On Robust Estimation of Likelihood Ratios: The ATVS-UPM System at 2003 NFI/TNO Forensic Evaluation

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

ATVS (Speech and Signal Processing Group)
Universidad Politécnica de Madrid (Spain)
jgonzalez@diac.upm.es, dramos@diac.upm.es, mgomar@atvs.diac.upm.es, jortega@diac.upm.es

Abstract

This paper summarizes the different algorithms developed in ATVS-UPM in order to submit a reliable Likelihood Ratio based forensic system, fully compliant with the bayesian framework for the analysis of forensic evidences, to 2003 NFI-TNO Forensic Speaker Recognition Evaluation. Once identified the main causes and consequences of the erratic estimation of Likelihood Ratios due to forensic conditions, mainly lack of data and mismatch between suspect and questioned speech, several algorithms are proposed and assessed using Switchboard data. Moreover, a new algorithm, TDLRA (Target Dependent Likelihood Ratio Alignment), guarantees efficiently the presumption of innocence for non-target speakers, which is a mandatory condition of any forensic system. The LR-based submitted system is then assessed with NFI-TNO forensic field data, showing an excellent performance in all evaluation conditions, preserving the presumption of innocence and providing a meaningful Likelihood Ratio for any questioned-speech/suspect-speech pair of the evaluation, which could be directly used for reporting to Court under this bayesian forensic framework.

1. Introduction

In the last decade, Forensic Science has shown that the proper way to submit results to Court from a scientific laboratory or forensic expert is what is known as the Bayesian Framework for the Analysis of Evidence [1][2]. This framework can be applied to Forensic Speaker Recognition (FSR) by means of automatic speaker recognition systems [3][4], or making use of classical phonetic-acoustic techniques [5].

Even though state of the art speaker recognition systems show extremely good discrimination abilities, as shown in yearly NIST evaluations, a step forward is needed to allow using those systems in a forensic environment. The adaptation process is not straight forward, specially when lack of data from the suspect (known as speech controls) or unmatched conditions (channel, noise, emotions, language, voice disguise, etc.) are present, which are usual situations in FSR.

Having those facts in mind, ATVS-UPM has focused his research in adapting his raw-score-based NIST-eval-type speaker recognition system [6] to be fully compliant with this bayesian framework, making original contributions for the computation of robust Likelihood Ratios (see [7] and Section 4), with a double objective: firstly, to provide a meaningful Likelihood Ratio (LR) for every questioned and suspect speech pair, avoiding or minimizing the big proportion of non-reporting cases present on forensic speaker recognition because of non-matching conditions or limited quality of the data, and secondly, to guarantee the presumption of innocence in all cases, keeping non-targets (innocents falsely involved as suspects) with LR scores smaller than one, that is, not supporting the prosecution hypothesis.

The paper is organized as follows. After this introduction, a short review on the relations of the bayesian framework for the analysis of the forensic evidence and classical bayesian theory is firstly presented. Secondly, we will focus on the problems for likelihood ratio estimation arising from lack of data, typically the absence of speech controls for within-source distribution estimation, and unmatching conditions between questioned and suspect speech. In the fourth section, algorithms for robust likelihood ratio estimation will be shown, focusing on a novel algorithm, TDLRA (Target Dependent Likelihood Ratio Alignment), which preserves non-targets for obtaining LR scores greater than one guaranteeing the presumption of innocence, critical in forensic applications. In order to show the effectiveness of these algorithms, different tests with and without the proposed robust algorithms will be shown using the Switchboard I Database. In section five, and as a summary of all aspects of the work reported in this paper, the participation of the ATVS-UPM forensic system in 2003 NFI-TNO Forensic Speaker Recognition Evaluation is described, assessing the adequacy of the proposed robust algorithms in a big (about 25000 tests) and real forensic field data evaluation. Finally, some conclusions are extracted, and a thorough reference list is included to fully understand the different topics addressed in this paper.

2. Bayesian Analysis of Forensic Evidence and Bayesian Decision Theory

In this section we show the relationship between the Bayesian Decision Theory under an error probability optimal framework regarding a two-class scenario, i.e., a verification task; and the Bayesian approach for the evaluation of evidence presented here for forensic purposes.

Automatic speaker verification as a bayesian decision is better understood if we start describing the more general classification problem, namely the identification task. Assumed all probability distributions involved are known, Bayesian Decision Theory give an optimal classification in the sense of minimum error probability. Given a set of N+1 hypotheses $\{H_0, H_1, H_2, ..., H_N\}$ being H_i = *"The speaker i is the author of the questioned speech"*, that will be referred as *classes*; a decision based in the maximization of the *a posteriori* prob-

This work has been supported by the Spanish Ministry for Science and Technology under projects TIC03-09068-C02-01 and TIC2000-1669-C04-01.

ability distribution given the observation (parameters extracted from questioned speech) leads, following the *Bayes Rule*, to a minimum classification error rule [8]:

$$\hat{H}\left(\mathbf{x}\right) = \arg\max_i p\left(H_i|\,\mathbf{x}\right) = \arg\max_i p\left(\mathbf{x}|\,H_i\right)p(H_i) \quad (1)$$

Viewed as a particularization of the identification task, speaker verification can be expressed as a bayesian hypothesis test when only two competing hypothesis are involved: H (*The speaker is the author of the voice*) and $\overline{H}$, the null hypothesis (*The speaker is not the author of the voice*). This leads, assuming uniform decision costs, to the following decision rule:

$$\frac{p\left(\mathbf{x}|\,H\right)}{p\left(\mathbf{x}|\,\bar{H}\right)} > \frac{\Pr(\bar{H})}{\Pr(H)} = \Theta \quad (2)$$

In other words, an score relative to an observation independent threshold based on *a priori* probabilities decides what decision to take.

