# Robust Estimation, Interpretation and Assessment of Likelihood Ratios in Forensic Speaker Recognition

Joaquin Gonzalez-Rodriguez [a,*], Andrzej Drygajlo [b],

Daniel Ramos-Castro [a], Marta Garcia-Gomar [c],

Javier Ortega-Garcia [a]

[a] *ATVS (Speech and Signal Processing Group), Escuela Politecnica Superior, Universidad Autonoma de Madrid, E-28049 Madrid, Spain*

[b] *Signal Processing Institute. Swiss Federal Institute of Technology Lausanne (EPFL), CH-1015, Lausanne, Switzerland*

[c] *Agnitio Biometrics, Centro de Empresas La Arboleda, Ctra. A-3 Km. 7. E-28031 Madrid, Spain*

---

**Abstract**

In this contribution, the Bayesian framework for interpretation of evidence when applied to forensic speaker recognition is introduced. Different aspects of the use of voice as evidence in the court are addressed, as well as the use by the forensic expert of the likelihood ratio as the right way to express the strength of the evidence. Details on computation procedures of likelihood ratios ($LR$) are given, along with the assessment tools and methods to validate the performance of these Bayesian forensic systems. However, due to the practical scarcity of suspect data and the mismatched conditions between traces and reference populations common in daily casework,

significant errors appear in $LR$ estimation if specific robust techniques are not applied. Original contributions for the robust estimation of likelihood ratios are fully described, including TDLRA (Target Dependent Likelihood Ratio Alignment), oriented to guarantee the presumption of innocence of suspected but non-perpetrators speakers. These algorithms are assessed with extensive Switchboard experiments but moreover through blind $LR$-based submissions to both NFI-TNO 2003 Forensic SRE and NIST 2004 SRE, where the strength of the evidence was successfully provided for every questioned speech-suspect recording pair in the respective evaluations.

*Key words:* Forensic Speaker Recognition, evidence, interpretation, robust Bayesian likelihood ratio, Tippett plots

## 1 Introduction

In the last decade, research has demonstrated that a probabilistic model based on the odds form of Bayes' theorem and likelihood ratio seems to be an adequate tool for assisting experts in forensic sciences to interpret evidence [1][10]. This framework can be applied to forensic speaker recognition (FSR) by means of automatic speaker recognition systems [17][12][8], or making use of classical phonetic-acoustic techniques [20] as is performed in other forensic areas [7][13]. Even though the state-of-the-art speaker recognition systems show sufficient between-speaker discrimination abilities in many modern applications, as shown in yearly NIST evaluations, a step forward is needed to allow those systems to be used in a real-life forensic environment. The adaptation

* Corresponding author. Tel.: +34-91-4973142; fax: +34-91-4972235
*Email address:* joaquin.gonzalez@uam.es (Joaquin Gonzalez-Rodriguez).

process is not straightforward, especially when lack of a sufficient amount of data from the suspect or mismatched conditions (channel, noise, emotions, language, voice disguise, etc.) between the suspected speaker and questioned recordings are present, which are usual situations in a forensic environment.

In consideration, ATVS has focused its research on adapting its raw-score-based NIST-eval-type speaker recognition system to be fully compliant with the Bayesian interpretation framework, and on developing original contributions to the robust estimation of likelihood ratios. The dual objective is:

- To provide a meaningful likelihood ratio ($LR$) for each questioned and suspect speech pair. This will reduce the significant proportion of non-reporting cases present in forensic speaker recognition because of non-matching conditions or limited quality of the data.
- To guarantee the presumption of innocence in an analyzed case by keeping $LR$ scores from potential suspected non-perpetrators speakers (non-targets) smaller than one. This corresponds to not implicitly supporting the prosecution hypothesis.

The paper is organized as follows. A brief presentation of the relationship between the Bayesian data-driven methodology to interpret voice as evidence in forensic automatic speaker recognition and the Bayesian decision theory used in speaker verification is presented in Section 2. In Section 3, we focus on the problems related to likelihood ratio estimation arising from a lack of a sufficient quantity of data (typically the absence of speech controls) for within-source distribution estimation, and mismatched recording conditions for questioned and suspect speech. Then, in Section 4, we discuss the need for robust $LR$ estimation illustrated by examples from the NFI-TNO 2003

Evaluation. Algorithms for robust likelihood ratio estimation are developed, focusing on a novel technique, TDLRA (Target Dependent Likelihood Ratio Alignment), which prevents potential suspected non-perpetrators speakers (non-targets) from obtaining $LR$ scores greater than one and this way guaranteeing the presumption of innocence, critical in forensic applications. In order to show the effectiveness of these algorithms, Section 5 presents the results of different tests using the Switchboard I Database. In this same section the participation of the ATVS forensic speaker recognition system in the 2003 NFI-TNO Forensic Speaker Recognition Evaluation is presented, which gives the assessment of proposed algorithms in a large (about 25000 tests) real-life forensic data evaluation. We also show some post-evaluation experiments that illustrate the effect of channel normalization techniques in forensic systems, as well as the results in the NIST 2004 Evaluation, which illustrate some problems of forensic automatic speaker recognition in the case of mismatched recording conditions. Finally, Section 6 concludes with a global summary.

## 2 Bayesian Interpretation of Evidence

### *2.1 Voice as Evidence*

A forensic expert has to interpret evidence material in the course of a criminal investigation. The forensic evidence represents the information extracted by the forensic scientist which has any relevant information to add to the judicial process. In the case of a questioned recording (trace), the evidence does not consist in speech itself, but in the quantified degree of similarity between the speaker dependent features extracted from the trace, and the speaker depen-

dent features extracted from recorded speech of a suspect, represented by their model [20]. In an automatic approach, this similarity measure is quantified by a similarity score. Thus, let $O = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_N\}$ be the observation sequence of feature vectors extracted from the speech data and $\lambda_s$ the suspect model generated using the suspect material. Under these circumstances we can state the following:

$$E = s\,(O, \lambda_s) \tag{1}$$

where $s$ represents the similarity score. In the case of statistical modelling, the distribution of the speech features of the suspected speaker can be represented by a Gaussian Mixture Model (GMM) [19]. In this case, the score can be computed by means of the well known log-likelihood ratio formula:

$$E = s\,(O, \lambda_s) = \log f\,(O\,|\lambda_s\,) - \log f\,(O\,|\lambda_{UBM}\,) \tag{2}$$

where $f\,(X\,|\lambda_s\,)$ and $f\,(X\,|\lambda_{UBM}\,)$ are the probability density functions for the suspect model and a Universal Background Model ($UBM$) [19], which are modeled as mixtures of Gaussians.

### *2.2 Bayesian Interpretation of Evidence*

Bayes' theorem is an important part of the process of the consideration of the odds. In fact, the theorem permits the revision, based on new information, of a measure of uncertainty about the truth or otherwise of an outcome or issue (such as a hypothesis). The essential feature of Bayesian inference is that it

permits one to move from prior to posterior probabilities on the basis of data or subjective assessments [21].

In forensic science, identity of source cannot be known with certainty, and therefore must be inferred [14]. The inference of identity can be seen as an individualisation process, from an initial population to a restricted class, or, ultimately, to unity [5]. Concerning the inference of identity in forensic speaker recognition, speaker verification and speaker identification (in closed- and open-set) techniques have been shown inadequate for assessing the evidence in this field [6]. These techniques implicitly use subjective thresholds, forcing the forensic expert to make yes or no decisions, which should be devolved upon the court.

In this paper, Bayes' theorem and a data-driven methodology to interpret evidence are adopted for speaker recognition. The discussion benefits from the extensive research work in other identification-of-the-source forensic fields (e.g. fingerprint, fibers, DNA or glass trace evidence) [1]. The odds form of Bayes' theorem shows how new data (questioned recording) can be combined with prior background knowledge (prior odds, province of the court) to give posterior odds (province of the court) for judicial outcomes or issues, as shown in Equation (3):

$$\frac{\Pr\left(H_0 \middle| E, I\right)}{\Pr\left(H_1 \middle| E, I\right)} = \frac{\Pr\left(E \middle| H_0, I\right)}{\Pr\left(E \middle| H_1, I\right)} \cdot \frac{\Pr\left(H_0 \middle| I\right)}{\Pr\left(H_1 \middle| I\right)} \tag{3}$$

where $\Pr\left(A \middle| B\right)$ denotes a conditional probability value, $H_0$ is the hypothesis "the suspected speaker is the source of the questioned recording" and $H_1$ is the hypothesis "another person (within a relevant population of individuals) is the

source of the questioning recording". $I$ represents the additional background information not related with the forensic evidence, such as number of suspects involved in the case, relationship between the suspect and the crime scene, etc. Equation (3) allows for revision based on new information of a measure of uncertainty (likelihood ratio of the evidence (province of the forensic expert)) which is applied to the pair of competing hypotheses.

