. This is done via MAP adaptation [36] of the global distribution $f_G(e) = N(\mu_G, \sigma_G)$ to the *suspect* ditribution $f_S(e) = N(\mu_S, \sigma_S)$, estimated from a set of M suspect target scores $X_S = \{x_{S1}, \ldots, x_{SM}\}$ obtained from the suspect speech involved in the trial. Therefore, an *adapted* within-source pdf $f(e|H_p, I) \equiv f_A(e) = N(\mu_A, \sigma_A)$ is obtained. See [34] for details.

Finally, WDP combines suspect target scores X_S with between-source distribution information to predict score variability not present in the suspect data. This is achieved by varying the within-source distribution variance to a value based on the between-source distribution. Formally, let $f_S(e) = N(\mu_S, \sigma_S)$ be the suspect distribution estimated from X_S target scores. $f(e|H_d, I)$ has been defined as the between-source distribution for a given forensic trial (see Equation 1). The objective within-source pdf after WDP is defined as $f(e|H_p, I) \equiv f_{WDP}(e) = N(\mu_{WDP}, \sigma_{WDP})$. Our goal is to compute the desired parameter σ_{WDP}, as $\mu_{WDP} = \mu_S$ will remain unchanged. First of all, we compute s_{low} , which will be the score which satisfies

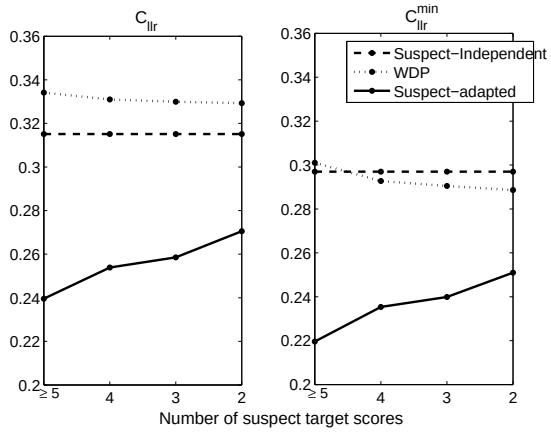

Figure 3: C_{llr} (a) and C_{llr}^{min} (b) of suspect-independent, WDP and suspect-adapted within-source computation with different amount of suspect scores (M).

[1] APE curves assess the probability of error for *all* values of the prior probabilities (expressed in prior log-odds using a logit function [21]) and costs involved in the decision process. Thus, we can represent both values in a single axis, because the priors and costs are related by a product in a Bayesian framework. See [21] for details.

$$\int_{s_{low}}^{\infty} f\left(e|\, H_d, I\right)\, de = \alpha \tag{3}$$

where $\alpha = 0.05$ in this contribution. Then, a gradient-descent algorithm for unconstrained optimization is used to compute σ_{WDP}, given that it is claimed to satisfy

$$\int_{-\infty}^{s_{low}} f_{WDP}\left(e\right)\, de = \alpha \tag{4}$$

Thus, $f\left(e|\, H_p, I\right)$ is predicted based on the information about the between-source distribution. Details may be found in [13].

4.2. Database and experimental protocol

The database and experimental protocol is fully described in [34] and briefly summarized here. Experiments have been performed using the evaluation protocol proposed in NIST 2005 SRE for the 8 conversation side training and 1 conversation side testing task (8c-1c, see [37] for details). Suspect target scores set X_S consists of all the target scores for each speaker from the whole score set in the evaluation, except the score used as evidence in each LR computation. We have only selected suspect vs. questioned speech comparisons having more than four suspect target scores, i. e., $M \geq 5$. A total number of 10.618 trials have been performed in this sub-condition. Background data, including global target score set X_G, has been extracted form NIST 2004 SRE database and protocol [15].

4.3. Results

First of all, we evaluate the effect of a lack of target suspect scores maintaining the rest of conditions by randomly selecting subsets of M scores from the total number of suspect target scores in each LR computation. Figure 3 shows the C_{llr} (calibration and refinement) and C_{llr}^{min} (refinement) values for all the techniques evaluated. It is shown that the overall performance (C_{llr}) of suspect-adapted within-source estimation clearly outperforms the rest of techniques for any value of M. On the other hand, WDP outperforms suspect-independent within-source estimation in discrimination power (C_{llr}^{min}) for $M \leq 4$. However, there is a calibration loss in the LR values computed using WDP, since the overall performance of the technique (C_{llr}) is the worst of all. This is mainly due to the fact that WDP is not considering calibration as an optimization objective (details in [13]), as will be noted bellow.