The main drawback of this approach relies in the fact that the likelihoods implied in the decision rule are almost never known. One way to handle with this situation in practice is to consider that the shape or kind of probability density functions involved is known. Then, a parametric model λ can be obtained with training data for each hypothesis likelihood involved in the rule (user and null hypotheses), which leads to:

$$\frac{p\left(\mathbf{x}|\,\lambda\right)}{p\left(\mathbf{x}|\,\lambda_0\right)} > \frac{\Pr(\bar{H})}{\Pr(H)} = \Theta \quad (3)$$

Introducing decision error costs $C_{FalseAlarm}$ and C_{Miss}, assigned respectively to a false alarm and a miss detection, the well known Likelihood Ratio decision rule is obtained:

$$\frac{p\left(\mathbf{x}|\,\lambda\right)}{p\left(\mathbf{x}|\,\lambda_0\right)} > \frac{C_{FalseAlarm}}{C_{Miss}}\frac{\Pr(\bar{H})}{\Pr(H)} = \Theta \quad (4)$$

When threshold Θ is selected based in a Cost Detection Function (CDF), prior probabilities and costs determine the decision to take, but the influence of the prior in the decision can be arbitrarily weighted by means of a variation in the costs assumed. In a forensic framework, an *acceptance/rejection* decision reporting, even when confidence measures are included, is hiding the weight of the prior assessed by the Court, since decision error costs can be arbitrarily selected by the forensic scientist. Moreover, there is no way to represent separately the influence of the prior and the decision error costs, since threshold election *merges* both weights involved in a single value.

The aim of bayesian forensic speaker recognition systems is completely different. Since identity in criminalistic is considered an *individualization* process, that is, a reduction from an initial population of suspects by means of a minimization in the probability of the null hypothesis for a given suspect, it can never be viewed as a *decision* process in the way expressed above, since threshold election in such a decision rule would be made by the forensic scientist, which can never decide any prior statement which belongs only to the Court [4]. Therefore, the aim of any forensic system can never be an acceptance/rejection decision, but the reporting to the Court of the Likelihood Ratio computed as defined in equation 6. This LR quantifies the weight of the forensic evidence in the Court decision process, that is, the degree of support or weaken to the prior statements once evidence E (*observation*) has been analyzed. Expressed with the odds form of the Bayes rule:

$$O\left(H|\,E,I\right) = \frac{\Pr\left(E|\,H,I\right)}{\Pr\left(E|\,\bar{H},I\right)}O\left(H|\,I\right) \quad (5)$$

Expressed in words, the Posterior odds = Likelihood Ratio x Prior odds, where the prior odds concern to the Court (background information relative to the case, I) and the likelihood ratio (LR) is to be computed by the forensic scientist or system:

$$LR = \frac{\Pr\left(E|\,H,I\right)}{\Pr\left(E|\,\bar{H},I\right)} \quad (6)$$

Equation 5 shows the way of reporting the weight of the forensic evidence by means of a clear separation between the Court competencies (prior statements and final a posteriori bayesian decision) and the forensic scientific duty (Likelihood Ratio computation and reporting).

3. Problems in LR Estimation

Speaker recognition under a forensic environment presents some peculiarities that will cause errors in the estimation of LR scores if special care is not taken to avoid them. In this section we analyze some of the most common of these problems, showing their origin and consequences over LR estimations.

Generally, in real forensic conditions the quality and quantity of the speech data the scientist can handle is far from ideal. This discouraging environment usually leads to strong mismatches between questioned and suspect speech and lack of data for proper distribution estimations. Moreover, in forensic science any system must guarantee the presumption of innocence, that is, non-targets (falsely accused to be the suspect) should not obtain LR scores greater than one, even if it leads to worse discrimination between targets and non-targets.

The problem of between-source estimation is related with the selection and number of available models of the adequate relevant population. Given that between-source distribution represents the random match probability distribution of the evidence within the relevant population, it must present the same characteristics of the suspect speech data (channel, language, length, session variability...), so a problem will exist when this matched reference population is not available [7].

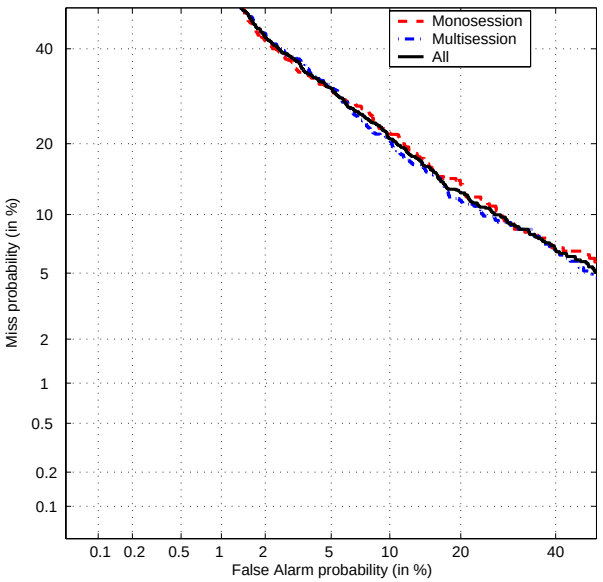

Figure 1: *ATVS-UPM@NIST'02-SRE system performance with Switchboard I data. Those reference raw scores are the basis for all algorithms in section 4.*