This hypothetical-deductive reasoning method, based on the odds form of Bayes theorem, allows the evaluation the likelihood ratio for the evidence ($E$)

$$LR = \frac{\Pr\left(E|\, H_0, I\right)}{\Pr\left(E|\, H_1, I\right)} \tag{4}$$

that leads to the statement of the degree of support for one hypothesis against the other, considering the circumstances of the case ($I$) and the result of the analysis of the questioned recording. In the remainder of this paper, explicit mention of the circumstances of the case $I$ is omitted in general from probability statements for ease of notation. The ultimate question relies on the evaluation of the strength of this evidence provided by an automatic speaker recognition method. In this case, the functions involved in $LR$ computation are *continuous* probability density functions, denoted as

$$LR = \left.\frac{f\left(x\,|H_0\right)}{f\left(x\,|H_1\right)}\right|_{x=E} = \frac{f\left(E\,|H_0\right)}{f\left(E\,|H_1\right)} \tag{5}$$

i. e., the $LR$ is a probability density function ratio evaluated in $E$ as defined in Equation 1.

It is important to remark that the $LR$ computed in Equation 5 is substantially

different from the log-likelihood-ratio formula used in Equation 2. The former relates priors and posteriors given hypotheses $H_0$ and $H_1$, and therefore its meaning is based on this inference process considering such hypothesis. In the latter, $H_0$ and $H_1$ are not considered, and the likelihoods computed are related to similarities between the test speech and the suspect and background model, respectively. Thus, the evidence score alone (Equation 2) cannot be used directly to relate priors and posteriors given $H_0$ and $H_1$.

### *2.3 Likelihood Ratio - Strength of Evidence*

The strength of the evidence is the result of the interpretation of the evidence, expressed in terms of the likelihood ratio of two alternative hypotheses. The procedure for the calculation and the interpretation of the evidence is presented in Figure 1. It includes the collection (or selection) of the databases (or data sets, because of the limited size of the suspect "database"), modeling and scoring from an automatic speaker recognition system, and the Bayesian interpretation. The likelihood ratio ($LR$) summarizes the statement of the forensic expert in the casework.

The Bayesian interpretation methodology used in this paper needs a two-stage statistical approach. The first stage consists in modelling multivariate feature data using Gaussian Mixture Models (GMMs). Figure 2 represents the statistical distributions of the GMM-based system scores obtained when $H_0$ and $H_1$ are true. Each of them models the probability density (pdf) of the scores when two recorded voices originate from the same person ($H_0$ is true) and from different persons ($H_1$ is true) [1].

[1] Similar 2-pdf plots can be easily obtained for every speaker in an assessment test,

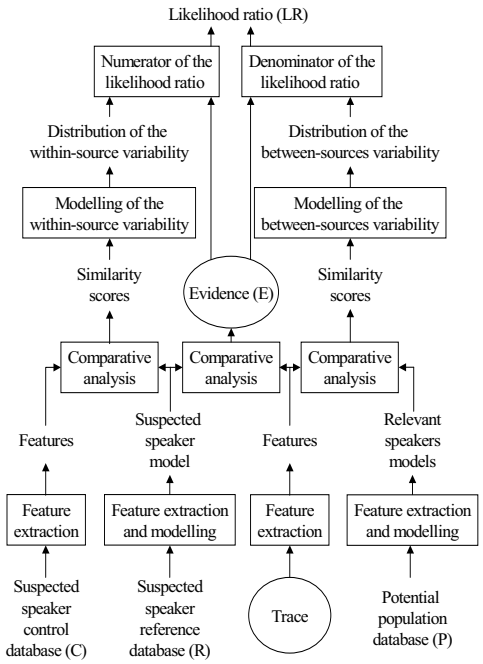



Fig. 1. $LR$ Computation Steps.

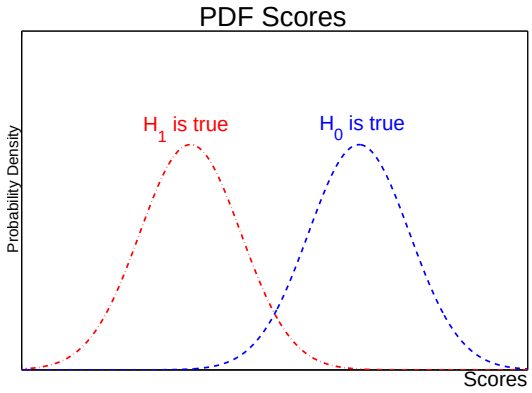



Fig. 2. PDFs of scores given by the GMM system.

The second stage transforms the data to a univariate projection based on modelling the similarity scores. The GMM method is not only used to calculate the evidence by comparing the questioned recording (trace) to the GMM of the

---

but they can also be all pooled together in a single 2-pdf plot using all $H_0$ and $H_1$ scores. In this latter case, the pooled scores are usually used to plot the performance of the system through DET plots [16].

suspected speaker (source), but also to produce data necessary to model the within-source (WS) variability of the suspected speaker (scores from suspect speech segments when tested with the suspect model), and the between-source (BS) variability of the potential population of relevant speakers (testing the questioned speech segment with the models from the potential population), given the questioned recording. Formally, $f\,(x\,|H_0\,)$ and $f\,(x\,|H_1\,)$ in Equation 5 represent the cited within-source variability and between-source variability. The interpretation of the evidence consists of calculating the likelihood ratio using the probability density functions (pdfs) of these distributions and the numerical value of evidence (Figure 3) [2].

The information provided by the analysis of the questioned recording (trace) leads to a specification of the initial reference population of relevant speakers (potential population) having voices similar to the trace, and, combined with the police investigation, to focus on and select a suspected speaker. Generally, the methodology used in this paper needs three databases for the calculation and the interpretation of the evidence: the potential population database (P), the suspected speaker reference database (R) and the suspected speaker

[2] Pooling of speakers or different traces from the same speaker is not possible in the computation of the likelihood ratio, as a specific between-source pdf is necessary for every single trace (questioned speech segment). In assessment experiments with many simulated suspects and $H_0/H_1$ hypothesis, it is observed that the $LR$ computation sustains a non-monotonic transformation from the original scores to the set of likelihood ratios for the same suspect-questioned speech pairs. As a result, if all likelihood ratios for the $H_0/H_1$ hypothesis are pooled (the final likelihood ratios, not the scores to compute every $LR$), different pdfs and then different DET plots will be obtained for scores and $LRs$ in the same assessment experiment.

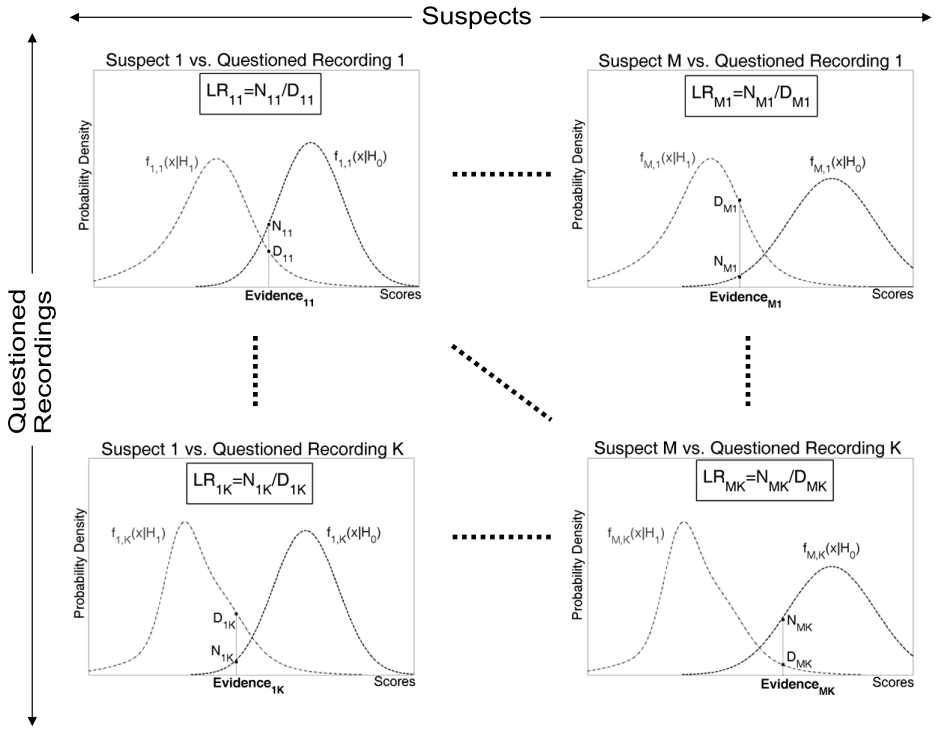



Fig. 3. Likelihood Ratio Computations from K questioned recordings and M suspect models. For simplicity in representation, computations are performed over all possible Questioned Recording-Suspect trials.

control database (C).