In Figure 4 we compare the perfromance of the different evaluated techniques under suspect data scarcity ($M = 2$). Results are presented in DET curves, Tippett plots, C_{llr} values and APE curves. DET curves in Figure 4(a) show that the discrimination capabilities of suspect-adapted within-source estimation are better for all operating points. However, this improvement is not so significant for low False Alarm rates (and DCF as defined by NIST [37], almost identical for the three systems). We also see that WDP slightly outperforms suspect-independent within-source estimation. In Figure 4(b) we also observe that suspect-adapted within-source computation performs better than the rest of approaches in terms of rates of misleading evidence. On the other hand, the numerical LR values computed by WDP under H_d are slightly higher than the rest, i. e., the support to the H_d hypotheses by the LR values computed using WDP is weaker. As it is noted below, this effect will lead to a calibration loss for the WDP technique.

Finally, C_{llr} values and APE curves are shown in Figure 4(c). It is observed that, although the discrimination capabilities (refinement) are better in WDP compared to suspect-independent within-source estimation, the calibration loss introduced by WDP leads to a poorer performance in terms of probability of error of the posterior decisions. This is because WDP aims at fixing the within-source distribution without considering the actual (and unknown) suspect data it claims to represent. Therefore, the predicted within-source pdf will not represent the actual distributions, and thus the technique will incur in a calibration loss. On the other hand, suspect-adapted within-source estimation, which is based on experimental data, clearly outperforms both techniques, also presenting a good calibration. These facts about the superiority of suspect-adapted within-source estimation over both suspect-independent within-source estimation and WDP are much clearer than in the DET plots of Figure 4(a).

5. Discussion

Experiments shown in this paper have illustrated the importance of the calibration of the LR values computed by a forensic system. Highly discriminant likelihood ratios *might* achieve a high performance in terms of probability of error of the posterior probabilities. However, a high calibration loss in the computed LR values may lead to arbitrarily high errors. Therefore, calibration is highly important in forensic reporting, because, in a Bayesian context, the fact finder will take decisions from the posterior odds inferred using prior odds and the LR values computed by the system.

This idea of calibration is strongly related to the rates of misleading evidence and the actual value of these misleading LR values showed in Tippett plots (see Figure 1). As C_{llr} is a logarithmic cost function, it penalizes the erroneous posteriors whatever they come from target of non-target trials. Thus, the optimization of C_{llr} implicitly leads to an optimization of both H_p and H_d rates of misleading evidences, but also of the numerical LR values being evaluated. Therefore, an optimization process focused on the rate of misleading evidence under H_d will lead to a calibration loss. While the reduction of the rate of misleading evidence and the limitation in such numerical LR values are important points in forensic systems, an optimal process for accomplishing this objective would imply the calibration of the system LR values and then the use of costs or utilities in the decision process by the fact finder, as has been recently proposed in forensic science [26].

6. Conclusions

In this paper we have emphasized the importance of calibration of LR values in Bayesian forensic speaker recognition following the rising needs being debated in the forensic science community. Questioning the infallibility of any forensic technique and demanding scientifically-sound methods for the admissibility of forensic evidence in the court are the main reasons for these new requirements. Some main guidelines for the use of forensic speaker recognition in courts may be drawn from this debate, such as the need of transparency, accuracy and testability for any technique to be admissible. This work has presented a methodology which copes with these interrelated requirements and therefore fulfills these needs. The transparency of the reasoning process under uncertainty is guaranteed by the use of the scientific and logical Bayesian framework for evidence analysis, as it happens in forensic DNA profiling. The