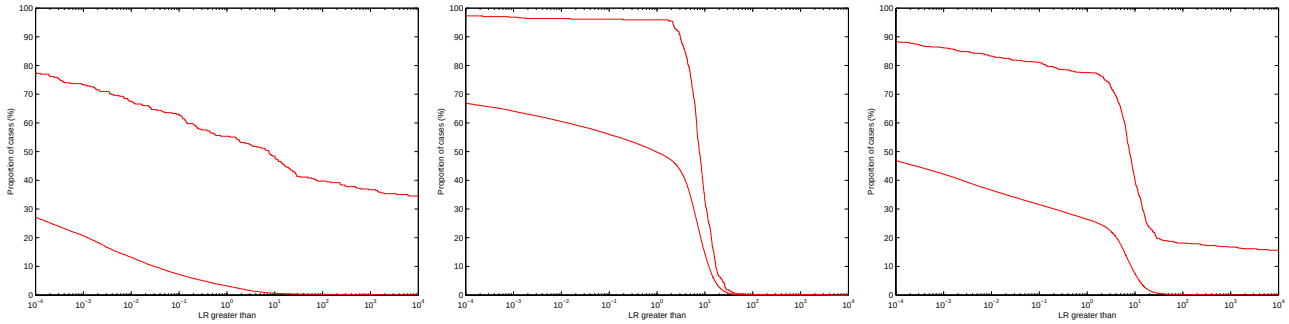

Figure 2: *Tippet Plots with Raw LR. 'Monosession' (left), 'multisession' (center) and 'all' (right) suspect Switchboard data.*

However, the biggest source of problems comes from inadequate estimations of the within-source underlying distribution, as the amount of speech controls usually available in forensics is not enough, both in quantity but specially in representation of unknown variability sources.

In the following section, we will try to give a thorough solution to all these problems, allowing the bayesian forensic system to be able of computing robust likelihood ratios whenever the raw-score speaker recognition system is able to perform a minimium discrimination between targets and non-targets (e.g., for any condition with EER under 30%).

4. Algorithms for Robust LR Computation

4.1. Questioned and Suspect Speech from Switchboard

All experiments reported in this section have been obtained using a 100 male speakers subset from SwitchBoard I parts 1, 2 and 3, which consists in landline telephone spontaneous conversational speech. Suspect models are obtained from 2 minutes of speech in two different groups:

- Single-session training (monosession): 50 speakers, suspect data is obtained from the same conversation.
- Multiple-session training (multisession): 50 speakers (different from above), 30 seconds of speech are obtained in each one of 4 conversations.

A user-dependent number of 30 seconds excerpts are used as questioned speech. Manual silence detection (labels) have not been used, and severe co-channel interference is present.

In order to test the robust LR estimation algorithms described below, a reference system is needed to provide raw scores to be used in the bayesian system later. Then, the ATVS-UPM UBM-MAP-adapted GMM system [6] as submitted to NIST'02 SRE has been used, with a UBM trained from 5 hours of male speaker data from SwitchBoard parts 4, 5 and 6. Three different experiments will be reported along this section, 'monosession' (50 spks), 'multisession' (50 spks), and 'all' (100 spks). Performance of this reference system in those three experiments is shown in figure 1.

4.2. Bootstrapping Suspect Data for Within-Source Distribution Estimation

In order to have as much scores as possible for within-source distribution estimation, a leave-one-out procedure is implemented, where N different segments/utterances are needed. Additionally, the absence of speech controls when a single recording from the suspect is available will also be solved. Two cases are possible, when a single suspect recording is available, being then divided into N uniform length segments, or when N different recordings are available. Then N different suspect temporal models are obtained for every speaker from N-1 segments/utterances each, obtaining N similarity scores which will be used for Within-Source Distribution Estimation (once the N scores are obtained, the temporal model are deleted and the suspect model is obtained from the N segments/utterances available). Those N scores will be an acceptable model of session variability in the multisession case, but a very optimistic one in the monosession case, as test speech is obtained from the same utterance than the model, giving an over-estimated model.

Figure 2 shows the performance of our bayesian forensic speaker recognition system when ML single-gaussian within-source estimation is performed directly from within-source modelling data, with a 30 seconds segments leave-one-out method. Between-source estimation has been obtained by means of comparing questioned speech with the relevant population (50 multisession 2 minutes-trained speakers), and again 32 gaussian EM-ML estimation is performed.

As shown, performance is poor, both for monosession, where targets unlikely score close to the optimistic model leading to low LR scores, and for multisession, where targets just obtain low LR scores which non-targets easily can also obtain due to small variance estimations from within source data.

4.3. Within-source Degradation Prediction (WDP)

Due to small within-source variances, evidences scoring higher than between-source estimation but lower than within-source estimation give erratic LR scores (unexpected high LR scores for non-targets and low LR scores for targets) even when a classical detection system would perform correctly. Within-source Degradation Prediction (WDP) is proposed [7] to stand for unknown degradation expected from unknown mismatch (channel, language, etc.), assuming no scores from impostors are expected higher than those in the between-source estimation. WDP is used differently in the two following cases (figure 4):

- Monosession data: within-source model, obtained from the same training data, determine the maximum scores that a target evidence could obtain.
- Multisession data: within-source model represent the known session variability in the speech controls, but unexpected lower scores are very likely due to different unmatched conditions.