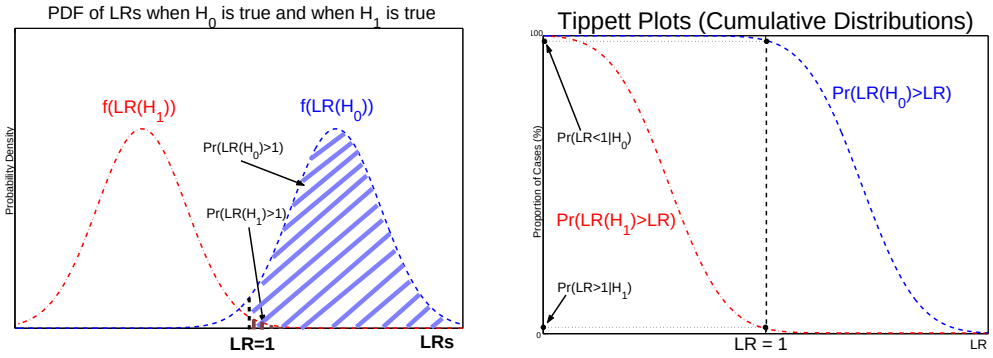
The potential population database (P) is used for modelling the variability of the speech of all the potential relevant sources, using the automatic speaker recognition method. P typically consists of a set of speaker models from a population of non-suspected individuals. In this sense, this population is constructed in the same way as the test-normalization (T-Norm) cohorts used in speaker verification [3]. It allows the evaluation of the between-sources variability given the questioned recording, which means the distribution of the similarity scores that can be obtained when the questioned recording is compared to the speaker models (GMMs) of the potential population database.

The calculated between-sources variability pdf $f(x\,|H_1)$ is then used to estimate the denominator of the likelihood ratio $f(E\,|H_1)$. There are no constraints about the method used in this estimation. In this contribution we have used maximum-likelihood estimation of a GMM distribution. Ideally, the technical characteristics of the recordings (e.g. signal acquisition and transmission) should be chosen according to the characteristics analyzed in the trace.

The suspected speaker reference database (R) is necessary to model the suspect's speech with the automatic speaker recognition method and can be recorded from the suspect speaker or obtained by other means. When recorded, speech utterances should be produced in a similar way to those of the potential population (P) database. The suspected speaker model obtained is used to calculate the value of the evidence, by comparing the questioned recording to the model. The suspected speaker control database (C) is necessary to evaluate the within-source variability, when the utterances of this database, which also can be recorded or obtained from available suspect speech, are compared to the suspected speaker model (GMM). C database consists of speech material coming from the suspect himself, and its utterances are compared with suspect speech material too, coming from the R database. Thus, we obtain target scores from the suspect himself, which are used to estimate a pdf modelling his within-source variability. This calculated within-source variability pdf $f(x\,|H_0)$ is then used to estimate the numerator of the likelihood ratio $f(E\,|H_0)$. Again, the method used in this estimation is not constrained. In this contribution we have used maximum-likelihood estimation of a single-Gaussian distribution. The elements of the C database should constitute utterances that are as equivalent as possible to the trace, according to the technical characteristics as well as to the quantity and style of speech.

Of great interest to the jurists is the extent to which the $LR$s correctly discriminate "same speaker and different-speaker" pairs under operating conditions similar to the case in hand. As was made clear in the US Supreme Court decision in the Daubert case (Daubert v. Merrell Dow Pharmaceuticals, 1993) it should be criterial for the admissibility of scientific evidence to know to what extent the method can be, and has been, tested. The principle for evaluation of the strength of evidence used in this paper consists in the estimation and the comparison of the likelihood ratios that can be obtained from the evidence $E$. On the one hand the hypothesis $H_0$ (the suspected speaker truly is the source of the questioned recording) is assumed and, on the other hand, the hypothesis $H_1$ (another speaker within a relevant population of individuals is the source of the questioned recording) is considered. The performance and reliability of an automatic speaker recognition method is evaluated by repeating the experiment described in the previous sections, with several speakers being at the origin of the questioned recording, and by representing the results using experimental (histogram based) probability distribution plots such as probability density functions and Tippett plots (Figure 4).

The representation of the results in Fig. 4 is the one proposed by Evett and Buckleton in the field of interpretation of forensic DNA analysis [9]. The authors have named this representation as "Tippett plots", referring to the concepts of "within-source comparison" and "between-sources comparison" defined by Tippett et al. These probability distributions are also known as reliability functions.




(a) (b)

Fig. 4. (a) PDF of $LRs$ in $H_0$ and $H_1$ hypothesis, and (b) Tippett plots (cumulative density functions of $LRs$ matching each of the two competing hypotheses).

### *2.5 Speaker verification and Speaker Forensic Analysis*

In this section we want to clarify the relations between speaker verification, associated to the use of thresholds, and forensic speaker analysis. In the latter case, the objective of the scientist is to help the Court in his decision process. However, speaker verification is the task of deciding, given a questioned recording, whether the suspected speaker is the source of this recording. This is a 2-class task. The two classes in forensic speaker verification are: $H_0$ - the suspected speaker is the source of the questioned recording, and $H_1$ - another person is the source of the questioned recording. The output of a speaker verification system is always a binary decision, independent of the technology used in the system. The implementation of the system will determine how this decision is generated. One way of generating the decision output can be accomplished by using Bayes' decision theory. The decision rule for the two-category classification problem can be determined by comparing the score obtained from a speaker recognition system (see Equation 2) with a decision threshold:

$$\text{chosen class} = \begin{cases} H_0 \text{ if } score \geq \theta_{score} \\ \\ H_1 \text{ if } score < \theta_{score} \end{cases}$$

where $\theta_{score}$ is the threshold.

The main drawback of the use of speaker verification in a forensic context is due to the subjective selection of the threshold by the forensic expert. The score is compared to the threshold and the suspected speaker is accepted or rejected as the source of the questioned recording. This interpretation of the evidence does not correspond to the concept of forensic individualization as it is widely accepted. Reporting a decision forces the scientist to ignore the prior probabilities of the case (see Equation 3), usurping the role of the Court in taking this decision. Even if these thresholds are computed from objective data, "*...the threshold is in essence a qualification of the acceptable level of reasonable doubt adopted by the expert [...]. Therefore, speaker verification is clearly inadequate for forensic purposes, because it forces the scientist to make decisions which are devolved upon the court*" [6].

Even if the $LR$ can be used for binary decisions (because a $LR$ can be used as a detector score), its main importance relies in its allowance to infer the posterior from the prior when $H_0$ and $H_1$ are being considered. This last property is the basis of the use of Bayesian inference applied to legal reasoning.

Forensic speaker recognition in real-world case-work presents some peculiarities that make errors appear in the estimation of $LRs$ if special techniques are not adopted to avoid them. In this section, we present some of these problems, showing their origin and consequences for $LR$ estimation.

Generally, in forensic conditions the quality and quantity of the speech data the forensic expert can handle is far from optimal. This specific environment usually causes strong mismatches between the questioned, suspect's and relevant population speech and lack of data for accurate distribution estimations [2][4]. Moreover, in forensic applications it is convenient, in order to guarantee the presumption of innocence, that suspected non-perpetrators speakers from the relevant population do not obtain $LR$ values greater than one, even if it leads to worse discrimination between suspected perpetrators (targets) and suspected non-perpetrators (non-targets) speakers.

The problem of between-source variability estimation is related to the selection and number of available models of the relevant population. Given that between-source variability distribution represents the random match probability distribution of the evidence within the relevant population, it should present the same characteristics as the suspect's speech data regarding transmission channel, language, etc. Hence, there are many problems to be solved when such a "matched" reference population is not available [12].

Another important source of problems comes from inadequate estimation of within-source variability distribution of the suspected speaker. It is possible in forensic investigations dealing with voice that the forensic expert has only

a single questioned recording and a single recording of a suspect. The task is normally to evaluate whether the voice in both these recordings comes from the same person. As a consequence, it is not always possible to evaluate the within-source variability of the suspect with this single recording. However, since this is a recurrent problem in forensic speaker recognition, it is necessary to define an interpretation framework for evaluating the evidence even in the absence of additional control recordings [4][12].

## 3 Algorithms for Robust LR Computation

In this section a number of original contributions will be presented for obtaining reliable *LRs* in the one-questioned-recording one-suspect-recording condition (as in the detection tasks in NIST SRE and NFI-TNO Evals). These techniques are also robust when limited or more extensive quantities of data are available.