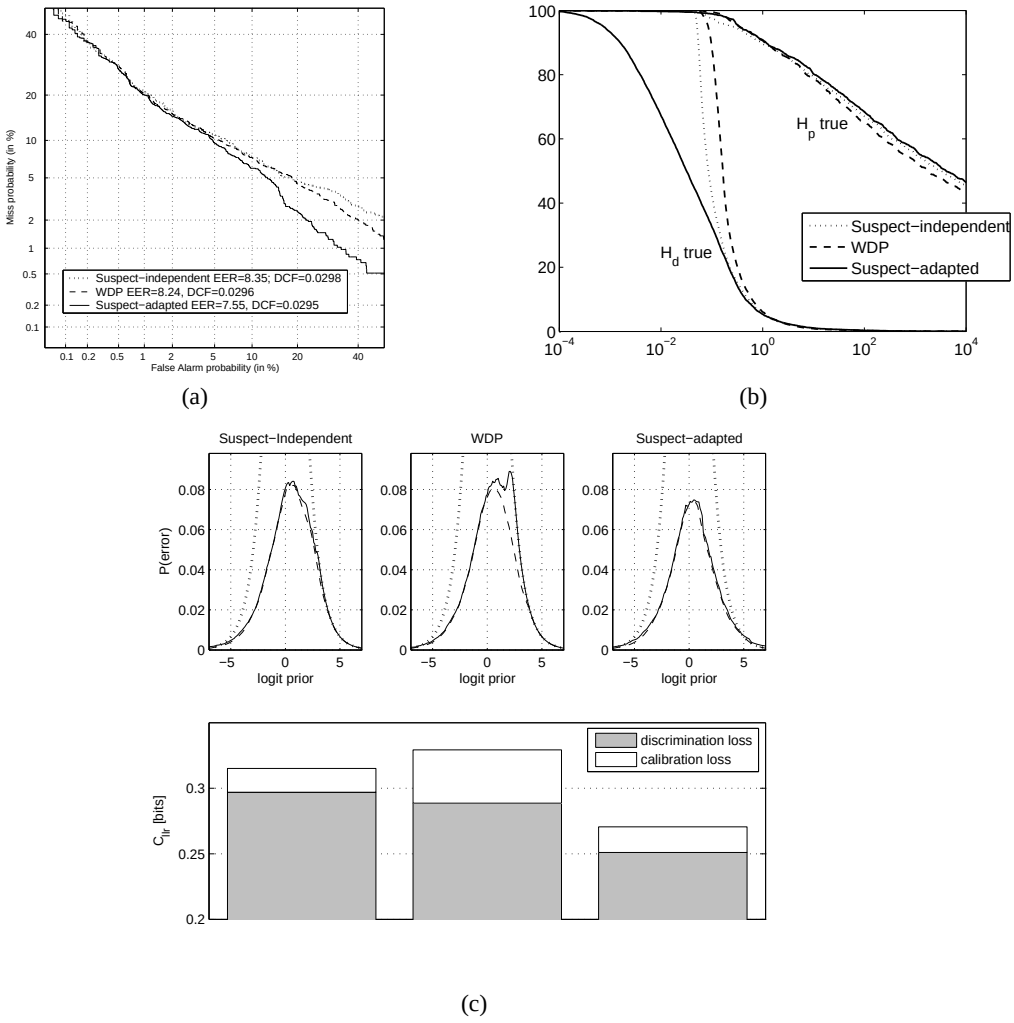

Figure 4: DET curves (a), Tippett plots (b) and APE curves (c) comparing suspect-independent, WDP and suspect-adapted within-source computation with scarce suspect data (M=2) in the selected subset from 8c-1c in NIST 2005 SRE.

discussion about the effects of a lack of calibration in automatic forensic speaker recognition systems has been supported by heuristic examples and experimental results. The conclusions from such discussion can be extended to any other forensic speaker recognition approach (semi-automatic, phonetic-acoustic, etc.) based on the Bayesian framework and reporting LR values. Several methods for the evaluation of forensic systems have been addressed, from classical techniques based on DET curves and Tippett plots to more recent application-independent approaches based on C_{llr} and APE curves. These two last metrics have been emphasized as a proper way of presenting results, as they show and highlight the calibration performance as a measure of reliability of the LR values computed by the forensic system. All these evaluation techniques, added to a clear and standard protocol such as those developed by NIST in their yearly SREs, give a method to perform proficiency tests in a controlled and transparent way. Therefore, the proposed methodology looks forward to fulfilling the needs of testability and standarization stated by the Daubert rules and demanded from forensic experts worldwide.

7. Acknowledgements

This work has been supported by the Spanish Ministry for Science and Technology under project TIC2003-09068-C02-01. The author D. R.-C. also thanks Consejeria de Educacion de la Comunidad de Madrid and Fondo Social Europeo for supporting his doctoral research. The authors also wish to thank Niko Brummer, Colin Aitken and Christophe Champod for fruitful discussions and useful comments.

8. References

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

[2] National Academy of Sciences Sackler Colloquium, "Forensic science, the nexus of science and the law," 2005, Presentations available at www.nasonline.org/site/PageServer?pagename= sackler_forensic_presentations.

[3] J. P. Campbell, "Linguistics and the science behind speaker identification," 2005, Available

at http://www.nasonline.org/site/PageServer?pagename=sackler_forensic_presentations.

[4] U.S. Supreme Court, "Daubert v. Merrel Dow Pharmaceuticals [509 U.S. 579]," 2003.

[5] S. A. Cole, "A history of fingerprinting and criminal identification," 2005, Available at http://www.nasonline.org/site/PageServer?pagename=sackler_forensic_presentations.

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

[7] I. W. Evett, "Towards a uniform framework for reporting opinions in forensic science casework," *Science and Justice*, vol. 38, no. 3, pp. 198–202, 1998.