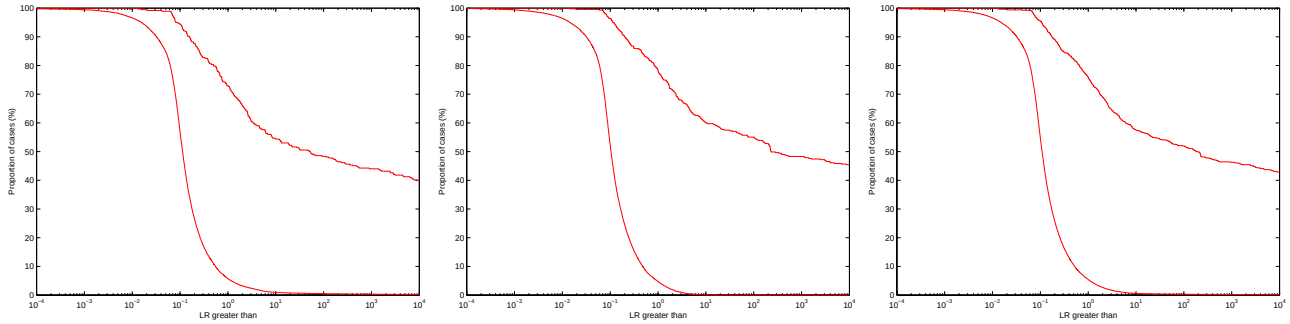

Figure 3: *Tippet Plots using WDP, WMVL and Outlier removal. 'Monosession', 'multisession' and 'all' suspect Switchboard data.*

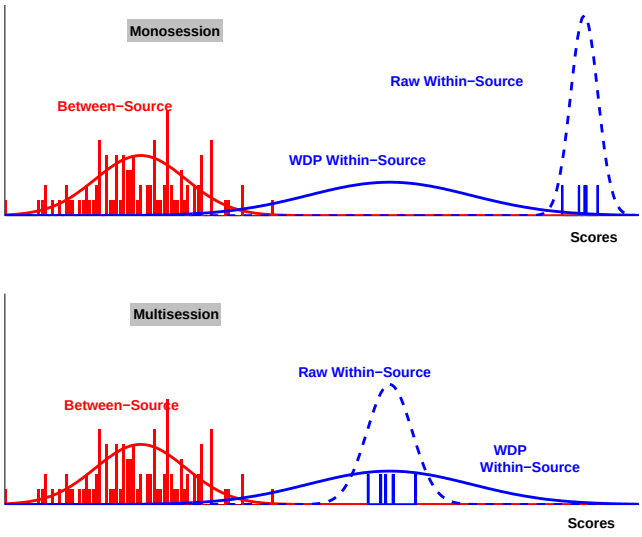

Figure 4: *Within-Source Degradation Prediction (WDP).*

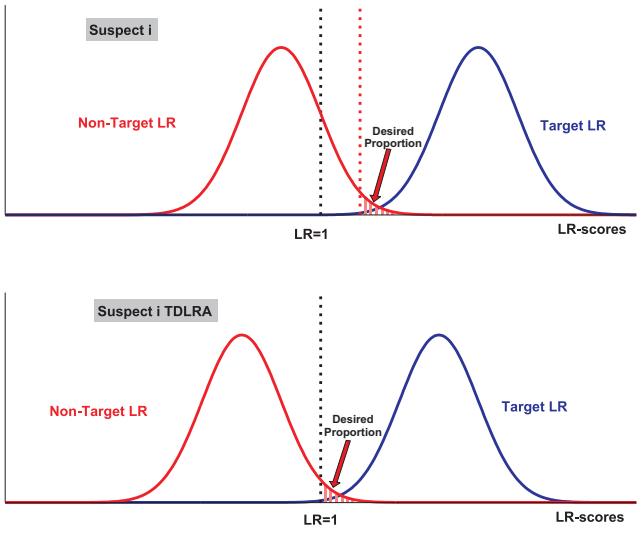

Figure 5: *Target Dependent LR Alignment (TDLRA).*

4.4. WMVL and Outlier Removal

In order to compensate for two types of estimation error, two complementary techniques are proposed:

- Within-source Minimum Variance Limiting (WMVL): even using WDP, very low estimated variances are still present due to highly coherent speech controls. Limiting the minimum variance of within-source estimations avoid those erratic LR scores.
- Outlier removal: Within-source model is usually estimated from few (less than five) scores. Then, singularities in one or two speech controls (very short true speech, laughing, noise peaks, etc.) can lead to inconsistent mean estimation.

In figure 3 the same three experiments as in figure 2 are reported, where the basic LR-based system has been significantly improved with the joint use of WDP, WMVL and outlier removal both for targets and non-targets in any of the tested conditions, and with an excellent,but still not ideal, 'presumption of innocence' performance.

4.5. Target Dependent LR Alignment (TDLRA)

In the previous section, different techniques have been proposed obtaining much more discriminant capabilities, but presumption of innocence is still not optimal. Recently, we have proposed the use of target dependent score alignment (TDSA) [9] for signature verification, which improves system performance. In this work, the objective will not be based in a cost detection function to be optimized per speaker but in guaranteeing the presumption of innocence for non-targets, which will be done in a speaker by speaker basis.

TDLRA works as follows. A set of non-target utterances with matched conditions relative to questioned speech is selected, and their respective LR values are computed for each suspect model, and modeled as a single-gaussian. Then, the LR correspondent to the proportion of non-targets that will be allowed to score above LR=1 (system configuration) is computed, and a simple normalization per speaker is performed. Figure 5 illustrates this technique.

TDLRA also has a positive effect in performance of the system when mismatching condition between the relevant population models and the suspect model are present. Since such a mismatch produce a misalignment in LRs obtained for each suspect, normalization between suspects performed by TDLRA leads to a better global system behavior.