### *3.1 Robustness in Forensic Speaker Recognition*

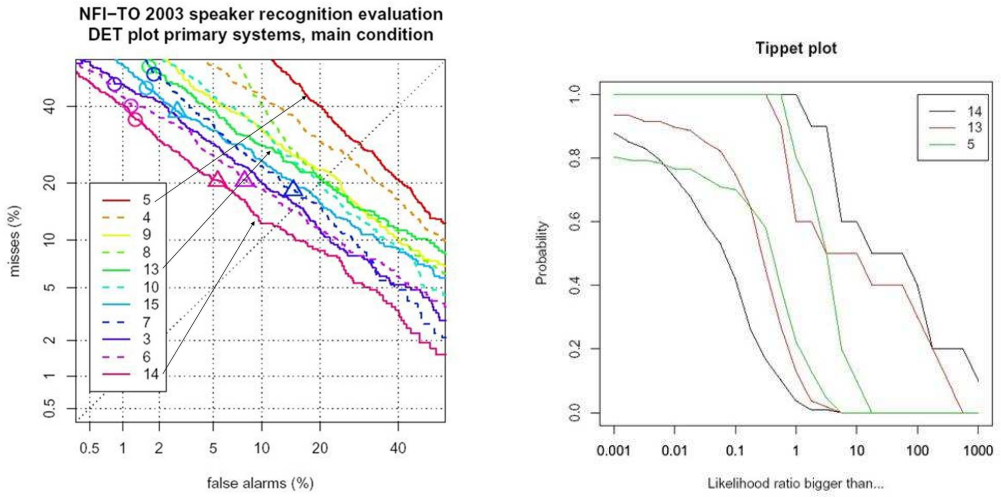
In this paper, *robustness* is understood as the capabilities of the forensic speaker recognition system to perform *reliable* likelihood ratio estimations in the described forensic environment, where severe mismatched conditions and limited suspect and trace data are usually present. Then, it is useful to compare the performance of the forensic system with the corresponding score-based system used to compute the *LR*. This is possible by means of plotting the *LRs* in a DET plot [16] as if they were used as detection scores, and comparing the result with the DET plot of the score-based system. In doing

so, we will observe that a large proportion of the discrimination ability of the system is lost due to the errors in $LR$ estimation described in section 3. Consequently, new techniques must be developed to compute more reliable likelihood ratios for forensic systems to be robust. The following example will aid in this discussion.

### *3.1.1 LR Computation in NFI-TNO 2003 Evaluation*

The NFI-TNO 2003 Forensic Speaker Recognition Evaluation is described in detail in section 4.1.3 (details on [15]). For the moment, a detailed explanation is not necessary. We only need to know that, in addition to the main evaluation task (detection), which is equivalent to the NIST yearly SREs but with new authentic Dutch forensic data, a new experiment was performed. In this experiment, namely experiment 6 ("Court Proof") in the NFI-TNO Eval, likelihood ratio computation was performed over the submitted score-based systems presented (Figure 5(a)), simulating a real forensic scenario. Likelihood Ratios were computed at NFI-TNO and Tippett plots were drawn to assess Forensic Speaker Recognition performance. These results are shown in Figure 5(b), showing the best and the worst score-based system, and one intermediate system.

The "Court Proof" task in the NFI-TNO Eval is then extremely interesting showing that any automatic speaker recognition system can be adapted to compute likelihood ratios. However, in this task the need for robustness in $LR$ estimation has been replaced with a large amount of suspect training data showing large speaker variability. Sometimes such a large number of (acknowledged) suspect recordings are impossible to be recovered in a forensic




(a) (b)

Fig. 5. (a) NFI-TNO 2003 Evaluation main task results and (b) Court Proof results. Source: [15].

scenario. The number of available (and "certified") suspect recordings/calls can be much smaller in some forensic field cases, and in the limit case we will have just one questioned recording and one suspect recording.

Therefore, forensic speaker recognition systems performance is good when there is a sufficient quantity, quality and variety of suspect speech material. In this case, within-source variability can be properly estimated having high generalization capabilities. However, in a general case, we will need specific robust estimation algorithms to cope with forensic scarce data, as shown in next sections.

### *3.2 Bootstrapping Suspect Data for Within-Source Distribution Estimation*

When there is no possible match between questioned speech and suspect data conditions, it is desirable to have as many scores as possible with high variability to estimate the within-source pdf. A leave-one-out procedure is described

here, 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 considered: monosession and multisession suspect data availability. In the former, a single suspect recording is available, being then divided into $N$ uniform length segments (after silence removal); in the latter, $N$ different recordings are available. In both cases $N$ different suspect temporary models are obtained for every speaker from $N-1$ segments/utterances each, obtaining $N$ similarity scores which will be used for within-source pdf estimation. Once the $N$ scores are obtained, the temporary models are deleted and the suspect model is obtained from the $N$ segments/utterances available. In this sense, in each resampling step of the algorithm, the left-out utterance will be the C database, and the remaining recordings will constitute the R database. Those $N$ scores will be an acceptable model of session variability in the multisession case (multiple suspect recordings), but a very optimistic one in the monosession case (single suspect recording), as test speech is obtained from the same utterance as the model, giving a highly biased within-source estimation. This estimate is void of channel and intersession variations and the variance will typically be underestimated. In the next sub-section, those $N$ scores will be considered in a different way depending on whether they come from monosession or multisession data.

### *3.3 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 $LR$

values (unexpected high $LR$ values for non-targets and low $LR$ values for targets) even when a classical detection system would perform correctly. Furthermore, it is widely known [19] that mismatched conditions between questioned speech and suspect recordings may imply variations in the score distribution ranges. So, when there are no speech controls (C database) that could match the questioned speech conditions in a reliable fashion (as happens in many real forensic speaker recognition applications), within-source modelling using non-matching suspect data can give a poor generalization performance, with the consequence that unreliable likelihood ratios may be obtained. Therefore, within-source variability should model all the possible situations in which the questioned speech can appear (regarding noise, time variability, speaker mood, communication channel, language, etc). Within-source Degradation Prediction (WDP) is proposed [12] to stand for the unknown degradation expected from unknown mismatch, assuming no scores from impostors are expected higher than those in the between-source estimation.

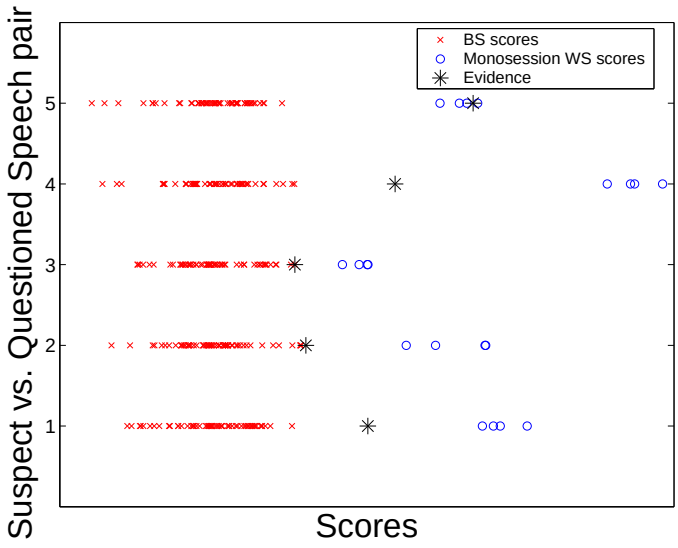



Fig. 6. Plot of between-source (BS), monosession within-source (WS) and Evidence scores in five sample cases extracted from the NIST SRE 2004 ATVS results.

Figure 6 is a very useful example for understanding the WDP motivation. We can see five single suspect-single questioned speech trials involving suspected perpetrators extracted from ATVS NIST SRE 2004 system scores. The plot shows the scores of questioned speech vs. models in the population (between-source scores), the monosession within-source scores obtained with the bootstrap procedure described above and the corresponding evidences for each suspect-questioned recording pair. The figure shows that as monosession WS scores represent the maximum matching between possible speech excerpts coming from the same suspect, evidence scores are in almost all cases below these values, but in any case within these limits (BS and monosession WS). WDP is then used differently in the two following cases:

- Monosession data: a single suspect recording is available. Scores used to estimate the within-source model are obtained via leave-one-out from a single suspect recording as described in previous section. In this case, these scores determine the maximum value that a target evidence could obtain, because they are obtained in the best matching conditions.
- Multisession data: several recordings are available. Scores used to estimate the within-source model represent the known session variability in the suspect data. However, recordings presenting mismatched conditions from the available suspect data can give different scores.