[8] C. Champod and I. W. Evett, "A probabilistic approach to fingerprint evidence," *Journal of Forensic Identification*, vol. 51, pp. 101–122, 2001.

[9] S. A. Cole, "More than zero: Accounting for error in latent fingerprint identification," *Journal of Criminal Law & Criminology*, vol. 95, no. 3, pp. 985–1078, 2005.

[10] D. Heath and H. Bemton, "Portland lawyer released in probe of Spanish bombings," *Seattle Times*, vol. May 21, 2004, Available at http://www.law.asu.edu/?id=8857.

[11] O. Ribaux, S. J. Walsh, and P. Margot, "The contribution of forensic science to crime analysis and investigation: Forensic intelligence," *Forensic Science International*, vol. 153, no. 2-3, pp. 171–181, 2006.

[12] W. M. Campbell, D. A. Reynolds, J. P. Campbell, and K. J. Brady, "Estimating and evaluating confidence for forensic speaker recognition," in *Proc. of ICASSP*, 2005, pp. 717–720.

[13] J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia, "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 331–355, 2006.

[14] J. F. Bonastre, F. Bimbot, L.-J. Boe, J. P. Campbell, D. A. Reynolds, and I. Magrin-Chanolleau, "Person authentication by voice: A need for caution," in *Proc. of Eurospeech 2003*, pp. 33–36.

[15] D. van Leeuwen, A. Martin, M. Przybocki, and J. Bouten, "The NIST 2004 and TNO/NFI speaker recognition evaluations," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 128–158, 2006.

[16] C. Champod and D. Meuwly, "The inference of identity in forensic speaker recognition," *Speech Communication*, vol. 31, pp. 193–203, 2000.

[17] P. Rose, "Technical forensic speaker recognition: Evaluation, types and testing of evidence," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 159–191, 2006.

[18] D. A. Reynolds, "An overview of speaker recognition technology," in *Proc. of ICASSP*, 2003, pp. 4072–4075.

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

[20] I. Cohen and M. Goldszmidt, "Properties and benefits of calibrated classifiers," in *Proc. of European Conference on Machine Learning ECML/PKDD*, 2004.

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

[22] D. A. Stoney, "What made us ever think we could individualize using statistics?," *Journal of the Forensic Science Society*, vol. 31, no. 2, pp. 197–199, 1991.

[23] C. Champod, "Identification/individualization: Overview and meaning of ID," *Encyclopedia of Forensic Science, J. Siegel, P. Saukko and G. Knupfer, Editors. Academic Press, London*, pp. 1077–1083, 2000.

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

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

[26] F. Taroni, S. Bozza, and C. G. G. Aitken, "Decision analysis in forensic science," *Journal of Forensic Sciences*, vol. 50, no. 4, pp. 894–905, 2005.

[27] F. Taroni, C. G. G. Aitken, and P. Garbolino, "De Finetti's subjectivism, the assessment of probabilities and the evaluation of evidence: A commentary for forensic scientists," *Science and Justice*, vol. 41, no. 3, pp. 145–150, 2001.

[28] J. Curran, *Forensic Applications of Bayesian Inference to Glass Evidence*, Ph.D. thesis, Statistics Department, University of Waikato, New Zealand, 1997.

[29] D. van Leeuwen and J. S. Bouten, "Results of the 2003 NFI-TNO forensic speaker recognition evaluation," in *Proc. of Odyssey 2004*, pp. 75–82.

[30] NIST, "2006 speaker recognition evaluation plan: http://www.nist.gov/speech/tests/spk/2006/index.htm," 2006.

[31] Niko Brummer, "Focal toolkit," Available in http://www.dsp.sun.ac.za/ nbrummer/focal/.

[32] D. Ramos-Castro, D. Garcia-Romero, I. Lopez-Moreno, and J. Gonzalez-Rodriguez, "Speaker verification using fast adaptive Tnorm based on Kullback-Leibler divergence," in *Proc. of 3rd COST 275 Workshop.*, 2005, pp. 49–52.

[33] F. Botti, A. Alexander, and A. Drygajlo, "An interpretation framework for the evaluation of evidence in forensic automatic speaker recognition with limited suspect data," in *Proc. of Odyssey 2004*, pp. 63–68.

[34] D. Ramos-Castro, J. Gonzalez-Rodriguez, Alberto Montero-Asenjo, and J. Ortega-Garcia, "Suspect-adapted MAP estimation of within-source distributions in generative likelihood ratio estimation," accepted in Odyssey 2006.

[35] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. A. Reynolds, "Sheeps, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," in *Proc. of ICSLP*, 1998.

[36] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[37] NIST, "2005 speaker recognition evaluation plan: http://www.nist.gov/speech/tests/spk/2005/index.htm," 2005.