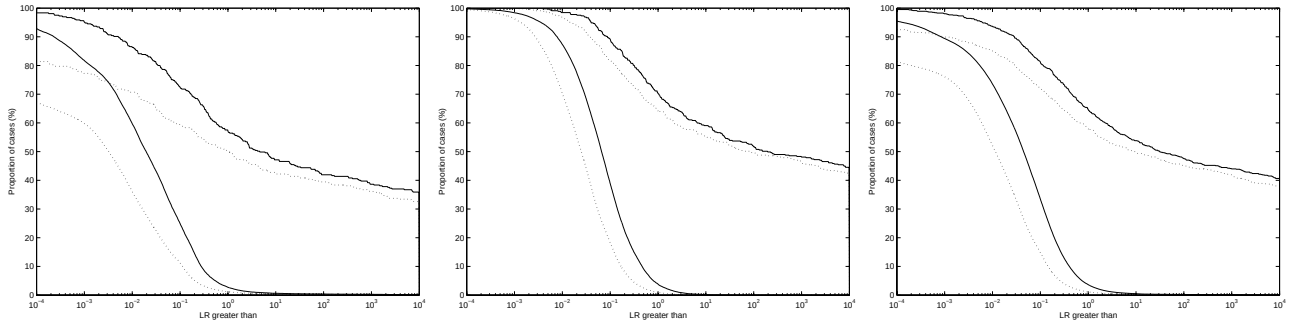

Figure 6: *Tippet plots for same system of figure 3 adding TDLRA. Two configurations for TDLRA are shown, with maximum Non-Target errors of 1% (dotted) and 3% (solid).*

Figure 6 presents system performance when TDLRA is used in the same Switchboard experiments, with two arbitrary values of 1% and 3% for the desired proportion of non target users supporting prosecution hypothesis. From these experiments, it is remarkable the degree of control over system performance with TDLRA technique, and specially controlling to keep the presumption of innocence for non-targets in the desired values.

5. ATVS-UPM Forensic System at NFI-TNO 2003 Evaluation

In order to determine the state of the art of text independent speaker recognition systems in a forensic context and the possibility of using the results of such systems for investigative purposes in police enquiries, the NFI-TNO Forensic Speaker Recognition Evaluation [10] was proposed in 2003 by the Netherlands Forensic Institute (NFI) and the Netherlands Organization for Applied Scientific Research (TNO). The speech material used in the NFI-TNO forensic speaker recognition evaluation was taken from real police investigations. This was done in order to obtain field data and to get as close as possible to a real forensic application. It consists of wire tapped cellular GSM to GSM telephone conversations recorded over a 23 month period. All speakers are males. The telephone line quality varies between recordings from excellent to moderate (extremes at the lower end were omitted). The telephone handsets used are unknown. The level and nature of background noises of the material varies and includes slight room reverberations, music in the background of the recording and in some cases background speakers (mostly sounds of children playing). Although the speaking style was constant (spontaneous speech; laughter, shouting and whispering was omitted) emotions varied between recordings from relaxed (frequent) to stressed (rare). The distribution of these parameters among speakers is not homogeneous. The range and distribution of recording dates between speakers varies. The material was edited by NFI in order to select single speakers and to make the material anonymous. Care was taken in editing so that no acoustic artifacts were introduced. Signalling noises in the telephone recordings were removed but speaking pauses were not edited out. The languages used are Dutch (79%), English (20%), Sranan Tonga (language spoken in Surinam) and Papiamento (language spoken at the Netherlands Antilles).

<site>	NDCF	%false-alarm	%miss
3	0.582	0.0035	0.5470
15	0.661	0.0277	0.3839
14	0.739	0.0535	0.2035
13	0.742	0.0097	0.6449
9	0.754	0.0178	0.5758
10	0.772	0.0022	0.7505
4	0.959	0.0001	0.9578
8	0.977	0.0085	0.8925
6	0.996	0.0793	0.2035
7	1.669	0.1485	0.1843
5	7.176	0.7119	0.0576

Figure 7: *Actual DCF in main NFI evaluation condition. ATVS-UPM-"raw" system is named sys9 here.*

5.1. Experiments

Added to the complexity of these forensic field data, the rules of the evaluation did not include any development data nor allowed the use of dutch data to optimize or adapt the submitted systems to the evaluation conditions. Several experimental configurations are proposed:

- Experiment 1 (Main Task): Dutch, 60 seconds training segments, 15 seconds test segments.
- Experiment 2 (Variation of parameters): Dutch, 30-120 seconds training segments, 7-30 seconds test segments, 1-4 sessions.
- Experiment 3 (Limited English Test): 60 seconds training segments, 15 seconds test segments.
- Experiment 4 (Cross-language test, Dutch test segments): 60 seconds training segments, 15 seconds test segments.
- Experiment 5 (Cross-language test, Non-Dutch test segments): 60 seconds training segments, 15 seconds test segments.
- Experiment 6 (Court proof): same speaker multiple models, test segments in order to estimate within-source speaker distribution.