Formally, let $f\left(x|\, H_0\right) = N\left(\mu_{WS}, \sigma_{WS}\right)$ and $f\left(x|\, H_1\right)$ the pdfs for the within- and between-source distribution for a given forensic trial. The WS probability density function is assumed to be Gaussian. The objective mapped pdf after WDP is defined as $f_{WDP}\left(x|\, H_0\right) = N\left(\mu_{WDP}, \sigma_{WDP}\right)$. Our goal is to compute the desired parameters $\mu_{WDP}$ and $\sigma_{WDP}$. First of all, we compute $s_{low}$ , which will be the score that satisfy

$$\int_{s_{low}}^{\infty} f\left(x \middle| H_1\right) dx = \alpha \tag{6}$$

where $\alpha$ is a design value set to 0.01 in this contribution. $s_{low}$ is called the *lower bound* for $f_{WDP}\left(x \middle| H_0\right)$. The computation of $\mu_{WDP}$ depends on the suspect data for the given trial. For the monosession case

$$\mu_{WDP} = \frac{\mu_{WS} + s_{low}}{2} \tag{7}$$

and for the multisession case

$$\mu_{WDP} = \mu_{WS} \tag{8}$$

Once $\mu_{WDP}$ is computed, a descent algorithm for unconstrained optimization is used to compute $\sigma_{WDP}$ (both in monosession and multisession cases), given that it is claimed to satisfy

$$\int_{-\infty}^{s_{low}} f_{WDP}\left(x \middle| H_0\right) dx = \alpha \tag{9}$$

The effects of $WDP$ on the WS pdf for both monosession and multisession data are shown in Figure 7.

### *3.4 WMVL and Outlier Removal*

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

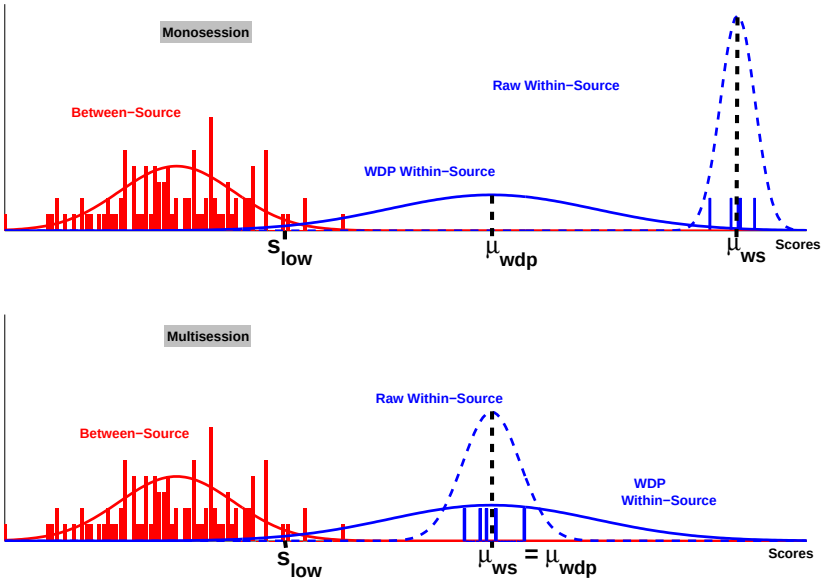



Fig. 7. Within-source Degradation Prediction (WDP).

- Within-source Minimum Variance Limiting (WMVL): even using WDP, sometimes very low estimated variances are still present due to highly coherent speech controls. WMVL simply limits the minimum variance of within-source estimations, avoiding possible erratic $LR$ values.
- Outlier removal: within-source model is usually estimated from very few (less than five) scores. As a result, singularities in one or two speech controls (very limited quantity of speech, laughing, noise peaks, etc.) can lead to inconsistent mean estimation. All scores falling under a defined, very low, between-source-pdf-dependent value are discarded as outliers. Let $s_{outlier}$ be this value. $s_{outlier}$ must satisfy:

$$\int_{-\infty}^{s_{outlier}} f\left(x|\, H_1\right) dx = \gamma \tag{10}$$

where $\gamma$ is a fixed system-defined value set to 0.25 in this contribution.

### *3.5 Target Dependent LR Alignment (TDLRA)*

In the previous section, different techniques have been proposed for obtaining much greater discriminant capability, but presumption of innocence is still not guaranteed, as a slight proportion of non-targets still obtain $LR$ scores greater than one. Recently, we have proposed the use of target dependent score alignment (TDSA) [11] for signature verification. This score normalization technique exploits user-dependent information to improve system performance in a cost detection sense. In the present work, the objective will not be based on a cost detection function to be optimized per speaker but in guaranteeing the presumption of innocence for suspected non-perpetrators speakers, which will be performed on a speaker by speaker basis. This objective will be satisfied by minimizing the number of suspected non-perpetrators speakers obtaining $LR > 1$.

The use of TDLRA in a forensic trial can be described as follows. Let $X_{LR-NT} = \{LR_{NT1}, ..., LR_{NTL}\}$ be the non-target $LRs$ obtained from the development set of non-suspected utterances and the given suspect model. First we estimate the $LR$ pdf for this non-target trials via maximum-likelihood assuming Gaussian distributions [3]. We denote this pdf as $f\left(x|\, X_{LR-NT}\right)$ [4]. Let $\beta$ be the desired proportion of suspected non-perpetrators with $LR > 1$. $\beta$ will be usually in the $1\% - 5\%$ range, and will be an a-priori defined value of the system.

[3] TDLRA models the non-target $LR$ distribution in a similar way as ZNorm models the impostor score distribution for a given target model [11].

[4] As this "impostor" $LR$ distribution is modelled with a single Gaussian, a finite error proportion will always be above $LR$=1, but the smaller this error proportion, the bigger will be the percentage of the targets with $LR$ below one.

We define $LR_{NT-\beta}$ as the $LR$ value that satisfies

$$\int_{LR_{NT-\beta}}^{\infty} f\left(x|\, X_{LR-NT}\right) dx = \beta \tag{11}$$

TDLRA technique is then applied to the $LR$ for the suspected speaker in the following way:

$$LR_{TDLRA} = LR - LR_{NT-\beta} \tag{12}$$

Figure 8 illustrates this technique. It is important to note that TDLRA is applied in operational conditions in an automatic mode.

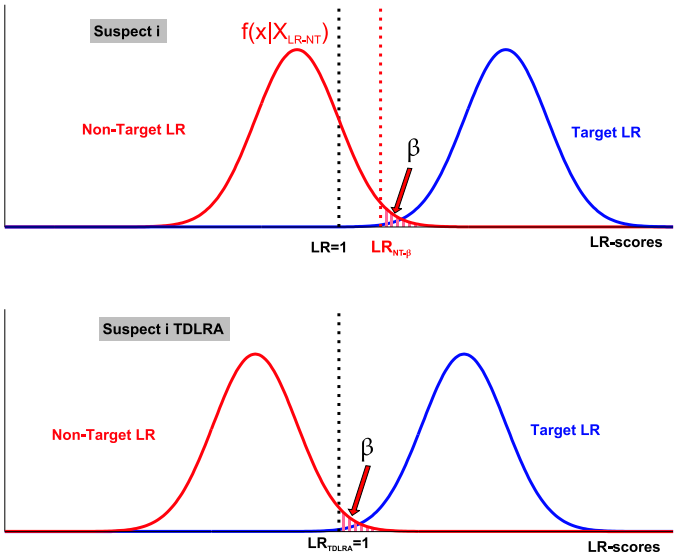



Fig. 8. Target Dependent Likelihood Ratio Alignment (TDLRA).

TDLRA also has a positive effect on the performance of the system when conditions are mismatched between the relevant population models and the suspect model. Since such a mismatch produces a misalignment in $LRs$ obtained for each suspect, normalization between suspects performed by TDLRA leads to a better global system behaviour.

# 4 Experiments

In order to assess the robustness of the proposed $LR$ computation algorithms described, a wide range of experiments is presented. They are based on trials constructed with extensively used public databases incorporating blind results and post-tuning experiments of the ATVS Bayesian forensic recognition system in international evaluations like the NFI-TNO Evaluation 2003 and NIST SRE 2004.

## *4.1 Databases and Experimental Framework*

Results based on experiments using three databases under very different and challenging conditions are presented. Below is a detailed description for each experimental framework used.

### *4.1.1 Switchboard I Experiments*

This framework was constructed using a 100 male speaker subset from Switch-Board I, which consists of landline telephone (different handsets) 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 a single conversation.
- Multiple-session training (multisession): 50 speakers (different from above), 30 seconds of speech from each call are obtained in each of 4 conversations.

A variable number of 30 second excerpts per suspected speaker was used as questioned speech. No manual silence detection was used either in training or test, and severe co-channel interference is present.

### *4.1.2 NIST SRE 2004*

Since 1996, the NIST yearly Speaker Recognition Evaluations (SRE) [18] have fostered research and development in speaker recognition. Each year, the corpus and the methodology try to focus on solving the main problems with state of the art technology. ATVS has participated in these evaluations since 2001, with its main focus in 2004 being to assess the robustness of the algorithms used in likelihood ratio computation.

NIST SRE 2004 has introduced many new challenges to the scientific community. First, a new database has been used, the MIXER corpus, which contains data recorded across different communication channels (landline, GSM, CDMA, etc.), using different handsets and microphones (carbon button, electret, earphones, cordless, etc.) and different languages (American English, Arabic, Spanish, Mandarin, etc.). A new *Fisher-Style* protocol for recording acquisition has improved randomness in the spontaneous component of the conversation. Moreover, the evaluation methodology is also new, giving the possibility of presenting a system to any task involving one training condition (10 seconds, 30 seconds, 1, 3, 8 and 16 conversation sides and 3 full conversations) and one test condition (10 seconds, 30 seconds, 1 conversation side and 3 full conversations). Each conversation side has an average duration of 5 minutes, having 2.5 minutes aprox. after silence removal. Although there were both genders in the corpus, no cross-gender trials were performed.

The mandatory task for any participant in NIST SRE 2004 was the *core* condition. It constituted trials having one conversation side for training and one conversation side for testing (1side-1side). The core task of the evaluation had more than 25000 trials. The (*Common Evaluation Condition*) was a subset of the core condition, having some special restrictions (all trials were in English, for example).

In the experimental framework described, strong mismatch between train and test data may appear, and careful selection of the parameters and components of the system (development data, population database, normalization sets, etc.) had to be made. Moreover, MIXER was a corpus never used before and no development data was provided to participants.

### *4.1.3 NFI-TNO Forensic SRE 2003*

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 [15] 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 performed to obtain field data and to emulate as close as possible 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 reverberation, music in the background of the recording and in some cases ambient speakers (mostly sounds of children playing). Although the speaking style was constant (speech containing 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 (spoken at the Netherlands Antilles).

### *4.2 Results*

#### *4.2.1 Switchboard I Results*

In order to test the proposed robust $LR$ estimation algorithms, a reference system is needed to provide raw scores to be used later in the Bayesian system. Then, the ATVS UBM-MAP-adapted GMM system as submitted to NIST'02 SRE was used, with a UBM trained from 5 hours of male speaker data from a different Switchboard-I (SWBI) partition. Three experiments will be reported in this section: 'monosession' (50 speakers), 'multisession' (50 speakers), and 'all' (100 speakers). Performance of this reference system in these three experiments is shown in Figure 9.

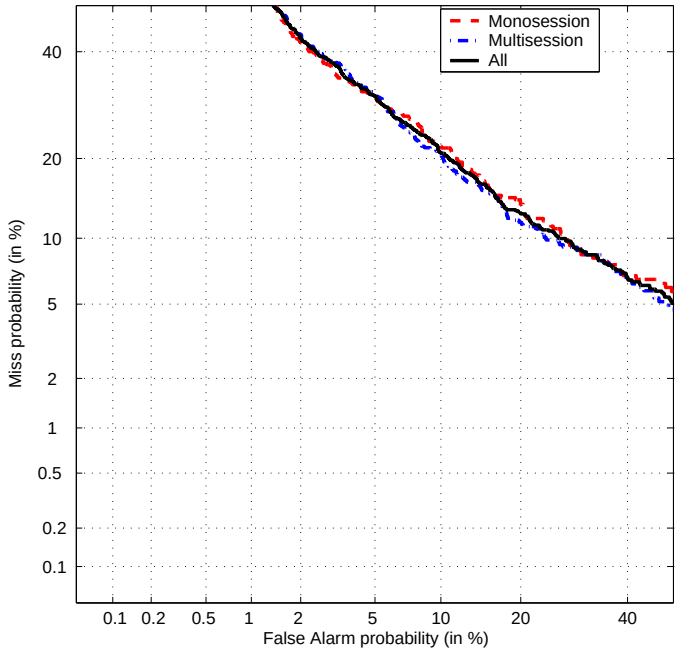
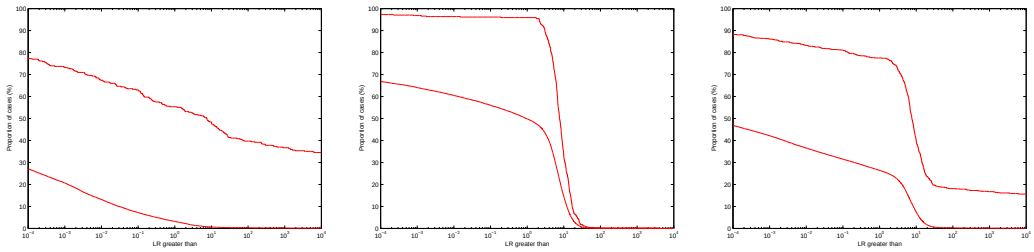



Fig. 9. ATVS NIST SRE 2002 system performance with Switchboard I data. The reference raw scores are the basis for all algorithms in section 4.

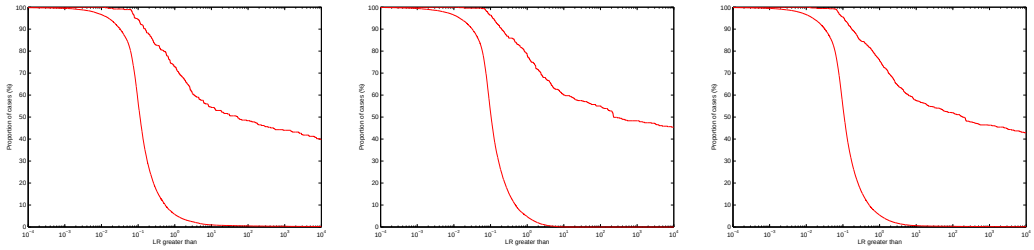
First we present $LR$ computations in the SWBI framework using bootstrap for obtaining within-source scores, but no other algorithm from those proposed. Figure 10 shows the performance of our Bayesian forensic speaker recognition system when ML single-Gaussian within-source estimation is performed directly from within-source modelling data using the leave-one-out method (N=4 for monosession experiments). Between-source estimation via 32 Gaussian EM-ML estimation was obtained by means of comparing questioned speech with the relevant population P (50 multisession 2 minutes-trained speakers).

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



(a) (b) (c)

Fig. 10. *Tippett Plots with Raw LR. (a) 'Monosession', (b) 'multisession' and (c) 'all' suspect Switchboard data.*

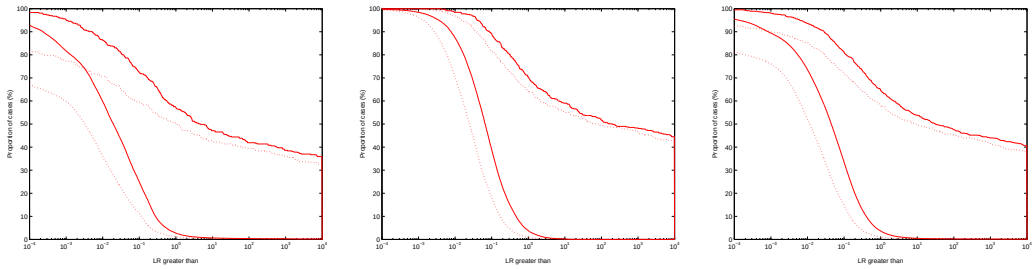


(a) (b) (c)

Fig. 11. Tippett Plots using WDP, WMVL and Outlier removal. (a) 'Monosession', (b) 'multisession' and (c) 'all' suspect Switchboard data.

In Figure 11 the same three experiments as in Figure 10 are reported, where the basic *LR*-based system has been significantly improved with the joint use of the proposed WDP, WMVL and outlier removal both for targets and non-targets in any of the tested conditions, showing an excellent, but still quite loose, 'presumption of innocence' performance.

Figure 12 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 the prosecution hypothesis. From these experiments, the degree of control over system performance with TDLRA technique is remarkable, especially the control over the presumption of innocence for non-targets in the desired values.



(a) (b) (c)

Fig. 12. Tippett plots for same system of Figure 11 adding TDLRA. Two configurations for TDLRA are shown, with maximum Non-Target errors of 1% (dotted) and 3% (solid).

### *4.2.2 NIST 2004 Evaluation Results and Experiments*

During the last two years ATVS has focused its research on robust $LR$ computation from raw scores, mainly starting from the core technology we used in the NIST SRE 2002. In order to participate in NIST SRE 2004, ATVS has tested some common channel normalization options such as RASTA filtering and Feature Warping.

In the NIST SRE 2004, ATVS's main objective was to assess the proposed robust $LR$ computation procedures and compare them with the score-based system performance of the rest of the participants. Therefore, ATVS submitted several systems to different evaluation conditions, as is shown in Table 1.