ATVS-UPM submitted two systems, a "raw" score NIST-eval type system (sys9-1, primary)[6] and a bayesian LR-based forensic system (sys9-2)[7]. In both cases, a GSM-coded version of Switchboard I Extended Data Database has been used both for background modelling (UBM) and reference population for LR computation. By the time of writing this paper, only the results of the primary systems are known. NFI have reported 14 participant sites. Of these, 13 sites submitted results,

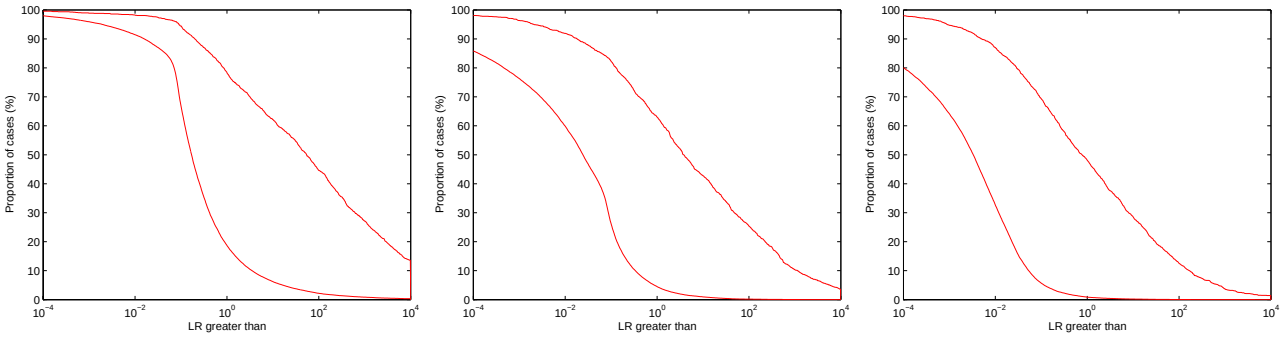

Figure 8: *ATVS-UPM forensic systems at NFI/TNO eval'03. Submitted (sys9-2a, left), same system with dutch population (sys9-2b, center) and TDLRA system with dutch pop. (sys9-2c, right)*

of which 11 were correct. Figure 7 shows actual DCF (Detection Cost Function) for the main condition of the evaluation.

However, we want to focus in our secondary system, sys9-2, the bayesian forensic one, where being fully compliant with the bayesian framework for the evaluation of evidence as defined in forensic science, we have been able to compute robust Likelihood Ratios for every single test file of the evaluation when compared to any suspect model. Note that no extra information is needed in our system even when having speech controls is a theoretical requirement for LR computation, or in other words, speech controls are directly obtained in our system from training speech with the leave-one-out procedure described above (as a special case, in experiment 6 -Court proof-speech controls in matched conditions are available and then no special requirement for robust LR estimation is present). The submitted forensic system performs all robust estimation techniques described above in section 4 with the exception of TDLRA, as additional dutch data was not allowed in the evaluation. Then, this secondary system can be directly used for evidential purposes in Court as a meaningful Likelihood Ratio (LR) is obtained with every test-file/suspect-model pair.

After the submission deadline and once the keys were distributed among participants, we have run again the evaluation making use of a dutch reference population (extracted from NFI/TNO field data), which was not permitted in the official evaluation. Anyway, this is a meaningful rule for a language independent evaluation but would be nonsense in a real forensic system as the use of matched data (language, channel,...) always improve system performance. Moreover, once matched data (in very general terms) is available, we have also run the evaluation with TDLRA, the robust LR estimation technique described above which maximally preserves the presumption of innocence of suspects, avoiding non-targets (innocents) obtaining LR scores greater than one.

The results of the submitted forensic system (sys9-2a) and the post-eval systems with dutch population (sys9-2b) and TDLRA with dutch population (sys9-2c) are shown in figures 8 and 9, respectively in the form of Tippet plots and DET curves. Remarkable results have been obtained with the forensic systems from two points of view: firstly, a meaningful LR score is obtained with the three systems for every test-file/suspect-model pair (both for targets and non-targets), as any LR has itself all the information needed in Court as shown in the bayesian framework for the analysis of evidence, while raw scores only give information when related to a threshold or user/impostor

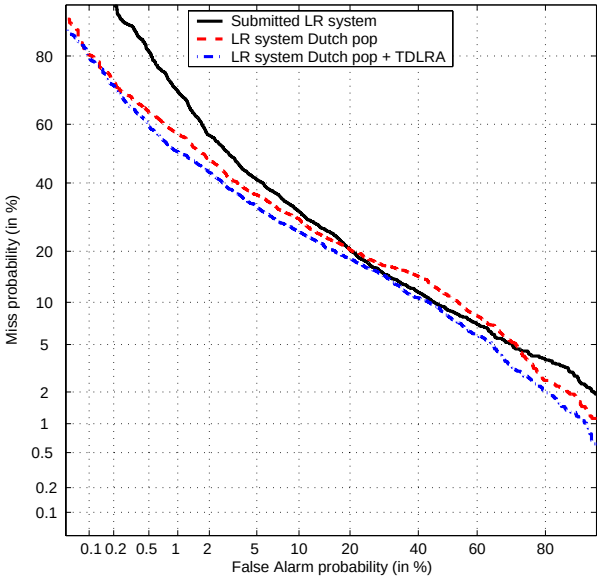

Figure 9: *ATVS-UPM forensic systems at NFI/TNO eval'03. Submitted (sys9-2a), same system with dutch population (sys9-2b) and TDLRA system with dutch pop. (sys9-2c)*

distributions, which should not be used in forensics as shown in [4]. And secondly, presumption of innocence is absolutely preserved when TDLRA is applied, where non-targets do not obtain LR scores greater than one, and about 50% of targets do obtain LR scores greater than one. Note that this a extremely complex evaluation condition, where more than 20.000 dutch files are tested with models obtained with 60 seconds from a single phone call and test files are just 15 seconds long, and even while 50% of targets are obtaining LR smaller than one, presumption of innocence is preserved in 100% of all cases (both for targets and non-targets).

Finally, we report for completion in figures 10 and 11 respectively the results of sys9-2b (submitted system with dutch population) and sys9-2c (TDLRA system with dutch population), in all 6 evaluation conditions, showing the excellent performance of TDLRA in any condition (different speech lengths and/or number of recording sessions for training, dutch/english/cross-language tests, etc.)