As it can be seen in the figure, ATVS submitted 5 systems to 3 different tasks. Systems 2 and 3 (score-based) used different parameter extractors, and system 1 (primary) was the fusion of both of them. Systems 4 and 5 (forensic) used scores computed with system 2 to perform robust $LR$ computation by means of WDP, WMVL and Outlier Removal. In System 5 TDLRA was also included.

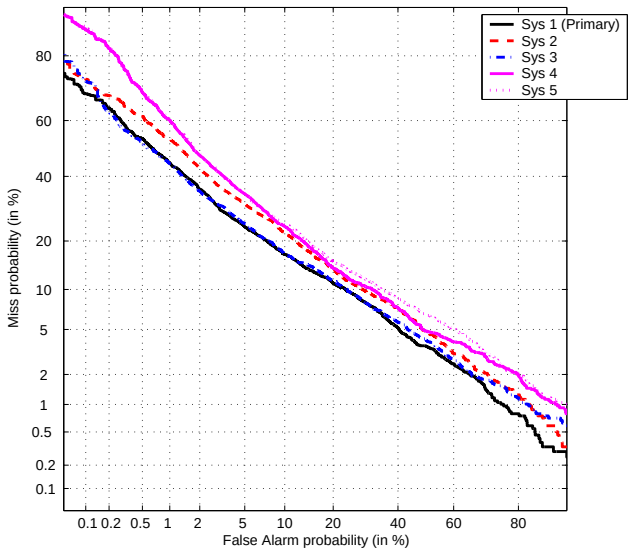
Table 1

ATVS systems in NIST 2004 SRE.

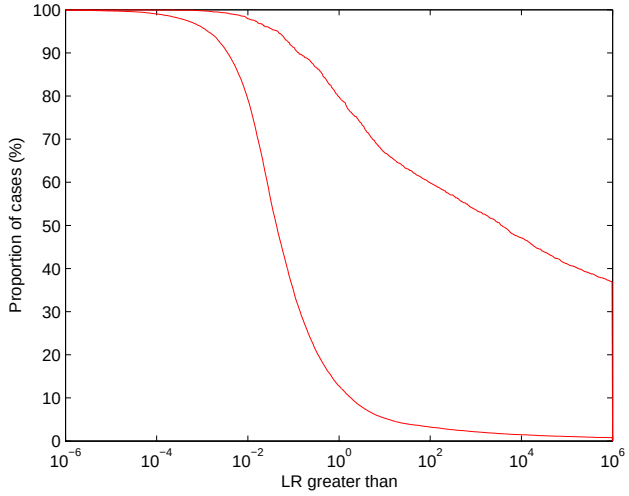
| Train: 1 Side | Test | | |
|---|---|---|---|
| | 1 Side | 30 Seconds | 10 Seconds |
| $Sys1 = Sys2 + Sys3$ | $X$ | | |
| $Sys2$ | $X$ | $X$ | $X$ |
| $Sys3$ | $X$ | | |
| $Sys4$ (Forensic LR) | $X$ | | |
| $Sys5$ (Forensic LR) | $X$ | | |

Figure 13(a) shows ATVS systems performance in NIST SRE 2004. It is interesting to note the difference between systems 2 and 4, and to analyze the Tippett plots resulting from $LR$ computation performed by system 4, as shown in Figure 13(b).

Although performance is good in all systems, we can see two undesired effects in both these plots. First, system 4 (forensic) presents a worse performance in low False Acceptance rate than its equivalent score-based, system 2. Second, Tippett plots from system 4 show a sub-optimal behavior in non-targets, even though separation between curves is acceptable. This effect is justified by the selection of the relevant population used. As illustrated in [12], and as will be presented later, a mismatch between target model and population cohort models in $LR$ computation gives a strong degradation in performance of trials involving suspected non-perpetrators. ATVS NIST SRE 2004 systems used two gender-dependent populations, but each one containing models only




(a)




(b)

Fig. 13. ATVS systems in NIST 2004 SRE. (a) Comparison of systems 1 to 5 (sorted) and (b) Tippett plots of system 4.

in English and belonging to different channels (landline carbon and electret handsets, CDMA and GSM). Therefore, the reference population in use was not adapted to each suspect model and this mismatch in channel and language between population and each target model may have caused the described degradation. In addition, Figure 13 reveals that TDLRA has produced no improvement in system performance. The reason is again related to matching.

TDLRA impostor sets used in NIST SRE 2004 were gender-dependent, but again they contained different channel and language utterances. Therefore, the mismatch once more biases the algorithm's impostor modelling, giving a poor estimate and therefore reducing its beneficial effects.

### *4.2.3 Results and Experiments in NFI-TNO 2003 Evaluation*

Adding to the complexity of the forensic field data available in NFI-TNO Evaluation 2003, the rules of the evaluation did not provided any development data nor allowed the use of Dutch data to optimize or adapt the submitted systems to the evaluation conditions. Several experimental configurations were proposed:

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

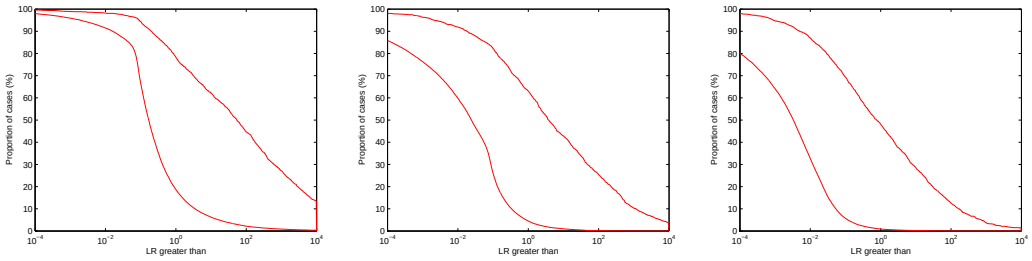
ATVS submitted two systems, a "raw" score NIST-eval type system (ATVS-1, primary) and a Bayesian $LR$-based forensic system (ATVS-2)[12]. In both

cases, a GSM-coded version of Switchboard I Extended Data Database was used both for background modelling (UBM) and the reference population for $LR$ computation.

However, we want to focus on our secondary system, ATVS-2, the Bayesian forensic one, fully compliant with the Bayesian framework for the evaluation of evidence. This system was able to compute robust likelihood ratios for every single test file of the evaluation when compared to any single suspect utterance. Note that no extra information is needed in our system even when having speech controls is a theoretical requirement for $LR$ computation. In other words, speech controls are directly obtained in our system from suspect 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 in that case there is no requirement for robust $LR$ estimation. 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, obtaining a meaningful likelihood ratio ($LR$) with every test-file/suspect-recording pair.

After the submission deadline and once the keys were distributed to 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 improves system performance. Moreover, once matched data (in very general terms) became available, we also ran the evaluation with TDLRA, the robust $LR$ estimation technique described above which maximally preserves the presumption of innocence of

suspects, preventing suspected non-perpetrators obtaining $LR$ scores greater than one.




(a) (b) (c)

Fig. 14. ATVS forensic systems at NFI/TNO eval'03. (a) Submitted (ATVS-2a), (b) same system with Dutch reference population (ATVS-2b) and (c) TDLRA system with Dutch reference population. (ATVS-2c)

The results of the submitted forensic system (ATVS-2a) and the post-eval systems using the Dutch population (ATVS-2b) and TDLRA with the Dutch population (ATVS-2c) are shown in figures 14 and 15, respectively in the form of Tippett plots and DET curves. Remarkable results have been obtained with the forensic systems from two points of view: firstly, a meaningful $LR$ value 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. And secondly, presumption of innocence is strongly preserved when TDLRA is applied, where a very small portion of non-targets obtain $LR$ values greater than one, and about 50% of targets do obtain $LR$ values greater than one. Note that this is an 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 obtain $LRs$ smaller than one, presumption of innocence is preserved in almost 100% of all cases.

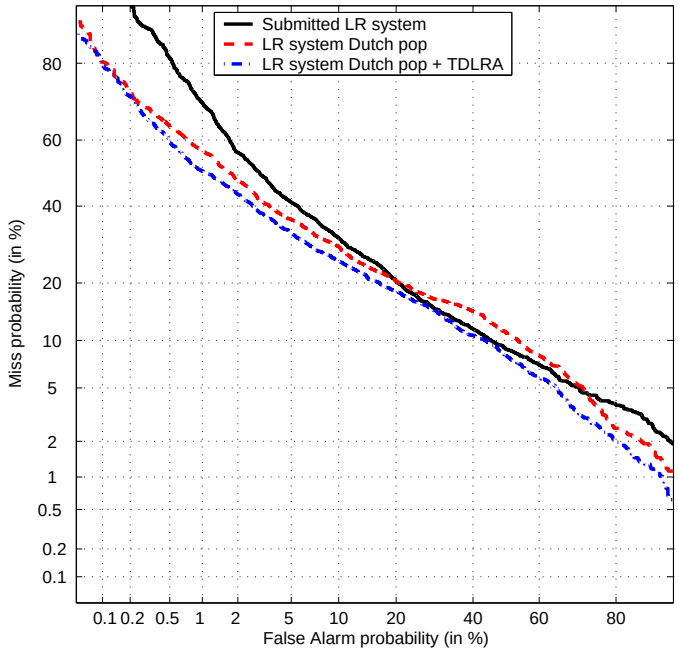
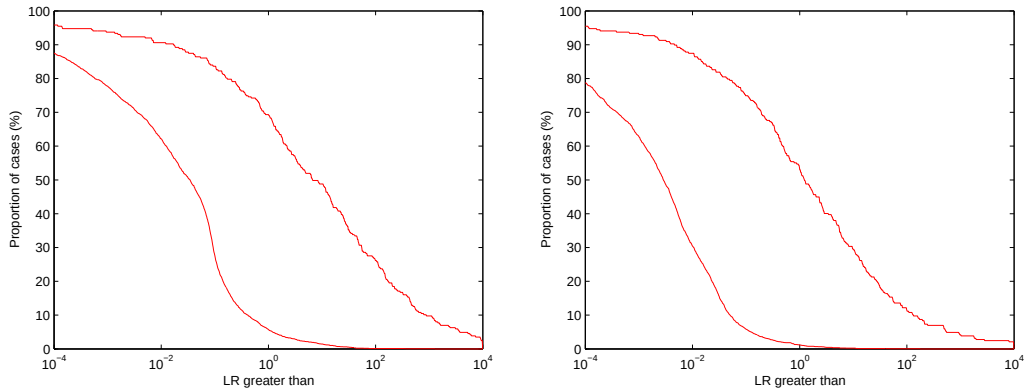



Fig. 15. ATVS forensic systems at NFI/TNO eval'03. Submitted (ATVS-2a), same system with Dutch population (ATVS-2b) and TDLRA system with Dutch pop. (ATVS-2c)

Figure 16 respectively show the results of ATVS-2b (submitted system with Dutch population) and ATVS-2c (TDLRA system with Dutch population), in evaluation condition 6 ("Court Proof"), showing excellent performance of TDLRA in this task.

After the improvements introduced in the ATVS score-based system during NIST SRE 2004 preparation, our new system was tested again with the NFI Eval primary condition data. In Figure 17, the performance of our $LR$-based forensic system is shown in terms of a Tippett plot for the NFI Eval primary data. Results show a much better performance even than those shown above in Figure 14 as better raw scores were available.




(a) (b)

Fig. 16. NFI/TNO eval'03 "Court Proof" condition. (a) Dutch population (ATVS-2b) and (b) TDLRA (ATVS-2c).

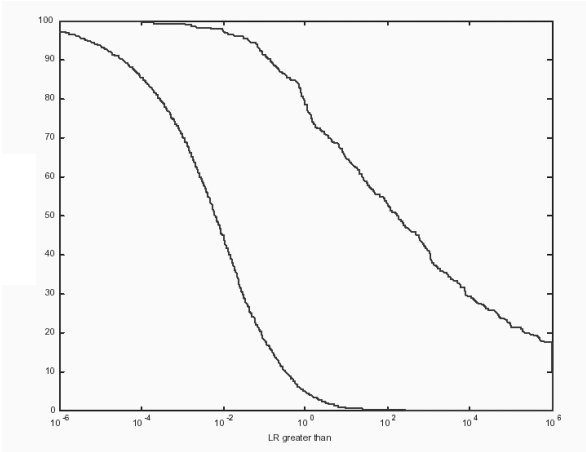



Fig. 17. Tippett plot of the ATVS $LR$ system obtained from warping+TNorm scores in the NFI Eval primary condition (Post-Evaluation experiments).

## 5 Conclusions

The Bayesian framework for the interpretation of evidence when applied to forensic speaker recognition has been introduced in this contribution. Different aspects of the use of voice as evidence in Court, as well as the use from the forensic expert of the likelihood ratio as the right way to express the strength

of the evidence have been reported. Details on computation procedures of likelihood ratios have been given, along with the assessment tools and methods to validate the performance of those Bayesian forensic systems. State-of-the-art speaker recognition systems are able to compute likelihood ratios when enough suspect data is available as has been proved in the "Court Proof" task of NFI-TNO 2003 Forensic Speaker Recognition Evaluation. However, due to the practical scarcity of suspect data and the mismatched conditions between traces and reference populations in daily casework, significant errors appear in $LR$ estimation if specific robust techniques are not used. Some original contributions have been proposed for obtaining robust likelihood ratios under real forensic conditions. The need for robust algorithms in this estimation process has been discussed, and its use justified, especially in cases where there is a considerable lack of data. We have proposed a set of algorithms, namely WDP, WMVL, WS Outlier Removal and TDLRA, that not only allow forensic systems to estimate robust likelihood ratios, but also make it possible for them to work in extreme circumstances, as happens when there is a single suspect recording and a single questioned call. The proposed algorithms have been assessed both with Switchboard landline telephone data, NFI-TNO GSM forensic field data and NIST SRE 2004 multichannel, multilanguage data. Special mention must be made of TDLRA (Target Dependent Likelihood Ratio Alignment), a novel algorithm that preserves the presumption of innocence for suspected but non-perpetrators speakers in all the reported Switchboard experiments and NFI-TNO Evaluation.

**Acknowledgements**

The authors wish to thank the members of the speech lab of Guardia Civil for fostering and supporting this research and for the continuous provision of new challenging field cases. This work was in part supported by the Spanish Ministry for Science and Technology under projects TIC2003-09068-C02-01 and TIC2003-08382-C05-01.
D. R.-C. also thanks Consejeria de Educacion de la Comunidad de Madrid and Fondo Social Europeo for supporting his doctoral research.

The authors wish to thank all five reviewers of this contribution for their extensive work, which has significantly improved the quality of the final manuscript.

## References

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

[2] A. Alexander, F. Botti, A. Drygajlo, "Handling Mismatch in Corpus-Based Forensic Speaker Recognition" in *Proceedings of Odyssey 2004, the Speaker and Language Recognition Workshop*, Toledo, Spain, pp. 69–74.

[3] R. Auckentaller, M. Carey and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, pp. 54–42, 2000.

[4] F. Botti, A. Alexander, A. Drygajlo, "An Interpretation Framework for the Evaluation of Evidence in Forensic Automatic Speaker Recognition with Limited Suspect Data" in *Proceedings of Odyssey 2004, the Speaker and Language Recognition Workshop*, Toledo, Spain, pp. 63–68.

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

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

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

[8] A. Drygajlo, D. Meuwly, A. Alexander, "Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition," in *Proceedings of Eurospeech 2003*, Geneva, Switzerland, 2003, pp. 689–692.

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

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

[11] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Target dependent score normalization techniques and their application to signature verification," in *Proc. of Intl. Conf. on Biometric Authentication, ICBA*,LNCS, 2004, Springer (to appear).

[12] J. Gonzalez-Rodriguez, D. Ramos-Castro, M. Garcia Gomar, and J. Ortega-Garcia, "On Robust Estimation of Likelihood Ratios: The ATVS-UPM System at 2003 NFI/TNO Forensic Evaluation" in *Proceedings of Odyssey 2004, the Speaker and Language Recognition Workshop*, Toledo, Spain, pp. 83-90.

[13] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro, and J. Ortega-Garcia, "Bayesian Analysis of Fingerprint, Face and Signature Evidences with Automatic Biometric Systems," *Forensic Science International*, 2005 (accepted).

[14] Q. Y. Kwan, *Inference of Identity of Source*, Department of Forensic Science, Berkeley University, CA, 1977.

[15] D. van-Leeuwen, J. Bouten, "Results of the 2003 NFI-TNO Forensic Speaker Recognition Evaluation" in *Proceedings of Odyssey 2004, the Speaker and Language Recognition Workshop*, Toledo, Spain, pp. 75-82.

[16] A. Martin et al. "The DET Curve in Assessment of Detection Task Performance" in *Proceedings of Eurospeech 1997*, Rhodes, Greece, pp. 1895–1898.

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

[18] M. Przybocki, A. Martin, "NIST Speaker Recogniton Evaluation Chronicles" in *Proceedings of Odyssey 2004, the Speaker and Language Recognition Workshop*, Toledo, Spain, pp. 15-22.

[19] D. Reynolds, T. Quatieri and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

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

[21] 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," in *Science and Justice*, 2001, 41(3), pp. 145-150.