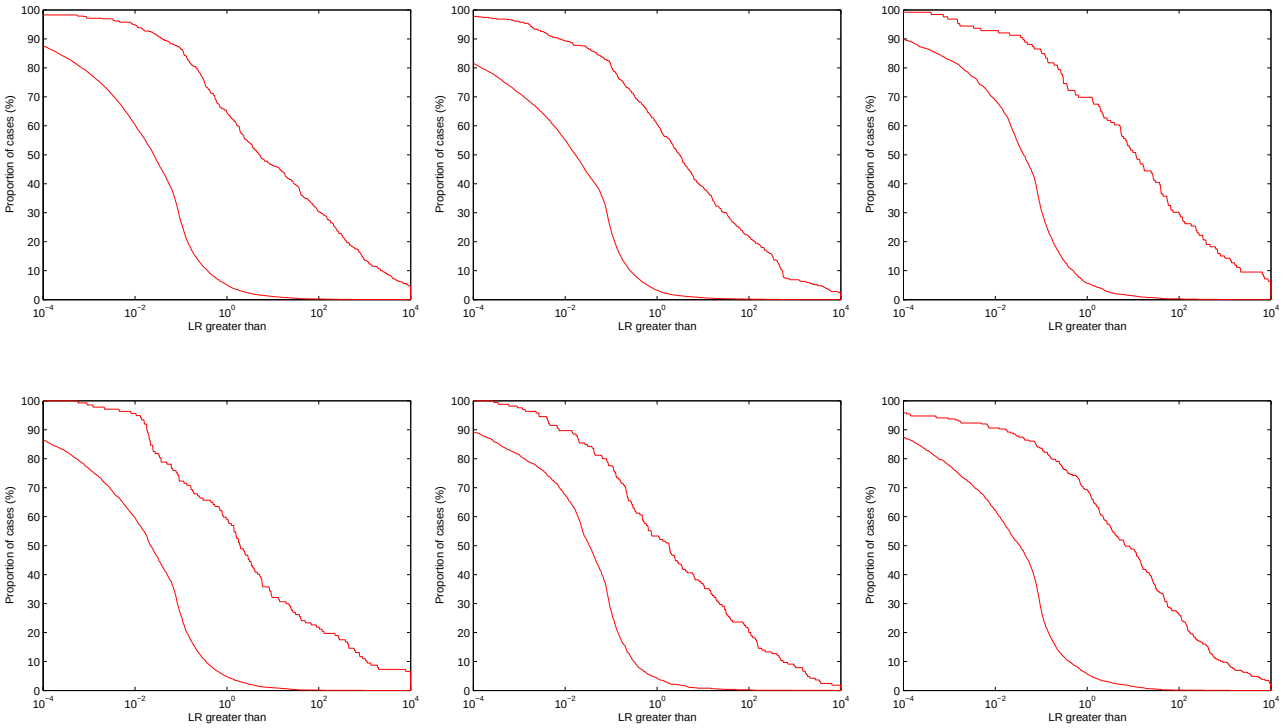

Figure 10: *NFI/TNO eval'03 with dutch population (left to right and up to down: experiments 1-6)*

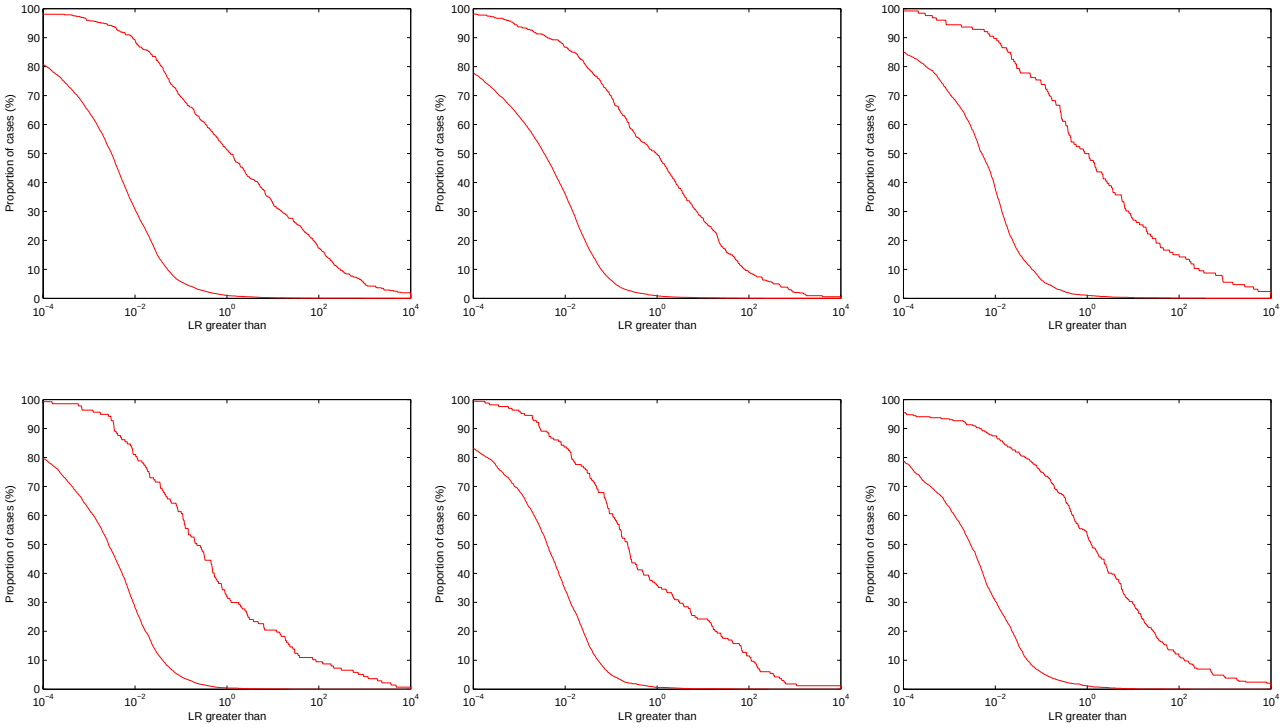

Figure 11: *NFI/TNO eval'03 with dutch population and TDLRA (left to right and up to down: experiments 1-6)*

5.2. Post Eval Experiments

During the last two years ATVS-UPM has focused his research in robust LR computation from raw scores, in our case from the basic technology we used in NIST Eval2002. In order to participate in NIST Eval 2004, ATVS has been testing some standard channel normalization options as RASTA filtering and Feature Warping.

Results are shown in figure 12 with NFI Eval data, where significant increase in performance have been obtained with the tested alternatives. Remarkably, the "warping+ tnorm" and "cmn+rasta+tnorm" alternatives give results equivalent to the best submitted system to the evaluation. Moreover, the LR system, obtained from the raw "warping+tnorm" scores holds the same detection performance.

Finally, in figure 13, the performance of our LR-based forensic system is shown properly in terms of a Tippet plot for the NFI Eval primary data. Results show even a much better performance than those shown above in 5.1 as better raw scores are obtained. TDLRA experiments with the same data will be shown at the conference.

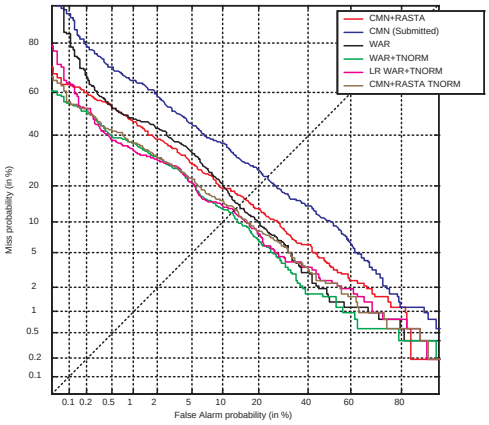

Figure 12: *Performance of ATVS-UPM system with different normalization techniques in NFI Eval primary condition.*

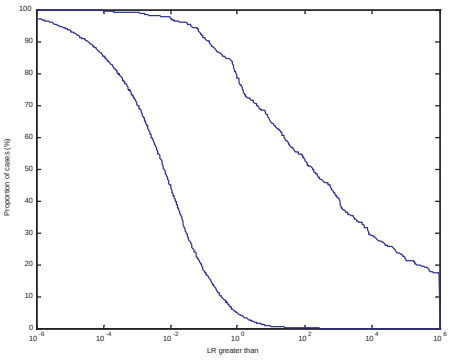

Figure 13: *Tippet plot of the ATVS-UPM LR system obtained from warping+tnorm scores in NFI Eval primary condition.*

6. Acknowledgements

Authors wish to thank all people from the speech lab of Guardia Civil for fostering this research and continuous new challenging field cases provision. D.R.C. also thanks "Consejería de Educación de la Comunidad de Madrid" and "Fondo Social Europeo" for supporting his doctoral research.

7. Conclusions

In this work some original contributions have been shown for the obtention of robust likelihood ratios in real forensic conditions. The proposed algorithms have been assessed both with Switchboard landline telephone data and NFI-TNO GSM forensic field data. A special remark must be made to TDLRA (Target Dependent Likelihood Ratio Alignment), a novel algorithm that preserves the presumption of innocence for non-targets in all the reported Switchboard and NFI-TNO experiments. Also remarkable is the fact that the system has performed robust LR estimation with blind data during official NFI-TNO Evaluation, having submitted a bayesian forensic system obtaining meaningful LR scores for the whole evaluation, even when no development data was available and dutch data was not permitted. The post-NFIeval experiments have allowed the use of dutch data, improving the submitted results, and also the use of novel TDLRA algorithm, needed of matched (dutch) data, obtaining a remarkable performance guaranteeing in all tests the presumption of innocence for non-targets, and a noticeable robust "detection" ability for targets, which allow for direct LR reporting to the Court within the forensic bayesian framework with the proposed system.

8. References

[1] C. G. G. Aiken, *Statistics and the Evaluation of Evidence for Forensic Scientists*, Wiley, 1995.

[2] I.W. Evett, "Towards a uniform framework for reporting opinions in forensic science casework," *Science and Justice*, vol. 38(3), pp. 198–202, 1998.

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

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

[5] P. Rose, *Forensic Speaker Identification*, Taylor & Francis Forensic Science Series, 2002.

[6] D. Garcia-Romero et al., "ATVS-UPM results and presentation at NIST'2002 speaker recognition evaluation," 2002.

[7] J. Gonzalez-Rodriguez, D. Garcia-Romero, M. Garcia Gomar, D. Ramos-Castro, and J. Ortega-Garcia, "Robust likelihood ratio estimation in bayesian forensic speaker recognition," in *Proceedings of Eurospeech*, Geneva, Switzerland, 2003.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, Wiley, 2001.

[9] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Target dependent score normalization techniques and their application to signature verification," in *Proc. ICBA*, 2004, (accepted).

[10] "http://speech.tm.tno.nl/aso/evalplan2003.pdf," .